KTH

Proceedings

RVK 02

RadioVetenskap och Kommunikation 02



Stockholm 10–13 juni 2002

PCC Personal Computing and Communication

U.R.S.I.

SNRV
Svenska Nationalkommittén för Radiovetenskap

VINNOVA
Verket för Innovationssystem

Naturvetenskap • Teknik • Medicin
Stiftelsen för Strategisk Forskning

Wireless@KTH
Center for Wireless Systems

Kungl. Vetenskapsakademien
The Royal Swedish Academy of Sciences

www.wireless.kth.se

rvk_02@wireless.kth.se

Proceedings
RadioVetenskap och Kommunikation 02

RVK 02

Stockholm 10–13 juni 2002

Editor: Jan-Olof Gustavsson

TRITA-S3-WS-0202
ISSN 1400-9137
ISRN KTH/WS/R--02/02--SE

Välkommen till RVK 02

Radiovetenskap och Kommunikation i Stockholm 10–13 juni 2002.

Det är nu 19de gången som Svenska Nationalkommittén för Radiovetenskap (SNRV) ordnar konferensen RVK, Nordens största radiokonferens. Första gången var redan 1949 i Stockholm. Vi är således tillbaka där den en gång startade.

Efter att under många år ordnat den i helt egen regi har konferensen de senaste åren ordnats tillsammans med andra aktörer inom radioområdet. Första gången var 1996 i Luleå då RVK ordnades tillsammans med Näringslivs- och teknikutvecklingsverket (NUTEK) och dess ramprogram för Telekommunikation. Samma upplägg användes 1999 då konferensen hölls i Karlskrona. Dessa arrangemang har varit en stor framgång inte minst genom den samlade manifestationen av svensk forskning inom radiovetenskap och telekommunikation som detta visat på.

Detta koncept har därför vidareutvecklats och i år anordnas RVK 02 tillsammans med både verket för innovationssystem (VINNOVA), som övertagit NUTEKs roll, och forskningsprogrammet Personal Computing and Communication, PCC. PCC är ett stort forskningsprogram som finansieras av Stiftelsen för Strategisk Forskning.

Programkommittén har i sitt arbete särskilt uppmärksammat spektrumfrågan och hur radiofrekvenserna skall utnyttjas på bästa möjliga sätt så att konflikter mellan olika tjänster undviks. Kommittén har också strävat efter att minska den obalans i antalet konferensbidrag som numera brukar finnas mellan grundvetenskapliga områden och tillämpningar såsom signalbehandling, kommunikationsnät och radiokommunikation. Valet av översiktsföredrag/keynotes skall ses mot denna bakgrund.

Konferensen inleds måndag den 10 juni då det blir 5 keynoteföredrag i gemensam plenarsession med internationella experter. Konferensprogrammet under tisdagen och onsdagen består av ca 30 temaföredrag samt 140 anmälda bidrag vilka presenteras som posters. Detta innebär att vi kan hålla både djup och bredd i programmet och att ett forum för mer informella kontakter mellan högskola, industri och samhälle finns tillgängligt. Då ordnar vi också en liten utställning och ger er tillgång till datorer och nätverk för Internetaccess. På torsdag blir det tillfälle till studiebesök.

Som vanligt arrangerar vi aktiviteter utöver luncherna. På måndag eftermiddag blir det ett välkomstparty och på tisdag kväll en konferensmiddag i studentikos och informell miljö. Här får alla rika tillfällen att lära känna varandra, umgås och ha trevliga diskussioner inom och utom radiovetenskapen.

Vi tackar KTH som står för de lokala arrangemangen med wireless@KTH, ett nyinrättat center för trådlös teknik, som huvudansvarig.

Vi tackar också alla andra som hängivet arbetat med RVK 02, både lokalt och nationellt, de som bidrar med vetenskapliga presentationer, övriga konferensdeltagare samt våra sponsorer och utställare vars insatser gjort konferensen möjlig.

STYRGRUPPEN • PROGRAMKOMMITTÉN • PROJEKTGRUPPEN

Som huvudansvariga för RVK 02

vill vi på Wireless@KTH hälsa er alla varmt välkomna till KTH i Kista och Stockholm.

Wireless@KTH är KTHs center för trådlösa system med ambitionen att stödja och verka för tvärvetenskaplig utbildning och forskning i samverkan med industrin. Centrat bildades för ett år sedan och spänner idag över ett flertal forskargrupper inom KTH, Stockholms Universitet och Handelshögskolan och har dessutom ett 20-tal partnerföretag. Huvudpartners är Telia och Ericsson.

Vi är glada och stolta över att ha fått uppdraget att arrangera RVK 02. Jag vill också passa på att tacka organisationskommiten och då framförallt Carl-Henrik Walde och Anna Almlöw som oförtröttligt jobbat med att få till stånd detta arrangemang.
Jag därför både hoppas och tror att ni skall trivas med RVK 02.

Välkomna!
Claes Beckman
Föreståndare Wireless@KTH

10 Radiosystem och personkommunikation

TISDAG 08.30 (F1/E31)

Radio Access Network Capacity for Future Personal Services 1
Peter Karlsson and Christian Bergljung

A Novel Direction-of-Arrival Estimation Algorithm for WCDMA 6
Buon Kiong Lau, Yee Hong Leung, Sören Andersson and Bo Hagerman

Analysis of Capacity Improvement For HiperLAN Type 2 Systems Using Switched Sector Antennas 11
P. Hafezi, P. Karlsson, F. Kristensen, C. Bergljung and M. Sandgren

Improving system performance in a WCDMA network using indoor pico base stations 16
Henrik Andersson, Robert S. Karlsson, Per Larsson and Patrik Wikström

On Signal Strength and Multipath Richness in Multi-Input Multi-Output Systems 21
Thomas Svantesson and Jon Wallace

On the geometry of signals for multipath fading channels 26
Tommy Svensson, Johan Lassing and Thomas Eriksson

Radio Resource Management Principles and Performance for WCDMA Release Five 31
Anders Furuskär, Stefan Parkvall, Magnus Persson and Maria Samuelsson

Unbiased Power Prediction on Broadband Channels 36
Torbjörn Ekman, Mikael Sternad and Anders Ahlén

Uplink Load Estimation in WCDMA 41
Erik Geijer Lundin, Fredrik Gunnarsson and Fredrik Gustafsson

14 PCC

TISDAG 08.30 (D1/D32)

The Algorithm/Hardware Co-Design Project 46
Viktor Öwall and John B. Anderson

Alternative Hardware Implementation of the Tailbiting BCJR Decoder 47
Karl Thorén, Viktor Öwall and John B. Anderson

MAP Criterion Image Coding 52
Tomas Eriksson, Mirek Novak and John B. Anderson

On the Performance of Incremental Redundancy Schemes with Turbo Codes 57
Carl Fredrik Leanderson, Giuseppe Caire and Ove Edfors

A PLL Clock Multiplier made from Digital Standard Cells 62
Thomas Olsson and Peter Nilsson

Distortion in Multistage Feedback Amplifiers 66
Martin L. Lantz and Sven Mattisson

On-chip signalling for future SOC systems 72
Imed Ben Dhaou and Hannu Tenhunen

5 Rymdfysik och radioastronomi

TISDAG 08.30 (D2/E35)

NEW ASTRONOMY WITH THE ODIN SATELLITE 77
Åke Hjalmarson

Radio studies of physics in space: New ideas and new facilities 78
Bo Thidé

A Novel SIS Mixer for Applications in Radio Astronomy 82
Vessen Vassilev, Christophe Risacher and Victor Belitsky

ALMA, THE ATACAMA LARGE MILLIMETRE ARRAY 87
R.S. Booth

Global Interplanetary Dust Distribution Measurements with the EISCAT and Arecibo HPLA radars 97
Asta Pellinen-Wannberg, Gudmund Wannberg, Assar Westman, Ingemar Häggström, John D. Mathews, Diego Janches and David D. Meisel

Incoherent scatter spectra for non-Maxwellian plasmas 101
Mikael Hedin and Ingemar Häggström

Plasma and electron line measurements with the EISCAT radars 106
Ingemar Häggström, Mikael Hedin and Björn Gustavsson

Analysis of HF enhanced 6300 Å airglow 109
Björn Gustavsson, Tima Sergienko, Urban Brändström, Thomas Leyser and Åke Steen

8 Elektroniska och optiska kretsar

TISDAG 10.30 (F1/E34)

Fundamental Potential and Limits of Fiberoptic Systems and Cellular Microwave Radio Systems in relation to Local Area Networks 114
Olle Nilsson

A 4.0 GHz Differential Frequency Synthesizer for HIPERLAN/2 in 0.8 μm CMOS 118
Håkan Magnusson and Andreas Kämpe

A Highly Integrated Single-Stage RF Down-Converter For Future Adaptive X-band Radar Systems 123
Robert Malmqvist, Andreas Gustafsson, Mattias Alfredsson and Aziz Ouacha

A Modular Asynchronous Wrapper 128
Jonas Carlsson, Weidong Li, Tormod Njølstad, Kent Palmkvist, Lars Wanhammar and Shengxian Zhuang

A tuned, inductorless, recursive filter LNA in CMOS 133
Christer Svensson, Peter Caputa and Stefan Andersson

Integrating self-calibrating image-rejection technique for 5.2 GHz HIPERLAN/2 receiver 137
Mikael Nylund

14 PCC

TISDAG 10.30 (D1/D34)

Mobile Wireless Access to Fixed Networks 141
Johan M Karlsson and Ove Edfors

A Procedure for Combining Positioning and History-Based Measurements in Dynamic Guard Channel Schemes 143
Roland Zander and Johan M Karlsson

A Structure for Work-Conserving Downlink Scheduling in Wireless FDD/TDM Based Systems 148
Torgny Holmberg and Johan M Karlsson

Evaluating a Performance Analysis of Slow FH Systems by Simulations 153
André Stranne, Fredrik Florén and Ove Edfors

Handover in Existing Wireless Systems and a Look Forward 158
Henrik Persson and Johan M Karlsson

The Need for an Alternative to IPv4-Options 163
Pierre Fransson and Andreas Jonsson

Hints and Notifications 167
Lars-Åke Larzon, Ulf Bodin and Olov Schelén

7 Radar och radiovågors utbredning

TISDAG 10.30 (D2/–)

Information Flow in Radar 172
Lars Falk

Radio Navigation and Radio Positioning in Sweden 176
Rolf Zetterberg

The Odin satellite's sub-mm measurements of ozone and related gases in the stratosphere –
Donal Murtagh

8 Elektroniska och optiska kretsar

TISDAG 13.30 (F1/E34)

Networks on Chip 178
Axel Jantsch

Layout Generation of Matched Capacitors 182
Emil Hjalmarson, Robert Hägglund, Per Löwenborg and Lars Wanhammar

Automatic Device Sizing in Analog Circuit Design 187
Robert Hägglund, Emil Hjalmarson and Lars Wanhammar

Optimal logic level pipelining for digit-serial implementation of maximally fast recursive digital filters 192
Oscar Gustafsson and Lars Wanhammar

Reduction of clock noise in mixed-signal circuits 197
Erik Backenius, Mark Vesterbacka and Robert Hägglund

14 PCC

TISDAG 13.30 (D1/D34)

Active Documents and Ubiquitous Computing and Communication –
Carl Gustaf Jansson, KTH

The Wireless IP Project 202
Mikael Sternad

Achieving Non-intrusive Notifications in a Ubiquitous Service Environment 207
Patrik Werle and Carl Gustaf Jansson

Authorization and Access Control System in Loosely Connected Distributed Environments 212
Matei Ciobanu Morogan

Dynamic Service Composition: Why and How 216
Sven Lämmermann

Thread Migration Protocol in the Mozart Systems 224
Dragan Havelka and Seif Haridi

A Tool for Real Time Simulations of TCP/IP over Wireless Fading Channels 229
Anna Ewerlid and Nilo Casimiro Ericsson

Effect of Prediction and Estimation Errors on Adaptive Modulation Systems for Wireless Channels 234
Sorour Falahati, Arne Svensson, Torbjörn Ekman and Mikael Sternad

Packet Transmission with Turbo Codes over Fading Channels 239
Bartosz Mielczarek and Arne Svensson

7 Radar och radiovågors utbredning

TISDAG 13.30 (–/E33)

A new remote sensing tool for forestry applications 244
Johan E.S. Fransson, Per-Olov Frölind, Anders Gustavsson, Gary Smith, Lars M.H. Ulander and Fredrik Walter

A Study of FMCW-radar for High Accuracy Measurements 249
Rahman Aljasmi and Jan Kjellgren

Autofocusing SAR imagery – some simulation results 254
Sune R.J. Axelsson

Deconvolution methods in radar array processing 259
Anders Nelander

Direction Finding for Electronic Warfare Systems Using the Phase of the Cross Spectral Density 264
Johan Falk, Peter Händel and Magnus Jansson

Mapping performance of curved-path SAR 269
Sune R.J. Axelsson

MIMO channel measurements in an urban street microcell 273
Jonas Medbo, Henrik Asplund, Mattias Törnqvist, David Browne and Jan-Erik Berg

Statistical modelling of the radar background response 278
Sune R.J. Axelsson

4 Elektromagnetiskt brus och störningar

TISDAG 15.30 (D2/E35)

Intentional Electromagnetic Interference (IEMI) Background and Status of the Standardization Work in the International Electrotechnical Commission (IEC) 283
M. W. Wik and W. A. Radasky

Protection of Aircraft against Electromagnetic Effects 287
Bo Wahlgren

Microwave Breakdown in Slots 291
D. Anderson, U. Jordan, A. Kim, M. Lisak, O. Lundén, M. Bäckström and I. Ferrer

Microwave Coupling into a Generic Object. Angular Pattern of Shielding Effectiveness and at Immunity Testing 295
Mats Bäckström, Jörgen Lorén and Magnus Höijer

Microwave shielding effectiveness of large mobile military systems and a low cost HPM-indicator 300
Kurt Eriksson, Björn Olsson, Karl-Gunnar Lövstrand, Barbro Nordström and Mats Bäckström

RFI in Operation and Control of Rail Vehicles 305
Jan-Olof Häggblad

14 PCC

TISDAG 15.30 (D1/D35)

Adaptive Antennas in Wideband Radio Access Networks 310
Björn Ottersten

A 20 MHz HiperLAN/2 MIMO Channel Model in NLOS Indoor Scenarios 311
Kai Yu, Mats Bengtsson, Björn Ottersten, Darren McNamara, Peter Karlsson and Mark Beach

Active-Reflect Antenna for Power Combining Unequally Spaced Arrays 316
Dhanesh G. Kurup, Anders Rydberg and Mohamed Himdi

Beamforming-based Estimators for Spatially Distributed Signals 319
Mikael Tapio

Optimised QR-decomposition Core for Adaptive Beamforming 324
Fredrik Edman and Viktor Öwall

Utilizing Channel Information in the Design of Space-Time Block Codes 328
George Jöngren, Mikael Skoglund and Björn Ottersten

Combatting Packet Loss Using Sinusoidal Extrapolation and Interpolation 333
Jonas Lindblom and Per Hedelin

Unsupervised Detection over Time-Dispersive Vector Channels 338
Anders Hansson and Tor Aulin

13 Kommunikationsnät och QoS

TISDAG 15.30 (F1/E31)

GMPLS and QoS: New Directions for the Internet 342
Arne A. Nilsson

Hybrid Modeling of Communication Networks Using Modelica 345
Daniel Färnqvist, Katrin Strandemar, Karl Henrik Johansson and João Pedro Hespanha

Education in Internet Techniques A project financed by the Knowledge foundation 350
Arne A Nilsson and Lars Magnus Lahne

Issues on Performance Measurements of TCP 352
Frida Gunnarsson, Fredrik Gunnarsson and Fredrik Gustafsson

Matching Multi-Fractal Process Parameters Against Real Data Traffic 357
Patrik Carlsson, Markus Fiedler and Arne A. Nilsson

Reducing Call Setup Delays Using SIP/SDP Compression 362
Hans Hannu, Jan Christoffersson, Stefan Forsgren, Stefan Wänstedt and Krister Svanbro

Scalability limitations of self-healing optical networks due to reliability constraints 367
Lena Wosinska

Wide Area Measurements of Voice Over IP Quality Revisited 375
Ian Marsh and Fengyi Li

3 Elektromagnetiska fält och vågor

ONSDAG 08.30 (F1/E33)

GEMS and SMART — Swedish pioneers in state-of-the-art electromagnetics 380
J. Gustafsson and U. Thibblin

A hybrid FDFD-BIE approachy to two-dimensional scattering from an inhomogeneous biisotropic cylinder 384
Martin Norgren

Dissipation in Frequency Selective Structures 387
Björn Widerberg

Identification of the current distribution at a radome from near-field measurements 392
Kristin Persson, Mats Gustafsson and Michael Andersson

Microwaves in a circularly bend waveguide of rectangular cross-section 396
Thomas Biro and Börje Nilsson

Microwaves in waveguides with curvature and varying rectangular cross-section 400
Börje Nilsson and Thomas Biro

Numerical techniques for mixed high-low frequency EM-problems 405
Stefan Hagdahl and Björn Engquist

Optimization of broadband transition from waveguide to microstrip 410
Thomas Rylander, Pär Ingelström, Jan Stake and Anders Bondeson

14 PCC

ONSDAG 08.30 (D1/D32)

Overall design of 4th generation wireless infrastructures (4GW) – A project overview & some lessons learned 415
Aurelian Bria, Maxime Flament, Fredrik Gessler, Olav Queseth, Rickard Stridh, Matthias Unbehaun, Jiang Wu and Jens Zander

Coverage Planning for Outdoor Wireless LAN Systems 420
Max Kamenetsky and Matthias Unbehaun

Digital Broadcasting and Mobile Cellular Networks to Provide Asymmetric Data Services – A Survey 426
Aurelian Bria

Diversity benefits in 60 GHz VCN wireless architecture 431
Maxime Flament and Arne Svensson

System Evaluation of Optimal Downlink Beamforming in Wireless Communication 436
Rickard Stridh, Mats Bengtsson and Björn Ottersten

TCP over WLAN and GPRS with Mobile IP 441
Jiang Wu

1 Elektromagnetisk mätteknik

ONSDAG 08.30 (D2/E35)

Accurate Measurements of small Antennas and Radiation from Mobile Phones in small Reverberation Chambers 446
Per-Simon Kildal and Yngve Hamnerius

Development of reverberation chamber for accurate measurements of mobile phones and mobile phone antennas 452
Charlie Carlsson, Per-Simon Kildal, Kent Rosengren and Jan Carlsson

Measured magnetic field distribution over a printed patch antenna compared to numerical calculations and transformation to far field distribution 457
Urban Lundgren, Jonas Ekman and Stewart Jenvey

Measuring fast electromagnetic transients 462
Åke Wisten and Per Mäkikaltio

Radar for measuring the content in storage tanks 467
Olle Edvardsson

The Radio-Aeronomy Receiver Systems at Onsala Space Observatory 471
Peter Forkman and Anders Winnberg

9 Signaler och signalbehandling

ONSDAG 10.30 (D1/E31)

Some Issues in Low Power Arithmetic for Fixed-Function DSP 473
Oscar Gustafsson and Lars Wanhammar

A Digital Down Converter for a Wideband Radar Receiver 478
Henrik Ohlsson and Lars Wanhammar

An efficient pipelined complex multiplier 482
Weidong Li, Shengxian Zhuang and Lars Wanhammar

Analog-to-digital Converter Error Correction using Frequency Selective Tables 487
Henrik Lundin, Tomas Andersson, Mikael Skoglund and Peter Händel

Estimation Error in Blind Time Error Estimation Algorithm for Time Interleaved A/D Converter System 491
Jonas Elbornsson, Fredrik Gustafsson and Jan-Erik Eklund

Narrow-Band and Wide-Band Frequency Masking FIR Filters with Short Delay 496
Linnéa Svensson and Håkan Johansson

Reducing Power Consumption with Selected Orthogonal Signals 501
Niklas Petersson, Albin Johansson, Per Ödling and Per Ola Börjesson

Noise radar using binary ADC 506
Sune R.J. Axelsson

13 Kommunikationsnät och QoS

ONSDAG 10.30 (F1/D34)

Modeling Inter-Domain Routing 511
Johan Nykvist and Lenka Carr-Motyčková

Hybrid ARQ Based on Serially Concatenated Block Codes Using Iterative Decoding for Real-Time Communication 517
Elisabeth Uhlemann, Tor M. Aulin, Lars K.Rasmussen and Per-Arne Wiberg

IP-based Distribution System for IEEE 802.11 Wireless LAN 522
Héctor Velayos and Gunnar Karlsson

Privacy Extensions for Stateless Address Autoconfiguration in IPv6 "Requirements for unobservability" 527
Alberto Escudero-Pascual

Simultaneous Routing and Resource Allocation in Wireless Networks 533
Lin Xiao, Mikael Johansson and Stephen Boyd

TCP over UMTS 538
Anders Dahlén and Per Ernström

TCP with Forward Error Correction 543
Henrik Lundqvist and Gunnar Karlsson

12 Antenner och antennsystem

ONSDAG 10.30 (D2/E34)

From Single to Dual Array Architectures A Systems Perspective 548
M.A.Beach, D.P. McNamara, P.N. Fletcher and P. Karlsson

Advanced Antenna Systems in WCDMA Networks – Downlink Capacity and Coverage Trade-offs 553
Bo Hagerman, Joszsef Barta, Sven Petersson and David Astély

Direction-of-Arrival Estimation in the Presence of Correlated Signals with Non-Ideal Uniform Circular Arrays 558
Buon Kiong Lau, Greg Cook and Yee Hong Leung

Model of an array antenna consisting of circular apertures on a truncated ground plane 563
Ola Forslund and Lennart Mattsson

Mutal coupling between dipoles on faceted conformal antenna arrays 568
Carina Marcus and Lars Pettersson

On the Capacity of Polarization Diversity Antennas in Satellite Communications 572
Sven Nordebo, Abbas Mohammed and Jonas Lundbäck

Patch antennas on inhomogeneous substrates 577
Thomas Rylander, Peter Waller, Gerhard Kristensson and Anders Derneryd

Test of Smart Antenna Properties in a Simulated Scattered Field Environment 582
Bo Olsson

The AMC-9 Ku-band Receive antenna 587
Maria Holm

9 Signaler och signalbehandling

ONSDAG 13.30 (D1/E31)

Advanced Antenna Systems for Cellular Networks 592
Bo Hagerman and David Astély

Array Mapping: Reduced Bias Transformation Matrix Design 596
Per Hyberg, Magnus Jansson and Björn Ottersten

Complex Spreading Sequence Design 601
Hai Huyen Dam, Hans-Jürgen Zepernick, Jörgen Nordberg and Sven Nordholm

Lower Symbol Error Probability Bounds for CPM Signaling over AWGN Channels 606
Zihuai Lin and Tor Aulin

Robust Cumulant-Based Blind Beamforming 611
Otto Tronarp and Jouni Rantakokko

Space-Time Code Design for Combined Channel Estimation and Error Protection 616
Jochen Giese and Mikael Skoglund

Sub-Band and Frequency Domain Double Talk Robust Algorithms 621
Jiaquan Huo, Sven Nordholm, Ka Fai Cedric Yiu and Kok Lay Teo

Subband Signal Separation 625
Jörgen Nordberg, Hai Huyen Dam and Sven Nordholm

Synchronization of Multicarrier Systems on the Twisted Copper Pair Channel 630
Daniel Landström, Niklas Petersson, Per Ödling and Per Ola Börjesson

Up- and Downlink Performance Evaluation Using Space and Polarization Measurements of UTRA-FDD Channels 635
Peter Eneroth, Christian Bergljung, Peter Karlsson, Mark Beach and Sze Ern Foo

Suppressed ambiguity in range by phase coded waveforms 640
Sune R. J. Axelsson

13 Kommunikationsnät och QoS

ONSDAG 13.30 (F1/D34)

A Research Agenda for Distributed Policy-based Management 644
Mads Dam, Gunnar Karlsson, Babak Sadighi Firozabadi and Rolf Stadler

A New Approach to Monitoring in IP Networks 647
Thomas Lindh

A distributed admission approach based on marking mechanism over Bluetooth best-effort network 652
Le Than Son, Henrik Schiøler and Ole Brun Madsen

On Router Control for Congestion Avoidance 657
Marine Auvert, Håkan Hjalmarsson, Karl H. Johansson and Gunnar Karlsson

PBAC: Probe–Based Admission Control 662
Ignacio Más and Gunnar Karlsson

Scheduling Algorithms for LET in Spatial Reuse TDMA 667
Jimmi Grönkvist

The design and application of an IP rate reservation scheme 672
Evgueni Ossipov and Gunnar Karlsson

The Effectiveness of Traffic Shaping in Networks with Small Buffers 676
György Dán and Viktória Fodor

On Delay Estimations in Random Medium Access Protocols 681
Erika Johansson and Ulf Sterner

2 Biologiska och medicinska aspekter

ONSDAG 13.30 (D2/E35)

Radiofrequency exposure from wireless equipment – European limits and test standards 686
Christer Törnevik

Recommended limits of electromagnetic exposure for the general public 689
Lars-Erik Paulsson

A detailed dosimetric study of an inner ear exposed to mobile phone radiation 692
R Kopecký, Y. Hamnerius, H. Nellström and M. Persson

Determination of RF exposure levels and compliance distances of cellular radio base stations 696
Lennart Hamberg, Nina Lövehagen, Martin Sieghahn and Christer Törnevik

Image reconstruction in a microwave tomograph for mammography 700
A. Danielsson, M. Persson and J. Howard

The importance of Specific Absorption Rate for subjective symptoms among mobile phone users 703
Jonna Wilén, Monica Sandström and Kjell Hansson Mild

Wave-Propagation from Medical Implants – Influence of Arm Movements on the Radiation Pattern 705
Anders J Johansson and Anders Karlsson

10 Radiosystem och personkommunikation

ONSDAG 15.30 (D1/D35)

Towards a 4G IP-based Wireless System Proposal 709
Tony Ottosson, Mikael Sternad, Anders Ahlén, Arne Svensson and Anna Brunström

An Algorithmic Technique for Construction of Multi Dimensional Signal Constellations 714
Jan-Erik Porath and Tor Aulin

Extrinsic Information Transfer Charts for Narrow-Band Multiple Access 719
Fredrik Brännström, Tor Aulin, Lars Rasmussen and Alex Grant

Multi-User Spatial Multiplexing – A Pragmatic Approach 724
Mats Bengtsson

Multiuser Diversity and Maximum Entropy Scheduling 729
Mathias Johansson

Performance Analysis of Impulse Radio Based on an Ultra Wideband Channel Model for Indoor Communications 734
Andres Alayon Glazunov

Synchronization of Code-Spread CDMA Systems 739
Fredrik Malmsten, Tony Ottosson and Erik G. Ström

Tabu Search Multiuser Detection in CDMA 744
Peng Hui Tan and Lars Rasmussen

Utilizing the channel correlation for MAI reduction in downlink multi-carrier CDMA systems 749
Anders Persson, Tony Ottosson and Erik Ström

3 Elektromagnetiska fält och vågor

ONSDAG 15.30 (F1/E33)

Nonlinear electromagnetic wave propagation 754
D. Anderson

Nonlinear self focusing of partially coherent light beams 758
L. Helczynski, M. Lisak, D. Anderson, B. Hall, R. Fedele and V.E. Semenov

Pulse dynamics in optical amplifiers in the presence of bandwidth limitation and nonlinear gain saturation 761
D. Anderson, P. Johannisson, M. Lisak, M. Marklund, J. Dudley, J. Harvey, R. Kruhlak, A. Peacock and E. Vanin

Suppression of Multipactor Breakdown in RF Equipment 765
D. Andersson, P. Ingvarson, U. Jordan, U. Jostell, A. Kryashev, G. Li, M. Lisak, V.E. Semenov and R. Udiljak

Suppression of nonlinear effects by phase alternation in dispersion managed optical fibre transmission 770
P. Johannisson, D. Anderson, M. Marklund, A. Berntson, M. Forzati and J. Mårtensson

Microwave breakdown thresholds in resonators and filters; the concept of effective diffusion length 774
D. Anderson, U. Jordan, L. Lapierre, M. Lisak, T. Olsson, J. Puech, V. Semenov and J. Sombrin

Författarregister 779

Radio Access Network Capacity for Future Personal Services

Peter Karlsson[1] and Christian Bergljung[2]

[1]Telia Research AB, Box 94, 201 20 Malmö, E-mail: Peter.C.Karlsson@telia.se
[2]Telia Research AB, Box 94, 201 20 Malmö, E-mail: Christian.S.Bergljung@telia.se

Abstract

It is important that wireless systems for mobile and nomadic users can fulfil the expectations and requirements of Internet Protocol services and applications. We compare the required infrastructure for a variety of radio access networks to provide capacity for expected traffic demands. Operation in specific licensed bands implies a limited capacity, but adding service operation in unlicensed and license exempt bands enhance the total capacity. In this paper we estimate the spectrum value to 10 euro per person and MHz, and propose enhanced capacity by means of a general dynamic frequency selection technique. Coexistence and spectrum sharing is more important than improving the capacity of each radio system by some percent. The ever increasing demand for bandwidth can be offered to the end users of wireless services if spectrum can be used not only for a specific system in a specific band. This implies a new type of spectrum engineering.

1 Introduction

The need and usage of bandwidth for telecommunications services have grown exponentially since telecommunications systems with few bits/s rates were launched in the mid 19th century. Another trend in the society that also grows exponentially is traveling. In the end of the 20th century a Swedish person traveled on average some 45 km per day [1]. These two mega trends in combination reveal that people need more bandwidth for telecommunication services but not at a fixed location, see figure 1. Hence it looks like we should sweep away the existing limitations of mobile and nomadic radio systems with a BROOM (e.g. BROadband On the Move project at Telia Research). Although the intense research in radio systems has solved many problems and improved the spectrum efficiency, the inherent problem is the limited amount of spectrum in combination with the variability of the radio channel.

Do people need all the bandwidth and are they willing to pay for it? We can of course be self-critical as engineers and researchers and try to avoid focusing on the technology. It seems, however, that there is not only a 'push' from telecom and datacom manufacturers and operators but also a 'pull' by the end users for more bandwidth on the move. New services can and will be developed based on the cost and availability of the link data rate and system capacity in the wireless access network. It seems that the gap between bandwidth for fixed and mobile services is not closing as would be required to fulfill the demands identified by the mega trends in the society. How can research in radio systems solve the bandwidth problem and reduce the fixed/mobile gap? Promising innovative technologies are being introduced, e.g. adaptive modulation/coding, multi-user detection and adaptive antennas both at the terminal and base station side, to provide an increased spectral efficiency. However, we must not only increase the spectrum efficiency in bits/s/Hz/site for each wireless technology, but also investigate how additional frequency bands could be used as discussed in [2]. This paper looks into some aspects of the radio access network capacity and discusses the value of spectrum and how spectrum management could be solved.

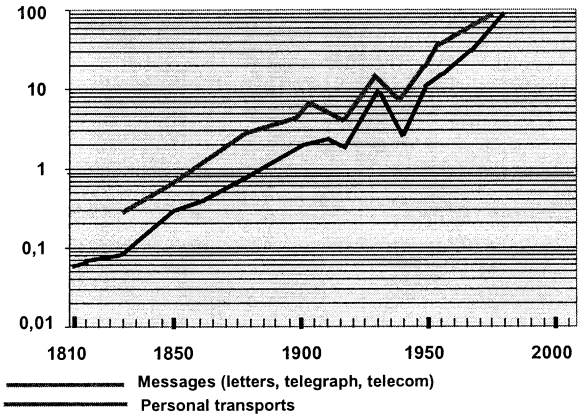


Figure 1: *Development of transport and message volumes 1810-1990. Source A. Grübler.*

2 Services and Data Rates

The second generation mobile systems all provide the same voice service to end users, although the data rates varies slightly. A global conference call with individual participants using, both fixed telephones and e.g. DECT, GSM and IS95 mobile terminals, is possible to operate.

The ambition now is to extend network capacity and make multimedia services access independent in a similar way. There is a growing interest in video conferences, email with attachments, audio and video download and picture upload. The Internet Protocol, and in particular version 6 (IPv6), could be the common denominator for most of such personal services. There is, however, no way to identify specific future killer applications, rather mention that we will have a manifold of applications enabled by broadband connectivity to the Internet [3].

Nevertheless, it is important to characterize the necessary capacity of a radio access network based on prediction of service categories and the data rate of each service. In this survey we have categorized future services into 5 groups based on what was used in UMTS Forum for spectrum requirement estimations [4]. Here the low rate services, like SMS and voice, have been omitted and instead we describe the following categories:

- MiMM (Medium speed interactive Multi-Media); Email, Video Conference
- HMM (High speed Multimedia); Web Browsing, File Transfer, Video Surveillance
- HiMM (High speed interactive Multi-Media); Web Hosting, Teleprescence
- VHMM (Very high speed Multi-Media) Spooling Video and Streaming Video
- VHiMM (Very high speed interactive Multi-Media); advanced Multiplayer Game.

The data rates and traffic of these services are given in Table 1. The values reflect what is expected to be required in a corporate environment in the period from 2006 to 2010 and the average activity factor summing up all services is close to 1.9 Mb/s per mobile terminal (MT) during busy hour.

Table 1: *Predicted service groups and traffic per mobile terminal in a corporate environment from year 2006.*

	Average data rate (Mb/s)	Peak Data rate (Mb/s)	Busy hr sessions (average)	Session Duration (s)	Activity Factor (Mb/s/MT)
MiMM	2	4	0,5	300	0,45
HMM	2	10	0,5	600	0,16
HiMM	5	8	0,3	1200	0,35
VHMM	7	20	0,1	1800	0,5
VHiMM	5	15	0,1	1500	0,4

Several analysts forecast that Internet traffic will continue to double every 12 months. This can be justified as a long-term trend in agreement with what is shown in Figure 1, even taking into account a dip from the current telecom recession, similar to what happened in the 1930s. The estimations shows that the data traffic will experience a 25 fold increase between 2001 and 2006 and exceed 10^{18} bits per day or 60 Tb/s of busy hour traffic in US [5]. This corresponds to an increase in the average traffic per US citizen from 8 kb/s to 200 kb/s. Our 1.9 Mb/s per user in a corporate environment year 2006 is well aligned with this Internet data rate growth. Average data rates in the range from 200 kb/s to 2 Mb/s should be possible to handle by radio access networks, since access independent services are generally asked for.

3 System Capacity

The capacity of each radio system is to a high degree dependent on the frequency bandwidth allocated. Let us make a capacity estimation and take Sweden as an example with some 9 million people and an area of 450000 km^2. We start with an impossible but, nevertheless, revealing scenario. We assume that a 2 Mb/s average service rate per user would be wanted already 2006 (see table 1) and one giant macrocell operating with an omni antenna without output power restrictions is deployed to reduce the cost of infrastructure. In this scenario 18000 GHz (assuming 1 b/s/Hz spectral efficiency) of spectrum would be required. But the Swedish area capacity requirement is not astonishing; on average just some 40 $b/s/m^2$ in this example.

The spectrum shortage problem for a low cost infrastructure is well known since the early days of mobile radio systems and high capacity layouts were instead proposed [6] and are now used in existing networks. Thus, if we make a cell splitting in this capacity estimation example, ending up in the order of today's GSM system, some 3000 base station sites can be used for a more realistic estimation.

High capacity is provided if all sites are equipped with 3 sectors and the maximum number of transceivers (TRX) is reused at each cell. The propagation path loss in combination with carrier to interference ratio for a sufficient end user quality are major limitation parameters of the capacity, due to the code/power restrictions in CDMA [7] and frequency reuse in FDMA/TDMA systems even if dynamic channel allocation is used [8]. When the basic parameters are taken into account, the capacity in b/s/Hz/site is typically in the range from 0.2 for GPRS to just above 0.45 for EDGE and WCDMA [9]. So at least some 2 [Mb/s] * 7.7 [Musers] / 0.5 [b/s/Hz/site] / 3000 [sites] = 10.3 GHz of spectrum would be needed to provide a 'dedicated' wireless Internet service for an 85% user penetration of the Swedish population. Of course, we must consider the traffic pattern and user activity factors as discussed above, but note that a 200 kb/s average rate in busy hour still would require at least 1 GHz of spectrum. The way forward is then an increase in cell splitting and number of base station sites. Operators have to trade infrastructure investments versus spectrum allocations.

We can also take Kista as an example with 24217 daytime workers on an area of 2.5 km^2 according to demo-

graphic statistics [10]. What is the radio access network capacity required to provide the Swedish wireless valley in Kista with a 1.9 Mb/s average wireless service during busy hours for each professional user/corporate employee as predicted for year 2006? With one central base station in Kista we would need approximately 1.9 [Mb/s] * 24217 [users] / 0.5 [b/s/Hz/site]/1site = 92 GHz of spectrum for a true mobile Internet in 2006 using the spectrum efficiency of the UTRA-FDD WCDMA standard.

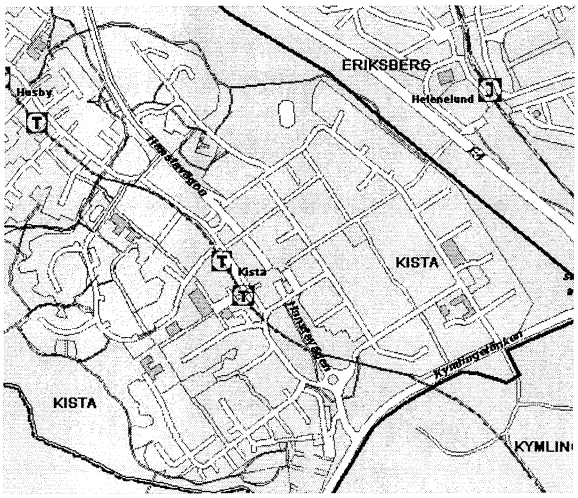


Figure 2: *Map of Kista; 24217 employees in 2.5km^2.*

Again the figure reveals the impossible scenario (even with lower activity factors); the UTRA-FDD certainly can fulfill the coverage of 2.5 km^2 with one base station site in Kista but there is an obvious capacity limitation. People tend to work and meet in clusters and most traffic is always generated in such hot spots. The expression 'Coverage is King' must seriously be taken into account and cellular systems have been developed to fulfil this basic requirement. However it is important to note that the area capacity requirement in this example is some 20 kb/s/m^2. Thus in such a high user density environment the area capacity is 500 times the average Swedish requirement (our first example).

The total UTRA-FDD spectrum allocation for all operators is only 2*60 MHz and it is obvious that a cell splitting is needed to provide the Swedish wireless valley with mobile Internet. Being more modest in average user service than the prediction above, we can rather assume that a 384 kb/s rate per link and an activity factor of 0.25 is relevant during busy hours around year 2006 in agreement with the 200 kb/s Internet traffic prediction. The actual maximum distance between base stations for the spectrum allocations and system efficiencies of standardised systems have been calculated for this Kista scenario. The radio link reliability for carrier to noise and interference ratio is set to 90% for each modulation at the cell border each system and a distance dependent path loss of 30*log10(d) is used. The results are shown in figure 3 and it is worth noting that the maximum BS separation to fulfil coverage *and* capacity requirements is within 30-200 m range both for cellular systems and WLANs. Note also that each BS site in this operator independent radio system comparison is 3 sectored and fully loaded with transceivers, i.e. 83 sites with 159 TRX each and 136 sites with 36 TRX each, for EDGE and UTRA-FDD, respectively.

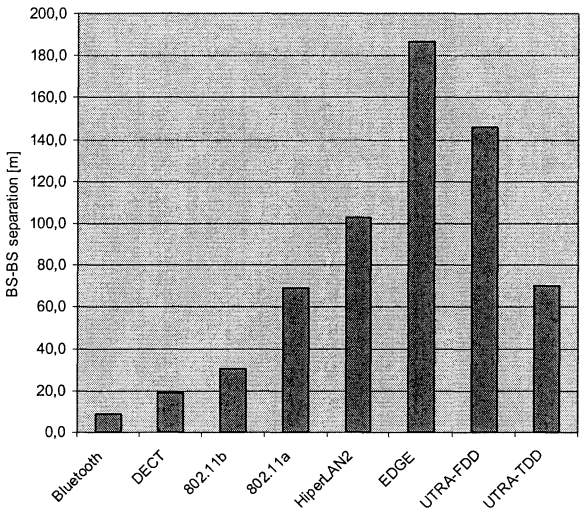


Figure 3: *Maximum BS to BS separation to fulfil wireless data capacity in Kista with different systems if all allocated spectrum can be used.*

The WLAN systems can do with only one TRX per site for the required capacity. Only Bluetooth and the 5 GHz WLAN systems IEEE 802.11a and HiperLAN2 are noise limited, while the cellular systems as well as IEEE 802.11b are capacity limited in this 200 kb/s/user scenario. Note in particular that the range difference between noise limited HiperLAN2 and capacity limited UTRA-FDD is only 20 m. Hence high capacity radio access and related infrastructure cost for wireless Internet in places like the wireless valley is still an interesting research topic. System combination and multi access advantages must clearly be taken into account.

3.1 Spectrum value

The radio access network capacity is mainly limited by the frequency allocation. Spectrum is internationally regulated and administrated by ITU-R under the UN. The world Radio Conference (WRC) discusses and resolves the spectrum allocations and system sharing principles every third year. CEPT handles the European countries and is working on spectrum engineering tasks and preparations for ITU-R. The radio frequency spectrum is, as indicated in the estimations in the previous section, a limited resource. The £22.5 billion paid in auction for UMTS spectrum in UK and the 80 billion euro paid in Germany are two extreme examples of the spectrum value. The license cost for the end user assuming a 50% penetration of UMTS service subscription is almost 1200 euro and 1900 euro in UK and Germany respectively. In Sweden the radio access network requirement is deter-

mined by a 99.98% population coverage defined by pilot tone reliability ends up in an infrastructure cost (mainly for base stations in rural areas) of 1000 euro per user assuming a 50% penetration. In all these cases the value of the spectrum can be discussed, nevertheless some 8-15 euro per user per MHz is invested. Let us then assume that the spectrum up to 6 GHz, where frequency allocations of standardized and proposed mobile system ends, has the same potential value for all persons. Thus the spectrum resource is worth approximately 24000 euro/person if the total population in Sweden is taken into account. Just think of the UHF terrestrial TV allocation of 380 MHz, alone worth some 6000 euro per person. Can we afford not optimizing spectrum management to handle the expected bandwidth and capacity needs?

4 Radio Access Enhancements

Almost all domains have been investigated for improving the radio link reliability and spectrum efficiency. The time domain and the frequency domain are enhanced by means of (sub-)optimum codecs, equalizers, Rake receivers etc. But then in the end most systems experience a similar performance in b/s/Hz/cell, irrespectively of frequency hopping TDMA, turbo coded CDMA or OFDM with coding over sub-carriers. Still, the spatial domain is not fully utilized, and if, only at the base station side. There are now several research results showing that several antennas both at the transmitter and receiver side, i.e. multiple input and multiple output (MIMO), increase the spectrum efficiency to several b/s/Hz/cell, see e.g. [11]. Different diversity and space-time coding schemes have been investigated and all add some complexity to the devices. However, the cost and complexity of additional antennas can be motivated by the improved spectrum efficiency.

Recently, laptop manufacturers have taken the terminal antenna into account and now provide solutions with two antennas as key elements to the integrated WLAN solution. This approach is also considered by standardization efforts in 3GPP, where MIMO is investigated for new releases. It is, however, almost strange that cellular mobile terminals, optimized for the variable radio environment, have not yet been equipped with several antenna elements. The arguments have always been that users want low cost devices. But customers using mobile services and the emerging wireless Internet have to pay for infrastructure and/or spectrum investments in addition to service development, operation and maintenance and terminals.

The intra system spectrum efficiency and access capacity could be enhanced by techniques mentioned above. We must also solve the spectrum management to meet the increasing traffic demand for the wireless Internet.

4.1 Dynamic Frequency Selection

The HiperLAN2 standard supports dynamic frequency selection (DFS) functionality, which is required both for intra-system performance and inter-system interference mitigation. Sharing with other radio systems, e.g. radars and satellites, is then possible and 455 MHz of licensed exempt spectrum in the 5 GHz band can be used. The DFS is an essential technique for operation in the 5 GHz band. Note the difference between the more general DFS and the dynamic channel allocation (DCA), which is used by DECT for intra system interference management in the specific frequency allocation. In addition to link adaptation, we can say that it is better to move a channel with less interference than to remain and decrease the coding/modulation mode. This kind of technique should be considered and incorporated in any new radio system proposal and thus allowing higher system throughput with operation in the least interfered channel.

4.2 Ultra wideband

Another method of spectrum management is ultra wideband (UWB). The idea here is to use frequency bands overlapping existing primary allocations by means of spread spectrum technique. After more than 3 years of investigations FCC announced on February 14, 2002 that UWB is allowed to operate in spectrum where other radio systems and services are allocated. The scare spectrum resources can then be used more efficiently allowing UWB devices to perform high-speed data transmissions.

However, the effects of widespread use of UWB are not yet fully known, and further research in spectrum management and sharing is needed. Interference from a large number of UWB devices could impact critical spectrum users, while DFS mechanism avoids channels with interference.

4.3 Resource Management

Management of radio resources can be made within a specific system or taking several radio access networks into account. Today GSM 900/1800 is seen as a single system by end users and resources are possible to handle when operators have allocations in both bands. Technically radio access networks should be possible to handle jointly for improved capacity.

Another solution is distributed resource management in Ad-Hoc networks but it is important to investigate the performance degradation in wireless environments [12]. Even if a multi-hop scenario could reduce the terminal output power the Ad-Hoc throughput per link is lower. The spectrum efficiency on top of the medium access control (MAC) layer varies significantly between existing centralized and distributed WLAN systems even if the same physical mode (PHY) is used. Comparison of the IEEE 802.11a and HiperLAN2 link throughput for different PHY modes is presented in Figure 4. The

throughput of the centralized system is roughly twice when averaging over the physical modes. The spectrum efficiency can be doubled without improving the physical layer at all. Coordination of MAC and physical layers in combination with frequency selection is important for the shared wireless medium.

Re-configurable multi mode terminals targeting dynamic spectrum management are key components towards the manifold of applications with different data rates.

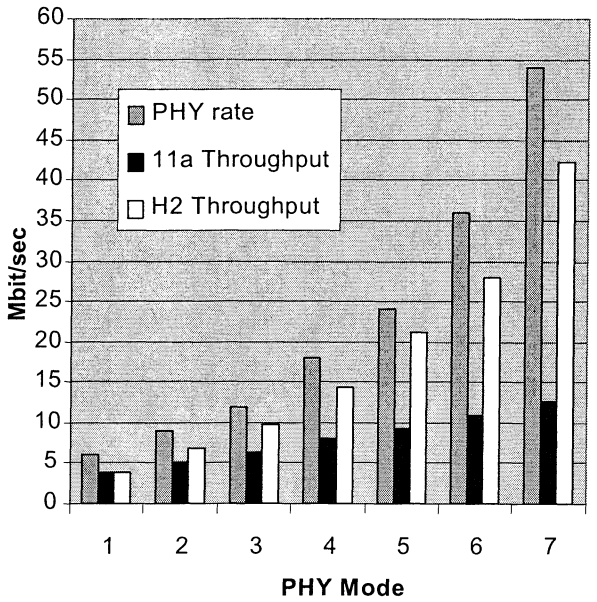


Figure 4: *Throughput above the MAC layer for a centralised (HiperLAN2) versus distributed (802.11a) resource control using similar physical layers.*

5 Conclusions

We have discussed the capacity of radio access networks for future personal services. Some of today's wireless systems have been used for comparisons of capacity and system layout in interference and noise limited scenarios. All access systems except WLANs operating in the 5 GHz band are capacity limited in urban areas like the wireless valley in Kista if expected Internet services to mobile terminals are considered. Frequency allocations are crucial and we estimated the value of the spectrum to some 10 euro per person and MHz. The spectrum value must be traded off the cost of increased system complexity. Thus, access capacity enhancements like multiple antennas are justified to provide the expected traffic increase to user terminals in the existing frequency allocations. The selection of appropriate frequency band for the end user by means of a general dynamic frequency selection is proposed for enhanced radio access network capacity. Principles of spectrum sharing should be further studied for mobile terminals to be able to operate in licensed, license exempt and unlicensed bands.

Acknowledgement

We acknowledge all contributions from people at the Wireless Solutions division at Telia Research AB.

References

[1] http://www.sika-institute.se/utgivning/ss2001_2r1.pdf.

[2] L. Wieweg and L. Sonesson, "Fair and square spectrum sharing principles", Proceedings of RVK 99, pp. 489 -494, RVK June 1999.

[3] M. W. Smith, H-T Leung, "Finding the killer application – the role of the Broadband Applications Laboratory", BT Technology Journal, Vol 20 No1, January 2002.

[4] ITU-R Report M.2023, Spectrum requirements for IMT-2000.

[5] http://www.insight-corp.com/atmip.html

[6] J-E Stjernvall, "High Capacity cellular layout", Proceedings of NRS 86, Nordic Radio Symposium, April, 1986.

[7] V. Veeravalli and A. Sendonaris, "The coverage-capacity tradeoff in cellular CDMA systems", IEEE Trans. On vehicular technology, Vol. 48, No. 5, September 1999.

[8] H. Eriksson, "Capacity improvement by adaptive channel allocation", In Proceedings IEEE Globecom'88, US, 1988.

[9] A. Furuskär, "Can 3G services be offered in existing spectrum?", Teknisk Licenciate degree, KTH, ISRN KTH/RST/R-01-04-SE, May 2001.

[10]http://www.usk.stockholm.se/internet/omrfakta/omrfakta.asp?omrade=s22603&typ=stadsdel

[11] R. Stridh, P. Karlsson, B. Ottersten, "MIMO Channel Capacity on a Measured Indoor Radio Channel at 5.8 GHz", Proceedings Asilomar Conference on Signals, Systems & Computers, October 2000.

[12] J. Zander, "Trends in Resource Management Future Wireless Networks", Proceedings of RVK 99, pp. 501-505, RVK June 1999.

A Novel Direction-of-Arrival Estimation Algorithm for WCDMA

Buon Kiong Lau[1], Yee Hong Leung[2], Sören Andersson[3], and Bo Hagerman[4]

[1]ATCRC, Curtin University of Technology, Australia. Email: vincent@atri.curtin.edu.au
[2]ATCRC, Curtin University of Technology, Australia. Email: leung@atri.curtin.edu.au
[3]Ericsson Research, Kista, Stockholm, Sweden. Email: Sören.S.Andersson@era.ericsson.se
[4]Ericsson Research, Kista, Stockholm, Sweden. Email: Bo.Hagerman@era.ericsson.se

Abstract

Despite the controversies that surround the introduction of the third generation (3G) mobile telecommunication systems, the massive investment by industry and the constant demand for new technology and services are expected to usher in its eventual success. WCDMA is a dominant standard in the 3G system designed to succede the highly successful GSM standard. A new technology proposed for WCDMA is adaptive antennas. Among the challenges involved in the technology are downlink beamforming and mobile user localization whereby a reliable direction-of-arrival (DOA) estimation algorithm could form part of the solution. In this paper, we propose a simple single-user approach to perform DOA estimation in WCDMA that is based on a well known space-time (2-D) rake receiver for uplink reception. We show that by incorporating some additional knowledge of the signal environment, promising improvement can be obtained. The algorithm is tested under the realistic COST259 channel model to further substantiate our claim.

1 Introduction

The long awaited rollout of WCDMA services in Japan in Oct 2001 marked the beginning of third generation mobile telecommunications services. Even though the initial consumer response did not match the industry's optimism, major telecommunication companies around the world are going ahead with their plans to introduce the new system. Aside from the pressure to recover the many billions of US dollars already invested into research and the purchase of spectrum licenses and equipment, the companies recognize the vast potential of the system in providing new revenue-generating services. The search is currently on for a killer application to encourage quick and widespread adoption of the technology.

The development and rollout of WCDMA also pose a new set of difficult technical challenges to engineers, as is clear from the recent Japanese experience. In comparison with second generation systems, WCDMA is much more advance in technology, and thus far more difficult to implement and maintain. Issues such as system capacity and quality of service will become increasingly important as the systems grow. The maturing adaptive antennas system (AAS) technology is favored to play an important role in the future enhancement and development of WCDMA systems [1].

The use of AAS in mobile telecommunications has received significant attention in recent years [1], [2], [3]. Commercial products incorporating the technology, such as Ericsson's RBS2205, are already available. The main advantage of AAS lies in its ability to exploit the spatial characteristics of the mobile channel to improve the reception of mobile signals. WCDMA systems, in particular, are inherently suitable for AASs because the capacity of CDMA systems is soft-limited by interference level and any suppression of interference would provide an instant capacity increase [1].

Currently, there exist a few different types of AAS architectures, ranging from the simple fixed multi-beam AASs for both uplink and downlink, to the fully-steerable beam AASs for downlink and combined spatial/temporal interference cancellation for uplink [4]. For steerable beam AASs, an explict estimation of the directions-of-arrival (DOA) of the signals of interest is required for downlink beamforming since the use of frequency division duplex (FDD) mode in WCDMA renders the use of estimated array weights based on uplink spatial signature ineffective for downlink beamforming [5]. Moreover, DOA is also important for emergency services and other radiolocation applications [6].

In this paper, we are concerned with the single-user approach to DOA estimation for WCDMA using uniform linear arrays (ULAs). In particular, we are interested in techniques that make use of information available to the AAS, e.g. [7], [8]. We propose a simple DOA estimation algorithm with an aim to *reuse* as much as possible the decoupled 2-D rake receiver structure in [9]. In essence, we make use of the pilot sequence and the root raised cosine (RRC) chip waveform in the Dedicated Physical Control Channel (DPCCH) [10] on top of the known desired user code in order to improve on the estimate of the desired signal covariance matrix. This improvement results in smaller biases and standard deviations in the DOA estimates caused by the non-whiteness of multiple

access interference (MAI). We also apply an estimate of the MAI-and-noise covariance matrix to whiten the signal-MAI-noise covariance matrix and show that a further improvement can be obtained. Among the benefits of this approach is that the number of array elements no longer limit the number and proximity of identifiable signal paths [7]. In fact, different users with the same DOA can be identified.

In this paper, we despread the received WCDMA signal and use Root-MUSIC [11] on the despreaded covariance matrix for DOA estimation. Many other DOA estimation techniques may be similarly applied to the covariance matrix, e.g. Root-WSF or MODE [12], or ESPRIT [3]. In the simulation study, we demonstrate the performance of our method in dynamic (with user mobility) simulations by the use of the realistic COST259 channel model [13]. We show that our method can significantly outperform the conventional approach.

2 WCDMA Uplink Signal Model

Here we give a summary of the WCDMA signal model [9], [10], [14]. The user data in WCDMA is carried on the Dedicated Physical Data Channel (DPDCH). More than one DPDCH may be multiplexed onto the in-phase (I-) and quadrature-phase (Q-) channels. The DPDCHs on each of the I- and Q-channels uses different orthogonal variable spreading factor (OVSF) codes for channelization. The spreading factor used in the DPDCHs differs according to the desired data rate. Each user also has a DPCCH with a constant spreading factor of 256 that is mapped onto the Q-channel. For DPCCH, one radio frame has 15 slots and each slot has 10 bits. The number of pilot bits is $N_p \in [3,\ldots,8]$ [10]. The DPDCHs with lower spreading factor (higher data rates) are allocated higher amplitude gains. The amplitude gain factors are: β_D for DPDCH, and β_C for DPCCH. In this paper, we consider only one DPDCH on the I-channel and one DPCCH on the Q-channel. In Fig. 1, we show the block diagram for uplink spreading.

The baseband representation of DPDCH [9] is

$$s_D(t) = \sum_{k=-\infty}^{\infty} b_D^{(k)} z_D(t - kT_D) \quad (1)$$

where

$$z_D(t) = \sum_{h=1}^{Q_D} c_D^{(h)} p(t - hT_c), \quad (2)$$

T_D is the symbol period of the data on DPDCH, T_c is chip period corresponding to the WCDMA chip rate of 3.84×10^6 chips/s, $c_D^{(h)} \in \{\pm 1\}$ is the hth chip of the OVSF spreading code, $b_D^{(k)} \in \{\pm 1\}$ is the kth data symbol and $p(t)$ is the impulse response of the RRC pulse shaping filter for the chip waveform on WCDMA. The roll-off factor of $p(t)$ is $\alpha = 0.22$. Similarly, the baseband representation of DPCCH is

$$s_C(t) = \sum_{k=-\infty}^{\infty} b_C^{(k)} z_C(t - kT_C), \quad (3)$$

where

$$z_C(t) = \sum_{h=1}^{Q_C} c_C^{(h)} p(t - hT_c), \quad (4)$$

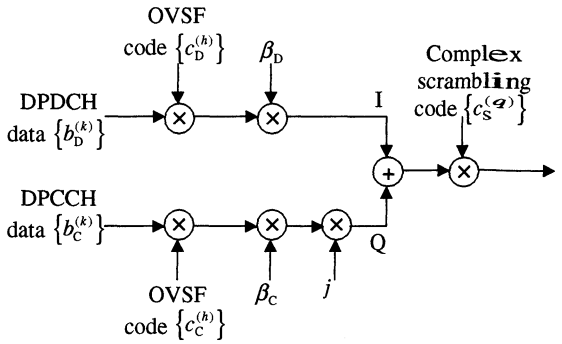


Fig.1. Block diagram for uplink spreading

and $b_C^{(k)}$, $c_C^{(h)}$, T_C and Q_C follows the notations of the DPDCH case.

The DPDCH and DPCCH are then appended onto the I- and Q-channels respectively of the transmitted QPSK signal. The complex scrambling code of length $Q_S = 38400$ chips (length of one radio frame) is also added. The scrambling code is aligned with the start of each radio frame. The transmitted baseband signal of the ith user equipment (UE) is then

$$s_i(t) = \left[\sum_{r=-\infty}^{\infty} s_S(t - rT_f)\right] \times \left[\beta_D s_D(t) + j\beta_C s_C(t)\right], \quad (5)$$

where

$$s_S(t) = \sum_{q=1}^{Q_S} c_S^{(q)} p_{T_c}(t - qT_c), \quad (6)$$

and

$$p_{T_c}(t) = \begin{cases} 1, & 0 \le t < T_c \\ 0, & \text{elsewhere} \end{cases} \quad (7)$$

$c_S^{(q)} \in \{\pm 1 \pm j\}$ is the qth chip of the complex scrambling code and $T_f = Q_s T_c = 10\text{ms}$.

The AAS of interest is a N-element ULA with inter-element spacing d and operating wavelength λ. For M UEs where $M = 1$ is the desired UE, the continuous time signal at the array output after the receiver RRC filter can be written as follows:

$$\mathbf{x}(t) = \sum_{\ell=1}^{L_1} \rho_{1\ell}\mathbf{a}(\theta_{1\ell}) s_1(t - \tau_{1\ell}) \otimes p(t) + \sum_{m=2}^{M}\sum_{\ell=1}^{L_m} \rho_{m\ell}\mathbf{a}(\theta_{m\ell}) s_m(t - \tau_{m\ell}) \otimes p(t) + \mathbf{n}(t) \quad (8)$$

where $\otimes$ denotes the convolution operation, $\theta_{m\ell}$, $\rho_{m\ell}$, and $\tau_{m\ell}$ are respectively the DOA, complex gain and delay of the ℓth path of the mth UE, L_m is the number of signal paths of the mth UE, $\mathbf{n}(t) = [n_1(t),\ldots,n_N(t)]^T$, $n_\eta(k)$'s are complex i.i.d Gaussian noise with power σ_N^2, and $\mathbf{a}(\theta_{m\ell}) = [a_1(\theta_{m\ell}),\ldots,a_N(\theta_{m\ell})]^T$,

$$a_\eta(\theta_{m\ell}) = \exp\left(j\frac{2\pi d}{\lambda}(\eta - 1)\sin\theta_{m\ell}\right). \quad (9)$$

Assuming negligible cross-correlation between desired signal and MAI.The covariance matrix is given by

$$\mathbf{R}_{SIN} = \mathbf{E}\{\mathbf{x}(t)\mathbf{x}^H(t)\} = \mathbf{R}_S + (\mathbf{R}_I + \mathbf{R}_N) = \mathbf{R}_S + \mathbf{R}_{IN} \quad (10)$$

where $\mathbf{R}_S$, $\mathbf{R}_I$, $\mathbf{R}_N = \sigma_N^2\mathbf{I}$ is the desired signal, MAI and noise covariance matrices, respectively.

3 Proposed Method

3.1 Oversampling

The use of the pilot sequence (i.e. a known signal) to estimate DOA in CDMA systems is well known [7]. Here, we use an oversampled (of factor $N_S > 1$) and normalized raised cosine (RC) pulse-shaped chip sequence to correlate with the received signal. We aim to recover the part of the signal covariance matrix that contains the direct path desired signal cluster (of local scattering). This is done by cross-correlating the received signal to obtain the time positions of the L_{c1} largest correlations, t_n's. We can then form

$$\mathbf{y}^{(n)} = \sum_{\kappa=1}^{N_E} \mathbf{x}\left(t - t_n - \kappa\frac{T_c}{N_S}\right)\tilde{s}_1^*\left(t - \kappa\frac{T_c}{N_S}\right) \qquad (11)$$

where $N_E = N_p N_S Q_C + \tilde{N} - 1$, and $\tilde{N}$ the length of the oversampled RC pulse used, $\tilde{s}_i(t)$ denotes the RC pulse convolved DPCCH chip waveform of ith user for the entire length of the known pilot sequence,. Therefore,

$$\hat{\mathbf{R}}_S^{(n)} \approx \frac{1}{N_E}\mathbf{y}^{(n)}\mathbf{y}^{(n)H},\ n \in \{1,\ldots,L_{c1}\}, \qquad (12)$$

where in our simulation study, we consider only one rake finger, i.e. one with the largest correlation.

3.2 MAI Whitening

To perform whitening for the nth rake finger, we need to estimate $\mathbf{R}_{IN}$ at that finger, i.e. $\hat{\mathbf{R}}_{IN}^{(n)}$. We start with obtaining an estimate of $\mathbf{R}_{SIN}^{(n)}$ as follows

$$\hat{\mathbf{R}}_{SIN}^{(n)} = \frac{1}{N_E}\sum_{\kappa=1}^{N_E}\mathbf{x}\left(t - t_n - \kappa\frac{T_c}{N_S}\right)\mathbf{x}^H\left(t - t_n - \kappa\frac{T_c}{N_S}\right). \qquad (13)$$

When more than one slot are used for DOA estimation, the covariance matrices of all the slots are averaged. From (8) and (10) we can then form

$$\hat{\mathbf{R}}_{IN}^{(n)} \approx \hat{\mathbf{R}}_{SIN}^{(n)} - \hat{\mathbf{R}}_S^{(n)}. \qquad (14)$$

We use $\hat{\mathbf{R}}_{IN}^{(n)}$ to whiten the MAI in $\hat{\mathbf{R}}_{SIN}^{(n)}$. The resulting $\hat{\mathbf{R}}_{SIN}^{(n)}$ is then used for DOA estimation. Note that (14) has an exact form for the case of chip rate sampling [9] (not involving $p(t)$), though in [9] it is used for uplink beamforming on a 2-D rake receiver.

4 WCDMA Simulator

In this section, we describe the WCDMA simulator developed for the evaluation of DOA estimation.

4.1 Simulator Structure

In the simulator, all UEs generate and transmit continuous data streams that follows the WCDMA radio interface specifications [10]. The long code is used for scrambling. The signals are passed through a COST259 channel and received at a base station (BS) equipped with an AAS. The AAS consists of a ULA where the array outputs are fed into a decoupled 2-D rake receiver [9].

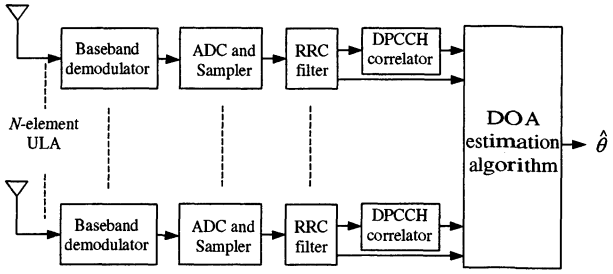


Fig. 2. Block diagram of simulator

The cross-correlation to search for the rake fingers is performed over $N_S = 2$ oversampled chip data. To model the asynchronous uplink, we assumed ideal sampling time for the desired UE and random sampling instants for the interfering UEs. Fig. 2 shows the block diagram of the parts of the receiver that are relevant to DOA estimation. Note that we do not require any additional information than those available to the 2-D rake receiver with $N_S = 2$.

4.2 Simulation Environment

In this study, we investigate a scenario with one desired UE and multiple interfering UEs in the 2GHz frequency band. The base station (BS) is operating in a 120° sector cell and employs a 4-element ULA with $d/\lambda = 0.5$. The simulation scenario is illustrated in Fig. 3.

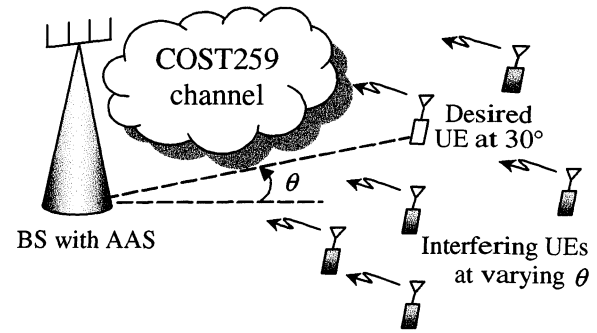


Fig. 3. Simulation scenario: a BS, a desired UE and multiple interfering UEs.

WCDMA uses both open (slow) and closed loop (fast) power control for all active UEs. Therefore, the power discrepancies among the UEs are expected to be small. Here we assume perfect power control such that the channel response of each UE in the system is normalized with respect to the reference antenna (i.e. antenna 1).

4.3 COST259 Channel Model

The COST259 channel model [13] is a versatile channel model devised under the COST259 project of the European Community. COST259 has been validated using measurements in the 1GHz to 2GHz range, but is expected to be applicable at least in the range 450MHz to 5GHz. It is a wideband directional channel model capable of providing channel impulse responses in both spatial (azimuth and elevation) and temporal domains.

Channel type	Parameter		
	Average number of clusters	Typical azimuth spread per cluster (degrees)	LOS cut-off distance (meters)
TU	1.17	10	500
BU	2.18	10	500
RA	1.06	5	5000
HT	2.00	5	5000

Table 1. Examples of COST259 Channel Parameters

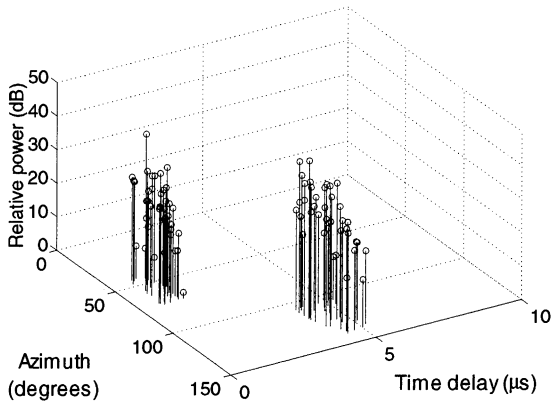


Fig. 4. Example of a BU channel realization with two clusters in the temporal/azimuthal plane.

The COST259 channel model incorporates, among others, the effects of path loss, fast fading and shadow fading. The environment currently implemented is macrocell. The channel types modeled are typical urban (TU), bad urban (BU), rural area (RA) and hilly terrain (HT), which are generalizations of the GSM channel models with the same names. Examples of the parameter settings are given in Table 1. An example of a BU channel realization with two clusters is shown in Fig 4.

5 Simulation Studies

In the simulation studies, for each of the COST259 channel types, RA, TU, BU, and HT, and a flat fading (FF) channel, we perform 300 Monte-Carlo experiments where in each experiment the scattering environment and UE distribution is changed. In each experiment, either 5 or 15 slots are used for DOA estimation where each slot contains 6 pilot bits. There are a total of 50 UEs: the desired UE (speech user) is situated at 30° and 250m from the BS while the 49 interfering UEs are randomly placed in the sector between the range 200m and 800m and each UE is moving at 3km/hr along the radial direction. We assume 30 speech users, 10 medium data rate users and 10 high data rate users where β_D/β_C of the speech, medium data rate and high data rate users are 1, 3 and 5, respectively. Out of sector MAI and ambient noise are modeled as Gaussian distributed noise. The signal-to-noise ratio (SNR) of each UE after the receiver RRC filter is 20dB. We apply root-MUSIC [11] in all cases for DOA estimation.

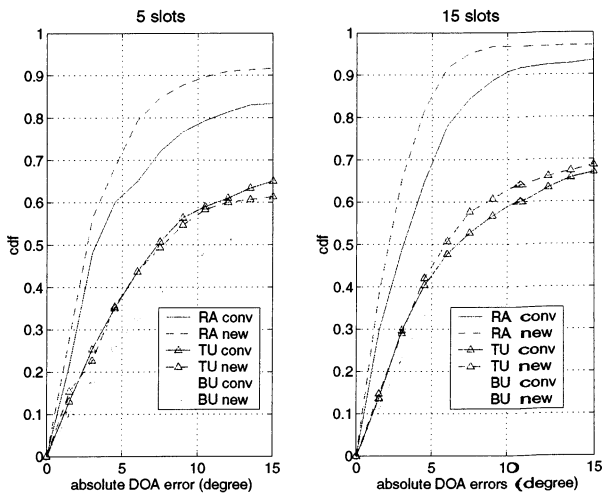


Fig. 5. Cdfs of absolute DOA error for RA, TU and BU

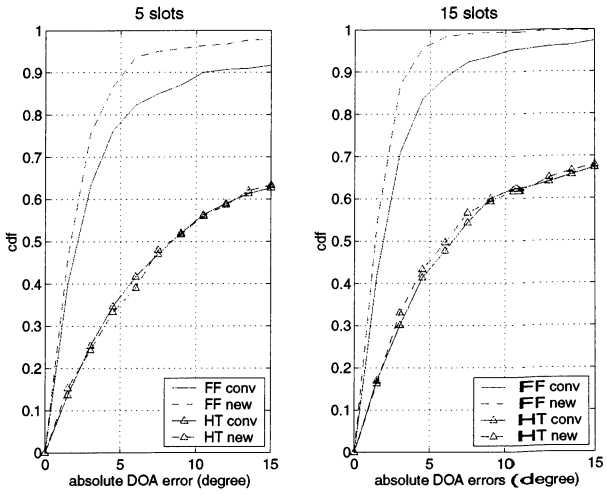


Fig. 6. Cdf of absolute DOA error for FF and HT

In the channel types considered, some experiments failed completely to locate the DOA of the desired UE and statistical measures such as mean and standard deviation are unsuitable for performance evaluation. Instead, the results are presented in terms of the cumulative distribution functions (cdf) of the absolute DOA estimation errors. Fig. 5 shows the results for RA, TU and BU for 5 and 15 slots and Fig. 6 shows the results for FF and HT. "conv" denotes the conventional method of using the desired UE chip waveform (over the pilot length) at chip rate to despread the received signal, i.e. $N_S = 1$ for (11)-(12) with $N_E = N_p Q_C$. $\hat{\mathbf{R}}_S^{(n)}$ is then used for DOA estimation. "new" denotes the new method with oversampling (for cross-correlation) and whitening, i.e. applying $N_S = 2$ to (11)-(14).

It is clear that in all cases, the use of more slots improves the DOA estimation. This gain is due to the slowly changing channel response for the UE speed of 3km/hr. In Fig. 5 and 6, we note that the performance of the new method is particularly promising in FF and RA. If we set the performance benchmark at 5° absolute error (in view of the beamwidth of ULA), the correctly resolved DOAs in both FF and RA are improved by 12 % for 5

slots, and 10% and 15% respectively for 15 slots. This is unsurprising since FF only has one line-of-sight (LOS) direct path and the one rake finger used is able to rake up all the desired signal power. The oversampling procedure provides more samples (and information) to the cross-correlation of (11) and thus improves the averaging process. This then reduces the level of MAI. Moreover, a good desired signal estimate also enable a good estimate of the MAI for whitening. Therefore, both oversampling and whitening procedures are effective. Likewise, RA has a high probability of LOS in the given scenario. It also has small angular and temporal spreads (see Table 1) that ensure a good estimate of the desired signal based on one rake finger.

For TU, BU and HT, performance differences are less well defined within the range of interest (at around absolute DOA error of 5°), with some cases giving marginal improvements (e.g. BU for 15 slots) while other cases only give similar performances to that of the conventional method. This is because of the significant angular and temporal spread for TU, BU and the high probability of non-LOS (NLOS) (see Table 1 and Fig. 4). In NLOS cases, even though the signal cluster of the local scatterers are still contained in a region around the nominal DOA, it has a smeared out energy spread in both the spatial and temporal domains. For the rake receiver, any signal path that is more than a sampling period from the rake finger of interest produces a virtual interfering UE. Also while HT has small angular and temporal spreads for each cluster and high LOS probability, it has a high probability of having two or more signal clusters. Once again, this smears out the energy and degrades the performance of one rake finger. A further challenge in these environments is the time-varying asymmetrical angular spread about the nominal DOAs. As such, many existing work on DOA in scattering environment cannot be applied as the focus is on symmetrical scattering sources and sources with a known angular distribution.

In view of these observations, we note that some improvements are expected especially for complex scattering environments, e.g. BU, if we take a few more rake fingers into account. Nevertheless, care is required in the choice of rake fingers due to the frequent occurence of more than one strong signal clusters in BU and HT where the DOAs of these clusters are significantly different from the nominal DOAs of the desired UE. Another interesting aspect for further work is the impact of non-ideal power control (residual "near-far problem").

5 Conclusions

The introduction of WCDMA presents many exciting technical challenges. Among them is the application of AAS technology in which DOA estimation plays an integral part. Apart from identifying user direction for downlink beamforming, it can also provide services on wireless communication networks such as public safety and enhanced emergency services, location sensitive billing, vehicle and fleet management, fraud detection, and Intelligent Transport Systems (ITS). In this paper, we proposed a technique that can be used to improve a simple single-user approach to DOA estimation in WCDMA. In the simulation studies, we showed that the new method gives promising gains over the conventional method under favorable environments and experiences no noticeable degradation in challenging environments.

References

[1] Y. J. Guo, et al. Advanced base station technologies for UTRAN. *IEE Electron. Commun. Eng. J.*, pp. 123-132, Jun. 2000.

[2] L. C. Godara. Application of antenna arrays to mobile communications, Part I: Performance improvement, feasibility, and system considerations. *Proc. IEEE*, 85(7), pp. 1029-1060, Jul. 1997.

[3] L. C. Godara. Application of antenna arrays to mobile communications, Part II: Beamforming and direction-of-arrival considerations. *Proc. IEEE*, 85(8), pp. 1193-1245, Aug. 1997.

[4] S. Andersson, B. Hagerman, H. Dam, U. Forssén, J. Karlsson, F. Kronestedt, S. Mazur, and K. J. Molnar. Smart Antennas for GSM and TDMA Systems. *IEEE Personal Communications*, 6(3), pp. 74-86, Jun. 1999.

[5] S. S. Jeng, et al. Experimental studies of spatial signature variation at 900MHz for Smart Antenna Systems. *IEEE Trans. Antennas Propagat.*, 46(7), pp. 953-962, Jul. 1998.

[6] J. J. Caffery, Jr. and G. L. Stüber. Overview of radiolocation in CDMA cellular systems. *IEEE Commun. Mag.*, 36(4), pp. 38-45, Apr. 1998.

[7] T. E. Biedka, et al. Direction finding methods for CDMA systems. *IEEE Proc. 30th Asilomar Conf. Signals, Syst., Comput.*, vol. 1, pp. 637-641, Pacific Grove, CA, Nov. 3-6, 1996.

[8] H. Park and H. Kim. A new direction finding scheme for DS/CDMA systems with array antenna. *Proc. IEEE MILCOM*, vol. 2, pp. 836-840, Los Angeles, CA, Oct. 2000.

[9] C. Brunner, M. Haardt and J. A. Nossek. On space-time RAKE receiver structures for WCDMA. *IEEE Proc. 33th Asilomar Conf. Signals, Syst., Comput.*, vol. 2, pp. 1546-1551, Pacific Grove, CA, 1999.

[10] ETSI. 3GPP. TS 25.211. Universal Mobile Telecommunications System (UMTS); Physical channel and mapping of transport channels onto physical channels (FDD). version 4.1.0, Jun. 2001.

[11] A. J. Barabell. Improving the resolution performance of eigenstructure-based direction-finding algorithms. *Proc. IEEE ICASSP*, Boston, MA, pp. 336-339. 1983.

[12] P. Stoica and K. Sharman. Novel eigenanalysis method for direction estimation. *IEE Proc.*, vol. 137, pt. F, pp.19-26, Feb. 1990.

[13] M. Correia, ed. *Wireless flexible personalized communications - COST 259: European co-operation in mobile radio research.* John Wiley & Sons, 2001.

[14] ETSI. 3GPP. TS 25.213. Universal Mobile Telecommunications System (UMTS); Spreading and modulation (FDD). version 4.1.0, Jun. 2001.

Analysis of Capacity Improvement For HiperLAN Type 2 Systems Using Switched Sector Antennas

P. Hafezi[1], P. Karlsson[1], F. Kristensen[2], C. Bergljung[1] and M. Sandgren[1]

[1]Telia Research AB, Box 94, 201 20 Malmö, E-mail: Pejman.x.Hafezi@telia.se

[2]Department of Electroscience, LTH, Box 118, SE-221 00 Lund, Sweden, E-mail: Fredrik.Kristensen@es.lth.se

Abstract

The advantages of the use of switched sector antenna in a wireless LAN system are investigated by simulating a deployment in a corporate environment based on the HiperLAN2 standard. The system performance measures for 3 Access Points serving 75 Mobile Terminals in an office environment compare the system throughput, percentage of blocked and dropped sessions for an omni directional case and one with a six sector antenna deployed at the Access Point. The emphasis has been on creating a scenario that places considerable demand on the system in order to investigate performance under high load conditions with both real time and non real time applications. The system throughput with switched sector antennas increases 30% while the output power in all links can be reduced more than 6dB compared to omni antennas.

1. Introduction

The HiperLAN Type 2 system (H/2) is a connection-based, high-speed wireless transmission system capable of supporting data rates up to 54 Mbps through the use of Orthogonal Frequency Division Multiplexing (OFDM) as the modulation scheme. H/2 is intended to complement currently available wireless access systems in hotspot areas (such as home, office, hotels, airports, and short range outdoors) allowing high capacity and throughput to the end users [1]. The specification standard for H/2, which has recently been finalised [2], allows for the support of 7 different combinations of constellation and code rates. This allows the system to exploit the variation in the radio channel quality to provide an appropriate rate of data transmission. Table 1 outlines the transmission modes supported by the H/2 system.

The medium access control (MAC) used in the H/2 is based on TDMA/TDD scheme where the Access Point (AP) controls all the transmissions over the air. Thus H/2 supports a mixture of real-time and non real-time services to each mobile terminal (MT). Provisions are also made in the standard for the use of multi-sector antennas by allowing use of up to seven elements at the AP [3]. Figure 1 depicts the frame format for both single and multi-sector antennas.

PHY-Mode	Modulation	Coding Rate	Bit Rate [Mbit/s]
1	BPSK	1/2	6
2	BPSK	3/4	9
3	QPSK	1/2	12
4	QPSK	3/4	18
5	16QAM	9/16	27
6	16QAM	3/4	36
7	64QAM	3/4	54

Table 1- Transmission Modes in HiperLAN2

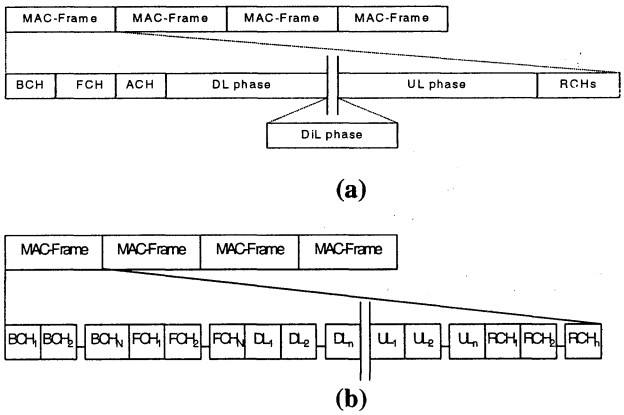


Figure 1 - MAC frame format for a) omni directional AP and b) multi-sectored AP [5]

Theoretically, from the system design point of view, the advantages offered by the use of switched multi-beam antennas could be summarised as

- The improvement in the multipath characteristics of the propagation channel through spatial filtering of the multipath components resulting in reduced RMS delay spread and improved Rician factor [4,6,12]. This reduction in the channel dispersion allows the link adaptation to be used more effectively, allowing support of higher transmission modes
- Improvement in the link budget and increase in the range of service area by taking advantage of the increased gain of the directional antennas employed. This however, is mainly applicable in the uplink due to regulatory eirp constraints.
- Capacity improvement by enhancing the transmission quality by reducing the interference

from the MTs not situated within the service area covered by a specific selected beam

Figure 2 depicts the calculated link throughput versus distance for omni and 6-sector antennas. It can be seen that the advantage of the use of sectored antenna is in it allowing a larger service area to be covered compared to the omni antenna. This is due to the additional azimuth gain offered by the directional antenna pattern and the reduction in the fluctuations in the propagation channel by switching sector. Additionally, considering MTs within a given environment, from Figure 2 it can be concluded that a large percentage of the MTs within the service area will benefit from an increase in the throughput. Only MTs close to the AP will experience a reduced throughput caused by the broadcast overhead, see figure 1b. Improved system throughput while maintaining a low packet delay in a radio network utilizing adaptive array antennas has been shown in e.g. [8].

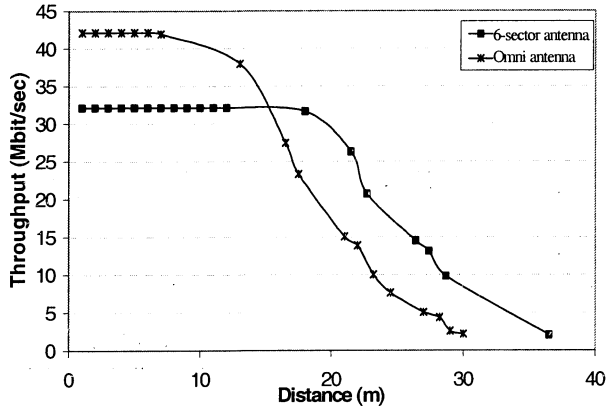


Figure 2 – Link throughput versus distance for omni and 6-sector antennas

In this paper the results from a H/2 system simulation analysis are presented to compare the system performance obtained using omni and switched 6-sector antennas at the APs in heavy traffic conditions.

In section 2 the simulation environment is presented together with the positions of the AP and MT terminals. Section 3 describes the parameter considerations for the system under investigation. A selection of performance comparisons is presented in section 4 and the conclusions are summarised in section 5.

2. Simulation Environment

The environment under investigation was the office of Telia Research AB situated on the third floor of the Scandinavian Centre in Malmö, Sweden. Extensive propagation measurements conducted in this office allows a realistic propagation model to be used in the simulations. The office can typically accommodate up to 75 users and the simulations presented in the subsequent sections consider 3 APs serving the 75 MTs. Figure 3 depicts the floor plan of the office area and positions of the APs and MTs.



Figure 3 – Layout of the simulation environment and the position of the APs (*) and MTS (.)

The Path Loss model used in the simulation is given by

$$L = \left(\frac{4.\pi.d}{\lambda}\right)^2 + L_E + L_D \qquad (1)$$

Where L_E represent the loss due to the attenuation through walls and floors, and L_D the additional loss due to slow fading effects. The choice of these parameters was influenced by the results of propagation measurements conducted in the environment [6] and those reported in [5]. The channel model used for the performance comparisons of omni and sectored antenna scenarios assume a non-LOS channel corresponding to the channel A in the standard specification [1] (with max. RMS delay spread of 50ns).

3. System Simulation Parameters

Some of the main considerations applied in the simulations are briefly described below.

Interference considerations: under high user density operation conditions (such as the traffic model considered in this study), the H/2 system is capacity and interference limited. Due to the unsynchronised nature of the TDD MAC frame, the interference can arise from both APs and MTs. This would imply that uplink and downlink transmission symmetry has to be considered which in turn would depend on the choice of the service class and will be discussed below. However, in general, the AP is given a higher weight due to the larger proportion of allocated transmission time.

In the simulations one MT or AP from the adjacent and coherent channels are randomly chosen as interferes for each AP/MT. It was assumed that similar systems would be running on the two floors above and below the office. The APs on the same floor would not use coherent frequency bands. The random frequency selection is made from one of the 19 available channels

(the max. available channels for indoor use in the 5 GHz band) and is updated every 20ms.

Service class considerations: three types of service classes, based on estimations in ITU-R for corporate environments beyond year 2005 [9], representing both real and non real time applications are considered in the simulations. Each service is modelled with different arrival rates and packet durations and chosen to be the most common and put different demands on the system:

- High MultiMedia (HMM), e.g., web browsing and file transfer and video surveillance.
- High Interactive Multimedia (HIMM), e.g., video conferencing, web hosting and client server. An Interactive service requires high bite rates in both uplink and downlink.
- Very High Multimedia (VHMM), e.g., streaming video and spooling video.

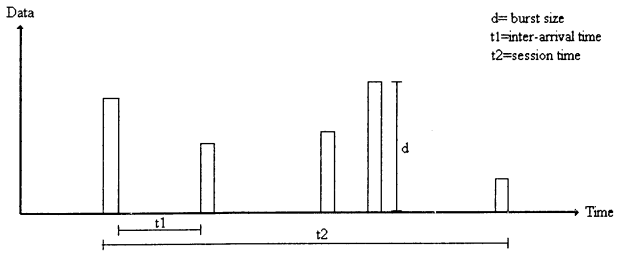


Figure 4 – Traffic model

The traffic model used in the simulations is depicted in Figure 4. This model is commonly used and describes the traffic in terms of bursts of data arriving with different intervals for the duration of a session. These parameters are calculated based on the following assumptions described in [10]:

- The arrival intensity of a session is modelled as a Poisson distribution with λ as stated in Table 2
- The duration of each session is determined to be exponential with a mean according to Table 2
- Burst sizes are approximated as lognormal distributed for all service classes, with a mean equal to:

Mean_Burst_Size=(Net_User_Bit_Rate)(Activity_Factor)* (Mean_Time_Between_Bursts)*

Table 2 summarises the parameters described above for each of the three service classes considered.

Each simulation run consists of a start-up phase and a simulation phase. In the start-up phase a random traffic situation is generated. In the simulation phase each MT generates new sessions with the intensity determined from Table 2, with the exception that an MT can only have one session of each service class running at the same time.

Parameter	HMM	HIMM	VHMM
λ	0.5	0.3	0.1
Mean session duration (s)	600	1200	1800
Net user bit rate (Kbps) DL / UL	2000 / 128	2500 / 2500	7000 / 128
Activity factor DL / UL	0.3 / 0.02	0.4 / 0.3	0.99 / ~0
% Real Time Session	0	50	50
Start up probability (%)	8.33	10	5

Table 2 –Traffic parameters for the three service types considered

To counteract interference H/2 system takes advantage of three techniques: Dynamic Frequency Selection (DFS), Power Control (PC) and Link Adaptation (LA) [2]. DFS allows the least interfered frequency channel out of the available channels to be selected by the AP. Implementation of DFS also eliminates the manual frequency planning when designing the system. The PC function in H/2 is centralized to the AP and its function is for the downlink (DL) transmission power to be adjusted to a level sufficient to achieve reliable communication between the AP and the most distant MT and for the uplink (UL) to meet the power target at the AP from each MT. LA allows the highest possible transmission rate to be selected from the available transmission modes (listed in Table 1) hence reducing the total transmission time.

The H/2 standard thus can handle intra system interference from other APs/MTs. Furthermore, since the 5GHz band designated for HiperLANs, is a license exempt band with many possible users who also share the spectrum with other systems (such as satellite and radar systems), the regulatory framework requires implementation of both DFS and PC for inter system interference reduction.

The simulation tool has implemented the three techniques and Table 3 summarizes some of the parameters used in the algorithms. To take into account external interferences, a test radar signal based on [7] was also taken into account to verify the DFS functionality.

4. Results

The comparisons between the performance measures for the omni and sectored antenna deployments at the APs are presented in terms of percentage of blocked sessions, sessions that are dropped due to low Carrier to Interference+Noise power ratio, C/(I+N), throughput per AP and the corresponding PER achieved.

Figure 5 shows the cumulative distribution function of recorded PER for the all the service types considered in

the simulation for both Omni and 6-sector AP antennas.

Simulation Parameter	Value
Number of available frequency channels	19
Antennas	Omni / 6-Sector
Max. AP/MT power	23 / 30 dBm
Standard deviation of slow log-normal fading	5.4 dB
Noise power	-93 dBm
Adjacent channel suppression	25 dB
DFS duration	10 s
DFS measure interval	1 s
Targets on PER	Real time sessions: 1 % Non-real time sessions: 5 %
LA update interval	200 ms

Table 3 – System simulation parameters.

Considering that the traffic model used in the simulations places a very high demand on the system resources, it can be seen that the use of sector antennas has resulted in an improved PER performance compared to the omni antenna hence reducing the need for the retransmission of the erroneously received data packets. This in turn would decrease the delay and so improves the system throughput performance (see also [11]) and supports real-time services.

Another point to consider is that the DFS implemented in the simulations is initiated every second during a 10 second period to calculate average interference values that will decide the frequency selection. During this time APs and MTs with low C/I were not dropped resulting in many errors being produced before a session was disconnected hence contributing to high PER. It is therefore possible to improve the PER by reducing the DFS initiation time. However, this is a time complex procedure and would require much computing resources.

From Figure 5 it can be seen that the use of sectored antennas at the APs have resulted in an improved PER for 90% of the simulated traffic. The high PER values obtained for sectored AP antennas in 10 % of the cases are due to the choice of the sector with highest carrier power C resulting in momentary reduction of the C/I ratio.

The PER performance versus distance for the omni and 6-sector AP antennas are compared in Figure 6. Both real and non-real time sessions are depicted. It can be seen that the 1% PER target for the real time sessions could not be achieved with the omni AP antenna while using sector antennas the achieved PER was below the target for MTs less than 6m away from the AP. For the non-real time sessions the maximum distance supported by the omni AP antenna to achieve a 5% PER target was 10m. Using sectored AP antenna this distance was increased to 16m.

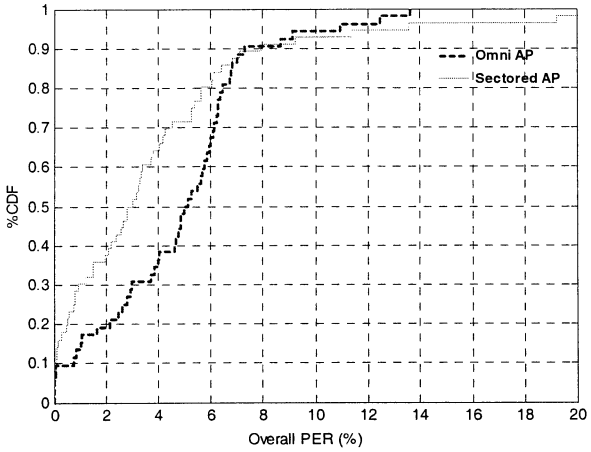


Figure 5 – Comparisons of the CDF of PER for overall traffic for omni and 6-sector AP antennas.

It has to be noted that the traffic conditions considered in the simulations place demands on the H/2 system similar to those achieved using 100BaseT connections and services will mainly experience a delay at longer ranges where the PER is higher than the target.

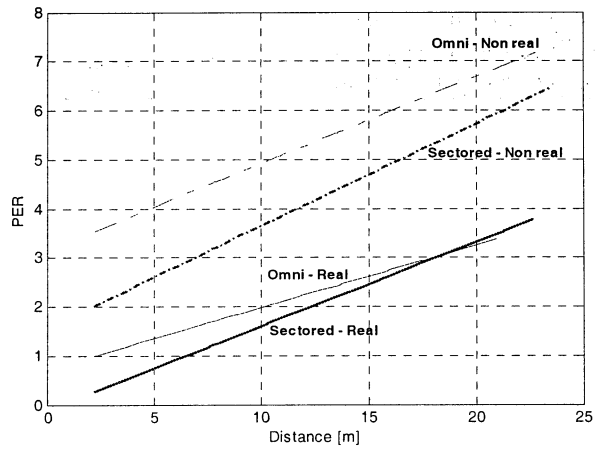


Figure 6 – PER in services vs distance for omni and 6-sector AP antennas.

Table 4 compares the system performance measures for the omni and sectored AP antenna scenarios. The throughput per AP indicates the average throughput for both uplink and downlink, in Mbps, through an AP during the simulation. Only received data that satisfied the target PER were considered. It can be seen that with the sectored AP antennas a throughput of 11.5 Mbps was achieved. This was an increase of about 2.8 Mbps or 32% compared to the omni AP antenna.

Comparisons are also made for the percentage of the dropped and blocked services. Dropped services are

those that have been disconnected from the system due to low C/I, too large packet buffers and if the service has not been allocated radio resources for a long time (11 seconds in the simulations). The results indicate that the percentage of the dropped HIMM services with sectored antenna at the AP was almost half those recorded with Omni antenna at the AP. Reduction in the number of dropped sessions were also observed for the demanding VHMM services.

Services that have been denied initial medium access due to the lack of radio resources are called blocked sessions. It can be seen that with sectored AP antennas, less than 1% of the services were blocked while this was nearly 5% for the Omni AP antennas. Switched sector antennas thus improve the Quality of Service according to the ITU-T QoS definition [E.800]: "The collective effect of service performance which determines the degree of satisfaction of a user of the service".

Using sectored antennas, a decrease of about 6dB was recorded in the average AP transmission power (DL) compared to Omni AP antenna transmission power. This was also observed in the uplink where the reduction in the MT transmission power was about 7dB. Thus, we want to emphasize that the higher system throughput was achieved with relaxed power consumption requirements of the MTs. The reduced output power also alleviates the inter system interference and thus improves co-existence and spectrum sharing in the 5 GHz band.

	Omni AP	**6-Sector AP**
Throughput / AP (Mbit/s)	8.7	11.5
% Services Blocked	4.5	0.7
% Services Dropped (1/2/3)	3.3/18.5/27.3	4.4/9.3/16.7
Average Power DL (dBm)	24.8	18.2
Average Power UL (dBm)	20.7	13.6

Table 4 – Comparison of system performance results for omni and 6-sector AP antennas

5. Conclusions

A H/2 system simulation tool was presented. With this tool the advantages of using a switched sector antenna in a wireless LAN system was investigated under heavily demanding traffic conditions similar to those expected in corporate environments 2005 according to ITU-R. It was shown that with correct choice of antenna sectors, considerable performance improvement over conventional omni antenna systems in terms of more than 30% increase in throughput per AP could be achieved. Reduction in the percentage of dropped and blocked services (both real-time and non real-time) and increase in the range can be achieved. The observed reduction in the average transmit power of 6 dB for UL and 7 dB for DL offer advantages of reducing the total interference in combination with increased throughput in the system as well as relaxing the power consumption requirements of the MTs.

6. References

[1] J. Khun-Jush, G. Malmgren, P. Schramm, J. Torsner, "HIPERLAN Type 2 for Broadband Wireless Communication", Ericsson Review, No.2, 2000.

[2] ETSI TS 101 475 V1.1.1, 'Broadband Radio Access Networks (BRAN); HIPERLAN Type 2 - Physical layer', Technical Specification 04/2000.

[3] ETSI TS 101 761-1 V1.1.1, 'Broadband Radio Access Networks (BRAN);HIPERLAN Type 2; Data Link Control (DLC) Layer Part 1: Basic Data Transport Functions', Technical Specification 04/2000.

[4] P. Karlsson, H. Börjeson, C. Bergljung, J. Pamp, 'Analysis and Results of Wideband Spatial Radio Channel Measurements at 5GHz – Impact on Multiple Antenna Systems', Millennium Conference on Antennas and Propagation, Davos – Switzerland, April 2000.

[5] COST Action 231, "Digital mobile radio towards future generation systems", EUR 18957, 1999

[6] P. Hafezi, M.A. Beach, A.R Nix, Y. Sun, 'Exploiting the Spatial Domain for High Bit Rate Indoor Communications', IEE Colloquium on Antennas and Propagation for Future Mobile Communications (Ref. No. 1998/219), pp. 12/1-12/6, London – UK, 1998

[7] J. Khun-Jush "Sample radar test signal for the testing of DFS in HIPERLAN Type 2", BRAN document no:BRAN23dzz, 2001-04-05

[8] R. Stridh, B. Ottersten, "Packet Data Throughput for Wireless Systems with Smart Antennas", IEEE VTC Proceedings, Amsterdam, September 1999.

[9] "5GHz RLAN Spectrum Requirement", ITU-R Document 8A-9B/86, October 2001.

[10] P. Wikström, "Real Time Video Traffic Modelling and Simulations in Third Generation Mobile Networks", Internal Technical report at Telia Research AB, 2000.

[11] A. Dahlen, P. Ernström, "TCP over UMTS", In RVK02 Proceedings, Stockholm, June 2002.

[12] A Doufexi, S. Armour, A. Nix, P. Karlsson and D. Bull, "Performance of HIPERLAN/2 using sectorised antennas", IEE Electronics Letters, Vol. 37, No.4, February 2001.

Improving system performance in a WCDMA network using indoor pico base stations

Henrik Andersson
Henrik.Andersson@kevab.com

Robert S. Karlsson
Robert.Karlsson@kevab.com

Per Larsson
Per.Larsson@kevab.com

Patrik Wikström
Patrik.Wikstrom@kevab.com

KEVAB, Kronborgsgränd 5, SE-164 46 Kista, Sweden
Phone: +46-8- 406 97 00, Fax: +46-8-406 97 30

Abstract

The bandwidth, quality of service, and indoor capacity demand is increasing in existing and future cellular networks. In emerging 3G technologies, like the WCDMA FDD system, new types of services are enabled by the higher data rates supported. It is expected that data services will generate more traffic than speech services.

We compare a WCDMA network with only macro sites to a system where pico base stations have been added in an indoor traffic hotspot.

Hotspot areas degrade the performance significantly for the macro cell network. We analyze how performance is improved by using indoor pico base stations. Indoor hotspot users have a negative effect on overall system performance of the whole network. We show that substantial performance gains can be made in the outdoor system as well as in hotspot traffic areas, when introducing pico base stations.

1 Introduction

In the first and second generation mobile networks speech has been the dominating application. For the emerging 3G technologies, like the WCDMA FDD system [1], new types of services are enabled by the higher data rates supported. It is expected that data services will generate more traffic than speech services.

Typically, in the initial roll-out of a WCDMA FDD system macro base stations will be used both for indoor and outdoor coverage. The network will be planned to support relatively high data rates indoors. After the initial roll-out there will be areas with poor coverage. In such situations pico base stations or other indoor solutions will be important. However, even in areas with sufficient coverage from the macro base network, pico base stations will play an important role. This is even more true in a WCDMA network than in a second generation network due to the strong correlation between capacity and coverage. The variety of services and their different demand of network resources will most likely also make the number of high traffic density areas larger. This is due to that a very low number of users can take up all resources in a cell (in second generation networks the correlation between the number of users and amount of traffic is stronger).

One fundamental property of a WCDMA FDD radio network is that the downlink output power is a common resource, at the base station, shared by all served users. Therefore, the more power needed for one user, the less power is available for the others. Due to, e.g., interference, building penetration and inner wall losses some indoor users will require a relatively large amount of output power when served by an outdoor macro base. In addition, it is in indoor environments one can expect the highest degree of high-data-rate-service usage. Altogether, some indoor users will have a negative effect on overall system performance. In this paper we analyze how performance can be improved by using indoor pico base stations.

2 Models

We investigate a macro cell network with omni base stations on a hexagonal grid (in the simulations wrap-around is used to avoid border effects). There is one indoor hotspot located as indicated in figure 1.

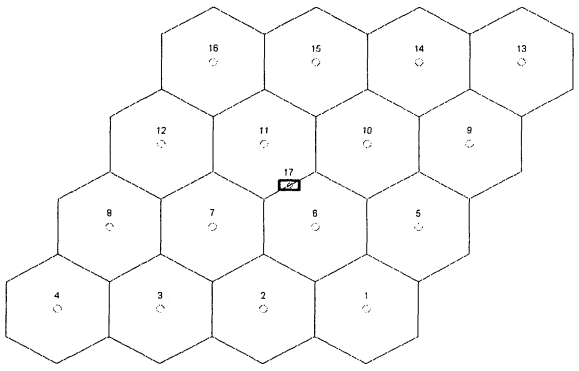


Figure 1. A macro cell network with base stations number from one to 16 and one hotspot located on the border between macro

bases 6 and 11. Base station 17 is a pico base station inside the building.

In our case the hotspot is located on the ground floor (50 x 100 m) of an office building. The floor consists of two corridors with several rooms (see pilot strength predictions on the floor in figure 4). The traffic is uniformly distributed over the macro cell network except in the hotspot where there is additional traffic. The analysis will be carried out with or without a pico base in the middle of the hotspot. The maximum output power levels is 20 W for the macro bases and 300 mW for the pico base. The performance is analyzed through radio network simulations.

2.1 Propagation models

Four propagation models are used to model the different paths exemplified in figure 2.

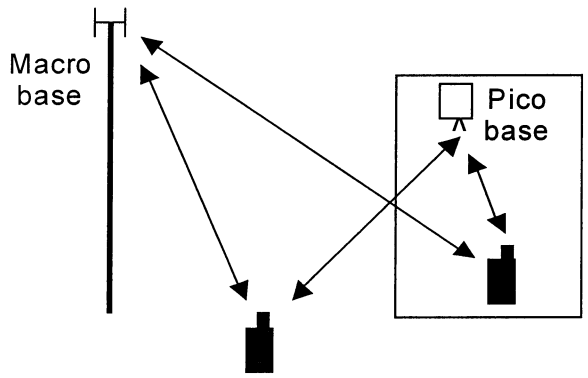


Figure 2. Different propagation paths modeled by different propagation models. Macro base to indoor mobile, macro base to outdoor mobile, indoor pico base to indoor mobile, and indoor pico base to outdoor mobile.

2.1.1 Macro to outdoor mobiles

This model is taken from [4] and it is applicable in urban and suburban areas with no high rise core. The buildings are of nearly uniform height. With an antenna height of 15 m the path loss, in dB, is

$$L = 128.1 + 37.6 Log(d) \quad (1)$$

where:
d = distance [km]

A log-normal shadow fading standard deviation of 10 dB is also considered.

2.1.2 Macro to indoor mobiles

It is here assumed that there are reflections from surrounding buildings so that the power levels are the same on each side of the building, see the figure 3 below (P_{ref}). The path loss from the macro base station to the reference points is calculated as above. The power level at a point inside the building is obtained by summing up the components from each side of the building including building penetration loss and indoor wall loss [2]. The building penetration loss is set to 12 dB and the loss due to inner walls is modeled by a loss per meter factor (0.8 dB/m).

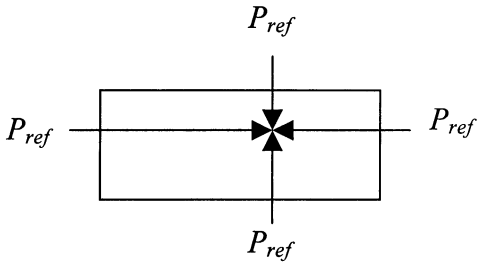


Figure 3. Model for calculation of path loss from surrounding macro cells.

The outdoor model gives the propagation loss at the ground floor. Higher up in the building the loss is less when using antennas mounted over the average roof top level.

2.1.3 Pico to indoor mobiles

The indoor propagation model is the multi-wall model, see e.g. [2], where the distance dependent part is the free space model and inner walls and floors add up in terms of additional loss. In dB we have

$$L = L_{FS} + L_c + \sum_{i=1}^{I} k_{wi} L_{wi} + k_f^{\left[\frac{k_f+2}{k_f+1} - b\right]} L_f \quad (2)$$

where:
L_{FS} = free space loss,
L_c = constant loss,
k_{wi} = number of penetrated walls of type i
k_f = number of penetrated floors,
L_{wi} = loss of wall type i,
L_f = loss between adjacent floors,
b = empirical parameter,
I = number of wall types.

According to [2] L_c shall be close to zero (in the simulations L_c is set to zero). The following values are given in [2].

L_{w1}	L_{w2}	L_f	B
3.4	6.9	18.3	0.46

Table 1. Parameter values, in dB, for the multi-wall model

The free space loss, L_{FS} , is given by the following equation,

$$L_{FS} = 32.4 + 20\log(d) + 20\log(f) \quad (3)$$

where:
d = distance [km],

f = frequency, [MHz].

In the simulations a wall loss of 3.4 dB, a frequency of 2000 MHz and a log-normal shadow fading standard deviation of 4 dB is used. Figure 4 shows predictions of the received pilot strength on a floor with a pico base station located in the middle (30 mW pilot).

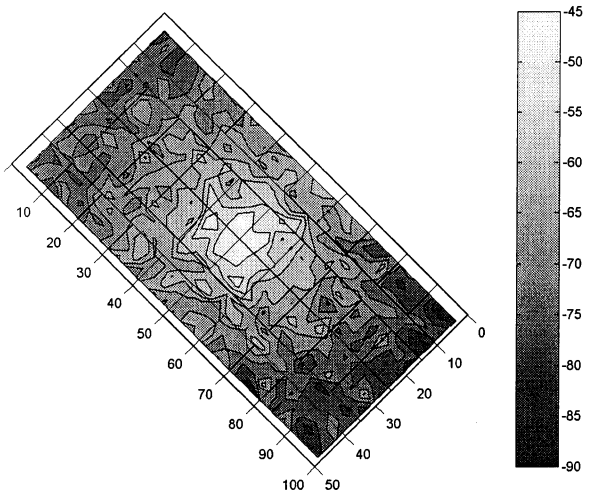


Figure 4. Received pilot strength predictions, in dBm, with a 300 mW pico base station in the middle of an office floor (30 mW pilot).

2.1.3 Pico to outdoor mobiles

This model comes also from [4] and gives the path loss, in dB

$$L = 148 + 40Log(d) \qquad (4)$$

A lognormal shadow fading standard deviation of 10 dB and a building penetration loss of 12 dB with a standard deviation of 8 dB is also added.

3 Performance Evaluation and Traffic Load

System performance is determined by examining many independent static traffic situations (snapshots). For each snapshot the mobiles positions are chosen randomly throughout the service area (more dense within the building). For each mobile station the serving base station is determined by looking at the received pilot signal from each base station and selecting the strongest one. If the pilot strength of the second strongest base station is within 3 dB of the strongest, a second link is added (soft handover). Table 2 shows some of the most important parameter used in the analysis.

	Pico BS	Macro BS
Max base station TX power	300 mW	20 W
Pilot power	30 mW	2 W
Power for the rest of the common channels	30 mW	2 W
Orthogonality factor (α)	0.94	0.60
Antenna gain	2 dBi	11 dBi
Cable losses	0 dB	3 dB
Body loss for speech users	3 dB	3 dB
Noise Figure (mobile)	7 dB	7dB

Table 2. Parameters used in the radio network simulations.

Once the serving base stations are selected an iteration is started to find the individual radio link power settings. We use the Distributed Constrained Power Control (DCPC) algorithm [5] for updating the powers. The transmission power in the n^{th} iteration for the i^{th} user is

$$P_i^{(n)} = \min\left[P_{\max}; \gamma_t \frac{P_i^{(n-1)}}{\gamma_i^{(n-1)}}\right] \qquad (5)$$

where $\gamma_i^{(n)}$ is the resulting SIR for the i^{th} user after the n^{th} iteration and γ_t is the SIR target. The DCPC works by balancing the SIRs, towards the target for all users, using the minimum necessary transmission powers. Iterations are done until convergence (when the maximum relative power change between iterations is less than 10^{-2}). Downlink received SIR for link i can be expressed as below.

$$\gamma_i^{(n)} = \frac{P_i^{(n)} / L_i}{(1-\alpha)\cdot Iown + Iother + \eta} \qquad (6)$$

where L_i is the path loss of link i, α is the orthogonality factor, Iown and Iother are the received interference from own and other cells respectively and η is the receiver noise.

The performance measure in these studies is the amount of satisfied users. A user is said to be satisfied when the quality is sufficient for the particular service. If E_b/N_0 is above E_b/N_0 target – 0.5 dB the user is here satisfied. Received E_b/N_0 in the downlink can be expressed as:

$$(E_b / N_0)_{DL} = \frac{W}{R} \cdot \gamma_i^{(n)} \qquad (7)$$

where W is the chip rate (3.84 Mcps) and R is the data rate.

For the numerical evaluation we have assumed a data rate of 12.2 kbps and a target E_b/N_0 of 8.7 dB modeling a speech service. The average traffic was 36 users per cell

in the macro network and an additional 36 users in the hotspot, voice activity was not included. The high number of users in the hotspot can be seen as equivalent to having a few hotspot high data rate users.

4 Results

A commonly used performance criteria is that there should be at least 95 % satisfied users. In figure 5 we have plotted the percentage of satisfied users for each base station in the system.

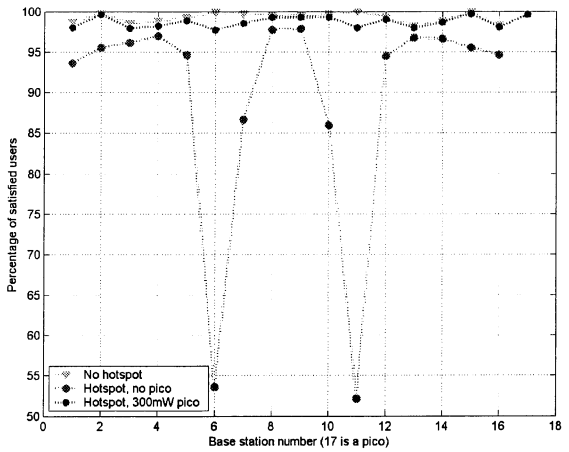


Figure 5 Percentage of satisfied users in each base station. Number 1-16 are macro layer stations and base station 17 is the pico cell placed in the hotspot.

The pico base station in the simulations use the same carrier as the macro layer.

We compare three cases:

1. *"No hotspot"*: Macro layer network with uniform traffic distributed on 16 macro base stations. A well balanced system with high quality.
2. *"Hotspot, no pico"*: Macro layer network and one hotspot with additional traffic. This results in an unacceptable quality degradation in the whole macro network. Not only the closest sites, even if these are most affected. The closest (lowest propagation loss) macro bases will try to serve the hotspot users. As a result the total downlink output power of those macro bases increases and the interference level in the entire network increases.
3. *"Hotspot, 300mW pico"*: Macro layer network and one hotspot with additional traffic. A pico base station is installed in the hotspot. The same carrier as for the macro layer is used. The pico base station relaxes the macro network by taking away some of the most demanding users. The result is that the quality is improved well above 95%.

Table 3 below lists the average number of satisfied users in the whole system, and located in the hotspot building, respectively.

	Average satisfied mobiles in the total system [%]	Average satisfied mobiles located in the hotspot [%]
No hotspot	98.8	-
Hotspot, no pico	85.1	9.2
Hotspot, 300 mW pico	98.0	97.6

Table 3. System and hotspot performance figures.

Case three above has been simulated with different hotspot loads. The hotspot load values are given as the load compared to the average macro cell load, i.e., load equal to one means that the amount of traffic in the hotspot is the same as the average amount of traffic in one macro cells (one carrier). Figure 6 shows the CDF of the output power from the pico base station for different loads. The macro network is loaded as in case 1 above. As can be seen the maximum pico power of 300 mW is not reach even for as high load as twice the macro load. In all these cases there are resources left in the pico base station and schemes to connect more mobiles to the pico can be used to save resources in macro cells (e.g., hysteresis when choosing serving cells).

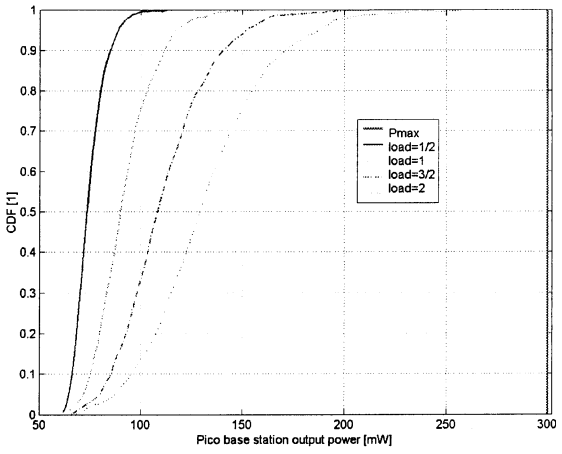


Figure 6. CDF over the pico base station output power for different hotspot loads.

5 Discussion

For systems with many hotspots it may be a good idea to dedicate one carrier for the pico layer to reduce interference from the macro layer.

In a real network (with reasonable good radio resource management algorithms) the system would, in case 2, obviously have blocked and dropped mobiles before it got as poor as shown in figure 5. However even

if we accept high blocking of traffic in the closest cells there will occur problems for mobiles moving into those highly loaded cells. If no proper handover can be made and such users are not dropped interference in the system will increase and available resources in other cells might run out. The problem spreads throughout the system and network quality degrades as in case 2. A lot of these possible problems can be solved by adding pico base stations in traffic hotspots to relax the macro layer (as in the case 3 above).

6 Conclusions

We have shown that hotspot areas can degrade the performance significantly for a macro cell network. Not necessarily only for the closest base stations but for the entire network. Introducing pico cells in hotspots will improve performance, for both indoor and outdoor users, even if the pico cells use the same carrier as the macro cell layer. The pico stations must not necessarily take all traffic in the hotspots, only enough to relax the macro layer.

7 References

[1] 3GPP Technical Specifications.
[2] COST 231, Final Report.
[3] 3GPP Technical Specifications, RF Scenarios, TR 25.942.
[4] UMTS 30.03, Selection procedures for the choice of radio transmission technologies of the UMTS , TR 101 112.
[5] Grandhi S. A., Zander J., Yates R., "Constrained Power Control", Wireless Personal Communications, vol. 1, no.4, pp. 257-70, 1995.

On Signal Strength and Multipath Richness in Multi-Input Multi-Output Systems *

Thomas Svantesson[1]†, Jon Wallace[2]

[1]Department of Signals and Systems, Chalmers University of Technology. E-mail: tomaso@s2.chalmers.se

[2]Dept. Electrical & Computer Engineering, Brigham Young University (BYU). E-mail: wall@ieee.org

Abstract

A common practice when analyzing systems using multiple antennas at both ends is to focus on the multipath properties and normalize the SNR out of the channel matrix. This paper studies the relationship between the signal strength and multipath richness using indoor measurements taken at the Brigham Young University campus. It is found that the SNR and the multipath richness can be strongly correlated. Hence, some caution is needed when normalizing out the SNR in system performance studies. Furthermore, for an unnormalized channel, the channel capacity usually rose when moving from NLOS into LOS since the loss in multipath was more than compensated for by an increase in SNR. A theoretical study also revealed that for moderate sized systems, the required SNR increase for a LOS channel to yield the same channel capacity as a NLOS channel is not very large.

1 Introduction

The wireless communication industry has experienced a tremendous growth during the last decades. However, the available bandwidth has not grown at the same pace, resulting in an increased cost for bandwidth. Hence, techniques that try to increase the bandwidth efficiency have become even more interesting than before. A new promising way of utilizing the spatial dimension of the channel is to employ multiple antennas at both the transmitter and receiver. These Multi-Input Multi-Output (MIMO) systems have been shown [1, 5, 6] to support much higher data rates than traditional single antenna systems while using the same amount of bandwidth. This increase is achieved by exploiting (instead of suffering from) the multiple paths that most signals will take between the transmitter and the receiver.

*This work was supported in part by the Swedish Foundation for Strategic Research, under the Personal Computing and Communications Program and the Hans Werthén Foundation.

†The author is currently visiting the Wireless Communications Group at Brigham Young University (BYU), USA.

Due to large gains in data rate and high bandwidth efficiency, MIMO systems have received much attention during the last years. Most of these papers have studied an upper bound on the achievable data rate, i.e. the channel capacity [4]. Since the large capacity gain by using MIMO systems is mainly due to the multipath richness, most authors have focused on this property by normalizing the Signal to Noise Ratio (SNR) out of the channel matrix. However, Non Line Of Sight (NLOS) scenarios with very rich multipath that would result in a high capacity, often experience low SNR. On the other hand, scenarios with Line Of Sight (LOS) usually have high SNR but low multipath richness. Thus, the multipath richness and the SNR are usually not independent and some care should be practiced when normalizing out the SNR in system performance studies.

The relationship between SNR and multipath richness will be investigated in this paper using indoor measurements taken at the BYU campus. The same measurements will also be used to study the channel capacity when taking into account both the SNR and the multipath richness. Finally, a theoretical investigation is conducted that tries to answer how much higher SNR a LOS channel requires to yield the same capacity as a rich NLOS channel.

2 MIMO Channel

Consider a communication system with N_T antennas that transmits independent data streams which are received by N_R antennas. For a narrowband system, the N_R received baseband signals $\mathbf{x}$ are related to the N_T transmitted signals $\mathbf{s}$ as

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{1}$$

where the $N_R \times 1$ vector $\mathbf{n}$ denotes additive noise and $\mathbf{H}$ is an $N_R \times N_T$ matrix that describes the channel. Here, the element $\mathbf{H}_{ij}$ represents the complex path gain from transmitter j to receiver i.

The channel capacity of a narrowband MIMO channel, assuming uniform power allocation and no chan-

nel state information at the transmitter, can be written as [4]

$$C(\xi) = \log\det\left[\mathbf{I} + \frac{\xi}{N_T}\mathbf{H}\mathbf{H}^H\right], \qquad (2)$$

where ξ denotes the SNR at each receive antenna. Most MIMO studies assume some kind of normalization of the $\mathbf{H}$ matrix, in order to focus on the multipath richness. Here, the channel capacity will be calculated using both an unnormalized $\mathbf{H}$ and a normalized channel matrix $||\mathbf{H}||_F = \sqrt{\mathrm{Tr}(\mathbf{H}\mathbf{H}^H)} = \sqrt{N_R N_T}$. These capacities will be referred to as normalized and unnormalized capacity.

The notion of multipath richness is less formal than capacity and there are several potential measures that could be used. Here, the concept of Effective Degrees of Freedom (EDOF) [3] will be used. This measure is based on the fact that for an $N \times N$ channel with rich multipath (fully decorrelated channel), a capacity increase of N bits is obtained when doubling the transmitted power. A correlated channel, i.e. a channel with less multipath, will exhibit a smaller capacity increase. Hence, a convenient measure of the multipath richness is the slope of the capacity curve defined as

$$\mathrm{EDOF} = \frac{\partial}{\partial\delta}C(2^\delta\xi)\Big|_{\delta=0} = \sum_{k=1}^{N}\frac{1}{1+\frac{N}{\sigma_k^2\xi}}, \qquad (3)$$

where σ_k denotes the singular values of the normalized channel matrix. Note that the EDOF is a real number in $[0, N]$. A LOS channel with one dominant propagation path will yield an EDOF close to one while a rich NLOS channel will be close to N. Unfortunately, the EDOF measure depends on the SNR since the number of independent transmission channels that rise above the noise floor depends on the SNR.

3 Measurement Setup

A narrowband custom made MIMO communications system designed and built at Brigham Young University (BYU) in Utah was used to collect measurements. The system was equipped with ten monopoles forming a uniform circular array at each end. However, since the elements were mounted over a ground plane the monopoles behave as dipoles and essentially have the same radiation patterns as dipoles. Furthermore, the elements were positioned in a circle with radius 0.86 wavelengths that approximately gives an element separation of a half wavelength. The operating frequency was 2.43GHz and the MIMO channel was sampled every 80ms. For a detailed description of the measurement equipment, see [5].

Measurements were taken within the Clyde building at the BYU campus, which is a building containing

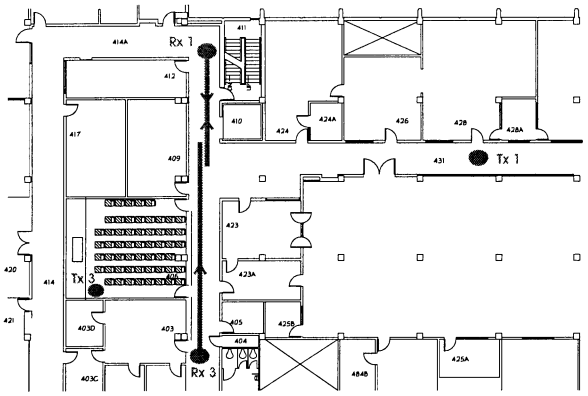


Figure 1: Locations within the Clyde building where the first and third measurement sets were taken. The direction in which the measurements were taken are indicated by arrows and the starting point of each set by a circle.

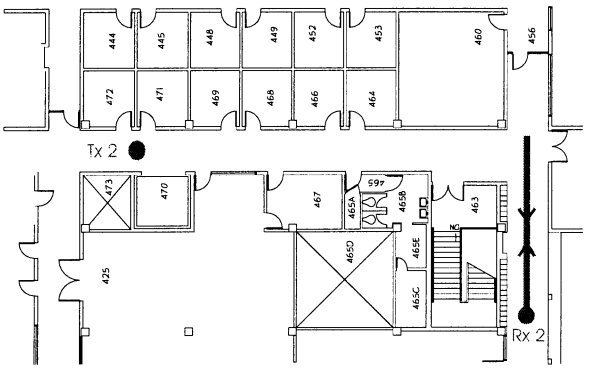


Figure 2: Location within the Clyde building where the second measurement set was taken.

classrooms. Since the aim of the data collection was to address the joint properties of signal strength and multipath richness, several measurements were taken where the receiver started in NLOS, passed a LOS situation, and continued into NLOS again. This was achieved by putting the transmitter in a corridor and letting the receiver move along another corridor that intersects the first. At the intersection, there were LOS conditions while NLOS conditions reappeared as the receiver moved away from the intersection.

Several measurements of this type of NLOS/LOS scenario were taken and in Figure 1, the layout of the first measurement set is shown. The layout of a similar experiment, the second measurement set, is shown in Figure 2. A slightly different experiment was carried out in the third measurement, see Figure 1. In this case, the transmitter was placed in a classroom and the receiver moved along an adjacent corridor. There were two doorways where the con-

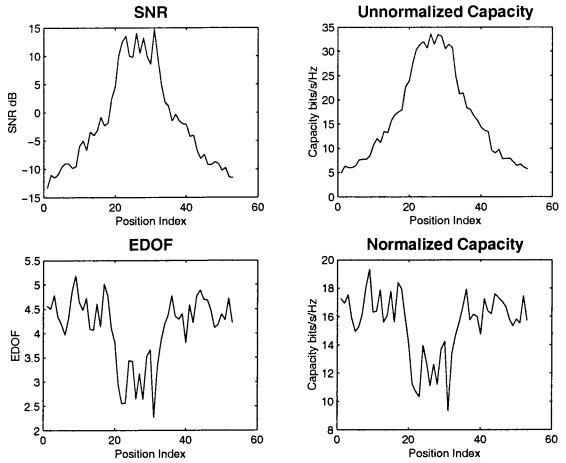


Figure 3: Capacity, SNR, and EDOF of the first measurement set.

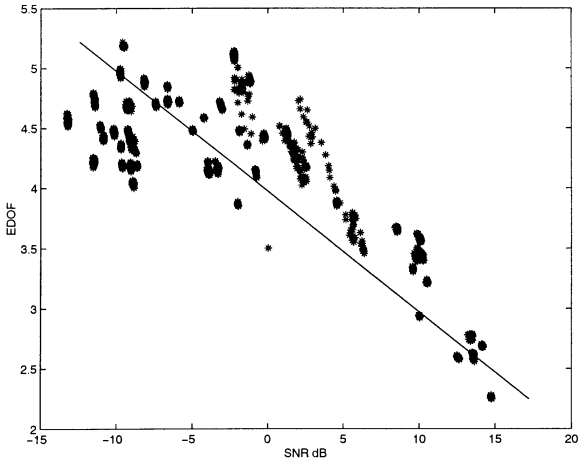


Figure 4: Scatter plot of EDOF and SNR for the first measurement set.

ditions became LOS, but most of the measurement path was NLOS. Many other measurements were also taken, but with similar results as these three sets.

4 Measurement Results

In this section the channel capacity, SNR, and EDOF will be shown for the measured scenarios. The normalized channel capacity and the EDOF will both be calculated for an SNR of 10dB. For the unnormalized channel capacity, the mean SNR over the entire measurement set will be set to 10dB to yield similar capacities as the normalized capacity where each matrix is normalized.

In Figure 3, the SNR, EDOF, and capacity are shown for the first measurement set. The channel was measured at 53 positions (separated by a foot each) along the corridor, see Figure 1. At each step, about 5s of data were recorded and the average SNR, capacity, and EDOF were calculated. As expected, a sharp increase in SNR is observed as the receiver moves into LOS. At the same time, the EDOF drops significantly and a drop in the normalized capacity is observed. The unnormalized capacity, on the other hand, experiences an increase since the drop in EDOF is compensated for and outweighed by an increased SNR. This simple measurement illustrates the importance of channel normalization. The fact that the SNR increase in a LOS environment may compensate for the loss of multipath has previously been reported in [2, 5].

A scatter plot of the SNR and the EDOF is shown in Figure 4. Each measurement during the 5s measurement interval is marked with a star and a least-squares fitted line is also shown. A strong dependency between SNR and EDOF is visible, as might have been expected. In fact, the correlation coefficient for the SNR and EDOF in this scenario is as high as 0.84. It is also interesting to note that at most positions along the measurement path, the result does not vary with time since the stars are quite focused. However, for three positions the stars are spread out indicating changing conditions over time. In these cases, there were people walking in the corridor between the transmitter and the receiver.

Very similar results were obtained for the second measurement set, see Figure 5 and 6. Again the SNR increase more than compensates for the drop in multipath richness and a high capacity is obtained in the unnormalized case. On the other hand, the normalized capacity experiences a significant drop due to the loss of multipath. Similar results were also obtained in other similar measurements at other locations within the Clyde building. Hence, it seems that when using the same transmitting power it is usually better to be in LOS with less multipath than in NLOS and more multipath. However, it should be noted that different buildings experience different conditions and the results presented here only represent this particular building.

The first and second measurement sets represents scenarios where the transition between LOS and NLOS is well defined. A slightly less defined transition from LOS to NLOS is the case of the third measurement set, see Figure 1, 7 and 8. Here, the transmitter was placed in a classroom and the receiver was moved along an adjacent corridor. There are two doorways that provide LOS conditions that give rise to the two peaks in the SNR and corresponding drops in EDOF. Again, the SNR increase more than compensates for the loss in multipath richness. However, the differences are slightly less obvious in this case since the

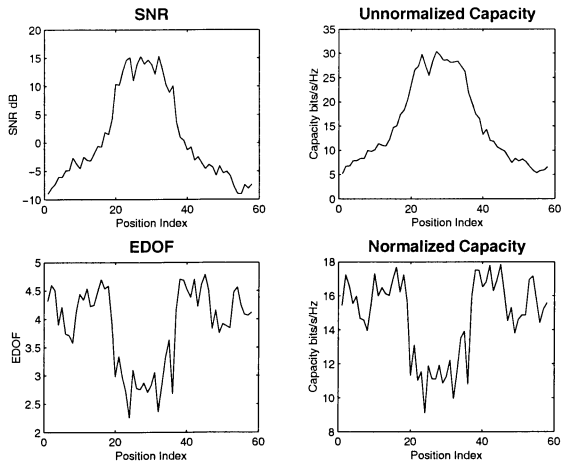


Figure 5: Capacity, SNR, and EDOF of the second measurement set.

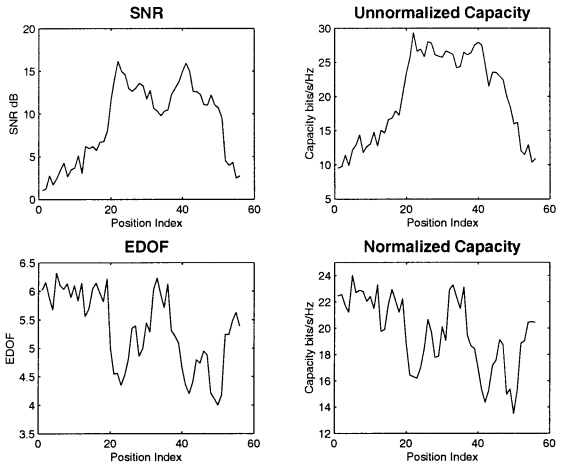


Figure 7: Capacity, SNR, and EDOF of the third measurement set.

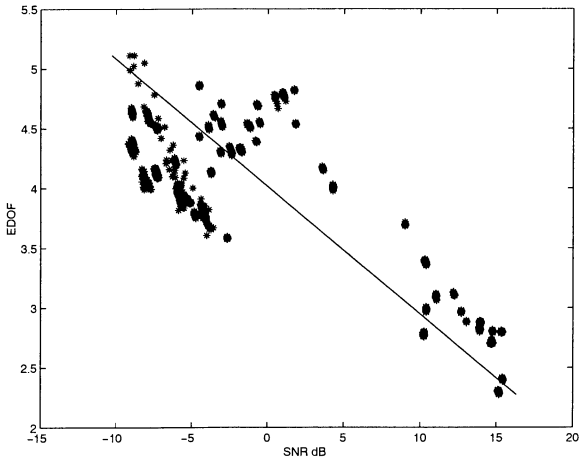


Figure 6: Scatter plot of EDOF and SNR for the second measurement set.

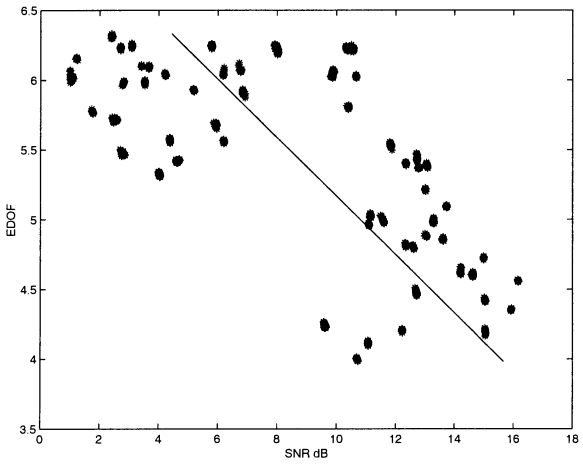


Figure 8: Scatter plot of EDOF and SNR for the third measurement set.

doorways are narrow and there is a significant amount of multipath even in the LOS instances. It should also be noted that in between the doorways there is only one wall that separates the transmitter and receiver which offers only about 5dB of attenuation.

5 Required SNR Increase

In all the indoor environments discussed in the previous section, a capacity increase resulted when entering LOS. The drop in multipath was more than compensated for by a significant increase in SNR. However, in these indoor environments there were still several propagation paths even in the LOS case. In other environments, such as outdoor environments, the LOS path may be more dominating than in the indoor cases studied here. In those cases the SNR increase may not be enough to compensate for the drop in EDOF. Hence, it is interesting to examine how much the SNR must be increased in LOS in order to yield the same capacity as a completely diffuse NLOS scenario.

The channel capacity for a LOS channel with only one propagation path, i.e. a completely specular channel, can be written in a much simpler form than the general form given in (2). By exploiting that the channel matrix $\mathbf{H}$ in this case is of rank one, the expression for the channel capacity can be reduced to [4]

$$C_s(\xi_s) = \log\left[1 + \xi_s N_R\right], \tag{4}$$

where the subscript in ξ_s denotes the SNR for this specular case. With this simple closed form expression, it is straightforward to calculate the required

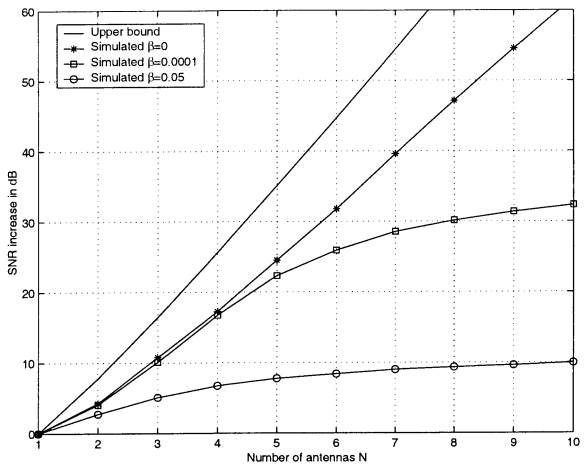


Figure 9: Required SNR increase for a completely specular and partly diffuse channel to reach the capacity of a completely diffuse channel with $\xi_d = 10$dB.

SNR for a specular channel to give the same average capacity as a diffuse channel which is denoted as $E[C_d]$. Here, $E[\cdot]$ represents expectation. By solving $C_s(\xi_s) = E[C_d]$ for ξ_s, the necessary SNR becomes

$$\xi_s = \frac{2^{E[C_d]} - 1}{N_R}, \tag{5}$$

where it is assumed that the capacity is calculated using the base two logarithm (bits). Unfortunately, no closed form expressions for the capacity exists in the general case. However, an upper bound on the SNR can be obtained by using an upper bound on the capacity that exploits Jensen's inequality and that log det is a concave function

$$E[C_d] \leq \log\det\left[\mathbf{I} + \frac{\xi_d E[\mathbf{H}_d\mathbf{H}^H]}{N_T}\right] = N_R \log\left[1 + \xi_d\right]. \tag{6}$$

Using (5) and (6), an upper bound on the SNR needed for a completely specular channel to yield the same average capacity as a diffuse channel becomes

$$\xi_s = [(1 + \xi_d)^{N_R} - 1]/N_R. \tag{7}$$

The required SNR increase ξ_s/ξ_d and the corresponding upper bound are shown for different number of antennas $N_T = N_R = N$ in Figure 9. Also shown is the SNR increase required in order for channels that are a combination of specular and diffuse $\mathbf{H}_{sd} = \sqrt{1-\beta}\mathbf{H}_s + \sqrt{\beta}\mathbf{H}_d$ to yield the same capacity as a completely diffuse channel. Here, the completely specular channel is denoted $\mathbf{H}_s$, the completely diffuse channel $\mathbf{H}_s$, and β is a mixing parameter $\beta \in [0, 1]$. All calculations were done using an SNR of the completely diffuse channel of 10dB. Surprisingly, the required SNR increase for moderate sized systems is not very large. For an 3×3, an SNR increase of only 10dB is required even for the completely specular channel $\beta = 0$. However, it should be mentioned that for higher SNR levels of the completely diffuse channel, the required SNR increase becomes larger. The upper bound is not tight but was used as a starting point for the numerical search required for the other curves that were calculated using Monte-Carlo simulations. It is also interesting to note that adding only a small fraction of diffuseness reduces the required SNR significantly. For example, adding 5% diffuseness to a completely specular channel corresponds to a 17dB lower required SNR increase for a 5×5 system.

6 Conclusions

Analyzing measurements of the MIMO channel taken at the BYU campus, it was found that the multipath richness and the signal strength usually not are independent. In fact, correlation coefficients as high as 0.8 were obtained. Thus, some caution is needed when normalizing out the SNR in system performance studies. Furthermore, for an unnormalized channel, the channel capacity usually rose when moving from NLOS into LOS since the loss in multipath was more than compensated for by an increase in SNR. Finally, an expression for the required SNR increase for a specular channel to yield the same capacity as a completely diffuse channel was derived. Surprisingly, the required SNR increase for moderate sized systems is not very large.

References

[1] J.P. Kermoal, L. Schumacher, P.E. Mogensen, and K.I. Pedersen. "Experimental Investigation of Correlation Properties of MIMO Radio Channels for Indoor Picocell Scenarios". In *Proc. IEEE VTC 2000 Fall*, Boston, MA, September 2000.

[2] D.P. McNamara, M.A. Beach, P.N. Fletcher, and P. Karlsson. "Capacity Variation of Indoor Multiple-Input Multiple-Output Channels". *Electronics Letters*, 36(24):2037–2038, November 2000.

[3] D-S Shiu, G.J. Foschini, M.J. Gans, and J.M. Kahn. "Fading Correlation and Its Effect on the Capacity of Multielement Antenna systems". *IEEE Trans. on Communications*, 48(3):502–513, March 2000.

[4] E. Telatar. "Capacity of Multi-Antenna Gaussian Channels". Technical report, AT&T Bell Labs, June 1995.

[5] J.W. Wallace and M.A. Jensen. "Measured Characteristics of the MIMO Wireless Channel". In *Proc. IEEE VTC 01 Fall*, pages 2038–2042, Atlantic City, NJ, USA, October 2001.

[6] P.W. Wolniansky, G.J. Foschini, G.D. Golden, and R.A. Valenzuela. "V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel". In *Proc. 1998 URSI Int. Symp. on Sign., Syst., and Elect.*, pages 295–300, 1998.

On the geometry of signals for multipath fading channels

Tommy Svensson, Johan Lassing and Thomas Eriksson

Chalmers University of Technology, Communication Systems Group, Department of Signals and Systems, SE-412 96 Göteborg, Sweden. E-mail: {tommy.svensson, johan.lassing, thomas.eriksson}@s2.chalmers.se.

Abstract

In this paper, we analyze the geometry of signals for multipath fading channels and their resistance to a non-linear transmitter. We do this by defining an orthonormal signal space as the eigenfunctions of the channel and the non-linearity is modeled as an amplitude clipping device. In addition, we make an analysis of the geometry of peak-to-average power reduction codes. We conclude that time localized signal spaces are more resistant to a non-linear transmitter and that PAPR reduction codes that provides a 3 dB reduction in PAPR seems to be possible with a negligible loss in the channel capacity of an amplitude restricted channel.

1 Introduction

In complex radio propagation environments a wideband signal becomes heavily distorted by multipath propagation and fading. In a slowly fading environment, OFDM has become a popular way to cope with the multipath environment. The reason is that equalization is easy to perform in the frequency domain, because the OFDM symbols are long in time and thus a guard time can be introduced to reduce the amount of multipath echoes with a small reduction in spectral efficiency. A problem with OFDM is the high peak-to-average power ratio, which causes the signal to be distorted by the power amplifier. This problem becomes even more prominent in the case of high carrier frequency communications, e.g. at the potential 60 GHz frequency band, since the power amplifier technology is less mature for these carrier frequencies. Hence, there is a trade-off to be made between channel equalization properties and adaptation to the amplifier non-linearity in the design of the signal space. In a fast fading environment, OFDM becomes less adapted to the channel since the fading cannot be considered to be constant over an entire OFDM symbol. Thus, there exists other signal spaces that should be more suited to such channel conditions.

In this paper, we analyze the relations above by a geometric model. We investigate the potential benefits of using more time localized signals than OFDM to reduce the distortion of a non-linear transmitter. The orthonormal signal space is defined by the eigenfunctions of the channel and the non-linearity is modeled as an amplitude clipping device. The degradation is measured as the signal-to-clipping noise ratio. We also make an analysis of the geometry of peak-to-average power reduction codes.

We conclude that time localized signal spaces are more resistant to a non-linear transmitter and that PAPR reduction codes that provides a 3 dB reduction in PAPR seems to be possible with a negligible loss in the channel capacity of an amplitude restricted channel.

2 System model

In the system model in Fig. 1 the data symbols m_i, $i = 0 \ldots N-1$ in the column vector $\boldsymbol{m}$ are transmitted after a linear transformation by the matrix M. The resulting vector $\boldsymbol{x}$ is a sampled complex baseband signal, which is passed through a memoryless and time-invariant transmitter device that may be non-linear. The transmitter is described by the matrix T, and the transmitted signal $\boldsymbol{s}$ propagates through a possibly time varying and frequency selective channel that is described by the channel matrix

$$H = \begin{bmatrix} h_0^{(0)} & 0 & 0 & \ldots & & 0 \\ h_1^{(1)} & h_0^{(1)} & 0 & \ldots & & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ h_{L-1}^{(L-1)} & h_{L-2}^{(L-1)} & \ldots & h_0^{(L-1)} & 0 \ldots & 0 \\ & \ddots & \ddots & & \ddots & \\ 0 & \ldots & h_{L-1}^{(N-1)} & h_{L-2}^{(N-1)} & \ldots & h_0^{(N-1)} \\ \vdots & & & \ddots & & \vdots \\ 0 & \ldots & & & 0 & h_{L-1}^{(N+L-2)} \end{bmatrix} \quad (1)$$

where $h_l^{(k)}$ is the l:th complex channel tap at time k. The signal is received by a matched filter receiver R. Finally, the decision symbols $\boldsymbol{d}$ are derived by using a matrix D. The decision regions for the data symbols will depend on how D is chosen. The model does not include thermal noise since we will study the geometry of the signal space only.

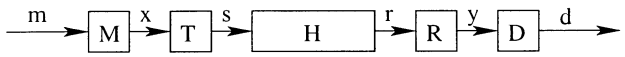


Figure 1: Model of the communication system.

3 Ideal transmitter

If we assume the transmitter device to be ideal, we have $T = I$ and

$$\boldsymbol{d} = DRHM\boldsymbol{m}. \tag{2}$$

The received signal is a linear combination of the column vectors of the channel matrix H. With receiver filters that are matched to the channel, the receiver matrix becomes $R = H^H$ and the decision variables $\boldsymbol{d}$ equals

$$\boldsymbol{d} = DH^HHM\boldsymbol{m}. \tag{3}$$

Since H^HH is a normal matrix, we can diagonalize it (or equivalently make a SVD of H)

$$H^HH = S\Lambda S^H \tag{4}$$

where Λ is a real diagonal matrix and S is an orthonormal matrix, whose columns are the eigenvectors of H^HH.

Now, if we want the decision device to be simple, we want the decision variables d_i to be scaled versions of the data symbols m_i. This is the case if we choose the mapper matrix $M = S$ and the demapper matrix $D = S^H$, since then

$$\boldsymbol{d} = \Lambda\boldsymbol{m}, \tag{5}$$

and thus $d_i = \lambda_i m_i$, $i = 0 \ldots N-1$. Hence, we should signal with the eigenvectors of the channel matrix H^HH.

A special case is a non-frequency selective channel $H = \Lambda_{h_t}$, where Λ_{h_t} is a diagonal $(N + L - 1) \times N$ matrix with main diagonal elements h_{t_i}, $i = 0 \ldots N - 1$. In this case, if we choose $S = I$ we get

$$\boldsymbol{d} = \Lambda_{h_t}^H\Lambda_{h_t}\boldsymbol{m} \tag{6}$$

which makes $d_i = |h_{t_i}|^2 m_i$, $i = 0 \ldots N-1$, i.e. PAM signaling diagonalizes this channel.

Another special case is a non-time selective channel $H = \Lambda_{h_f}F_N$, where Λ_{h_f} is a diagonal $(N + L - 1) \times N$ matrix with main diagonal elements h_{f_i}, $i = 0 \ldots N-1$ and F_N is the normalized DFT matrix of order N. If we choose $S = F_N^H$ we get

$$\boldsymbol{d} = \Lambda_{h_f}^H\Lambda_{h_f}\boldsymbol{m} \tag{7}$$

and $d_i = |h_{f_i}|^2 m_i$, $i = 0 \ldots N-1$. Since $M = F_N^H$ and $D = F_N$, OFDM diagonalizes this channel.

4 Non-ideal transmitter

Many channels can be regarded as slowly time varying, and a common assumption is that the channel is static over the data block to be transmitted. In this case, OFDM is the basis that diagonalizes the channel. However, in a real system these signals are distorted by a non-linear transmitter device, i.e. the amplifier, unless a large back-off is used. The non-linearity spreads the data symbol energy into other signaling dimensions, making the decision variables in the receiver to be dependent. In OFDM the symbol energy is uniformly distributed over the entire signal. This makes all symbols in $\boldsymbol{m}$ to affect the signal amplitudes in $\boldsymbol{x}$ and the signal envelope varies heavily. In contrast, when PAM signaling is used only one symbol at a time affect the signal amplitude. Hence, when PSK is used as the data alphabet the envelope is constant.

We will assume that the energy is one for the data symbols in each dimension, and that each data symbol is transmitted by a normalized orthogonal basis function. Furthermore we model the non-linear transmitter as a clipping device, where the signal is clipped at amplitude level A.

4.1 Clipping energy of OFDM

A common measure of the envelope variation of a signal is the peak-to-average power ratio (PAPR) [1]. This may be a poor measure, since a few signals with occasionally very high envelope values will give a high PAPR even if the energy is small in the peaks. A better measure of the implication of the non-linear distortion is the amount of energy that is removed from the signal. The reason is that the clipping introduces self-interference.

By making some simplified assumptions on the transmitted waveforms in an OFDM system, it is possible to get an indication of the level of noise introduced by the clipping device. To this end, we assume that the transmitted signals are formed from two independent, zero-mean Gaussian variables of variance 1/2 each, so that the amplitude of the time samples of the transmitted waveforms are all Rayleigh distributed. This assumption is valid, since the time-samples are the result of a superposition of many independent realizations of random experiments. If we define a clipped waveform $s_A(t)$ as

$$s_A(t) = \begin{cases} s(t) & |s(t)| < A \\ As(t)/|s(t)| & |s(t)| \geq A \end{cases} \tag{8}$$

we may calculate the power in the error signal $s_e(t) = s(t) - s_A(t)$ as a function of A from the samples $s[k]$ of the time waveforms. Regardless if we use oversampling of the waveforms or not, the power in the error

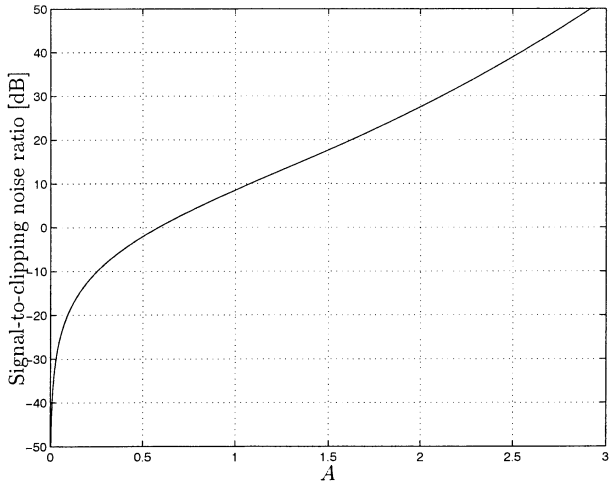


Figure 2: Analytical signal-to-clipping noise ratio as a function of the clipping level A for OFDM.

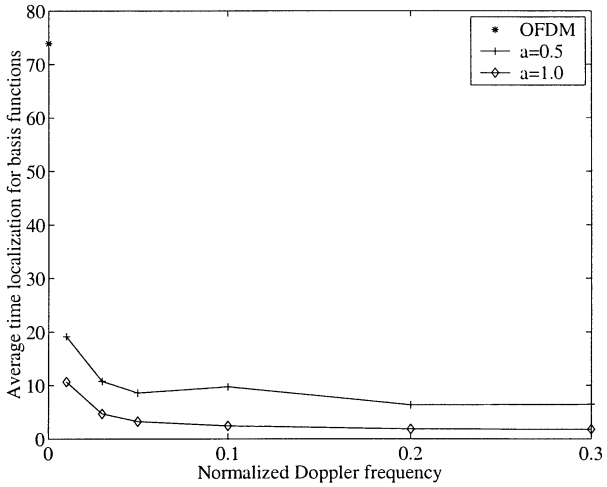


Figure 3: Average time localization for the basis functions S of a multipath Rayleigh fading channel $N = 256$ and $L = 10$ with different Doppler frequency f_m and power decay parameter a. For each f_m and a, the time localization is calculated as an average of 50 channel realizations. As a reference, the time localization for OFDM is shown.

signal after clipping is given by

$$E[|s_e[k]|^2] = \int_A^\infty (x - A)^2 f_X(x) dx \qquad (9)$$

where

$$f_X(x) = 2xe^{x^2} \qquad (10)$$

Carrying out the integration gives the expression

$$E[|s_e[k]|^2] = e^{-A^2} - A\sqrt{\pi}\,\mathrm{erfc}\,(A) \qquad (11)$$

If we now evaluate the ratio $E[|s_A[k]|^2]/E[|s_e[k]|^2]$, where

$$E[|s_A[k]|^2] = \int_0^A x^2 f_X(x) dx + A^2 \int_A^\infty f_X(x) dx = 1 - e^{-A^2}, \qquad (12)$$

we get an indication of how much the signal is distorted from the clipping operation. The result is seen in Fig. 2, where the signal-to-clipping noise ratio is defined as $E[|s_A[k]|^2]/E[|s_e[k]|^2]$.

4.2 Time localized signals

In order to study more general basis functions than OFDM, we calculate the orthonormal basis S for a multipath Rayleigh fading channel [1, p. 180]. The L channel taps are separated by the sampling time, and the power delay profile is assumed to be exponential with decay parameter a

$$|h_l^{(k)}|^2 = \exp(-al), \text{ for } l = 0 \ldots L-1 \text{ and } \forall k. \qquad (13)$$

Hence, the larger a the more relative power is concentrated to the direct path. Furthermore, the time variation is modeled by using Jakes' model for the Doppler spread [2, p. 20] with f_m as the normalized Doppler frequency.

Let us define the degree of time spread of a basis function as the standard deviation for the energy distribution over the N time samples, and the time localization of a basis as the average time spread of the basis functions.

In Fig. 3 the time localization of the basis S as a function of Doppler frequency f_m is shown for two different values of the power decay parameter a. As shown, the basis functions becomes more time localized as the direct path becomes stronger and also as the channel becomes more time-varying.

An interesting question is how more time localized signals are affected by a non-linear transmitter. Is the signal-to-clipping noise ratio increased? In Fig. 4 the basis functions S are used to transmit QPSK data with unit energy. The signals are clipped at two different amplitude levels, $A = 1.5$ and $A = 1.1$. As is seen, the signal-to-clipping noise ratio is improved for more time localized signals.

If we choose a signal space that is more time-localized than the signals that diagonalizes the channel, we gain in signal-to-clipping noise ratio, but instead the decision symbols $\boldsymbol{d}$ becomes dependent, i.e. cross-talk between the signal dimensions. Hence, there is a trade-off to made in the signal design between signals S, which diagonalizes a channel, and signals that are resistant to a non-linear transmitter. For a time-varying channel, we could instead of OFDM choose a

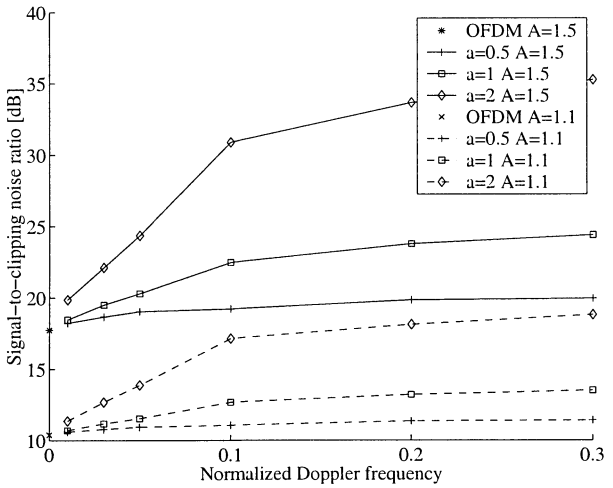


Figure 4: Simulated signal-to-clipping noise ratio for signals carrying QPSK data symbols and using the basis functions S of a multipath Rayleigh fading channel $N = 256$ and $L = 10$ with different Doppler frequency f_m and power decay parameter a. Two levels of the clipping amplitude A is shown. For each f_m and a, the energy loss is calculated as an average of 1000 channel realizations. As a reference, the simulated signal-to-clipping noise ratio for OFDM is given.

signal space that is more time-localized than S, but with the same cross-talk energy and larger signal-to-clipping noise ratio.

4.3 Aspects on PAPR reduction codes

Another way to reduce the self-interference is to not allow the signals with too high envelope values to be transmitted. This will reduce the information rate. Let us view the basis signals in the time domain as vectors in an n-dimensional space. One way to estimate the rate R of a channel code that selects only these allowed signals is to calculate the remaining volume of the n-dimensional signal space, when all the not allowed code words are removed.

Consider a communication system model in which the transmitted signal is a vector in n-dimensional space, having energy E_n in each dimension. The transmitted vector is disturbed in each dimension by a Gaussian random variable having zero mean and variance $N_0/2$, so that the received vector is the superposition of the transmitted vector and a noise vector.

With this simple model we can easily motivate the capacity formula for the Gaussian channel derived by Shannon. The channel capacity theorem is a consequence of the sphere packing bound as the dimension n tends to infinity, and therefore we calculate the number of spheres of radius $\rho = \sqrt{nN_0/2}$ that can be fitted inside a sphere of radius $r = \sqrt{nE_n + nN_0/2}$.

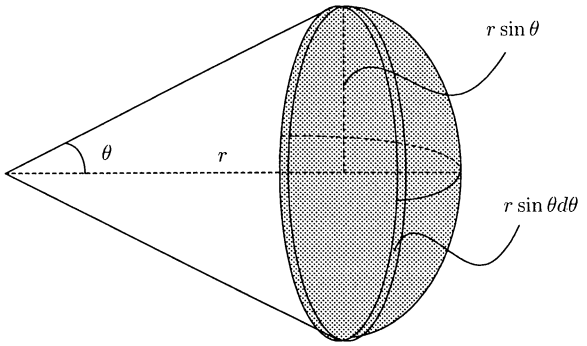


Figure 5: n-dimensional spherical segment. The shaded region is the volume $V_s(\theta)$.

This is easily done,

$$M = \frac{V_n r^n}{V_n \rho^n} = \left(\frac{E_n + N_0/2}{N_0/2}\right)^{n/2} = \left(1 + \frac{2E_n}{N_0}\right)^{n/2} \tag{14}$$

where V_n is the volume of the n-dimensional unit sphere. Now, since M is the number of available signal points, which for a system signaling with rate R bits per dimension equals 2^{nR}, we have

$$M = 2^{nR} = \left(1 + \frac{2E_n}{N_0}\right)^{n/2} \quad \Rightarrow$$
$$R = \frac{1}{2}\log_2\left(1 + \frac{2E_n}{N_0}\right) \tag{15}$$

which is recognized as the channel capacity for the Gaussian channel. Note that this is not a proof of the capacity theorem, although it can be shown that this rate actually is achievable, while no higher rate is, as n approaches infinity.

If we now say that the possible transmitted signals must have an energy no greater than n (i.e., $E_n = 1$), the possible transmitted vectors are all contained within the n-dimensional sphere of radius $\sqrt{n}$. If we denote the volume of this sphere by V_1 and consider the rate to be R_1, each signal point can be assigned a volume of $v = 2^{-nR_1}V_1$ from the original signal sphere. Now, say that no individual dimension is allowed to assume values greater than a certain threshold A, so that the original signal sphere will be distorted. If the threshold A is not too small (the exact value will be stated below) the regions that are cut away by this clipping will be perfect spherical segments. The volume of these spherical segments can be calculated exactly. Referring to Fig. 5, the n-dimensional spherical segment is built up from spheres of dimension $n-1$ having a radius $r\sin\theta$ and height $r\sin\theta d\theta$. Integrating over these spheres will give us the volume of an n-dimensional spherical seg-

ment having a half-cone angle θ_1 as

$$V_s(\theta_1) = V_{n-1} r^n \int_0^{\theta_1} (\sin\theta)^n d\theta = \frac{r^n \pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \int_0^{\theta_1} (\sin\theta)^n d\theta \qquad (16)$$

We may bound this integral by the following expression [3]

$$V_s(\theta_1) = \frac{r^n \pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)} \int_0^{\theta_1} (\sin\theta)^n d\theta \leq \frac{r^n \pi^{\frac{n-1}{2}}}{(n+1)\Gamma\left(\frac{n+1}{2}\right)} \frac{(\sin\theta_1)^{n+1}}{\cos\theta_1} \qquad (17)$$

where the left side is asymptotic to the right side as $n \to \infty$. The relation between the angle θ_1 and the clipping level A is $\theta_1 = \arccos\frac{A}{\sqrt{n}}$ and as long as $A \geq \sqrt{n/2}$, the clipped regions will be perfect spherical segments, and we may use the above formulas to estimate the volume removed by the clipping device. Therefore, let us calculate the remaining volume as a function of θ_1:

$$V'(\theta_1) = 1 - \frac{2nV_s(\theta_1)}{V_n r^n} \geq 1 - \frac{2n}{n+1}\frac{V_{n-1}}{V_n}\frac{(\sin\theta_1)^{n+1}}{\cos\theta_1} \qquad (18)$$

If we make the substitution $\theta_1 = \arccos\frac{A}{\sqrt{n}}$, we get

$$V'(\theta_1) \geq 1 - \frac{2n}{n+1}\frac{V_{n-1}}{V_n}\frac{\sqrt{n}\left(1-\frac{A^2}{n}\right)^{\frac{n+1}{2}}}{A} \qquad (19)$$

This expression can be calculated for reasonable values of n. As n grows large, we may evaluate the quotient V_{n-1}/V_n by using Sterling's asymptotic formula [4] for the Γ-function. This gives,

$$V'(\theta_1) \approx 1 - \sqrt{\frac{2}{\pi}}\frac{n^2}{n+1}\frac{(1-\frac{A^2}{n})^{\frac{n+1}{2}}}{A} \qquad (20)$$

where the above relations are valid as long as $A \geq \sqrt{n/2}$. Inserting the minimum value for A in equation 19, gives the expression

$$V'(\theta_1) \geq 1 - \frac{n}{n+1}\frac{V_{n-1}}{V_n} 2^{-\frac{n}{2}+1} \qquad (21)$$

This expression is shown in Fig. 6, from which it is clear that this clipping eventually removes no volume at all. From this we can state that a 3 dB reduction in PAPR is possible (since clipping occurs at $1/\sqrt{2}$ from the peak level) with a negligible rate loss.

5 Conclusion

We have shown that it is actually an advantage if the channel is time-varying over the data block, since then more time-localized signals diagonalizes the channel and such signals are less distorted by a non-linear transmitter device. Hence, there is a potential to use more time localized signals. We also discussed that there is a trade-off to made in designing signals that diagonalizes a channel and the resistance to a non-linear transmitter. Of course, there are practical limitations in using signals that are dynamically adapted to the channel, since the channel has to be predicted, and one of the most important advantages of OFDM is the existence of the fast Fourier transform (fft).

Furthermore, we conclude that PAPR reduction codes that provides a 3 dB reduction in PAPR seems to be possible with a negligible loss in the channel capacity of an amplitude restricted channel.

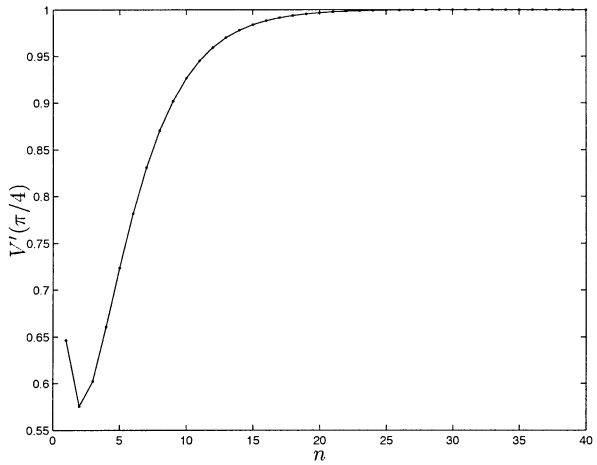


Figure 6: The remaining volume after clipping spherical segments from the n-dimensional sphere with $A = \sqrt{n/2}$.

References

[1] R. Van Nee and R. Prasad, *OFDM for wireless multimedia communications*, Artech House Publishers, London, 2000.

[2] W. C. Jakes, Ed., *Microwave Mobile Communications*, IEEE Press, 1993, reissue.

[3] C. E. Shannon, "Probability of error for optimal codes in a gaussian channel," *Bell System Technical Journal*, vol. 38, pp. 611–656, May 1959.

[4] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing.*, Dover Publications, Inc., New York, 1972.

Radio Resource Management Principles and Performance for WCDMA Release Five

Anders Furuskär[1, 2], Stefan Parkvall[1], Magnus Persson[1] and Maria Samuelsson[1]
1) Ericsson Research, Sweden [@era.ericsson.se] 2) Wireless@KTH, Royal Institute of Technology, Sweden

Abstract

In release 5 of the WCDMA standard, support for peak bitrates for data services in the order of 10 Mbit/s, together with reduced delays and increased system throughput is introduced. This paper discusses some radio resource management principles associated with the high-speed packet data bearers introduced in release 5. The areas included are scheduling for WWW services, provision of bearers for streaming services, as well as the simultaneous use of high speed packet data bearers and conversational bearers. Simulation results indicate that with appropriate radio resource management, all the above challenges may be handled with high capacity.

1 Introduction

Wideband CDMA (WCDMA) is rapidly emerging as the leading global 3G standard, enabling high quality and capacity Mobile Internet access. Release 99, the first release of the WCDMA specifications, fully supports the IMT-2000 requirements, including data rates up to 2 Mbit/s and support of both packet and circuit switched services [1]. With the introduction of release 5 of the specifications in the spring of 2002, WCDMA packet data support is further enhanced. Peak data rates in the order of 10 Mbit/s together with lowered roundtrip delays and increased capacity provide a further boost for wireless data access.

WCDMA release 5 employs three fundamental principles in order to provide these high data rates, as well as to improve the system capacity. These all rely on rapid adaptation of the transmission parameters to the rapidly varying radio conditions: fast link adaptation, fast hybrid ARQ, and fast scheduling of a shared channel. The basics of these principles and their operation in WCDMA release 5 have previously been described in e.g. [2] and [3]. This paper extends the above studies by presenting and evaluating basic principles for scheduling of WWW users, taking Transmission Control Protocol (TCP) interaction into account, provision of streaming bearers and simultaneous use of multiple services. First, a short overview of the packet data enhancements provided by release 5 of the WCDMA specifications is given. Next, models and assumptions used are discussed, followed by a presentation of the employed principles and performance results for the WWW, streaming and integrated services studies. Finally, conclusions are drawn.

2 WCDMA Release 5 High-Speed Packet Data

In release 5 of the specifications, a new transport channel targeting packet data transmissions, the HS-DSCH (High-Speed Downlink Shared CHannel), is introduced, supporting three basic principles: fast link adaptation, fast hybrid ARQ, and fast scheduling. These three principles rely on rapid adaptation to changing radio conditions and the corresponding functionality is therefore placed in the Node B instead of the RNC. The HS-DSCH corresponds to a shared channelization code resource, shared among the users primarily in the time-domain. In Fig. 1, the code resource, consisting of multiple codes of spreading factor 16, is illustrated. A part of the code tree is allocated for the HS-DSCH, while the remaining part is simultaneously used for other channels, e.g., dedicated channels used for speech services. Allocation of the shared code resource among the users is done on the basis of the 2 ms HS-DSCH TTI (Transmission Time Interval). The use of a short TTI allows for tracking fast fading, improves the granularity in the scheduling process, and reduced roundtrip delays. Simultaneous transmission to multiple users is possible by using distinct parts of the codes allocated for HS-DSCH transmission, although the primary way of sharing the code resource is in the time domain.

Fast link adaptation is the process of rapidly adapting the transmission formats to the instantaneous channel

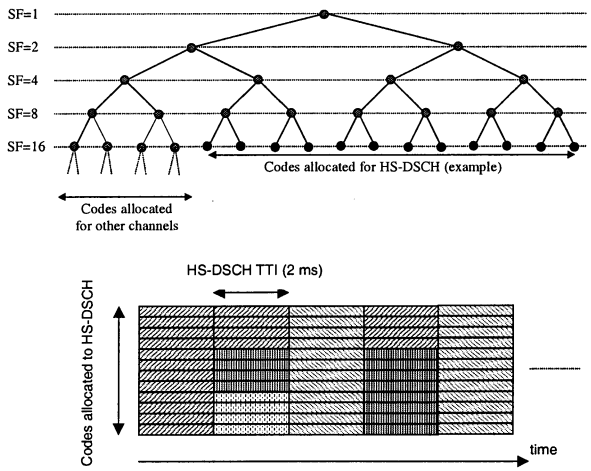


Fig. 1: In the code domain, a part of the code tree is set aside for the HS-DSCH. In the figure, an example with 12 codes reserved for HS-DSCH transmission and simultaneous transmission to up to three users is shown.

conditions. In addition to QPSK, higher order modulation in form of 16QAM is supported, providing improved spectral efficiency in bandwidth limited scenarios. By adjusting the code rate and the modulation scheme used, the momentary data rate can be matched to the instantaneous radio conditions and hence maximize the utilization of the radio channel. This is briefly illustrated in Fig. 2, where an example of link throughput is depicted. Knowledge of the instantaneous radio conditions are provided by regular measurement reports from the UE.

Fast hybrid ARQ implies rapidly requesting retransmission of erroneously received data entities and using soft combining of multiple transmission attempts. By combining soft information from multiple transmission attempts, the number of retransmissions needed, and thus the delay, will be reduced. Hybrid ARQ with soft combining also adds robustness against link-adaptation errors and is closely related to the link adaptation mechanism.

Fast scheduling is in control of allocating the shared resources among the different users and operates on the 2 ms TTI basis. It is a key element in the design and to a large extent determines the overall behavior of the system. In addition to traffic statistics, the scheduler also takes the instantaneous radio conditions into account in the scheduling process. Note that information of the radio conditions is required by the link adaptation scheme as well. Maximum system throughput is obtained by assigning all available radio resources to the user with the currently best radio conditions, while a practical scheduler could include some degree of fairness.

3 Models and Assumptions

To illustrate the feasibility of providing a multitude of services to subscribers using the HS-DSCH in a cellular environment, a number of system simulations have been carried out. These are based on the link results used to generate Fig. 2.

The simulation environment consists of 12 sectors with 3 sectors per site and 1.5 km site-to-site distance deployed in an ordinary hexagonal layout. Wrap-around is used to prevent border effects. Standard models for pathloss (29+35*log(R)), shadow fading (standard deviation of 8 dB) and multipath fading (ITU Pedestrian A or 3GPP Typical Urban) are used.

The WWW traffic model captures the characteristics of web browsing sessions, basically consisting of log-normal distributed object sizes and with exponential distributed reading time between each object (5 s in average). The average object size is about 17 kbyte. The TCP data flow is controlled with a *sliding window* protocol. In the slow start phase, the window size is gradually increased. When the first IP-packet loss occurs, TCP decreases the window and enters the "congestion avoidance" phase, in which the window is increased at slower rate. The maximum segment size is set to 1460 bytes, initial window size is set to 1. For further TCP details see e.g. [4]. The video streaming model [5] emulates an H.263 RTP/UDP/IP based video streaming application

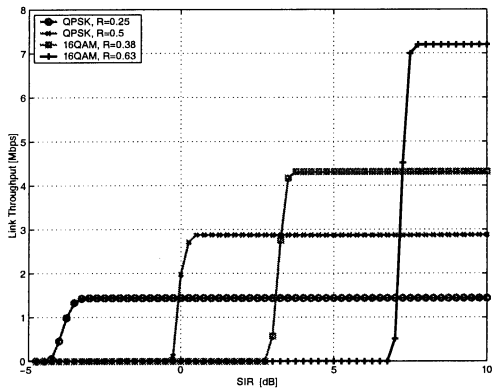


Fig. 2: Illustration of link throughput for the initial transmission attempt using four different combinations of modulation schemes and code rates, selected based on the instantaneous radio conditions. Note that the HS-DSCH is not limited to any particular code rates and these curves only serve as an illustration of the principle.

with an average source rate of about 54 kbps with a codec frame rate of 10 frames/s. Poisson arrivals of streaming sessions (clips) with 20 s mean clip duration are assumed. The voice traffic model has a constant information bit-rate of 12.2 kbps.

The modulation and coding scheme is selected based on UE C/I measurement. The Hybrid ARQ signaling from the UE is assumed to be error-free but subject to delays. The uplink is not explicitly simulated. The downlink common control channels is allocated 20% of the base station power. The remaining part is allocated to HS-DSCH, HS-DSCH associated DPCH and speech DPCH.

The one way delay between server and radio access network is assumed to be to 50 ms and no packet losses are assumed to occur between server and radio access network.

4 Performance Results

This section presents performance results comparing different schedulers, as well as demonstrating the viability of providing streaming services and mixing HS-DSCH-based services with other WCDMA services. It should be noted that algorithms used are not optimized, and that the results therefore should be viewed as feasibility indicators rather than as absolute capacity figures.

4.1 WWW TCP Throughput Results for different Schedulers

The characteristics of different schedulers have already been evaluated in e.g. [2] and [3]. In this paper these studies are extended to also include effects of TCP. The scenario evaluated assumes no packet losses between the server and radio access network so it is mainly the slow start phase that affects the web browsing performance.

Three types of schedulers have been evaluated: C/I, Round Robin (RR) and Proportional Fair (PF) scheduling algorithms.

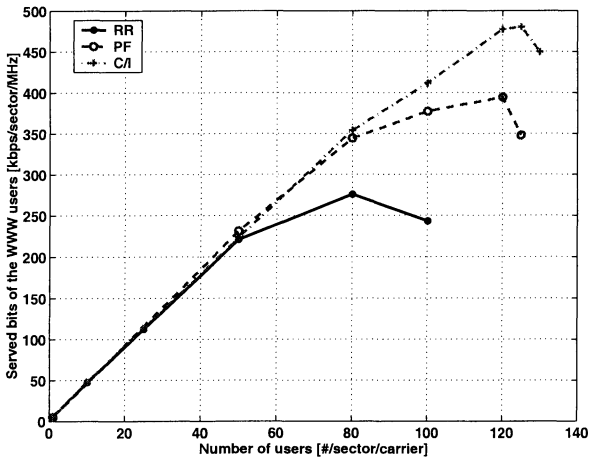


Fig. 3. WWW with TCP: served traffic vs. number of users, Pedestrian channel.

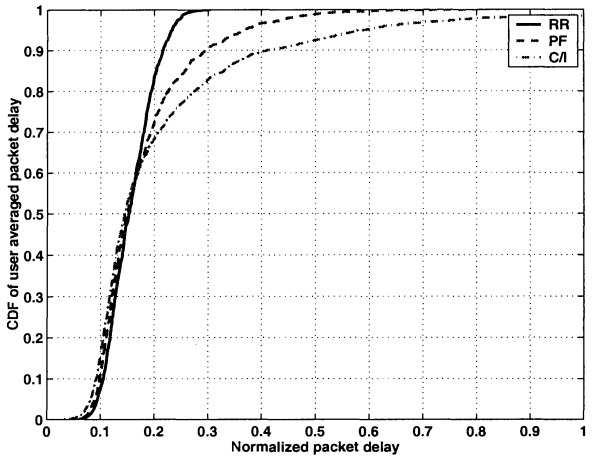


Fig. 4. WWW with TCP: CDF of normalized delay, Pedestrian channel.

The *C/I* scheduler directs transmission to the user with the momentarily best channel conditions, with a delay of 2 TTIs, allowing for the highest possible data rate at each instant and thus maximizing the overall throughput. However, the price paid for the high overall throughput is the potentially large variations in service quality among the user population.

The *round robin* scheduler cycles through the list of active users and thus is fair in the average sense. However, as the round robin scheduler does not base its decision on the varying channel quality, the throughput performance suffers. Hence, a scheduling algorithm that both takes channel quality into account and at the same time introduces some degree of fairness may be desirable.

The *proportional fair* scheduler schedules the user with the currently highest ratio between instantaneous C/I and average transmission rate. This algorithm takes advantage of the short-term channel variations while at the same time maintaining the same long-term throughput for all users.

In summary, the preferred scheduler depends on the desired trade-off between system throughput and fairness for high loads. The system throughput of HS-DSCH is shown in Fig. 3 for a WWW service including TCP effects. The corresponding values of system throughput without TCP are 300 (round robin), 500 (proportional fair) and 570 (C/I) kbps/sector/MHz respectively. As seen in the figure the C/I scheduler performs best in terms of maximum served traffic. When applying TCP the served traffic is decreased about 15-20% for the C/I and proportional fair cases and about 10% for the round robin case.

The user quality in terms of normalized packet delay, measured at the different schedulers capacity limits, is shown in Fig. 4. It can be seen that the tail of the normalized delay could be substantially larger for C/I compared to round robin or proportional fair scheduling algorithms. Note that the capacity limits differ between the schedulers though. For equal traffic loads, the C/I scheduler typically results in the lowest normalized delay.

The conclusion is that the loss in system throughput when introducing TCP is about 10-20% depending on the scheduler chosen. Further, with TCP the evaluated schedulers provide fairly similar user quality, in contrast with [3] (no TCP effects were included) where C/I was outstanding. However, the C/I scheduler is still a good choice if some of the users can tolerate higher delays.

4.2 Video streaming Capacity Results

The fraction of satisfied users versus offered load for a streaming service on HS-DSCH is shown in Fig. 6. In this study the system capacity is defined as the offered load at which the percentage of satisfied users equals 95%. A user is satisfied if no re-buffering occurs due to starved receiver buffer and the initial buffering time is less than a predefined time, T_{limit}.

The required initial receiver buffer size (Fig. 5), which corresponds to 5 seconds at a constant transmission rate, has to be reached before the client starts to play the streaming clip. T_{limit} is set to 5 seconds plus a margin due to variations in the transmission rate.

The streaming capacity has been evaluated for both round robin and C/I schedulers. Fig. 6 shows that the C/I scheduler gives 40% better capacity than round robin. The results also show that the capacity is insensitive to different values on T_{limit}. The conclusion is that the streaming service performs well over HS-DSCH but the quality is varying within the cell.

4.3 Mixed Service Principles and Capacity Results

For multiple services, capacity is defined as the maximum aggregate load for which all services' quality requirements are sustained. The quality requirements are expressed in two steps; on system level, a Fraction of Satisfied Users (FSU) requirement of 95% for both voice and WWW is assumed. A voice user is assumed to be satisfied if it is not blocked and its voice quality is acceptable, which in turn requires an average FER of at most 1%. For web browsing, exceeding a Circuit Switched Equivalent (CSE) bitrate of either 50 or 250 kbps is required to be satisfied.

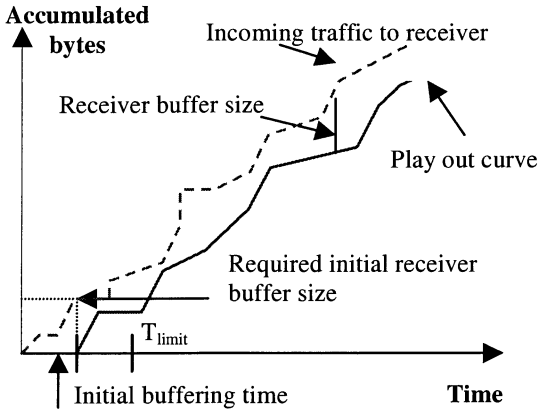


Fig. 5. Streaming quality requirements. The initial buffering time must be less than a predefined time, T_{limit} and the receiver buffer size must always be larger than zero during play out.

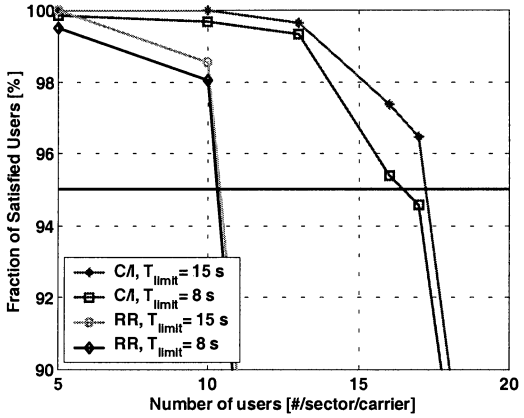


Fig. 6. Streaming capacities for round robin and C/I scheduling. Two values of T_{limit}, 8 and 15 seconds, are also compared. Typical Urban channel

With the above multiple service capacity definition, capacity is always limited by the least robust service. Hence, an unbalance in quality between service groups reduces overall capacity, whereas a resource allocation that balances quality, i.e. makes all services equally robust, maximizes capacity. In our case, as described in [3], this quality balancing may in many cases simply be achieved by controlling the HS-DSCH power. For example, if the voice service limits the overall capacity, the HS-DSCH power may be limited in order to reduce the interference caused to the voice users, and thereby improve their quality. A simple way of limiting the HS-DSCH power is to not allocate it *all* the remaining available base station power, but rather the remaining power up to a threshold P_{th}. Thus, $P_{HS\text{-}DSCH} = \max(P_{th} - P_{other}, 0)$, where P_{other} is the total power used excluding the HS-DSCH.

The effect of this HS-DSCH power limiting is depicted in Fig. 7, which shows the fraction of satisfied voice and WWW users as a function of WWW traffic load, for a fixed voice load of 5 Erlang/MHz/sector. With P_{th}=20 W, the interference caused by the HS-DSCH renders poor voice quality before the WWW quality gets unacceptable. As a result, the voice service limits the overall capacity to about 55 kbps/MHz/sector. With P_{th}=15 W on the other hand, a higher traffic load can be sustained by the voice bearers. The price paid for this is of course a poorer WWW quality. This does not affect the overall capacity however, since the load tolerable by WWW is still as high as the load tolerable by voice. The overall capacity has now increased to some 80 kbps/MHz/sector, which corresponds to a gain of about 40%.

By repeating the above analysis for different service mixes, joint voice and WWW capacity regions may be derived, as depicted in Fig. 8. The areas delimited by the curves correspond to feasible traffic mixes, for which acceptable fractions of satisfied users for both service types are sustained. The bitrate requirement for WWW users is here 50 kbps. Further, two different HS-DSCH power thresholds are compared, 15 W and 20 W. It is seen that P_{th} = 15 W is better when the voice load is high, whereas P_{th} = 20 W is better for lower voice loads. The higher tolerable HS-DSCH power for low voice loads is due to the higher power available per voice user for lower number of voice users.

With a higher bitrate requirement of 250 kbps (Fig. 9), WWW becomes the limiting service for all service mixes, and the best performance is achieved with P_{th}=20 W. Thus, here even better performance could be achieved by limiting the total power budget of the voice bearers, in order for the WWW bearers to tolerate a higher interference. The above results somewhat contradict those presented in [3], where lower HS-DSCH power thresholds were found appropriate. The reason for this is the use of more robust voice bearers in this study.

It may be noted that with appropriate thresholds the capacity in fact exceeds the reference of 'preserving relative individual capacities', which corresponds to a straight line between single service endpoints. The high WWW capacities seen already for voice loads close to the single service voice capacity are due to that voice bearers even at these traffic loads do not always use the full power resource. On the contrary, in favorable fading situations for voice users, a large fraction of the base station power may still be available for the HS-DSCH, yielding significant WWW capacities. This phenomenon is particularly evident with the heavily fading Pedestrian channel type used. Noticeably, at the voice-only capacity limit, the average base station output power is only about 13 W.

It may be suspected that channel types yielding non-orthogonal intra-cell interference and/or less fading, would lead to different results than the above. The capacity evaluation is therefore in addition to the Pedestrian channel also done for a 3GPP Typical Urban channel. The results are shown in Fig. 10. Despite lower WWW capacities and a higher voice capacity in the single service endpoints, the results are quite similar to the Pedestrian case. An HS-DSCH power threshold of 20 W typically outperforms the 15 W threshold. However, it appears that the high WWW capacities close to the voice-only endpoint are now less evident. This is explained by the lower power left by the voice bearers with the Typical

Urban Channel, as compared to the more fading Pedestrian channel. With the Typical Urban channel the average base station power at the voice-only capacity limit is about 17 W.

Finally, in cases where the voice service limits overall capacity, not only the amount of interference introduced by the HS-DSCH, but also its rapid variation may affect the voice bearers. Tests with smooth ramping of the HS-DSCH power however show insignificant gains, indicating that the voice bearers can indeed cope with the interference variations caused by the HS-DSCH.

5 Conclusions

With release 5 of the WCDMA specifications, support for peak rates in the order of 10Mbit/s is introduced. This is enabled through the use of fast link adaptation, fast hybrid ARQ and fast scheduling. WCDMA is hence evolved well beyond the initial third generation requirements. Through appropriate management of radio resources (scheduling), the new high-speed packet data bearers perform well with TCP for WWW services. It is also possible to provide bearers for streaming services with high capacity. Finally, integration of the high-speed packet data bearers with other WCDMA bearers is straightforward.

References

[1] E. Dahlman et al., "WCDMA – The Radio Interface for Future Mobile Multimedia Communications", *IEEE Transaction on Vehicular Technology, vol. 47, no. 4, pp. 1105-1118. November 1998.*

[2] S. Parkvall, et al., "The Evolution of WCDMA Towards Higher Speed Downlink Packet Data Access", *in proceedings of IEEE VTC 2001 (Spring), May 2001.*

[3] S. Parkvall et al., "evolving WCDMA for Improved High-Speed Mobile Internet", in *proceedings of FTC'2001, Beijing, China, November2001.*

[4] J. Peisa et al., "TCP Performance over HS-DSCH", *in proceedings of IEEE VTC 2002 (Spring), May 2002.*

[5] H. Nyberg et al, "A Streaming Video Traffic Model for the Mobile Access Network" ", in *proceedings of IEEE VTC 2001 (Fall), USA, October 2001.*

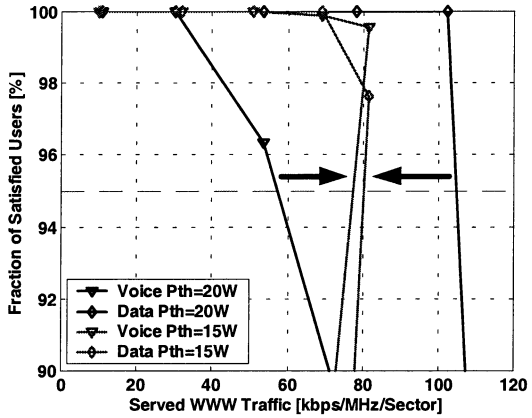


Fig. 7. Example of quality versus traffic load behavior and resource balancing.

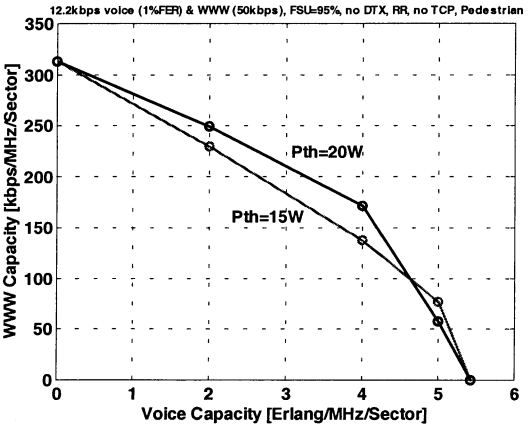


Fig. 8. Joint voice and WWW capacity regions for different HS-DSCH power thresholds, 50 kbps CSE bitrate requirement, Pedestrian channel.

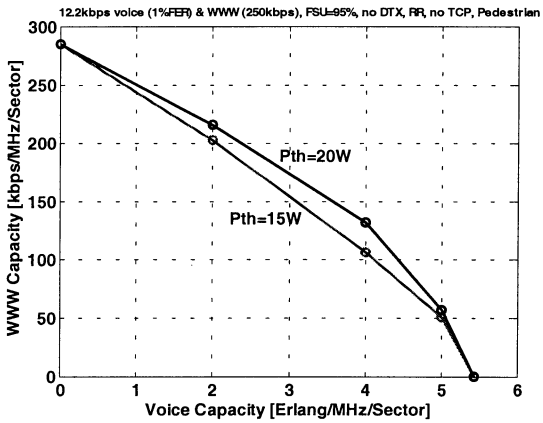


Fig. 9. Joint voice and WWW capacity regions for different HS-DSCH power thresholds, 250 kbps CSE bitrate requirement, Pedestrian channel.

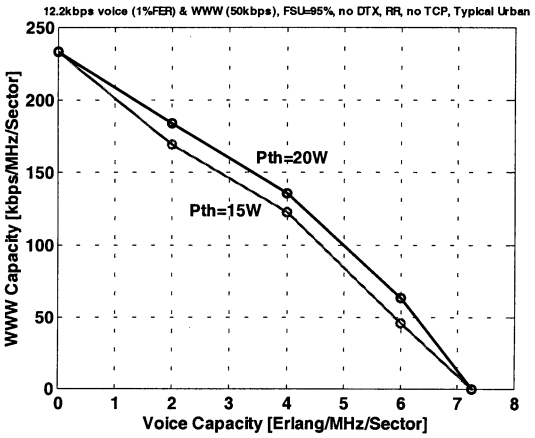


Fig. 10. Joint voice and WWW capacity regions for different HS-DSCH power thresholds, 50 kbps CSE bitrate requirement, Typical Urban channel.

Unbiased Power Prediction on Broadband Channels

Torbjörn Ekman[1], Mikael Sternad[2] and Anders Ahlén[2]

[1] Institutt for informatikk, Postboks 1080 Blindern, 0316 Oslo, Norway. E-mail: torbjorn.ekman@signal.uu.se

[2]Signals and Systems, Uppsala University, PO Box 528,SE-751 20 Uppsala, Sweden.

E-mail: mikael.sternad@signal.uu.se , anders.ahlen@signal.uu.se

Abstract

Prediction of the rapidly fading envelope of a mobile radio channel enables a number of capacity improving techniques like fast resource allocation or fast adaptive modulation. Some power prediction algorithms are based on linear prediction of the taps of the complex impulse response and then forming the power prediction as the sum of the absolute square of the prediction of the taps in the impulse response. This will render a biased power predictor, that generally underestimates the power. We propose a bias compensated power predictor and derive the optimal prediction coefficients for the Rayleigh fading channel. The importance of efficient noise reduction of channel samples that are used as inputs to the predictor is also emphasized. A performance evaluation of the prediction algorithm is carried out on measured broadband mobile radio channels.

1 Introduction

Prediction of the coefficients of mobile radio channel is of interest for a range of applications such as power control, adaptive resource allocation, as well as adaptive coding and modulation [1], [2]. Power control in e.g. WCDMA requires short-term prediction over only a small fraction of the distance between two dips in the short-term fading pattern. Radio resource allocation and planning would require accurate and more long-term prediction, the longer the better. Different methods have been proposed to cope with this problem [3], [4].

Here we will predict the received power of a mobile radio channel by prediction of the individual complex taps in the channel impulse response. The power can then be estimated by adding the squared magnitudes of the predicted taps as in [3]. That will, however, give a biased power predictor, which underestimates the power. We propose a compensation for the bias that improves the power prediction performance.

Estimated snapshots of the channel impulse response are used as input to a set of channel tap predictors. The SNR of these samples is improved by using Wiener smoothers, which improves the total performance. Our approach to prediction of the received power will be evaluated on measured broadband data.

2 The Measured Channels

2.1 The Estimation Method

Measurements obtained by Ericsson Research in urban and suburban areas, at the carrier frequency 1880 MHz and with a 6.4 MHz baseband sampling rate, have been utilized. The sampling rate of the channel is 9.14 kHz whereas the highest Doppler frequency is about 160 Hz. Thus we have a highly oversampled channel [6].

The channel can be modeled as a time varying FIR-filter, with taps denoted by $h_k(t)$. Let n denote a discrete time index at the symbol rate. For each measurement location, the time-varying complex-valued channel taps $h_k(n)$ in

$$s(n) = \sum_{k=1}^{n_k} h_k(n)u(n-k) + w(n) \qquad (1)$$

were estimated. Here, $s(n)$ is the received baseband signal sampled at 6.4 MHz, $u(n)$ is a known filtered transmitted sequence, obtained from back-to-back measurements of the receiver connected to the transmitter, and $w(n)$ is noise.

A channel length $n_k = 120$, corresponding to a time span of 18.75 μs, was found to be sufficient at all measurement locations. The channel was assumed time-invariant over blocks of 700 samples (channel sampling period $T =$ 109.4 μs $\approx$ 0.01 wavelength), an assumption that introduces negligible errors at these fading rates [5]. Block least squares estimates of the channel taps were calculated, resulting in 1430 channel samples, indicated by the time index t,

$$\bar{h}_k(t) = h_k(t) + v_k(t)\ ;\ k = 1, \ldots, 120\ ;\ t = 1, \ldots, 1430$$

covering 156 ms at each location. Here, $h_k(t)$ denote the noise-free tap samples, while the estimation error $v_k(t)$ is zero mean noise that is uncorrelated with $h_k(t)$.

2.2 Noise Reduction

We exploit the oversampling of the channel to reduce the noise with an FIR Wiener-smoother. The covariances of the channel taps $h_k(t)$ are in the filter design represented by a zero order Bessel function of the first kind (Jakes model) and the noise $v_k(t)$ is assumed to be white. The

Jakes model is a good description of the average short term fading of the channel taps, although the different realizations will deviate significantly from it. For channel taps $h_k(t)$ estimated by a least squares procedure, here based on blocks of 700 samples, the estimation error $v_k(t)$ will be close to white when considering consecutive snapshots.

For a given smoothing lag and filter order, a FIR Wiener smoother is parameterized by the SNR and the Doppler frequency. These entities can be estimated from the power delay profile and the average Doppler spectrum [6].

In the measurements at hand, the Doppler frequency normalized by the channel sampling frequency is on the order of 10 mHz. This corresponds to 100 samples per traveled distance measured in wavelengths. Smoothing (use of future channel samples) provides much better noise reduction performance than filtering only. The performance improves with an increasing smoothing lag. When using a smoother with 128 coefficients on a Jakes model with 10 mHz maximal Doppler shift, almost full performance is obtained already at a smoothing lag of 10 and the performance saturates at a smoothing lag of about 20.[1]

The use of a smoothing lag in the noise reduction introduces a corresponding delay in the data used by the predictor. This will affect the attainable predictor performance. To circumvent this problem, a bank of noise reduction smoothers are used, that have smoothing lags from zero up to a lag that is sufficient to obtain close to optimal noise reduction. Here, five smoothers with lags 0, 2, 5, 10 and 20 are used. The channel predictor, described in Section 3.3 below, can then always use regressor variables with appropriate time lags. These regressors have been given the best, at that time, available noise reduction.

2.2.1 The Wiener smoother

The correlation function for the channel in the Jakes model is given by the zero order Bessel-function of the first kind,

$$r_J(\tau) = \mathrm{J}_0(2\pi f_D \tau) \,, \tag{2}$$

where f_D is the Doppler frequency.

A FIR Wiener smoother with N coefficients and smoothing lag ℓ has the coefficient vector

$$\mathbf{w}^{[\ell]} = (\mathbf{R}_J + \mathbf{I}\sigma_e^2)^{-1}\mathbf{r}_J^{[\ell]} \,, \tag{3}$$

where $\mathbf{R}_J$ is the covariance matrix for a Jakes model channel, $\mathbf{r}_J^{[\ell]} = [r_J(-\ell) \ldots r_J(N-1-\ell)]^T$ while σ_e^2 is the variance of the channel estimation error normalized by the channel tap variance. The FIR-Wiener smoothing estimate with lag ℓ is thus

$$\tilde{h}_k^{[\ell]}(t-\ell|t) = \bar{\varphi}_k^H(t)\mathbf{w}^{[\ell]} \,, \tag{4}$$

[1] The performance of the FIR-smoother is similar to the IIR-smoother proposed in [4]. This FIR-smoother needs on the order of hundred of coefficients whereas the IIR-smoother needs only on the order of ten coefficients. Still, the FIR-smoother is here preferred due to numerical sensitivity of the coefficients in the IIR-smoother.

where $\bar{\varphi}_k^H(t)$ is a vector with current and delayed samples of the estimated channel

$$\bar{\varphi}_k(t) = \left[\bar{h}_k(t)\ \bar{h}_k(t-1) \ldots \bar{h}_k(t-N+1))\right]^H \,. \tag{5}$$

3 Channel Predictors

We first discuss prediction of complex channel taps $\bar{h}_k(t)$. In the considered measurements, different taps are only weakly correlated, so they will be predicted individually. We then consider tap power predictors that are quadratic functions of the noise reduced taps.

3.1 Linear Prediction of Complex Taps

Even though the mobile radio channel changes dramatically over short traveled distances, the parameterization of the amplitude and phase is slowly time varying. We can thus assume that the optimal linear predictor coefficients change slowly and without abrupt changes. In this section, we study the performance of a linear FIR-predictor for prediction of a complex time series with stationary statistics. This is a valid approximation for a block based method using channel statistics collected over a few meters.

Consider a flat Rayleigh fading channel. It can be modeled as a correlated complex circular Gaussian stochastic variable with zero mean, that is here denoted $x(t)$ instead of $h_k(t)$. Furthermore, assume the channel estimation error, here denoted $v(t)$, to be circular complex Gaussian distributed with zero mean.

In a vector formulation of the FIR-predictor, the signal (tap) can be predicted from past noisy observations as

$$\hat{x}(t+L|t) = \boldsymbol{\varphi}^H(t)\boldsymbol{\theta} \,, \tag{6}$$

where $\boldsymbol{\theta}$ is a vector with prediction coefficients and the components of the regressor $\boldsymbol{\varphi}(t)$ consist of smoothed observations of $x(t)+v(t)$ with different delays. The components of $\boldsymbol{\varphi}(t)$ will thus also be complex circular Gaussian stochastic variables. From the Wiener-Hopf equations, the MSE of a predictor like (6) is minimized by $\boldsymbol{\theta} = \boldsymbol{\theta}_c$ with

$$\boldsymbol{\theta}_c = \mathbf{R}_\varphi^{-1}\mathbf{r}_{x\varphi} \,, \tag{7}$$

where $\mathbf{r}_{x\varphi} = \mathrm{E}\{x(t)\boldsymbol{\varphi}(t-L)\}$ is the cross-correlation between the signal and the regressor, and $\mathbf{R}_\varphi = \mathrm{E}\{\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^H(t)\}$ is the correlation matrix for the regressors. With $r_x = E|x(t)|^2$ the minimal prediction MSE is

$$\mathrm{E}|x(t+L) - \hat{x}(t+L|t)|^2 = r_x - \boldsymbol{\theta}_c^H\mathbf{R}_\varphi\boldsymbol{\theta}_c \,. \tag{8}$$

3.2 Quadratic Power Prediction

3.2.1 Biased power prediction

From past and present noisy observations, the power of $x(t)$ at time $t+L$, that is $|x(t+L)|^2$, is to be predicted.

We first examine the use of $|\hat{x}(t+L|t)|^2$, with $\hat{x}(t+L|t)$ given by (6), as a power predictor,

$$\hat{p}_b(t+L|t) = |\hat{x}(t+L|t)|^2 = \boldsymbol{\theta}^H \boldsymbol{\varphi}(t)\boldsymbol{\varphi}^H(t)\boldsymbol{\theta} \,. \quad (9)$$

This predictor utilizes quadratic functions of the regressor variables and is thus a nonlinear predictor. This predictor for the power is used for channel prediction in e.g. [3] and [4]. For any $\boldsymbol{\theta}$, the average power prediction error will be

$$\mathrm{E}\{\varepsilon_{p_b}(t)\} = \mathrm{E}\{|x(t)|^2 - |\hat{x}(t|t-L)|^2\} = r_x - \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta}, \quad (10)$$

where $r_x = \mathrm{E}\{|x(t)|^2\}$ is the variance (power) of the channel. The absolute square of the complex prediction is thus a *biased power predictor*. With $\boldsymbol{\theta} = \boldsymbol{\theta}_c$, this bias becomes equal to the variance (8) of the optimized complex prediction error. The power of $x(t)$ is thus underestimated.

3.2.2 Unbiased power prediction

To avoid underestimation of the power, we propose an unbiased power predictor, that modifies (9) with a compensation for the bias, as

$$\hat{p}(t+L|t) = |\hat{x}(t+L|t)|^2 + r_x - \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta} \,. \quad (11)$$

The prediction error $p(t) - \hat{p}(t|t-L)$ will then have zero mean. The modification is especially noticable when the complex channel is hard to predict, i.e. when $\boldsymbol{\theta} \to 0$. The unbiased prediction then becomes the average power r_x, whereas the biased predictor will predicted zero power. The proposed unbiased power predictor thus offers a graceful transition from perfectly predicted channel states to situations where only the average power is known.

3.3 Unbiased Power Predictors for Frequency Selective Channels

Tools from the preceding sections can now be combined. The linear regression

$$\hat{h}_k(t+L) = \boldsymbol{\varphi}_k^H(t)\boldsymbol{\theta}_k \,, \quad (12)$$

represents an L step ahead FIR-prediction of a channel tap $h_k(t)$. Here, $\boldsymbol{\theta}_k$ is a vector of predictor coefficients. The regression vector $\boldsymbol{\varphi}_k(t)$ consists of smoothed samples (4) of tap k up to time t, with a sub-sampling factor of m, where the largest available smoothing lag is used:

$$\boldsymbol{\varphi}_k(t) = \left[\tilde{h}_k^{[0]}(t|t)\ \tilde{h}_k^{[m]}(t-m|t) \ldots \right. \\ \left. \ldots \tilde{h}_k^{[(M-1)m]}(t-(M-1)m|t)\right]^H . \quad (13)$$

Here M denotes the order of the predictor (12). If $(M-1)m$ is larger than the largest available smoothing lag, then the corresponding smoothed signal with the largest smoothing lag is used instead.

The unbiased power predictor (11) is then given by

$$\hat{p}_k(t+L|t) = |\hat{h}_k(t+L|t)|^2 + r_{h_k} - \boldsymbol{\theta}_k^H \mathbf{R}_{\varphi_k} \boldsymbol{\theta}_k \,, \quad (14)$$

where $r_{h_k} = \mathrm{E}\{|h_k(t)|^2\}$ and $\hat{h}_k(t+L|t)$ is given by (12).

Each tap of a frequency selective channel can be regarded as a flat fading channel. For a Rayleigh fading tap, that is $h_k(t) \sim \mathbf{CN}(0, r_{h_k})$, an optimal choice of $\boldsymbol{\theta}_k$ is then given in Appendix A as

$$\boldsymbol{\theta}_k = \mathbf{R}_{\varphi_k}^{-1} \mathbf{r}_{h_k \varphi_k} \,, \quad (15)$$

where $\mathbf{R}_{\varphi_k}$ is the covariance matrix for the regressor to tap k and $\mathbf{r}_{h_k\varphi_k} = \mathrm{E}\{h_k(t)\boldsymbol{\varphi}_k(t)\}$. Somewhat remarkably, this is the same adjustment as (7), which optimizes the L-step prediction of the complex tap.[2]

An unbiased predictor for the total power is finally obtained by summing contributions from each significant tap

$$\hat{P}(t+L|t) = \sum_{k=1}^{n_k} \left(|\hat{h}_k(t+L|t)|^2 + r_{h_k} - \boldsymbol{\theta}_k^H \mathbf{R}_{\varphi_k} \boldsymbol{\theta}_k \right) . \quad (16)$$

If all channel taps are unpredictable, then the channel power will be predicted by its average $\sum_k r_{h_k}$. In receivers that use only a subset of the taps, the available power is the sum over this subset. The predicted available power is then the sum over the predictions for this subset, whereas the other taps act as interference.

4 Results

4.1 Simulations

The expected performance for prediction of a flat Rayleigh fading channel with Jakes spectrum will indicate the performance on true measured channels. We thus first examine the Jakes model to obtain reasonable choices for the delay spacing (the sub-sampling factor m measured in wavelengths) in predictors for one tap. The channel correlation is then given by, $r_x(\tau) = r_x \mathrm{J}_0(2\pi f_D \tau)$. The delay spacing and the prediction range are measured in wavelengths, as fading is a spatial phenomenon. The time-frequency product $f_D\tau = v\tau/\lambda$, where v is the speed of the mobile, is the traveled distance measured in wavelengths.

For a channel corrupted by white noise, the theoretically obtainable prediction NMSE can be calculated, given the prediction range, the order of the predictor and the delay spacing. This can also be done for a predictor using smoothed regressors. Our finding is that when using the smoothed regressors the performance is rather insensitive to the choice of delay spacing, as long as it is in the range of 10 to 20 samples per traveled wavelength.

Figure 1 shows how predictable a Jakes channel is with and without smoothed regressors. The predictor has eight coefficients and a robust choice of delay spacing, that gives close to optimal performance over SNRs $E|h_k(t)|^2/Ev_k(t)^2$ ranging from 10 to 50 dB.

[2] All $\boldsymbol{\theta}_k$ that satisfy $\boldsymbol{\theta}_k = e^{i\omega}\boldsymbol{\theta}_c$, with $\omega \in [0\ 2\pi]$, are solutions, since the phase becomes irrelevant when the absolute square is taken.

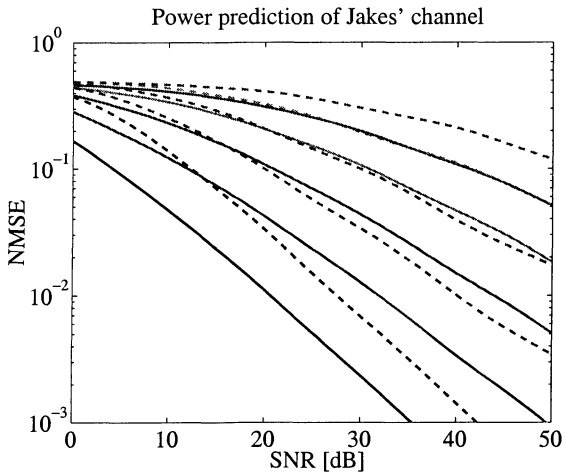


Figure 1: The prediction NMSE of an unbiased quadratic power predictor for a Jakes channel as a function of SNR. Solid lines use smoothed regressors, while dashed lines use noisy regressors. The uppermost curves corresponds to a prediction range L of 0.5 wavelengths. The prediction range decreases to 0.1 for the two consecutive lower curves.

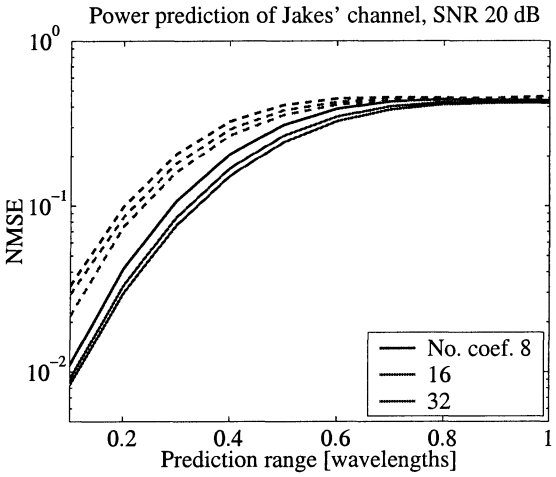


Figure 2: The prediction NMSE at 20dB SNR of an unbiased quadratic power predictor for a Jakes channel as a function of the prediction range. Solid and dashed lines are using smoothed and noisy regressors respectively. The uppermost curves correspond to a predictor with 8 coefficients, the next to 16 and the lowest to 32 coefficients.

An increase of the number of coefficients in the predictor results in only a minor decrease in the NMSE, see Figure 2. The major gain is attained by the use of noise reduction.

4.2 Power Prediction on Measured Channels

As only noisy observations of the channel taps are available, the true tap is taken as the smoothed tap obtained with smoothing lag 20. This is a good approximation of the true taps for SNRs above 0 dB. A total of 37 measurement locations are used in the evaluation. Only measurements where the average channel-to-estimation-error power ratio is above 10 dB are used.

The algorithms are designed for channels with stationary statistics. If abrupt changes are detected new prediction parameters would have to be estimated. To avoid such situations in the evaluation, a few channels where the average power changed by more than 3 dB from the first quarter of the data to the last quarter, were ruled out.

4.2.1 Quadratic power prediction

For each tap the estimated sample covariances $\hat{\mathbf{R}}_{\varphi_k}$ and $\hat{\mathbf{r}}_{h_k\varphi_k}$ are used to obtain the prediction coefficients (15) of the unbiased power predictor (14). The prediction of the total power is obtained as in (16). The same holds for the biased power predictor (9) for comparison.

Independent of the prediction range, 900 samples are used in the target vector for the training. The validation interval was on the order of 400 down to 100 samples, depending on prediction range and Doppler frequency.

The delay spacing was taken as the robust choice, as obtained from simulations using the Jakes model, depending on the Doppler frequency but not on the SNR. A predictor using eight coefficients was used, as simulations show that little can be gained by increasing the number of coefficient above this value if noise reduction is performed in a separate step, cf Figure 2.

4.2.2 Average and last sample prediction

The two simplest power predictors, the average power and the last sample predictor, are also included in the evaluation. The average power is estimated using the same training interval as for the more advanced predictors and it is evaluated on the same validation interval.

4.2.3 Performance

The average NMSE of the power prediction algorithms in Figure 3 shows that the unbiased power predictor provides the best performance for all prediction horizons measured over the validation set. As the horizon L increases, its NMSE slowly approaches that of the average predictor from below. The biased power predictions and the last sample predictions become slightly worse than using the average power at large prediction ranges.

5 Conclusions

Using smoothed noise reduced regressors result in a significant increase of the prediction performance on channels described by the Jakes model. The combined effect of improved noise reduction that eliminates the lag in the regres-

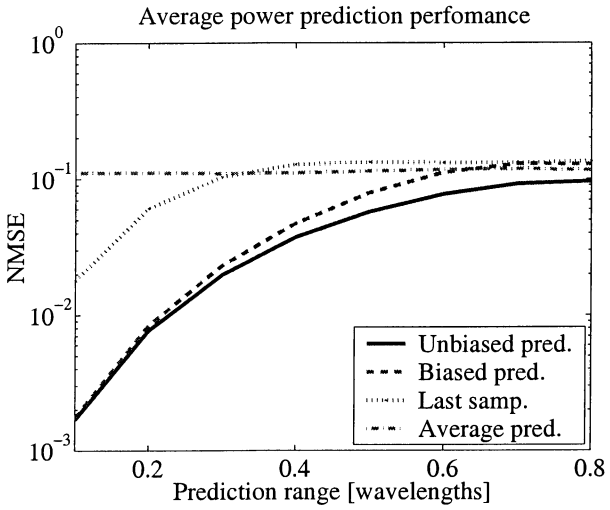


Figure 3: The average power prediction normalized MSE evaluated at 37 measurement locations.

sors and unbiased power estimation has resulted in significant performance improvements, as compared to our previous results reported in [4]. The proposed unbiased power predictor offers a graceful transition from perfect channel state information to just knowledge of the average power. This is demonstrated in evaluations of the algorithm on measured channels.

A Appendix

For convenience, the time index is dropped for $x(t)$, $\hat{x}(t|t-L)$ and $\varphi(t)$. The covariance of $x(t)$ is $\mathrm{E}\{|x|^2\} = r_x$ and the covariance for the complex prediction is

$$r_{\hat{x}} = \mathrm{E}\{|\hat{x}|^2\} = \mathrm{E}\{\boldsymbol{\theta}^H \boldsymbol{\varphi}\boldsymbol{\varphi}^H \boldsymbol{\theta}\} = \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta} \,. \quad (17)$$

The cross-correlation between x and the prediction $\hat{x}$ is

$$r_{x\hat{x}} = \mathrm{E}\{x\hat{x}^*\} = \mathrm{E}\{\boldsymbol{\theta}^H \boldsymbol{\varphi} x\} = \boldsymbol{\theta}^H \mathbf{r}_{x\varphi} \,. \quad (18)$$

When using the biased power predictor (9) the MSE for the power prediction error in (10) becomes

$$\begin{aligned} \mathrm{E}\{|\varepsilon_{p_b}|^2\} &= \mathrm{E}\{(|x|^2 - |\hat{x}|^2)^2\} \\ &= \mathrm{E}\{|x|^4 - 2|x|^2|\hat{x}|^2 + |\hat{x}|^4\} \,. \end{aligned} \quad (19)$$

Examine the three terms of (19) separately. As both x and $\hat{x}$ are circular complex Gaussian the first and third terms are $2r_x^2(0)$ and $2r_{\hat{x}}^2(0)$ respectively. For the same reason we can express the second term as

$$\begin{aligned} \mathrm{E}\{|x|^2|\hat{x}|^2\} &= \mathrm{E}\{|x|^2\}\mathrm{E}\{|\hat{x}|^2\} + \mathrm{E}\{x\hat{x}^*\}\mathrm{E}\{x^*\hat{x}\} \\ &= r_x r_{\hat{x}} + |r_{x\hat{x}}|^2 \,. \end{aligned} \quad (20)$$

The MSE (19) for the biased power predictor is thus

$$\begin{aligned} &\mathrm{E}\{|\varepsilon_{p_b}|^2\} = \quad (21) \\ &2r_x^2 - 2\boldsymbol{\theta}^H \mathbf{r}_{x\varphi}\mathbf{r}_{x\varphi}^H \boldsymbol{\theta} - 2r_x \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta} + 2|\boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta}|^2 . \end{aligned}$$

This is the criterion that should be minimized to obtain the coefficient vector $\boldsymbol{\theta}$ that provides the best power predictor of the type (9).

The MSE for the unbiased power predictor is obtained from (21) by subtracting the square of the bias,

$$\begin{aligned} \sigma_{\varepsilon_p}^2 = \mathrm{E}\{|\varepsilon_p|^2\} &= \mathrm{E}\{(|x|^2 - |\hat{x}|^2)^2\} - |r_x - \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta}|^2 \\ &= r_x^2 - 2\boldsymbol{\theta}^H \mathbf{r}_{x\varphi}\mathbf{r}_{x\varphi}^H \boldsymbol{\theta} + |\boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta}|^2 \end{aligned} \quad (22)$$

A.1 Optimal coefficients

To find the $\boldsymbol{\theta}$ that minimizes $\sigma_{\varepsilon_p}^2$ in (22), we set the partial derivative of the MSE with respect to $\boldsymbol{\theta}$ equal to zero

$$\begin{aligned} 0 &= \frac{\partial \sigma_{\varepsilon_p}^2}{\partial \boldsymbol{\theta}} \quad (23) \\ &= -2\boldsymbol{\theta}^H \mathbf{r}_{x\varphi}\mathbf{r}_{x\varphi} + \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta}\boldsymbol{\theta}^H \mathbf{R}_\varphi + \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta}^H \mathbf{R}_\varphi \boldsymbol{\theta} \,, \end{aligned}$$

which result in the equation

$$\boldsymbol{\theta}^H (\mathbf{r}_{x\varphi}\mathbf{r}_{x\varphi}^H - \mathbf{R}_\varphi \boldsymbol{\theta}\boldsymbol{\theta}^H \mathbf{R}_\varphi) = 0 \,. \quad (24)$$

Here, $\boldsymbol{\theta} = 0$ is the trivial solution. The other solutions are given by

$$\boldsymbol{\theta}\boldsymbol{\theta}^H = \mathbf{R}_\varphi^{-1} \mathbf{r}_{x\varphi}\mathbf{r}_{x\varphi}^H \mathbf{R}_\varphi^{-1} \,, \quad (25)$$

with one obvious solution as $\boldsymbol{\theta} = \boldsymbol{\theta}_p$, where

$$\boldsymbol{\theta}_p = \mathbf{R}_\varphi^{-1} \mathbf{r}_{x\varphi} \,. \quad (26)$$

These are the same coefficients that are optimal for the complex prediction (6).

References

[1] S. Chua and A. Goldsmith, "Variable-rate variable-power MQAM for fading channels," *IEEE Vehich. Tech. Conf.*, Atlanta, GA, May 1996, pp 815-819.

[2] N.C. Ericsson, "Adaptive modulation and scheduling of IP-traffic over fading channels," *IEEE Vehicular Technology Conference- VTC'99-Fall*, Amsterdam, September 19-22, 1999, pp 849-853.

[3] A. Duel-Hallen, S. Hu and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Processing Magazine* vol. 17, pp 62-75, May 2000.

[4] M. Sternad, T. Ekman and A. Ahlén, "Power prediction on Broadband Channels," *IEEE Vehich. Tech. Conf.- VTC'2001-spring*, Rhodes, Greece, May 2001.

[5] T. Ekman, "Analysis of the LS Estimation Error on a Rayleigh Fading Channel," *IEEE Vehich. Tech. Conf.- VTC'2001-spring*, Rhodes, Greece, May 2001.

[6] T. Ekman , "Prediction of Mobile Radio Channels," *Licentiate thesis*, Uppsala University, Dec. 2000.

Uplink Load Estimation in WCDMA

Erik Geijer Lundin[1], Fredrik Gunnarsson[2] and Fredrik Gustafsson[3]

[1]Linköpings universitet, 581 83 LINKÖPING, SWEDEN. E-mail: geijer@isy.liu.se

[2]Linköpings universitet, 581 83 LINKÖPING, SWEDEN. E-mail: fred@isy.liu.se

[3]Linköpings universitet, 581 83 LINKÖPING, SWEDEN. E-mail: fredrik@isy.liu.se

Abstract

All cellular radio systems depend on well performing radio resource management algorithms to efficiently utilize available resources. These algorithms make their decisions based on some sort of resource quantity. In the uplink of a WCDMA system, the total received power at the base stations is a natural choice of such a quantity. Unfortunately this quantity cannot be measured with enough accuracy to be used by resource management algorithms. One way of getting around this problem is to estimate a closely related quantity, namely the uplink noise rise. In this paper, we propose and evaluate four different methods to estimate the noise rise. The estimates are insensitive to users' bit rate and use data readily available in the system. The best performing method estimates, on average, the uplink noise rise with an error of less than 1 dB for practical load scenarios.

1 Introduction

Crucial for obtaining good utilization of available radio resources are well operating radio resource management algorithms. Such algorithms rely on knowledge of some sort of resource quantity. Examples of such quantities are the number of active users in a cell, which is an example of capacity measures sometimes referred to as *hard capacity* [1], and total received uplink power which is an example of a *soft capacity* measure. The main difference is that the latter considers the effects of the radio environment and of multiple cells. Soft capacity is used for uplink admission control in [1, 2, 3, 4]. In this paper we will partially focus on the uplink noise rise, i.e., total received interference power, I over background noise power, N. Another quantity, closely related to the uplink noise rise is the relative uplink load of the system. In [1, 2, 5, 6, 7] the uplink load, L is related to the uplink noise rise Λ by

$$L = 1 - \frac{1}{\Lambda} \Leftrightarrow \Lambda = \frac{I}{N} = \frac{1}{1-L} \qquad (1)$$

Clearly, $L = 0$ implies a noise rise of 1 and hence the total uplink interference is simply the background noise, and as L tends to one, the uplink interference power tends to infinity. As can be seen in Figure 1, it is not recommended having a working point at a too high load level since at high load levels even a small increase in uplink load corresponds to a substantial interference power increase. Such an increase may, under some circumstances, require unpleasant actions. A noise rise above 4 or 5 dB is often not considered feasible given power control and maximum transmission power constraints. A noise rise of 4 dB corresponds to an uplink load of 0.6, which is relatively far from what can be thought of as a theoretical max, somewhere just below one. The relation can for example be useful for an operator wanting to maintain the noise rise of the system below some fixed level. He can then use the simple one-to-one relation and find the corresponding uplink load. This level can be used in various resource management algorithms independently of users' data rate, path loss coefficients and users' location.

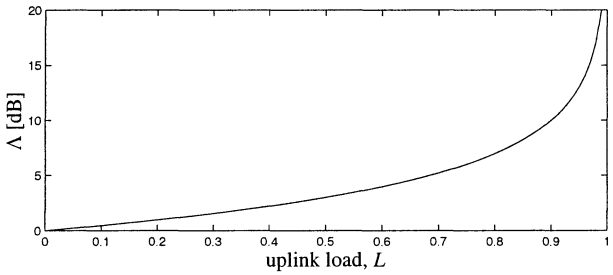


Figure 1: Nonlinear relation between uplink load, L and uplink noise rise.

An overview of the rest of this paper is as follows. Section 2 presents relevant models and notation. These are utilized in Section 3, where we derive estimates for the uplink noise rise and discuss the strength of the estimates. Section 4 discusses some of the sources of the estimation errors. In Section 5 the estimates are evaluated using simulations. Finally, Section 6 gives some conclusive remarks.

2 System Model

Consider a scenario in the uplink of a cellular radio system where we have B base stations (different sectors at the same cell is herein denoted as different base stations) and M active mobiles. Denote the transmission power of user i with $p_i(t), i \in \{1, 2, ..., M\}$. The signal power used by user i is attenuated by a factor $g_{i,j}(t)(< 1)$ before arriving at base

station $j, j \in \{1,2,...,B\}$. The received carrier power of user i at base station j is then simply $C_{i,j}(t) = p_i(t)g_{i,j}(t)$. This signal is added to the other users' signals and background noise in the base station to form the *total interference power* (total received power):

$$I_j^{tot} = N_j + \sum_{i=1}^{M} p_i(t)g_{i,j}(t) \quad (2)$$

Assume that mobile i is connected solely to base station j_i. The perceived quality is related to the *carrier-to-interference-ratio* (CIR) defined by $\gamma_i(t) = C_{i,j_i}(t)/(I_{j_i}^{tot}(t) - C_{i,j_i}(t))$. Power control in WCDMA aims at maintaining $\gamma_i(t)$ equal to a user dependent *target CIR*, $\gamma_i^{tgt}(t)$ [8]. These targets are continuously reconsidered by *radio resource management algorithms*. For presentational ease we will from here on use the *carrier-to-total-interference-ratio* (CTIR), $\beta_i(t) = C_{i,j_i}(t)/I_{j_i}^{tot}(t)$ instead of $\gamma_i(t)$[1]. In the system considered here, a mobile may be connected to several base stations at different physical sites (soft handover) or to several base stations at the same site (softer handover). Utilizing maximum ratio combining of the received signals in softer handover yields:

$$\beta_i(t) = \sum_{k \in K_i} \frac{g_{i,k}(t)p_i(t)}{I_k^{tot}(t)} \quad (3)$$

where K_i is the set of base stations user i is connected to. Conversely, in the case of soft handover *selection combining* is used and consequently the signal with the best quality measure is chosen, i.e.

$$\beta_i(t) = \max_{k \in K_i} \frac{g_{i,k}(t)p_i(t)}{I_k^{tot}(t)} \quad (4)$$

Note that both soft and softer handover are used concurrently in a true system as well as in the simulator. Maximum in (4) will be taken over the results from the local softer handover combinings if only a subset of the active links are in softer handover.

3 Uplink Load Estimates

In this section we derive estimates for the uplink load and state some fundamental requirements. All the estimates use only data readily available in a WCDMA system. Throughout this section the time index t will be suppressed for notational ease when not needed.

3.1 Approximations and Estimates

In order to derive the estimates a number of assumptions have to be made. First, assume that the expression for β_i found in (3) when in softer handover is approximately true for soft handover as well. The assumption allows us to write

$$\beta_i = p_i \sum_{k \in K_i} \frac{g_{i,k}}{I_k^{tot}} \Leftrightarrow p_i = \frac{\beta_i}{\sum_{k \in K_i} \frac{g_{i,k}}{I_k^{tot}}} \quad (5)$$

[1]Note the relation $\beta_i = \gamma_i/(1+\gamma_i)$.

Inserting the above expression for p_i into (2) and assuming perfect power control ($\beta_i = \beta_i^{tgt}$) yield

$$I_j^{tot} = N_j + \sum_{i=1}^{M} g_{i,j}p_i = N_j + \sum_{i=1}^{M} g_{i,j} \frac{\beta_i^{tgt}}{\sum_{k \in K_i} \frac{g_{i,k}}{I_k^{tot}}} \quad (6)$$

In case of soft handover, we get

$$I_j^{tot} = N_j + \sum_{i=1}^{M} g_{i,j}p_i = N_j + \sum_{i=1}^{M} g_{i,j} \frac{\beta_i^{tgt}}{\max_{k \in K_i} \frac{g_{i,k}}{I_k^{tot}}} \quad (7)$$

At this stage there are two different ways of proceeding in order to solve for I_j^{tot}.

- Assume that I_k^{tot} approximately equals I_j^{tot}, i.e. we assume the uplink interference power to be approximately the same in all neighboring cells. With this approximation, (6) may be rearranged into

$$I_j^{tot} = \frac{N_j}{1 - \sum_{i=1}^{M} \beta_i^{tgt} \frac{g_{i,j}}{\sum_{k \in K_i} g_{i,k}}} \quad (8)$$

  Comparing this expression to (1), the sum over the users may be identified as an estimate of the uplink load in cell j, L_j. Hence, estimates of the uplink load and noise rise are

$$\hat{L}_j^1 = \sum_{i=1}^{M} \beta_i^{tgt} \frac{g_{i,j}}{\sum_{k \in K_i} g_{i,k}} \Leftrightarrow \hat{\Lambda}_j^1 = \frac{1}{1 - \hat{L}_j^1} \quad (9)$$

- Assuming $N_j = N_k$, dividing equation (6) by N_j and substituting $\frac{I_j^{tot}}{N_j}$ with Λ_j yield

$$\Lambda_j(t) = 1 + \sum_{i=1}^{M} \beta_i^{tgt}(t) \frac{g_{i,j}(t)}{\sum_{k \in K_i} \frac{g_{i,k}(t)}{\Lambda_k(t)}} \quad (10)$$

  By assuming $\Lambda_k(t)$ to approximately equal $\Lambda_k(t-1)$ this too becomes a practically tractable estimate. (the "2" indicates the second estimate)

$$\hat{\Lambda}_j^2(t) = 1 + \sum_{i=1}^{M} \beta_i^{tgt}(t) \frac{g_{i,j}(t)}{\sum_{k \in K_i} \frac{g_{i,k}(t)}{\hat{\Lambda}_k^2(t-1)}} \quad (11)$$

  According to equation (1) yet another estimate of L can be found from (11):

$$\hat{L}_j^2 = 1 - \frac{1}{\hat{\Lambda}_j^2} \quad (12)$$

If we, instead of assuming softer handover, assume soft handover is used everywhere we get β_i according to (4) instead of (3) and hence the sum over k in (5) will be substituted with taking maximum. The corresponding estimates would then be

$$\hat{L}_j^3 = \sum_{i=1}^{M} \beta_i^{tgt} \frac{g_{i,j}}{\max_{k \in K_i} g_{i,k}} \Leftrightarrow \hat{\Lambda}_j^3 = \frac{1}{1 - \hat{L}_j^3} \quad (13)$$

and

$$\hat{\Lambda}_j^4(t) = 1 + \sum_{i=1}^{M} \beta_i^{tgt} \frac{g_{i,j}}{\max_{k \in K_i} \frac{g_{i,k}}{\hat{\Lambda}_k(t-1)}}$$

$$\hat{L}_j^4 = 1 - \frac{1}{\hat{\Lambda}_j^4} \tag{14}$$

4 Sources of Estimation Error

A number of different effects and approximations degrade the performance of the load estimate. These include

- the soft(er) handover assumption
- imperfect power control
- the assumption regarding relation between the uplink interference in all cells, alternatively the background noise and the dynamics of the uplink interference
- path gain measurement errors according to [9] and sparsely sampled path gain measurement reports
- the assumption that unknown path gains are equal to zero
- TX rise (see below)

The effects on the estimation error are explored in Section 5.

4.1 TX Rise

In this section we define and discuss *TX rise*. For simplicity, assume that each user is allowed to connect to no more than one base station at a time in this section. TX rise is the result of the fast power control adjusting each user's transmission power to compensate for fast fading dips in the path gain between mobile and connected base station. The mobile uses a high power during deep fades. This may yield a considerable interference addition to neighboring cells since a dip to the connected base station does not always imply an, as deep, dip to the neighboring base stations.

As an example, we consider mobile 1, which is connected to base station 2, and study its interference contribution to base station 3. Due to power control, each user's transmission power is approximately inversely proportional to the momentarily path gain between the user and the controlling base station, i.e.,

$$p_1 \propto \frac{1}{g_{1,2}}$$

Base station 3 will receive the signal $p_1 g_{1,3}$ which is proportional to the ratio between the momentarily path gains, i.e.,

$$C_{1,3} \propto \frac{g_{1,3}}{g_{1,2}}$$

The path gain reports provided by the users do not include fast fading (the actual reported value is low pass filtered; $F\{g_{i,j}\}$). Since the estimates, unlike fast power control, use users' path gain reports there will be a part of the uplink interference power that will not be counted for in the estimates. Essentially, we can observe $\frac{F\{g_{1,3}\}}{F\{g_{1,2}\}}$ but not $\frac{g_{1,3}}{g_{1,2}}$, which is related to the actual interference contribution in base station 3. In the top graph of Figure 2 we have plotted the ratio of low-pass filtered versions of the path gain measurements (solid). Of course this will not detect the high peaks in the ratio of the momentarily path gains (dashed). Perhaps momentarily sharp peaks are tolerable by base station 3. A better model could therefore be to consider the low-pass filtered load contribution which is essentially proportional to $F\{\frac{g_{1,3}}{g_{1,2}}\}$. However, as indicated by the bottom graph in Figure 2, the load contribution is still underestimated. In fact, this exemplifies that

$$\frac{E\{g_{i,j}\}}{E\{g_{i,k}\}} \neq E\{\frac{g_{i,j}}{g_{i,k}}\}$$

The solid line in the bottom graph of Figure 2 thus repre-

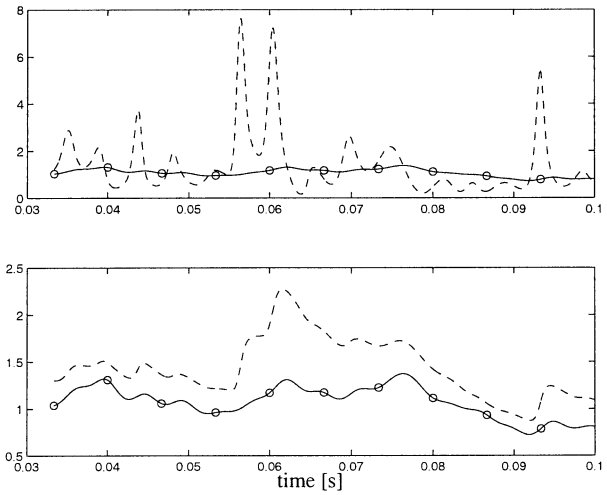


Figure 2: top: Ratio between filtered measurements (solid) and momentary $\frac{g_{1,3}}{g_{1,2}}$ (dashed) bottom: as above but comparing to a filtered ratio, $F\{\frac{g_{1,3}}{g_{1,2}}\}$

sents the quantity used in our estimates of the average noise rise while the dashed line represents the quantity that corresponds to user i's actual contribution to the average uplink noise rise in base station j.

5 Simulations

Simulations have been used to investigate the error of the different estimates at different load levels. Recall the discussion in the introduction, where we pointed out the one-to-one relation between uplink load and noise rise. It is therefore irrelevant if we discuss performance in terms of load or noise rise. We will choose the one which best

shows the interesting facts. To maintain simulation times to a moderate level, we have used two types of streaming users; 64 kbps and 192 kbps. The simulated environment is a suburban situation with 21 base stations, slow moving users (5 km/h), path loss coefficient $\alpha = 3.5$, shadow fading ($d_0 = 100$ m, $\sigma = 6$ dB), fast fading and RAKE receivers. We have used uniformly distributed users throughout the simulations. Measurements of path gains are subject to measurement errors according to [9], i.e. a 90% confidence interval of 1.5 dB. To keep signaling to a minimum only path gain measurement to the five strongest base stations are reported by each user. In [10] it is shown that $\hat{L}^1$ and $\hat{L}^2$ are practically independent of the type of path gain reporting, periodic or event-based, and therefore only event-based reports are considered. To isolate the sources of estimation error considered here, the users have no upper transmission power limit.

5.1 Single Service

Figure 3 shows the performance of the four different estimates when using 64 kbps streaming users. As each user has no more than one link, assuming soft or softer handover does not make any difference in this scenario. The major gap between the lines is due to TX rise, since fast fading is not simulated in the runs corresponding to the dashed lines. From this figure we can conclude that TX rise does give a considerable contribution to the total error. Also, all four estimates produce an underestimate of the uplink noise rise, mainly due to that measurements of all path gains are not available or slightly inaccurate. In Figure 4 a more realis-

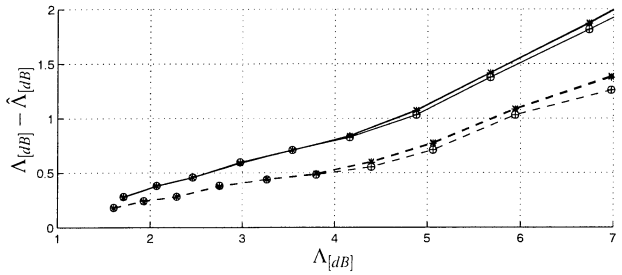


Figure 3: Noise rise estimate error with 64 kbps users. One link per user. Dashed: No fast fading. $\hat{L}^1$: 'o', $\hat{L}^2$ 'x', $\hat{L}^3$: '+', $\hat{L}^4$: '*'

tic scenario is shown where each user is allowed up to three handover links at a time. First of all, assuming softer handover ($\hat{L}^1$ and $\hat{L}^2$) gives larger average error than assuming soft handover ($\hat{L}^3$ and $\hat{L}^4$). Under these circumstances, i.e., almost equal load in all cells, the errors from all four estimates are low in average and has comfortably small standard deviation. Assuming $I_k^{tot}(t) = I_k^{tot}(t-1)$ (i.e., $\hat{L}^2$ and $\hat{L}^4$) provides estimates with higher average error but with lower standard deviation. In these estimates, we use the previous estimate to estimate each user's contribution to the noise rise. By looking at equation (11), which is repeated below, we realize that underestimating the noise rise at time

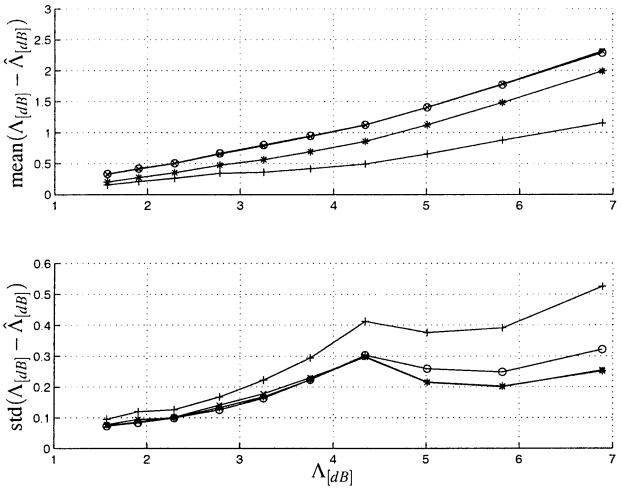


Figure 4: Noise rise estimate error with 64 kbps users. Up to three links per user. Upper: mean. Lower: standard deviation. $\hat{L}^1$: 'o', $\hat{L}^2$ 'x', $\hat{L}^3$: '+', $\hat{L}^4$: '*'

$t-1$ results in underestimating every term in the sum that constitutes the noise rise at time t.

$$\hat{\Lambda}_j(t) = 1 + \sum_{i=1}^{M} \beta_i^{tgt} \frac{g_{i,j}}{\sum_{k \in K_i} \frac{g_{i,k}}{\hat{\Lambda}_k(t-1)}}$$

The lowest line in the upper graph of Figure 4 represents $\hat{L}^3$. We can see that it estimates the noise rise with an error less than 1 dB on average for as high true noise rise as 6 dB. The slight increase in error for $\hat{L}^3$ as the noise rise increases, may be explained by looking in the load domain. As mentioned in the introduction, even a small difference in load gives a substantial increase in noise rise at high load levels. This also means that a small error in the uplink load estimate results in a large error in the noise rise estimate when operating at high load levels. In Figure 5 estimated load is plotted versus true load computed according to equation (1) where Λ is given by the simulator. In this domain we see no dramatic increase in estimation error as the load of the system increases. All the lines in the figure are quite straight but do not have the correct incline. This can best be explained by studying the left half of equation (13). Ideally, the estimate is a sum over all the users in the network. The fact that the line representing $\hat{L}^3$ in Figure 5 does not have the correct incline thus corresponds to that each user's contribution is underestimated. The behavior is more visible when looking at the other estimates, especially $\hat{L}^1$ and $\hat{L}^2$. Also $\hat{L}^2$ and $\hat{L}^4$ have a sum over, ideally, all the users, shown in (11), and hence the same reasoning applies. To summarize, there are two reasons why the lines in Figure 4 deviates from a straight line for high load levels:

- even a minor error at high loads produces a nonnegligible error in the noise rise
- the sum representing the load estimate has a larger

error due to many users, all with a minor underestimated contribution.

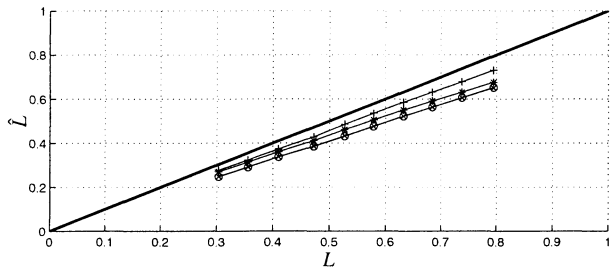


Figure 5: Uplink load estimate error when using 64 kbps users. $\hat{L}^1$: 'o', $\hat{L}^2$ 'x', $\hat{L}^3$: '+', $\hat{L}^4$: '*'

5.2 Multiple Services

In this section we will illustrate the estimates' insensitivity to users' data rate, represented by β_i^{tgt}. To not make the figure too cluttered, we have chosen to exemplify with $\hat{L}^3$ and $\hat{L}^4$. The load estimates with two different types of users are depicted in Figure 6. Both estimates are practically independent of the users' data rate. This is a great benefit compared to hard capacity, where each service has to be translated into a number related to the load contribution.

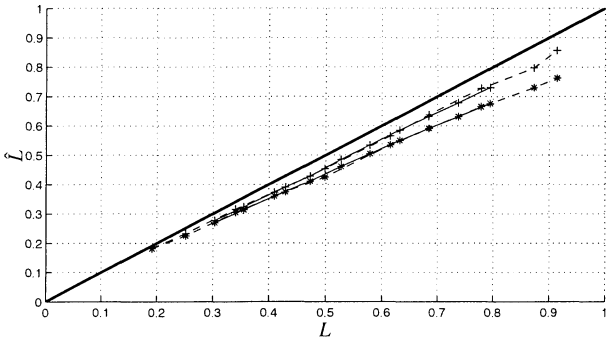


Figure 6: Uplink load estimate error when using 64 kbps users (solid) and 192 kbps users (dashed). $\hat{L}^3$: '+', $\hat{L}^4$: '*'

6 Conclusions

Uplink noise rise is considered to be a relevant uplink load measure in WCDMA, but it is difficult to measure accurately in practice. In this work, we have proposed and evaluated four different estimates of the uplink noise rise of a WCDMA system. The estimates have different assumptions regarding soft handover and interference power dynamics and correlation between cells. We have shown that, given readily available data, the estimates provide good average accuracy even for high interference levels. The estimate assuming soft handover and that the interference power is approximately constant over a time period of ten milliseconds provides a stable estimate which has low average error in all scenarios simulated, below 1 dB for moderate interference levels. Furthermore, all estimates inherently handle multi-service scenarios well. Due to the low variation in the single estimate errors, a rather simple correction algorithm should provide even better estimates. The above mentioned properties of the estimates make them well suited for monitoring of the current radio network load as well as for radio resource management.

References

[1] H. Holma and J. Laakso. Uplink admission control and soft capacity with MUD in CDMA. In *Proc. IEEE Vehicular Technology Conference*, Amsterdam, the Netherlands, Sep 1999.

[2] C. Y. Huang and Yates R. D. Call admission in power controlled CDMA systems. In *Proc. IEEE Vehicular Technology Conference*, Atlanta, GA, USA, May 1996.

[3] Y. Ishikawa and N. Umeda. Capacity design and performance of call admission control in cellular CDMA systems. *IEEE Journal on Selected Areas in Communications*, 15(8), 1997.

[4] E. Geijer Lundin. Uplink admission control based on estimated interference in WCDMA systems. Master Thesis LiTH-ISY-EX-3123, Dept. of Electrical Engineering, Linköpings universitet, Sweden, 2001.

[5] N. Wiberg and A. Gioia. Uplink packet access control in WCDMA. In *Proc. IEEE Vehicular Technology Conference*, Tokyo, Japan, May 2000.

[6] J. Shapira. Microcell engineering in CDMA cellular networks. *IEEE Transactions on Vehicular Technology*, 43(4), 1994.

[7] W. Rave, T. Köhler, J. Voigt, and G. Fettweis. Evaluation of load control strategies in an UTRA/FDD network. In *Proc. IEEE Vehicular Technology Conference*, Rhodes, Greece, May 2001.

[8] F. Gunnarsson. *Power Control in Cellular Radio System: Analysis, Design and Estimation.* PhD thesis, Linköpings universitet, Linköping, Sweden, April 2000.

[9] 3GPP Technical Specification Group Radio Access Network. Requirements for support of radio resource management. Technical report 3G TS 25.133, 2000.

[10] F. Gunnarsson, E. Geijer Lundin, G. Bark, N. Wiberg. Uplink admission control in WCDMA based on relative load estimates. *Proc. International Conference on Communications, New York*, April 2002.

The Algorithm/Hardware Co-Design Project

Viktor Öwall[1] and John B.Anderson[2]

[1]Electroscience Dept., Lund University, Box 118, SE-221 00 Lund, Sweden.

[2]Information Technology Dept., Lund University, Box 118, SE-221 00 Lund, Sweden.

E-mail: [1]vikt@es.lth.se, [2]anderson@it.lth.se.

Abstract

We summarize the progress so far in the PCC project Algorithm/Hardware Co-Design, one of six projects within the Swedish National research project Personal Computing & Communication (PCC).

Introduction

Future wireless systems will have traffic that is much wider band and more complex than today's. This Project studies some key physical-layer issues: Error protection, signal compression, data security, and most of all, how to provide wider bandwidth at an affordable cost. All of these require efficient signal handling by algorithms and processors that are an order more complex than before. Future silicon implementation processes will have a much higher transistor count, higher clock rates, and lower power consumption per computation. Therefore, it will be possible to integrate today's system much more efficiently and at a lower cost. However, experience shows that the growth in algorithm complexity will be even higher than the increase in number of transistors, resulting in continued challenges in hardware implementation. Handling the complexity of integrated circuits with several hundred millions of transistors with today's methodologies will be a bottleneck in future design processes. This is the motivation for new implementation strategies to reach efficient solutions. One research area that is long overdue is hardware/algorithm co-optimization. Today algorithms and hardware aspects are studied separately with little or no communication across the interdisciplinary boundaries. Instead, these processors should be designed by a close cooperation between signal and algorithm theorists and circuit implementation experts. A good design evolves through a back and forth exchange between these sides.

The Project goal is to develop advanced processors for picture and video compression, packet error protection, CDMA, and security maintenance. These should respond to quality of service demands and make use of soft information, both of which exist in future wireless systems. A core processor in many of these, the BCJR algorithm, is to be implemented as a chip via the algorithm/hardware codesign methodology. Three project work packages are now complete, one will soon finish, and two are half completed.

The work package in tailbiting codes for packets (Per Ståhl with Rolf Johannesson) came to an end in December 2001. It discovered many new tailbiting codes, codes that equal and sometimes exceed the abilities of any other short packet protection codes with the same parameters. Security algorithms was complete in May 2002 (Fredrik Jönsson with Thomas Johannesson). It developed cryptographic primitives that provide protection against unauthorized intrusion and looked at efficient implementations. The code-division multiple-access package (Ola Wintzell with Kamil Zigangirov, with partial support from Ericsson Company) finished in February 2002; this package delivered studies of pulse-position systems, convolutional code-based CDMA, and a complete Shannon theory of this kind of CDMA.

Scheduled to finish in September 2002 is the turbo coding work package (Carl-Fredrik Leandersson with Torleiv Maseng and Ove Edfors). It has completed work on new interleavers and has begun a study of soft information in list-Viterbi decoding. The two remaining work packages are half complete and their progress is reported in two papers that follow. One looks at new low-rate methods of coding images based on the model–residual approach (Tomas Eriksson with Mirek Novak and John B. Anderson). In mid 2002, basic software modules for soft-information picture coding had been completed and a new MAP-based method had been successfully demonstrated. The core of this method is a BCJR processor. In the other work package, hardware implementation aspects of the BCJR algorithm are being investigated (Karl Thoren with John B. Anderson and Viktor Öwall). A complete study of precision had been performed and some new computation procedures have been proposed.

Alternative Hardware Implementation of the Tailbiting BCJR Decoder

Karl Thorén[1], Viktor Öwall[2], John B. Anderson[3]

[1]Dept. of Applied Electronics, Lund University, Box 118, SE-221 00 Lund. E-mail: Karl.Thoren@es.lth.se.
[2]Dept. of Applied Electronics, Lund University, Box 118, SE-221 00 Lund. E-mail: Viktor.Owall@es.lth.se.
[3]Dept. of Information Technology, Lund University, Box 118, SE-221 00 Lund. E-mail: john_b.anderson@it.lth.se.

Abstract

This paper introduces an alternative implementation of the feedforward tailbiting BCJR decoder. Logarithmic MAP decoders have some difficulties when adding linear values in the logarithmic domain. The max-log MAP algorithm, which is an efficient decoder with low complexity, approximates this addition as the largest of the two added values. The algorithm introduced in this paper has been developed with consideration for hardware implementation. The algorithm works both with logarithms of 2 and in the linear domain, which solves the problems the logarithmic MAP decoders suffers from.

1 Introduction

The algorithm presented in this paper introduces an alternative way to implement the BCJR algorithm, named after its inventors Bahl, Cocke, Jelinek and Raviv [1]. The BCJR algorithm uses matrix operations, which will give rise to both multiplications and additions. Since multiplications are rather complex operations, the most commonly used algorithms avoid this problem by storing the data as logarithmic values. A multiplication of linear values corresponds to an addition in the logarithmic domain, which is a function with lower complexity. However, an addition in the linear domain does not have a simple corresponding function in the logarithmic domain. The basic concept of the algorithm introduced in this paper, is to store values in both the linear and the logarithmic domains. By doing so, the multiplication can be reduced to a bitshift, whereas the addition will still be an addition. We have chosen to call this system with both logarithmic and linear values semi-log.

The BCJR algorithm, is a symbol-by-symbol soft-output MAP decoder for trellis codes. It exists in at least two basic versions, the original algorithm for terminated codes [1], and a version for tailbiting codes [2]. This study focuses on the implementation of the BCJR applied to feedforward tailbiting convolutional codes, although the conclusions here apply to codes with feedback polynomials as well. Tailbiting convolutional codes are a powerful class of trellis codes suited especially to packet transmission. The MAP decoding method adds to this the ability to take in and put out soft information.

The BCJR algorithm is based on two linear vector recursions, one forward through the trellis and one backward. These have the form $\alpha_t = \alpha_{t-1}\,\Gamma_t$ and $\beta_t = \Gamma_{t+1}\,\beta_{t+1}$, where α and β are vectors, Γ is a sparse matrix, and t is the index of the trellis stage. In the original BCJR algorithm the recursions start from vectors with a 1 in the terminating states; in the tailbiting algorithm, the TB-BCJR, they start from the right and left principal eigenvectors of the product of the trellis Γ matrices. A more efficient tailbiting BCJR version, called the Wrap Algorithm [2], does not explicitly find the eigenvectors, but instead performs a few extra forward and backward recursions before and after the main body. As the recursions propagate, the numbers in the α and β vectors will decrease. A system to scale up the vectors must be used, otherwise the number representation will eventually not manage to represent the small values.

The system in which the algorithm is used in this paper, is presented below in a simplified picture, see figure 1. The sender is transmitting a signal with energy E_b, which will be distorted in the channel by additive white Gaussian noise with zero mean and power spectral density, $N_0/2$. Since a digital receiver has finite resolution, the signal will be quantitized into an approximative level. The probability of the signal being in that level is then calculated for both the scenario that a one and that a zero was sent. Some systems use the

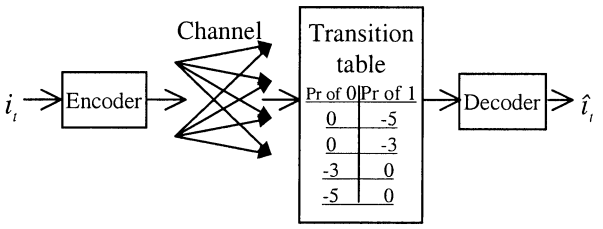


Figure 1. *Simplified picture of a soft channel decoding system. The values in the transition table are logarithms of the channel probabilities.*

standard technique, which creates a transition table by taking the logarithm of the probabilities; then subtract the smallest element from the table, scale it and finally round off the values to integer numbers. These will be inserted into the decoder, which finally will give an estimate of the message as an output.

Several fundamentally different implementation techniques exist. For the terminated codes the max-log MAP algorithm [3] is an attractive alternative, which gives a high ratio between decoding gain and low complexity compared to other MAP implementations [4]. As the name implies, the algorithm works in the logarithmic domain. This will make the linear multiplications, into mere additions in the logarithmic domain. Due to the difficulty of translating linear additions between two values to logarithmic operations, the max-log MAP decoder will choose the largest value as an approximation of the two added together. The cost of simplifying the algorithm is thus that some noise will be introduced to the system. The max-log MAP algorithm has the property that if the inputs to the decoder are integer numbers, the output will be that as well. Neither additions nor the max operations used in the algorithm adds any bits of lower significance.

2 The semi-log approach

Storing values in both the logarithmic domain and the linear domain, gives some advantages in performing multiplications are obtained. Assume that a value x_{log2} is the result of performing the $\log_2$ operation on x. Say that we have y_2 which is the fixed-point binary representation of the variable y and we would like to do the operation $x \cdot y$. If x_{log2} is an integer this could be done by simply applying a bitshift to the left on y_2 by x_{log2} bits, written as $y_2 << x_{log2}$. A negative x_{log2} would result in a bitshift to the right. Since the BCJR algorithm calculates with probabilities, which are between zero and one, the $\log_2$ numbers will be negative and thus the shifts will always be to the right. The advantage with bitshifts is that the operation has quite low complexity, e.g. it can be implemented as either several consecutive shifts or as a barrel-shift.

3 Recursions

The complexity of calculating the recursions is equal for both $\alpha_t = \alpha_{t-1} \Gamma_t$ and $\beta_t = \Gamma_{t+1} \beta_{t+1}$, so we will only consider the β_t case here. Assume that the memory, m, of the encoder is 2, then β_t could be calculated as

$$\beta_t = \begin{bmatrix} \Gamma_{t+1}(0,0) & \Gamma_{t+1}(0,1) & 0 & 0 \\ \Gamma_{t+1}(1,0) & \Gamma_{t+1}(1,1) & 0 & 0 \\ 0 & 0 & \Gamma_{t+1}(2,2) & \Gamma_{t+1}(2,3) \\ 0 & 0 & \Gamma_{t+1}(3,2) & \Gamma_{t+1}(3,3) \end{bmatrix} \cdot \begin{bmatrix} \beta_{t+1}(0) \\ \beta_{t+1}(1) \\ \beta_{t+1}(2) \\ \beta_{t+1}(3) \end{bmatrix}, \quad (1)$$

where $\beta_{t-1}(i)$ and $\Gamma_t(i,j)$ are elements in the vector and the matrix respectively.

An element in the β_t vector would be calculated as

$$\beta_t(0) = \Gamma_{t+1}(0,0) \cdot \beta_{t+1}(0) + \Gamma_{t+1}(0,1) \cdot \beta_{t+1}(1) \cdot \quad (2)$$

Using the max-log MAP algorithm this becomes

$$\beta_t(0)_{\log} = \max\left(\Gamma_t(0,0)_{\log} + \beta_{t+1}(0)_{\log}, \Gamma_t(0,1)_{\log} + \beta_{t+1}(1)_{\log}\right). \quad (3)$$

To calculate one element in α_t, or β_t which has the same complexity, this would require 2 additions + 1 comparison. Totally for calculating both α_t and β_t using a memory m encoder, this would require 2^{m+2} additions and 2^{m+1} comparisons. For decoding the entire message, this should be multiplied by the length of the message, and if the wrap algorithm is used the length of the pre-wrap must be added to the length of the message.

For the fixed-point semi-log decoder, one element in β_t using a memory 2 encoder is calculated as

$$\beta_t(0) = \beta_{t+1}(0) << \Gamma_{t+1}(0,0)_{\log_2} + \beta_{t+1}(1) << \Gamma_{t+1}(0,1)_{\log_2}, \quad (4)$$

which is 2 bitshifts and 1 addition. Generalized to a memory m code the entire α_t and β_t would require 2^{m+2} bitshifts and 2^{m+1} additions.

It is necessary to do scaling of the vectors periodically during the recursions, since the probabilities represented in the α and β vectors are decreasing for each step in the recursions. If one would ignore this all the elements in the vector would finally be so small that they could not be represented by the numerical system.

In digital design, scaling with powers of 2 is hardware-efficient, since it only applies a bit-shift to the digital word. If leading zeros are detected in all elements of the α_t or the β_t vector, it is just to bitshift the vector elements to the left, so that the most significant bit (MSB) has a 1 in at least one of the elements. If we limit ourselves to do a bitshift with only one bit at each recursion, it is only necessary to investigate the MSB in the elements. Are all MSBs zero, apply a bitshift to the elements.

4 Output

Since there exist so many different applications for the BCJR algorithm, we have chosen to stop at calculating the λ_t vector. The elements in the λ_t vector contain the probability that the encoder was in a certain state, whereof half of the states correspond to that a 0 was sent and the other half to that a 1 was sent. For feedforward convolutional codes, we form the row vector $\lambda_t = \alpha_t \bullet \beta_t$, in which $\bullet$ means componentwise multiplication. A

similar quantity exists for feedback encoders. The first element in the λ_t vector is calculated as

$$\lambda_t(0) = \alpha_t(0) \cdot \beta_t(0) \cdot \qquad (5)$$

Using this definition, the logarithmic algorithms will only perform an addition instead of the multiplication. However, for the semi-log algorithm, the α_t and β_t, vectors are linear and thus an multiplication is performed to calculate λ_t.

In simulations of different implementations of feedforward tailbiting BCJR decoders, the BCJR decoder has been used to give a hard decision output. By comparing the sum of each half of the elements in the λ_t vector, the hard decision has been made in favor of the most likely input into the encoder. The results from these simulations are shown in figure 2 and 3. The figures clearly show that the semi-log MAP decoder needs long word size to get a good performance, mainly due to multiplications. There are several optional ways of both reducing the complexity of the multiplications and working around the need of long wordlength, but this is beyond our scope.

One drawback of the semi-log algorithms is that the transition-table must consist of integer numbers. Comparing the figures one notices that the semi-log decoder performs better than the max-log MAP decoder in certain intervals in the channel with 8 soft levels. As the number of soft levels increases the max-log decoder can take advantage of not having rounded off transition values.

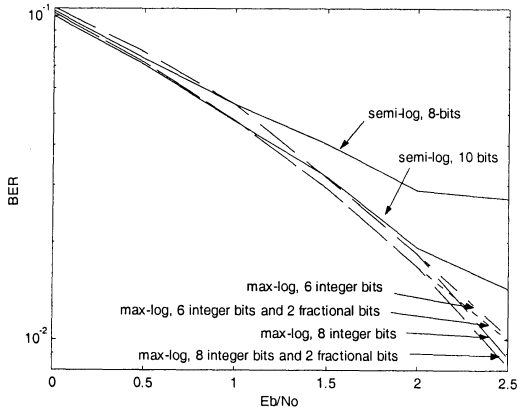


Figure 2. *Simulation with a receiver detecting 8 soft levels, ranging from amplitude –1 to +1, which are fed into the decoder. A (7,5) encoder is used and 40 information bits are transmitted. Solid lines are the semi-log algorithm, top one using an 8-bit wordlength decoder, and the next one uses 10-bit wordlengths. The remaining lines are max-log MAP simulations. The dotted line it is simulated using a 6-bit integer part and a 2-bit fractional part (hidden mostly), the dash-dotted line is a 6-integer bit simulation, the dashed lines are first a 8-integer bit decoder and second a decoder using 8-bit integer part and a 2-bit fractional part.*

5 Robustness of the BCJR-Algorithm

An issue that turns out to be related to the word size is how sensitive the decoder is to incorrect information about the channel's SNR. To test the robustness in changing the decoders' parameters we have done a series of simulations for the Gaussian channel and the BSC.

Common for all studied decoders is that the BER is quite stable when different SNRs are told to the decoder. However, when the internal SNR is set to values less than 0 dB the behavior of the decoder changes and it can fail to perform well, see figure 4 and 5. Therefore we have chosen two scenarios with little in common, either in the generator matrix or the length of code words. The dotted line is the normal waterfall curve and the rings are the channels this curve has been simulated for, varying from –3 to 6dB. The solid nearly flat lines, however, show the performance of the decoder if an inaccurate E_b/N_0 is told to it.

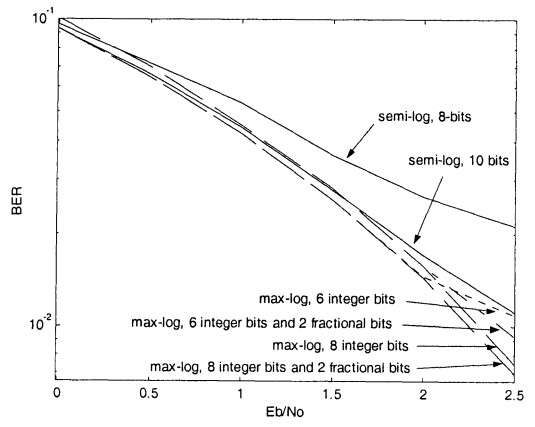


Figure 3. *Simulation with a receiver detecting 32 soft levels ranging from amplitude –2 to +2. Same notation used as above.*

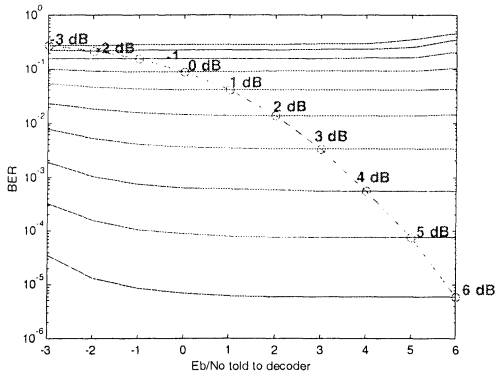


Figure 4. *Robustness of the (5,7) decoder when 100 information bits are sent.*

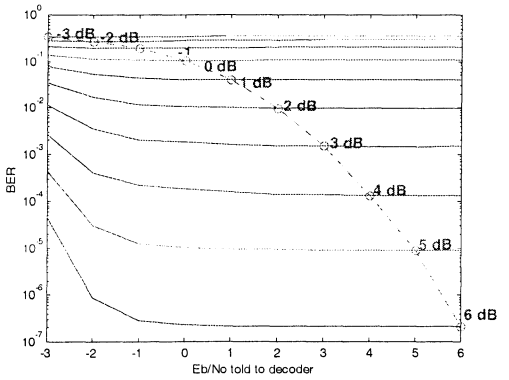


Figure 5. *Robustness of the (66,62) decoder when 48 information bits are sent.*

A better channel turns outs to need a greater wordlength, i.e., a more complex decoder. Otherwise the performance will differ too much from the infinite precision case [5].

Since the tailbiting BCJR decoder is quite stable to incorrect SNR information, it is interesting to make a tradeoff between performance and wordlength by misrepresenting the channel's SNR to the decoder. Giving the decoder a poorer SNR than is actually the case will reduce its wordlength requirement. Figure 6 shows how the wordlength can be reduced in a decoder using linear numbers. However, this is not the semi-log MAP algorithm used here.

Simulations of semi-log MAP decoders, where the E_b/N_0 has been mismatched to reduce the wordlength, have been performed for the same channels as in the previous simulations. As with the pure BCJR decoder, the semi-log implementation is quite robust to being misinformed about the channel true E_b/N_0, see figure 7 and 8.

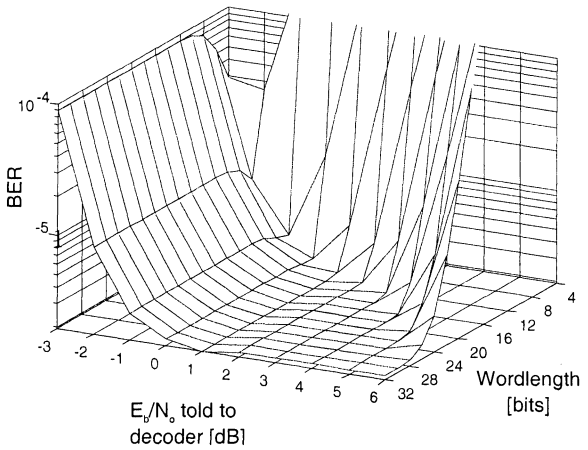


Figure 6. *The actual E_b/N_0 in the channel is 6dB. This would require a very long wordlength, but by adjusting the SNR told to the decoder, the wordlength can be reduced dramatically to around 6 bits.*

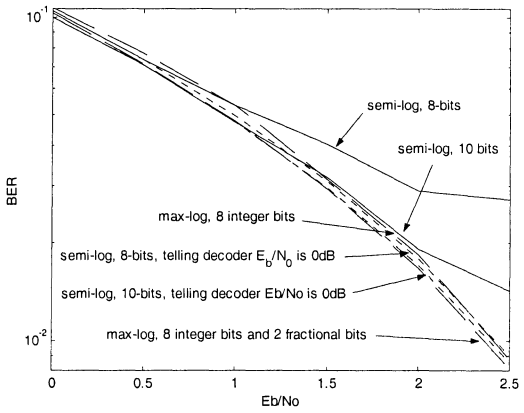


Figure 7. *The simulation is made with a receiver with 8 soft levels ranging from amplitude –1 to +1. A (7,5) encoder is used and 40 information bits are transmitted. Solid lines are the semi-log algorithm; the top one uses a 8-bit wordlength decoder, and the next one uses 10-bit wordlength. The dashed lines are max-log MAP decoders, the first one uses 8-integer bits and the second uses a 8-bit integer part and a 2-bit fractional part. The dotted lines are the semi-log decoders, using 8-bits and 10-bits respectively.*

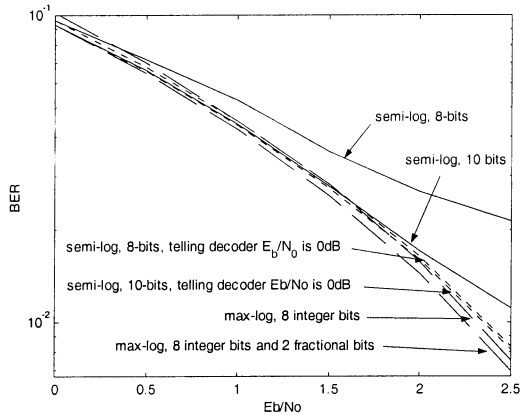


Figure 8. *The simulation is made as in figure 7, with a receiver giving 32 soft levels ranging from amplitude –2 to +2 to the decoder. Same notation used as above.*

6 Conclusions

The fixed-point semi-log TB-BCJR decoder is a good competitor with the max-log MAP algorithm. It uses the simpler operation bitshift instead of addition and removes the problem of adding linear values in the logarithmic domain and the effects this introduce. However, these advantages come with the cost that linear representation does not have the same dynamic range as logarithmic representation and thus have to use a greater wordlength.

The tailbiting BCJR decoder calculates with probabilities, and its numbers can decrease rapidly. This will require a very large wordlength in the decoder. To avoid this it is possible to use hardware efficient bit-shifts to scale the numbers. The tailbiting BCJR decoder is very robust to being misled about the channel characteristics. This can be used to make a tradeoff between a small change in error correcting performance and a huge reduction in complexity of the decoder.

References

[1] L. R. Bahl, J. Cocke, F. Jelinek and J. Raviv, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate", *IEEE Trans. Information Theory,* March 1974, pp. 284–287.

[2] J. B. Anderson and S. M. Hladik, "Tailbiting MAP Decoders", *IEEE Journal on Selected Areas in Communications,* Feb 1998, pp. 297-302

[3] P. Robertson, E. Villebrun and P. Hoeher, "A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain", *Proc. IEEE Int. Conf. Commun.*, pp. 1009-1013, June 1995.

[4] C.F. Leandersson, O. Edfors, T. Maseng and T. Ottosson, "On the Performance of Turbo Codes and Convolutional Codes of Low Rate" *IEEE Vehicular Technology Conference*, 1999

[5] K. Thorén, V. Öwall, J.B. Anderson, "Precision Issues in the Implementation of BCJR Decoders", *Proc. IEEE Int. Symposium on Information Theory*, Washington, p. 289, June 2001.

MAP Criterion Image Coding

Tomas Eriksson[1], Mirek Novak[2] and John B. Anderson[3]

[1,2,3]Dept. of Information Technology, Lund University, Box 118, SE-221 00 Lund, Sweden.
E-mail: [1]tomas@it.lth.se, [2]mirek@it.lth.se, [3]anderson@it.lth.se.

Abstract

BCJR based source coding of image residuals was recently introduced. In this paper, a joint source–channel coding system is formed from a trellis representation of the residual signal. The BCJR algorithm is applied to find the MAP encoding. MAP and minimized squared error encoding are compared.

1 Introduction

In a recent paper [1] we introduced the use of the maximum *a posteriori* probability (MAP) criterion in the context of image coding. The proposed method consisted of splitting the image into two components, a model and a residual image, and to code them separately with different methods. The possibility of using a trellis structure to encode the noise-like residual part of the image was also discussed. It was concluded that the original image could be represented by a set of polynomial basis images for the model and the BCJR algorithm [2] could be used to encode the residual part.

In this paper we discuss further whether the MAP path in the trellis, the one corresponding to the highest probability conditioned on the observed information, can be used to represent the observed signal. Once a trellis representation of the residual signal is obtained, the next step is to form a joint source–channel encoding system. It is natural to let the design be of that founded on the same trellis structure.

Related work can be found in e.g. [3, 4]. In [3] the concept of trellis coded quantization (TCQ) was introduced. TCQ was developed by utilizing the similarities of source coding and modulation for digital communications. The work is based on alphabet constrained rate distortion theory and uses the minimum squared error (MSE) as the distortion measure. Hence the well-known Viterbi algorithm (VA) is used in the encoding procedure. Based on these ideas, convolutional channel coding was introduced in the context of TCQ by Belzer *et al.* [4]. Again, the VA is used and the metric is the MSE.

The novelty of this work is the use of the BCJR algorithm and the MAP criterion in the source coding procedure. Since TCQ is strongly connected to the VA and MSE, we do not use the TCQ language, despite other similarities, such as the use of a trellis.

2 Trellis coding using the BCJR algorithm

Consider the encoding of a source sequence $\boldsymbol{\xi}= \xi_1\xi_2 \ldots\xi_N$. All ξ_t represent one or two (real) symbols. A trellis encoder would produce a sequence of labels, $\boldsymbol{L}= L_1L_2\ldots L_N$, where each label is a function of the present and the next trellis state, i.e., $L_t = f(S_t, S_{t+1})$. All labels have the same range.

Denote by $\boldsymbol{l}$ a realization of the sequence $\boldsymbol{L}$. If $g(\boldsymbol{l})$ represents a mapping from the space of trellis labels to the space of real numbers, the quantization error $\boldsymbol{q}$ can be written as

$$\boldsymbol{q} = \boldsymbol{\xi} - g(\boldsymbol{l}) \tag{1}$$

We now introduce the MAP as a distortion measure. An advantage of this metric in image coding is that MAP allows single outliers. MSE does not allow this, and instead tries to find a compromise resulting in the smallest *average* distortion. When using a MAP encoder, shot-noise can appear. However, shot-like noise can easily be removed by a median filter. With the MAP criterion, minimizing $\boldsymbol{q}$ in Eq. (1) is equivalent to the following. Given $\boldsymbol{\xi}$ find an outcome $\boldsymbol{l}$ of $\boldsymbol{L}$ such that

$$\Pr\{\boldsymbol{l} \mid \boldsymbol{\xi}\} \cdot \Pr\{\boldsymbol{\xi}\} \tag{2}$$

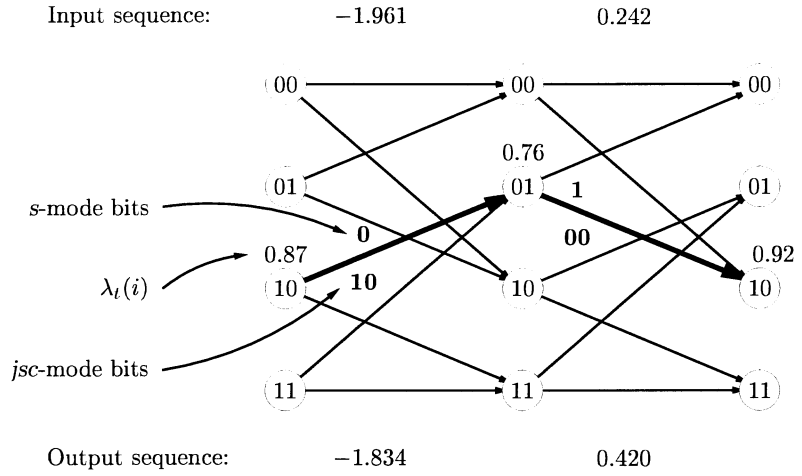
is maximized.

The BCJR algorithm [2] outputs the probability that a trellis state i is reached at time t, given the observed input sequence $\boldsymbol{\xi}$. If S_t is the state at time t, we can write this probability as

$$\Pr\{S_t = i \mid \boldsymbol{\xi}\} \tag{3}$$

The algorithm does not compute this probability directly, but instead computes

$$\begin{aligned} \lambda_t(i) &\triangleq \Pr\{S_t = i, \boldsymbol{\xi}\} \\ &= \Pr\{S_t = i \mid \boldsymbol{\xi}\} \cdot \Pr\{\boldsymbol{\xi}\} \end{aligned} \tag{4}$$



State S_t	State S_{t+1}	s v	jsc l	rec. y
00	00	0	00	y_0
00	10	1	11	$-y_0$
01	00	0	11	$-y_0$
01	10	1	00	y_0
10	01	0	01	y_1
10	11	1	10	$-y_1$
11	10	0	10	$-y_1$
11	11	1	01	y_1

Figure 1: Trellis section showing path maps and branch symbols according to the example. The table shows the path map and trellis branch symbols for the $(7, 5)$ convolutional code.

Since the values of $\boldsymbol{l}$ are defined by a description of the path from state i at time t to state j at time $t+1$ in the trellis, we have a well defined mapping, $\{\boldsymbol{l}\} \rightarrow \{S_t = i\}$ between labels and states; i.e., Eqs. (2) and (4) are equal and hence the BCJR algorithm can be utilized for our purpose. From the λ sequence, the path in the trellis is determined as the transitions between the states with highest probability at each time instant, given that these are legal transitions in the trellis. This path is the representation of the source sequence $\boldsymbol{\xi}$.

Once the shape of the trellis is set, the labels should be chosen in such a way that they are pseudo-randomly distributed on the trellis branches. This is important since the sequence to be encoded has the property of being pseudo-randomly distributed.

The residual encoder can be used in two modes, as a pure source encoder (s-mode) or as a joint source–channel encoder (jsc-mode), cf. Figure 1. For both modes, the vector containing the encoded sequence is denoted $\boldsymbol{c}_i$. When representing the residual signal, each ξ_t corresponds to one or more of the real symbols.

In the s-mode we only consider source coding. For the coding rate one bit per source sample, we simply let the trellis path map $\boldsymbol{l}$ represent the source sequence. Therefore we restrict ourselves to path maps with two possible branches entering and leaving each state. To accomplish lower coding rates, multiple source samples are associated with each trellis branch.

When used as a joint source–channel encoder, jsc-mode, convolutional codewords $\boldsymbol{v}$ are sent over the transmission channel. The codewords correspond to the trellis branch symbols produced by a convolutional encoder. However, no separate channel encoder is required since the output sequence can be found directly in the trellis structure used by the BCJR algorithm. As in the s-mode, we first determine the path map $\boldsymbol{l}$, and then let the branch symbols $\boldsymbol{v}$, defined by a function $h(\boldsymbol{l})$, represent the source. The function $h(\boldsymbol{l})$ can be for example a convolutional encoder, which naturally incorporates the property of distributing the codewords in a pseudo-random manner.

Let $\phi(\boldsymbol{v})$ define the decoding function. If each $\boldsymbol{v}_t$ represents a branch symbol pair, we have the mapping $\phi : \{\boldsymbol{v}_t\} \rightarrow \{-y_1, -y_0, y_0, y_1\}$. The mapping from codewords to reconstruction levels is chosen such that the Hamming distance of the labels should reflect the separation properties of $\phi(\boldsymbol{v})$. Hence we have

$$\begin{aligned} 10 &\mapsto -y_1 \\ 11 &\mapsto -y_0 \\ 00 &\mapsto y_0 \\ 01 &\mapsto y_1 \end{aligned}$$

In similar way, each (four-bit) label is used to produce two real values for $R = 1/4$ codes.

The reconstruction values of the Lloyd-Max pdf optimized quantizer are used [5]. We assume that the source is Laplacian and hence $y_0 = 0.420$ and $y_1 = 1.834$.

Example

As an example, let the input sequence to be coded be $\{\ldots, -1.961, 0.212, \ldots\}$. The $(7, 5)$ convolutional code is used to determine the trellis and label structures, cf. Figure 1. The BCJR algorithm now uses one forward and one backward recursion to find the λ values [6]. From the λ values the encoded sequences $\{\ldots, 0, 1, \ldots\}$ for the s-mode and $\{\ldots, 10, 00, \ldots\}$ for the jsc-mode are determined. A corresponding decoder would produce $\{\ldots, -1.834, 0.420, \ldots\}$ as the output sequence.

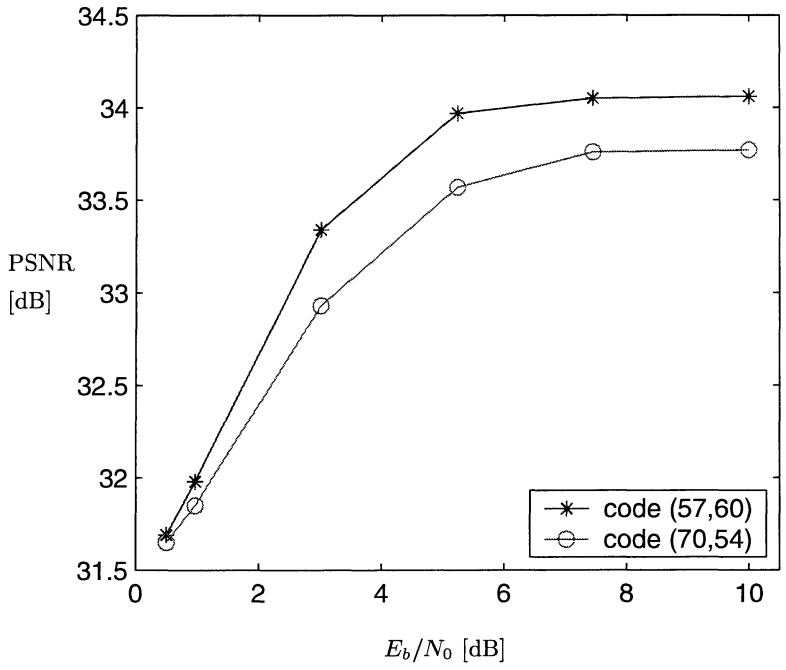


Figure 2: Test image Lena—image PSNR as a function of channel SNR.

3 Results

Different test images have been encoded in the *s*-mode with 1 and 0.5 bpss (bits per source sample) in the residual coding procedure. Figure 3 shows a comparison between the proposed method and two methods based on MSE, i.e., DCT with 24 uniformly quantized coefficients and blocksize of 8×8 pixels and standard vector quantization with a blocksize of 4×4 pixels. One should remember that when printed on paper, the quality of all images is degraded.

Figure 3 shows a segment of the reconstructed test image Lena. The original image, Figure 3 (a), contains a number of thin short diagonal lines in the hatband. As seen in Figures 3 (b)–(c), these lines are very hard to preserve using a method based on MSE. However, using the MAP criterion, Figure 3 (d), fine lines like the striped texture on the woman's hat are clearly visible. Also, blocking artifacts in the images of Figures 3 (b)–(c) can be noticed. The proposed method does not show this brick pattern. Thus the perceptual quality of the image is improved.

In the joint source–channel coding system the total residual rates become $1 \cdot 2 = 2$ bpss for $R = 1/2$ codes and $0.5 \cdot 4 = 2$ bpss for $R = 1/4$ codes. Although the encoding rates are the same, simulations show that $R = 1/2$ codes yield better performances even at low SNRs. See Figure 2 for rate $R = 1/2$ codes.

References

[1] T. Eriksson, M. Novak and J.B. Anderson, "BCJR Based Source Coding of Image Residuals," in *Proc. Nordic Radio Symposium (NRS 2001)*, Nynäshamn, Sweden, pp. 63–66, April 2001.

[2] L.R. Bahl *et al.*, "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate", *IEEE Trans. Inform. Theory,* vol. IT-20, pp. 284–287, March 1974.

[3] M.W. Marcellin and T.R. Fischer, "Trellis Coded Quantization of Memoryless and Gauss-Markov Sources," *IEEE Trans. Comm.,* vol. 38, No. 1, pp. 82–93, Jan. 1990.

[4] B. Belzer, J.D. Villasenor and B. Girod, "Joint Source Channel Coding of Images with Trellis Coded Quantization and Convolutional Codes," in *Proc. 1995 Int'l Conf. Image Proc. (ICIP'95),* vol. 2, pp. 85–88, Oct. 1995.

[5] N.S. Jayant and P. Noll, *Digital Coding of Waveforms,* Prentice-Hall, Englewood Cliffs, NJ, 1984.

[6] J.B. Anderson and S.M. Hladik, "Tailbiting MAP Decoders," *IEEE Journ. Sel. Areas Comm.,* vol. 16, No. 2, pp. 297–302, Feb. 1998.

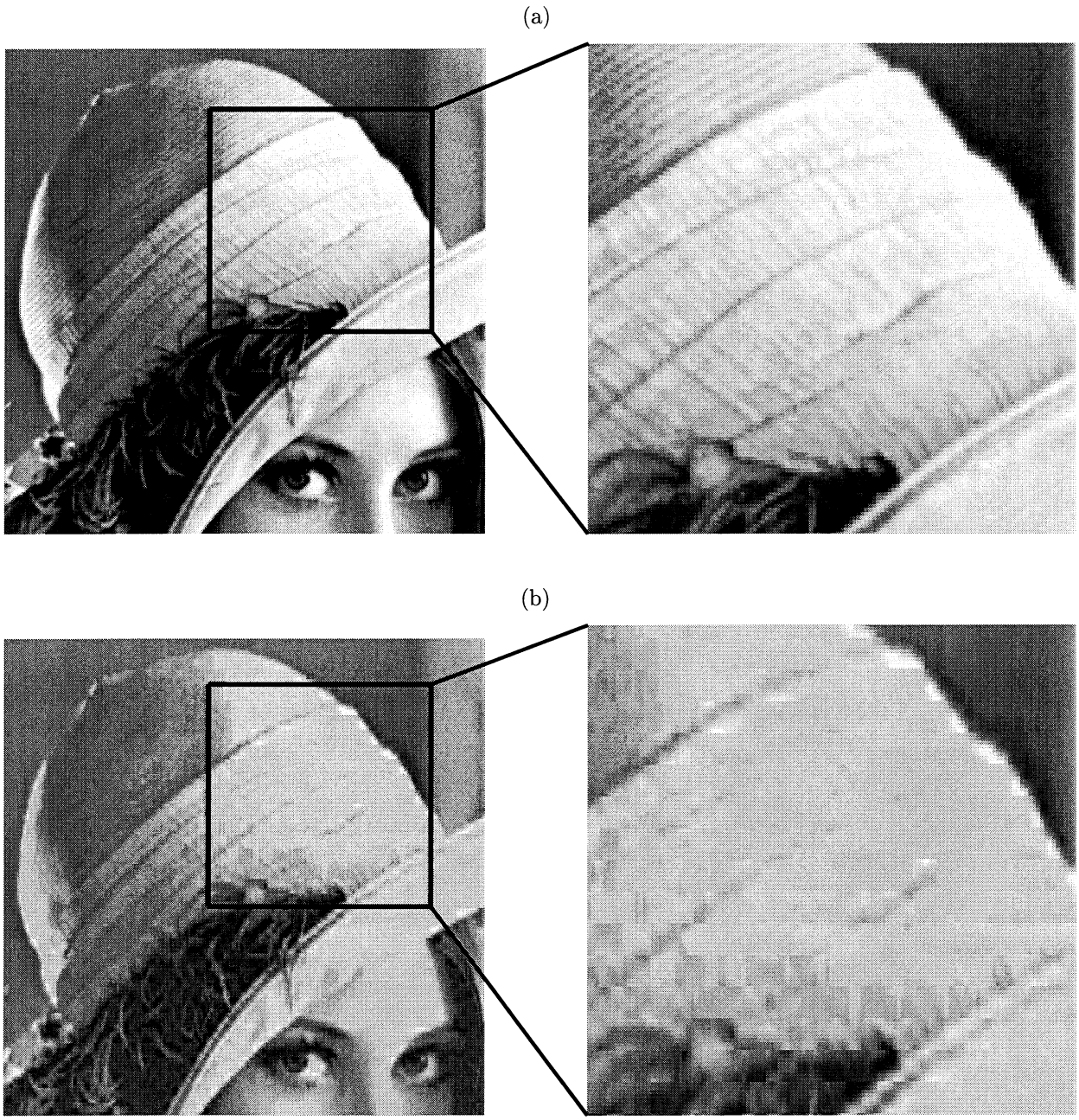
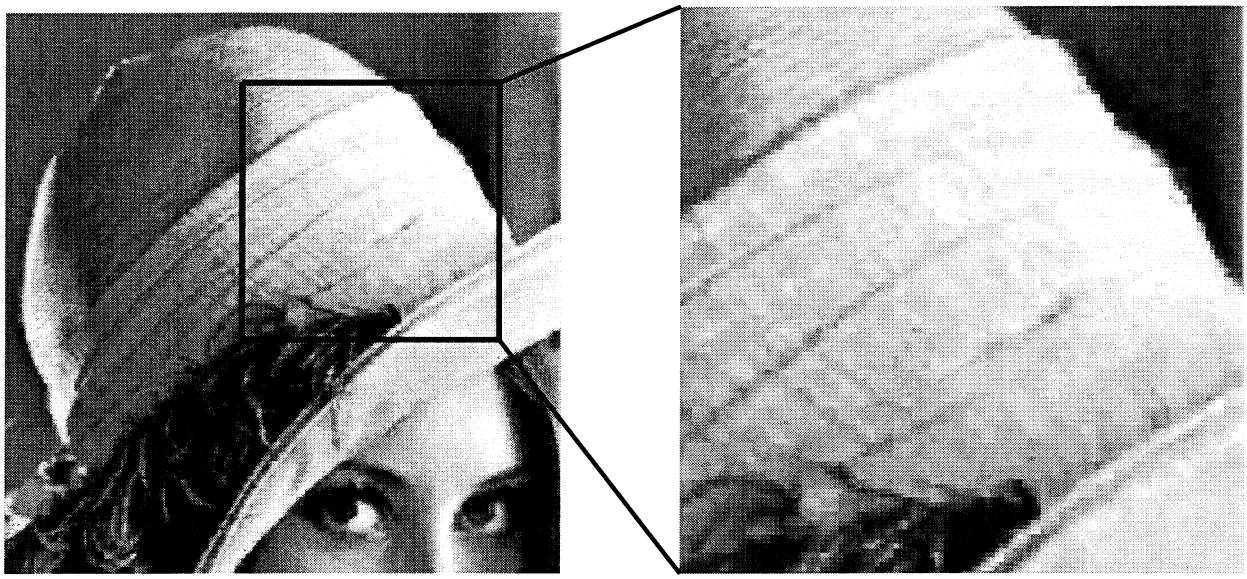


Figure 3: (a)–(b): Test image `Lena`—segment of reconstructed image; (a) original, (b) Quantized DCT (blocksize 8×8, 1.13 bpss, PSNR 30.9 dB).

(c)



(d)

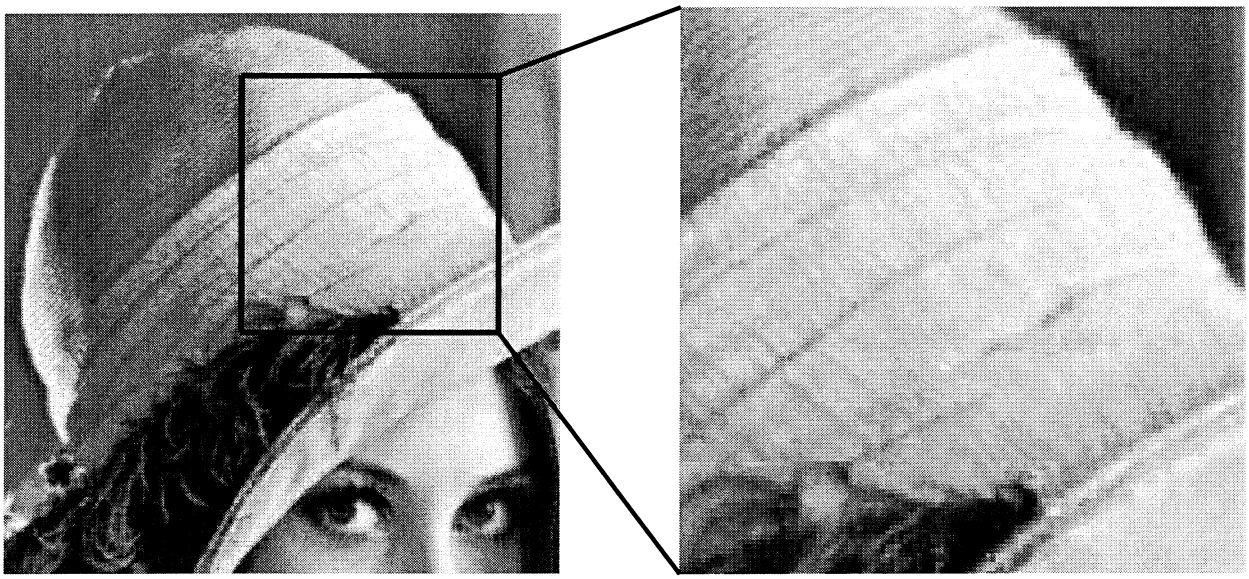


Figure 3: (c)–(d): Test image `Lena`—segment of reconstructed image; (c) Standard VQ (blocksize 4×4, 0.81 bpss, PSNR 31.1 dB), (d) MAP criterion using the BCJR algorithm (blocksize 8×8, 1 bpss, PSNR 34.31 dB).

On the Performance of Incremental Redundancy Schemes with Turbo Codes

Carl Fredrik Leanderson[1], Giuseppe Caire[2] and Ove Edfors[1]

[1]Dept. of Electroscience, Lund University, Lund, Sweden. E-mail: cfl@es.lth.se.
[2]Mobile Communications Group, Institute Eurecom, Sophia-Antipolis, France. E-mail: caire@eurecom.fr.

Abstract

A semi-analytical bound on the frame error performance of rate compatible punctured Turbo (RCPT) codes on the block-fading Gaussian collision channel is presented. Based on this bound, it is concluded that incremental redundancy schemes with RCPT codes have the potential to yield significant throughput improvements compared to similar schemes based on convolutional codes.

1 Introduction

Incremental redundancy techniques, are subject to an increasing interest in order to obtain transmission with low error probability in packet-oriented high-speed wireless links. This is due to the ability of incremental redundancy schemes to achieve high throughput by combining forward error correction (FEC) and automatic repeat request (ARQ) [1].

Information theoretic considerations in [2] indicate that incremental redundancy protocols have the potential to yield simple and efficient error correction for slotted Gaussian multiple-access channels. In particular it is demonstrated that it is possible to achieve higher throughput with incremental redundancy protocols than slotted ALOHA or packet combining [3] protocols. However, a challenging issue is to find practical codes with high performance and low complexity suitable for incremental redundancy protocols.

In contrast to conventional ARQ techniques, incremental redundancy protocols require channel codes that are rate compatible. Rate compatible punctured convolutional (RCPC) codes with application to ARQ were first considered in [4] and subsequently this concept has been extended to larger classes of convolutional codes [5] as well as Turbo codes [6]. In particular, Type-II Hybrid ARQ schemes [1] based on convolutional codes have been studied in [7, 8]. Rate compatible Turbo codes were first introduced in [9] and a significant amount of work on Turbo codes in Type-II Hybrid ARQ schemes has been performed [10–13]. Most of these investigations have been limited to the additive white Gaussian noise (AWGN) channel. There are, however, indications that Turbo codes are effective also on other channels, such as for instance multiple-access channels [14].

Motivated by the increasing deployment of wide-band packet data transmission in mobile multiple-access systems, where interference may be a significant source of performance degradation, we study the performance of RCPT codes on the block-fading Gaussian collision channel. Initial studies of the performance of RCPC codes on this channel have been performed in [15]. This paper extends these investigations to RCPT codes, focusing on the frame error rate (FER) in an incremental redundancy scheme, given a certain number of (re-)transmissions M.

2 System Description

A slotted multiple-access communication system is considered. There are in total N_u users in the system and these users are active according to a Bernoulli distribution with probability p in every time slot. The load of the system is $G = N_u \cdot p$. We consider a static protocol where the transmission probabilities of the individual users are constant in time. The transmission probabilities of individual users are thus not affected by the amount of decoding failures for their previously transmitted packets. A subset of interferers $K(s)$ out of a total of $N_i = N_u - 1$ potential interferers are simultaneously active in time slot s according to the binomial distribution with parameters (p, N_i). The transmitted signal of a user, as well as the interfering signals, are subject to block-independent Rayleigh fading with complex envelope $c_{k,s}$ for each active user k. In the system model, there is also a noise contribution due to thermal noise. The decoder of a particular user k performs single-user detection and treats this noise as well as the interference from the other users of the same time slot as AWGN and sees a channel according to

$$y_{k,s,l} = c_{k,s} x_{k,s,l} + v_{k,s,l}. \tag{1}$$

The transmitted channel symbols $x_{k,s,l}$ from user k in time slot s at symbol position l are assumed real valued and the analysis is restricted to binary antipodal modulation. The interference-plus-noise component $v_{k,s,l}$ includes Gaussian background noise and the interference caused by the other users active in slot s.

The signal-to-noise-and-interference ratio (SNIR) of the signal in (1) corresponding to user $k = 1$ at transmission slot j can be written as

$$\beta_{1,j} = \frac{c_{1,j}^2 E_1}{N_0 + \sum_{k \in K(j)} c_{k,j}^2 E_k}, \tag{2}$$

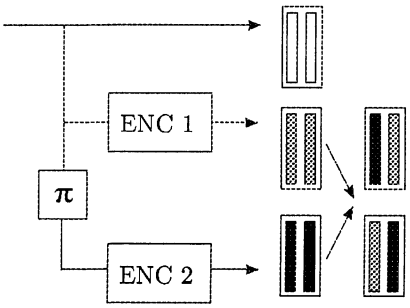


Figure 1: Systematic RCPT encoder.

where E_1 and E_k are the channel symbol energies of the desired user and the interfering users, respectively, and N_0 is the single-sided noise power spectral density. For ease of exposition and with no loss of generality since the system is symmetric, it is sufficient to consider only one particular user $k = 1$. Hence, the user index in the SNIR will in the sequel be omitted letting $\beta_j = \beta_{1,j}$.

The probability density function (PDF) of β_j for this channel model has been derived in [15]. In the system under consideration it is assumed that each transmitter-receiver pair employs an incremental redundancy scheme for transmission over the channel. The transmitters are assumed to have no channel state information. However, perfect channel state knowledge is assumed at the receivers. At the receiver side, maximum-likelihood (ML) detection is assumed available. Although this is not the case for Turbo codes of reasonable size, iterative decoding techniques at hand have been observed to provide near ML-performance [6].

In order to provide a series of incremental-redundancy packets, we consider systematic rate $1/M$ RCPT codes. For the sake of analytical tractability we consider RCPT codes with random interleavers, although improved performance can be achieved with careful interleaver design [16].

The structure of the systematic RCPT codes used in the incremental redundancy system studied is illustrated in Figure 1. The encoder consists of two component encoders with an intermediate interleaver and yields one systematic bit stream and two coded bit streams. These bitstreams are used to compose subpackets of equal length in order to provide incremental redundancy. The systematic bit stream provides the first subpacket. The coded bit streams are punctured in a cyclical manner such that coded bits from both encoders are multiplexed into both the first and the second subpacket of coded bits, *cf.* Figure 1. This enables the decoder to invoke iterative Turbo decoding, already upon reception of the first retransmission. This yields improved performance compared to the approach in [10], where the subpacket of the first retransmission contains explicitly coded bits from the first component encoder which prohibits Turbo decoding until both the subpackets with coded bits have been received.

A higher number of subpackets from RCPT codes can be achieved with different strategies: through 1) the use of low-rate component encoders and appropriate puncturing/multiplexing into subpackets, or 2) the use of multiple component encoders with intermediate interleavers. Due to the expected increase of decoding complexity when using multiple encoders and corresponding multiple decoders the first approach appears to be the most attractive one.

In the Turbo codes under consideration, tailbits are used to force the first encoder back to the zero state in order to terminate the trellis in a known state, which yields a small rate-loss. In our computations and simulations this is reflected in a decrease of the energy of the transmitted symbols.

3 Performance Analysis

A crucial component of the performance analysis is the fact that different subpackets transmitted over the block-fading Gaussian collision channel are subject to different channel states. Therefore, in order to be able to apply distance based bounding techniques derived for the AWGN channel, it is necessary to compute the multivariate distance distribution of the RCPT codes under consideration. For simplicity, the analysis in this paper is restricted to the case of binary antipodal modulation.

The multivariate distance distribution offers the joint multiplicity of Hamming distances between different subpackets belonging to the same sequence of information bits. For a particular mothercode codeword, it is possible compute an equivalent squared Euclidean distance to the all-zeros codeword by computing the scalar product of the Hamming distances and the channel state vector of the different subblocks. By linearity of the codes under consideration, the same set of Euclidean distances to surrounding codewords apply for every codeword and it is therefore sufficient to compute the multivariate Hamming weight distribution.

The multivariate Hamming weight distribution of the punctured component codes of an RCPT can conveniently be computed using the recursive method introduced in [17]. In order to calculate the corresponding weight distribution of an RCPT code one has to take into account the effect of the interleaver. Calculation of the weight distribution resulting from a particular interleaver is today an operation that is far too complex for interleaver lengths of practical interest.

In [18], a tractable method for calculation of the Turbo code weight distribution was presented, based on ensemble averaging over all interleavers of a given length. The method is based on a probabilistic device called a *uniform interleaver*. Using the concept of a uniform interleaver, the expected weight distribution of a Turbo code where the interleaver is chosen as a random permutation can be calculated. Even though the weight distribution of a particular Turbo code does not coincide with the ensemble average weight distribution calculated in this way, the concept of uniform interleaving enables insight into the impact on

the weight distribution from different choice of interleaver length and component code parameters such as memory, rate, generator polynomials and puncturing. For instance, improvements of the component codes with respect to minimum distance properties will directly improve the minimum distance distribution of the corresponding, uniformly interleaved, Turbo code.

We will briefly describe how the multivariate Turbo code weight distribution for random choice of interleaver can be enumerated using the concept of uniform interleaving in conjunction with weight-enumerating functions. Define the multivariate weight enumerator of a component code with an input sequence of K bits (information bits and tailbits) as

$$T(I, D_1, \ldots, D_{M-1}) \triangleq \sum_{i \geq 0} \sum_{d_1 \geq 0} \cdots \sum_{d_M \geq 0} t_{i,d_1,\ldots,d_M} I^i D_1^{d_1} \cdots D_{M-1}^{d_M}, \quad (3)$$

where the exponents i and $d_1, \ldots, d_{M-1}$ represent the input sequence weight and the set of output sequence weights. The number of codewords with subpacket weights $d_1, \ldots, d_{M-1}$, generated by input sequences of weight i is contained in $t_{i,d_1,\ldots,d_M}$. The variables I and $D_1, \ldots, D_{M-1}$ are dummy variables that enable a convenient enumeration of the weight distribution. For instance, multiplication of two weight enumerators corresponds to convolution of the weight distributions. This is a useful property when deriving the multivariate weight distribution for a Turbo code with uniform interleaver, since it corresponds to an M-dimensional convolution of the component code weight distributions. The weight enumerator in (3) is also referred to as the multivariate *input-redundancy weight enumerating function* (IRWEF).

The Turbo codewords are combinations of two component code codewords that both result from the same input-sequence weight i. The *conditional weight enumerating function* (CWEF)

$$T_i(D_1, \ldots, D_{M-1}) \triangleq \sum_{d_1 \geq 0} \cdots \sum_{d_M \geq 0} t_{i,d_1,\ldots,d_M} D_1^{d_1} \cdots D_{M-1}^{d_{M-1}}, \quad (4)$$

enumerates the number of codewords of different subpacket weights $d_1, \ldots, d_{M-1}$, corresponding to a certain input-sequence weights $i = 0, 1, \ldots, K$. Given a particular weight of the input sequence, the weight distributions of the two component codes can be represented in terms of the CWEFs as $T_i^{C_1}(D_1, \ldots, D_{M-1})$ and $T_i^{C_2}(D_1, \ldots, D_{M-1})$.

Using a uniform-interleaver, it is assumed that a certain input-sequence of weight i is permuted with equal probability into all other sequences of weight i. The average CWEF of Turbo code weight distributions, over the ensemble of all interleavers of length K, can be then be calculated as

$$T_i^{TC}(D_1, \ldots, D_{M-1}) = \frac{T_i^{C_1}(D_1, \ldots, D_{M-1}) T_i^{C_2}(D_1, \ldots, D_{M-1})}{\binom{K}{i}} = \sum_{d_1 \geq 0} \cdots \sum_{d_M \geq 0} t_{i,d_1,\ldots,d_M}^{TC} D_1^{d_1} \cdots D_{M-1}^{d_{M-1}}, \quad (5)$$

where $1 \Big/ \binom{K}{i}$ is the probability that a weight-i sequence is mapped to a specific weight-i sequence by the uniform interleaver for $i = 0, 1, \ldots, K$.

Having computed the multivariate weight distribution of the RCPT under investigation, using the concept of uniform interleaving, it is possible to calculate upper bounds on the average error probability over the ensemble of interleavers after ML-decoding upon the $1, 2, \ldots, M$th received subpacket. The union bound (see for example [19]), the Viterbi bound [20] and the tangential sphere bound [21] have been considered in order to upper bound the ML-decoding performance of Turbo codes.

The tangential sphere bound by Poltyrev [21], yields the tightest upper bound of the three. However, unfortunately, the tangential sphere bound requires significantly more computational effort than the other two. This causes relatively complex computations in the case of the block-fading Gaussian collision channel, since it is necessary to solve an optimization problem for each simultaneous instance of multiple channel states for the M (re-) transmissions, due to the multiplicative nature of the fading. Asymptotically, for high SNRs, the union bound gives a tight upper bound on the error performance, but the union bound is in practice useless below the cutoff rate since it is very loose in the corresponding region of SNRs. Especially for Turbo codes, operating well below the cutoff rate, the tangential sphere bound and the Viterbi bound are more suitable.

Using the coefficients of the multivariate weight enumerator in (5), the tangential sphere bound can be modified to the case of the block-fading Gaussian collision channel according to

$$P(e|\beta_1, \ldots \beta_M) \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z_1^2/2} \cdot \left\{ 1 - \bar{\gamma}\left(\frac{n-1}{2}, \frac{r_{z_1}^2}{2}\right) + \sum_{i,d_1,\ldots,d_M : i\beta_1 + d_1\beta_2 + \ldots + d_M\beta_M < \alpha_{i,d_1,\ldots,d_M}^2} t_{i,d_1,\ldots,d_M}^{TC} \cdot \left(Q(\lambda_{i,d_1,\ldots,d_{M-1}}(z_1)) - Q(r_{z_1})\right) \cdot \bar{\gamma}\left(\frac{n-1}{2}, \frac{r_{z_1}^2 - \lambda_{i,d_1,\ldots,d_{M-1}}^2(z_1)}{2}\right) \right\} \cdot dz_1,$$

with

$$r_{z_1} = \left(1 - \frac{z_1}{\sqrt{n_1\beta_1 + \ldots + n_M\beta_M}}\right) r,$$

$$\lambda_{i,d_1,\ldots,d_M}(z_1) = r_{z_1}/r \cdot \frac{\sqrt{(n_1\beta_1 + \ldots + n_M\beta_M)(i\beta_1 + \ldots + d_{M-1}\beta_M)}}{\sqrt{(n_1 - i)\beta_1 + \ldots + (n_M - d_{M-1})\beta_M}},$$

$$\alpha_{i,d_1,\ldots,d_{M-1}}(z_1) = r\sqrt{1 - \frac{i\beta_1 + \ldots + d_{M-1}\beta_M}{n_1\beta_1 + \ldots + n_M\beta_M}},$$

$$\bar{\gamma}(a, x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt, \quad a, x > 0 \quad \text{and}$$

$$n = n_1 + \ldots + n_M,$$

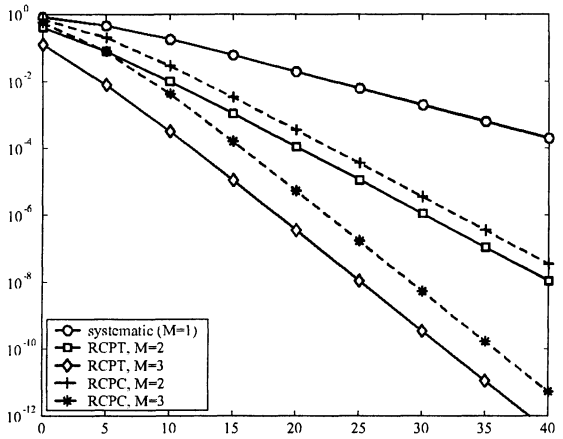


Figure 2: FER performance of systematic RCPT codes vs. systematic RCPC codes with $M = 1, 2$ and 3 (re-)transmissions, on the block-fading channel.

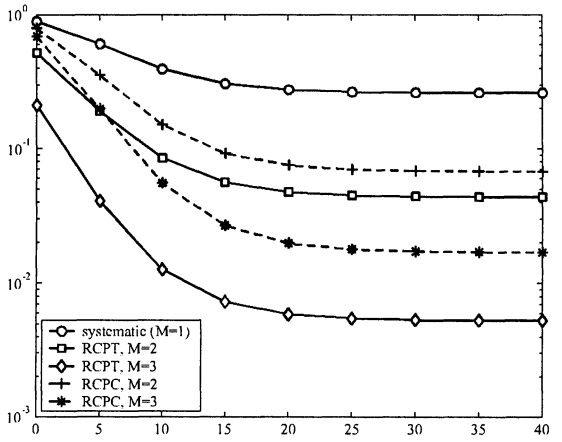


Figure 3: FER performance of systematic RCPT codes vs. systematic RCPC codes with $M = 1, 2$ and 3 (re-)transmissions, on the block-fading Gaussian collision channel with $N_u = 10$ users and transmission probability $p = 0.05$ ($G = 0.5$).

where n_1 is the maximum Hamming weight of the systematic subblock and $n_2, \ldots, n_M$ are the maximum Hamming weights of the remaining $M-1$ redundancy subblocks.

In order to achieve a tight bound there is an associated optimization problem to solve, *i.e.* finding the optimal cone radius r. This is time consuming since it includes finding the root to a relatively complicated integral equation with high requirement of precision in order to achieve a tight bound. The optimization problem can be stated in the following form

$$\frac{\sqrt{\pi}\Gamma(\frac{n-2}{2})}{\Gamma(\frac{n-1}{2})} = \sum_{i,d_1,\ldots,d_{M-1}\,:\, i\beta_1+\ldots+d_{M-1}\beta_M<\alpha_k^2} t^{TC}_{i,d_1,\ldots,d_M} \cdot \int_0^{\theta_{i,d_1,\ldots,d_M}} \sin^{n-3}\phi d\phi,$$

where

$$\theta_{i,d_1,\ldots,d_{M-1}} = \cos^{-1}\left(\lambda_{i,d_1,\ldots,d_{M-1}}(z_1)/r_{z_1}\right).$$

An upper bound on the FER of a particular coding scheme on the block-fading Gaussian collision channel with M (re-)transmissions can be obtained from the modified tangential sphere bound above and the joint PDF of the channel SNIRs $f(\beta_1, \ldots, \beta_M)$. Since the channel states of different transmissions are independent, the joint PDF can be factorized as

$$f(\beta_1, \ldots, \beta_M) = f(\beta_1) \cdot \cdots \cdot f(\beta_M),$$

with [15]

$$f(\beta_j) = e^{-\beta_j a} \cdot \left(\frac{1+\beta_j - p\beta_j}{1+\beta_j}\right)^{N_i-1} \cdot \frac{p \cdot N_i + a \cdot (1+\beta_j)(1+\beta_j - p\beta_j)}{(1+\beta_j)^2},$$

where $a = 1 \Big/ \frac{E_b}{N_0}$ is the inverse channel SNR. The upper bound on the FER can now be written as

$$P(e) \leq \int_0^\infty \cdots \int_0^\infty P(e|\beta_1, \ldots, \beta_M) \cdot f(\beta_1, \ldots, \beta_M) \cdot d\beta_1 \cdot \ldots \cdot d\beta_M. \quad (6)$$

The integrations in (6) are performed by Gaussian numerical quadrature (see for example [22]), where the semi-infinite integrals intervals are reduced to finite integrals using the bounding technique proposed in [7].

In Figure 2, the tangential sphere bound on the FER performance of a systematic RCPT code and a systematic RCPC code on the block-fading channel (no interference) is compared. The information sequence corresponds to 102 bits (including tail bits). The systematic RCPT code uses component codes with feedforward generator 5_8 and feedback generator 7_8, *cf.* Figure 1. The systematic RCPC code uses feedforward generators 7_8 and 5_8. This yields in total $M = 3$ subpackets available for incremental redundancy transmission for each information sequence.

Figure 3 shows the same performance comparison for the case of the block-fading Gaussian collision channel with channel load $G = 0.5$.

4 Conclusions

This paper has presented a performance evaluation technique for rate compatible punctured Turbo (RCPT) codes in incremental redundancy schemes over the block-fading Gaussian collision channel. A conditional upper bound on the frame error rate (FER), given the channel states of different packet (re-)transmissions over this channel, was established by modifying the tangential sphere bound by Poltyrev to apply for multivariate weight distributions. The expectation of this upper bound was then, by knowledge of the channel probability density functions, computed by multidimensional numeric integration.

It was demonstrated that RCPT codes have the potential to yield improved FER performance over the block-fading Gaussian collision channel, compared to rate compatible punctured convolutional (RCPC) codes. This improvement in FER performance translates into increased throughput of incremental redundancy schemes based on RCPT codes relative to RCPC codes.

References

[1] S. Lin and D. J. Costello, *Error Control Coding: Fundamentals and Applications.* Prentice-Hall, N.J., 1983.

[2] G. Caire and D. Tuninetti, "ARQ protocols for the Gaussian collision channel," *IEEE International Symposium on Information Theory, Sorrento, Italy*, Jun. 2000.

[3] D. Chase, "Code combining – a maximum-likelihood decoding approach for combining and arbitrary number of noisy packets," *IEEE Transactions on Communications*, vol. 33, pp. 385–393, May. 1985.

[4] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Transactions on Communications*, vol. 36, pp. 389–400, Apr. 1988.

[5] P. Frenger, P. Orten, T. Ottosson, and A. Svensson, "Rate-compatible convolutional codes for multirate DS-CDMA systems," *IEEE Transactions on Communications*, vol. 47, pp. 828–836, Jun. 1999.

[6] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near shannon limit error-correcting coding and decoding: Turbo codes," *In Proc. International Conference on Communications, Geneva, Switzerland*, vol. 2, pp. 1065–1070, May 1993.

[7] E. Malkamäki, *Performance of Error Control over Block Fading Channels with ARQ Applications.* PhD thesis, Helsinki University of Technology, 1998.

[8] S. Falahati, "Convolutional codes for wireless packet data systems," licentiate thesis, Chalmers University of Technology, Gothenburg, Sweden, 2000.

[9] A. S. Barbulescu and S. S. Pietrobon, "Rate compatible Turbo codes," *Electronics Letters*, vol. 31, pp. 535–536, Mar. 1995.

[10] D. N. Rowitch and L. B. Milstein, "Rate compatible punctured Turbo (RCPT) codes in a Hybrid FEC/ARQ system," *In Proc. Globecom'97*, pp. 55–59, 1997.

[11] J. Hámorský, U. Wachsmann, J. Huber, and A. Cizmár, "Hybrid automatic repeat request scheme with Turbo codes," *In Proc International Symposium on Turbo Codes, Brest, France*, pp. 247–250, Sep. 1997.

[12] J. Li and H. Imai, "Performance of hybrid-ARQ protocols with rate compatible Turbo codes," International Symposium on Turbo Codes , Brest, France, 1997.

[13] K. R. Narayanan and G. L. Stüber, "A novel ARQ technique using the Turbo coding principle," *IEEE Communications Letters*, vol. 1, pp. 49–51, Mar. 1997.

[14] R. J. McEliece, "Are Turbo-like codes effective on nonstandard channels?," *Information Theory Society Newsletter*, vol. 51, pp. 1–8, Dec. 2001.

[15] C. F. Leanderson and G. Caire, "The performance of incremental redundancy schemes based on convolutional codes in the block-fading Gaussian collision channel," *submitted to IEEE Transactions on Wireless Communications*, 2001.

[16] J. Hokfelt, O. Edfors, and T. Maseng, "A Turbo code interleaver design criterion based on the performance of iterative decoding," *IEEE Communications Letters*, vol. 5, pp. 52–54, Feb. 2001.

[17] X. Huang, N. Phamdo, and L. Ping, "Recursive method for generating weight enumerating functions of trellis codes," *Electronics Letters*, vol. 37, pp. 773–774, Jun. 2001.

[18] S. Benedetto and G. Montorsi, "Design of parallel concatenated convolutional codes," *IEEE Transactions on Communications*, vol. 44, pp. 591–600, May 1996.

[19] J. Proakis, *Digital Communications.* John Wiley and Sons, 1995. ISBN 0-07-051726-6.

[20] A. Viterbi and A. Viterbi, "Improved union bounds on linear codes for the input-binary AWGN channel, with applications to Turbo codes," *IEEE International Symposium on Information Theory, ISIT'98, Boston, USA*, p. 29, Aug. 1999.

[21] G. Poltyrev, "Bounds on the decoding error probability of binary linear codes via their spectra," *IEEE Transactions on Information Theory*, vol. 40, pp. 1284–1292, Jul. 1994.

[22] A. H. Stroud, *Approximate Calculation of Multiple Integrals.* Prentice-Hall,Englewood Clifs, New Jersey, 1971.

A PLL Clock Multiplier made from Digital Standard Cells

Thomas Olsson[1] and Peter Nilsson[2]
[1].Dept. of Electroscience, Lund University, Lund, Sweden. E-mail: Thomas.Olsson@es.lth.se
[2].Dept. of Electroscience, Lund University, Lund, Sweden. E-mail: Peter.Nilsson@es.lth.se

ABSTRACT: **A fully integrated digital PLL used as a clock multiplying circuit is designed. The PLL has no off-chip components and it is made from standard cells found in most digital standard cell libraries. It is therefore portable between processes as an IP-block in netlist format. Using a 0.35µm standard CMOS process and a 3.0V supply voltage, the PLL is designed for a frequency range of 170MHz to 360MHz and occupies an on-chip area of about 0.07mm^2.**

1. INTRODUCTION

PLLs are widely used circuits for clocking digital IP-blocks. Traditionally, a PLL is made as a partly analog IP building block. However, integrating an analog PLL in a digital noisy environment is difficult. In addition, the analog PLL is also sensitive to process variations and must therefore be redesigned for each new process.

Robust and easy implemented fully digital clock multipliers without phase locking are proposed in [1] and [2]. These clock multipliers produce a fixed number of cycles for each period of an external reference clock signal followed by an idle margin. For many digital applications, using such simpler clock generators is a possible solution. However, for a number of applications, using for instance synchronous on-chip communication, a PLL is necessary to ensure correct functionality.

In this paper, an implementation of a digital PLL made of digital building blocks is presented. The digital PLL is designed as an IP-block implemented as a netlist linked to a standard cell library. For most digital applications, a standard cell description of the PLL simplifies the design, since the design including the PLL becomes portable between technologies.

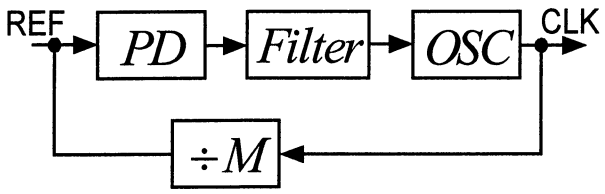


Figure 1. PLL block structure.

The block structure of a PLL, shown in figure 1, is divided into a phase detector, a loop filter an oscillator and a frequency divider.

For the digital PLL, the phase detector is a slightly modified standard type IV detector [3], the loop filter consists of a counter and a digital recursive filter, and the oscillator is a digitally controlled oscillator (DCO).

2. DESIGN OVERVIEW

Figure 2 shows the block structure for the digital PLL. The phase detector is controlling the counter, which is clocked by "clk_out".

A decoder transforms the output from the digital filter (W) to control signals (C) for the DCO. The output of the DCO is divided by the multiplication factor and then compared in phase to the reference. The signal "UPDATE" from the phase-detector is used to clock the registers and to reset the counter.

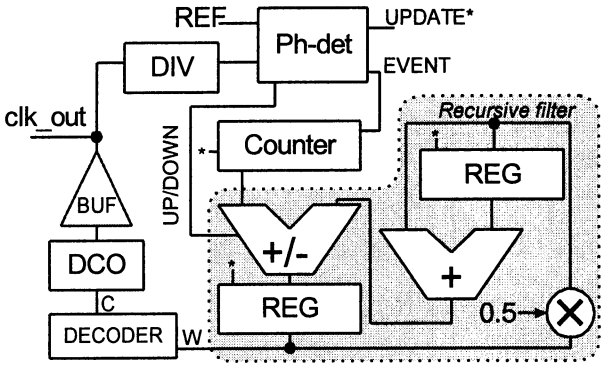


Figure 2. Block structure of digital PLL.

3. DIGITALLY CONTROLLED OSCILLATOR

Local clock generators based on ring oscillators have many advantages such as robustness, small size and low power consumption. The ring oscillator in its simplest form consists of an odd number of inverters connected in a circular chain. Such a circuit has no stable operation point and will therefore oscillate. The ring oscillator frequency is determined by the propagation time through the chain of inverters.

There are many methods to manipulate the ring oscillator frequency. The most straightforward technique is to change the propagation delay by changing the number of inverters. Other techniques are to use current starved inverters [4] or a delay line of controllable capacitors [5].

The difficulty when making an all standard cell PLL is often to implement an oscillator with high resolution. Often, a variable number of inverters are used for implementing a variable delay. However, this results in delay steps of several hundred ps which gives an inaccurate and unstabile phase lock for high frequency applications.

The oscillator for the digital PLL is a 7 stage ring oscillator with one inverter replaced by a NAND-gate for shutting down the ring oscillator during idle mode. To change the frequency of the ring oscillator, a set of 21 inverting tri-state gates are connected in parallel with each inverter (see figure 3). When the tri-state gates are enabled additional current drive is added to each inverter stage. The 126 tri-state gates are controlled by a 126 bit vector, C, which is decoded from a 7 bit control word, W. The vector C is all ones for W=0 and all zeros for W=125 to 127. For W<126, the number of zeros in C is equal to W.

In table 1, the number of enabled tri-state gates in parallel with each inverter is given for varying control word (W).

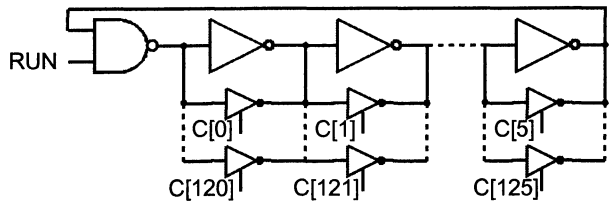


Figure 3. DCO.

W	Inv1	Inv2	Inv3	Inv4	Inv5	Inv6
0	21	21	21	21	21	21
1	20	21	21	21	21	21
2	20	20	21	21	21	21
|	|	|	|	|	|	|	|
|	|	|	|	|	|	|	|
123	0	0	0	0	1	1
124	0	0	0	0	0	1
125	0	0	0	0	0	0
126	0	0	0	0	0	0
127	0	0	0	0	0	0

Table 1. Number if enabled tri-state gates.

Simulated period time versus digital control word for the DCO at 3.0 V supply voltage is shown in figure 4. The plot in figure 4 has a slope of between 10 ps/bit and 55ps/bit. It is important to keep the slope low since the slope sets the resolution of the DCO.

A negative slope must also be avoided, since this might cause the PLL to be unstabile. The decoding from W to C makes negative slope impossible since adding current drive by enabling a tri-state gate can not slow down the oscillator.

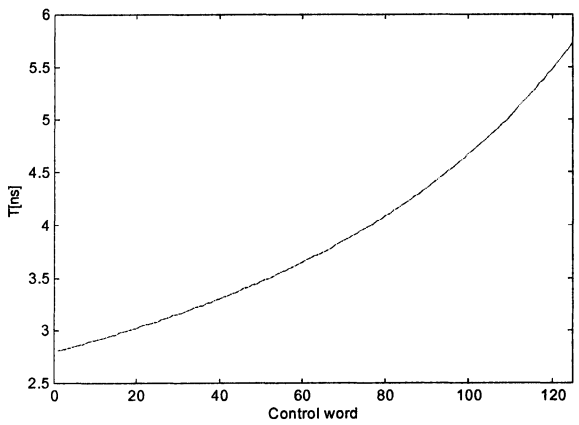


Figure 4. Simulated period time vs. digital control word (W).

4. PHASE DETECTOR

The type IV phase detector has in its original configuration two outputs controlling the oscillator frequency: One for signaling "UP" and one for signaling "DOWN" for the duration of the phase error. A slight modification is done to produce one signal "UP" or "DOWN" and one signal "EVENT" showing the length of the phase error. The modified phase detector is shown in figure 5.

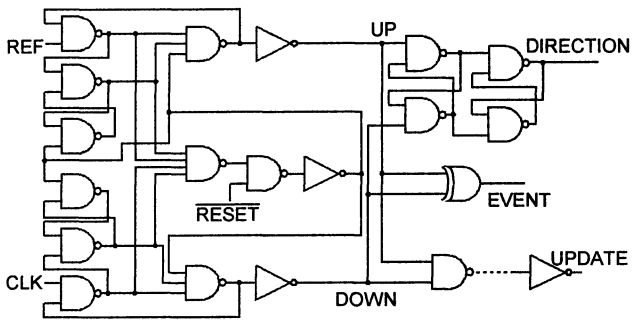


Figure 5. Phase detector.

Due to internal delay of the phase detector, the pulses at node "UP" and "DOWN" are longer than the actual phase error. There is therefore always a pulse at both nodes "UP" and "DOWN". For the "EVENT" signal in figure 5, the internal delay is canceled out. Thereby, a more accurate measurement of the phase error is achieved. This PLL implementation works as a synchronous digital circuit, which thus needs a clock pulse for updating all registers. At the end of each phase error a short pulse is produced at the node "UPDATE". This pulse, which is as long as the internal delay of the phase detector, is further delayed and used as a system clock. Figure 6 shows a simulation of the "UP", "DOWN", "EVENT" and "UPDATE" pulses. For the simulation, the phase error is set to 1.40ns, which gives an "EVENT" signal of about 1.42ns. The pulse

"UPDATE" is about 800ps independently of the phase error.

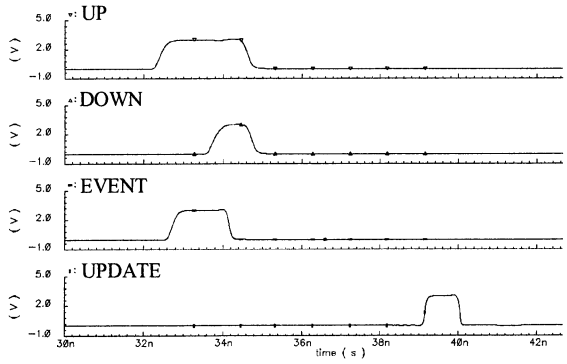


Figure 6. The "EVENT" and "UPDATE" signals.

5. LOOP FILTER

The loop filter consists of a counter and a digital recursive filter. The counter measures the phase error and the result from the counter is used to update the control word for the NCO. In order to get a high resolution when measuring the phase error, the counter is preferably clocked at a frequency higher than the output from the NCO. To avoid using a very high frequency for this purpose, clock flanks instead of complete pulses are counted during the phase error. This enables use of the output from the NCO instead of implementing an extra oscillator.

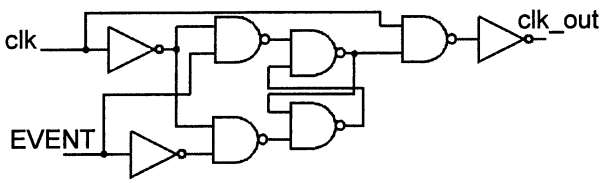


Figure 7. Clock-gating circuit.

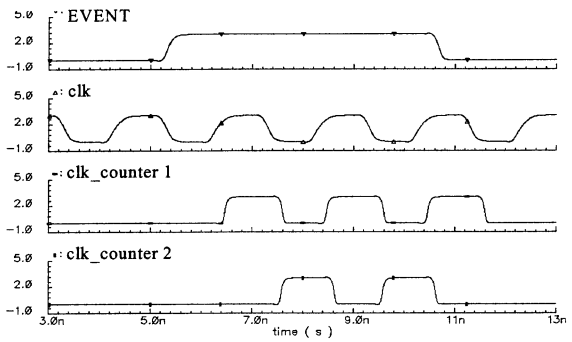


Figure 8. Gated clock pulses.

The circuit of figure 7 is used for gating a clock signal when "EVENT" is low. The clock pulses at the output "clk_out" are always complete pulses. Two three-bit counters are equipped with the circuit of figure 7. One counter is clocked with the gated version of the clock and the other counter is clocked with the gated version of the inverted clock. Figure 8 is a simulation showing the "EVENT" signal, the clock signal and the clock bursts counting up the two counters. The result from the two counters is then added to get the number of clock flanks during the phase error.

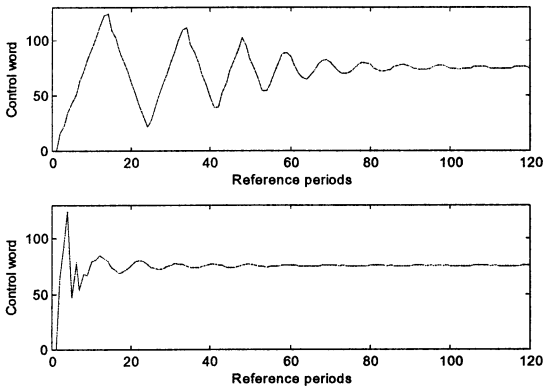


Figure 9. Step response.

Since the counters are only three-bit wide, they will often saturate during the initial phase of an impulse response. To get a faster impulse response and thereby a shorter lock time for the PLL, the output from the counters is multiplied by 4 whenever both counters are saturated. The effect of this is shown in figure 9, where the lower plot shows the improvement in step response.

Figure 10 shows the recursive filter, which is used for stabilizing the PLL feedback. The signal "W" is the digital control word to the NCO and the signal "Counter" is the sum of the two tree-bit counters.

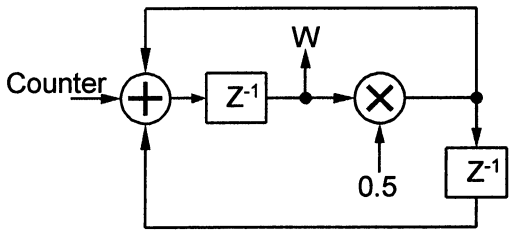


Figure 10. Recursive filter.

Figure 11 is a plot showing the digital control word when phase lock is achieved. The plot of figure 11 is a zoom in version of the lower part of figure 9. Since an exact frequency never is found, the control word will oscillate between two values during phase lock. This oscillation gives a phase noise, which is equal to the resolution of the NCO.

Since simulating a PLL is extremely time-consuming, a MATLAB model of the digital PLL is used for the simulations of figure 9 and 11.

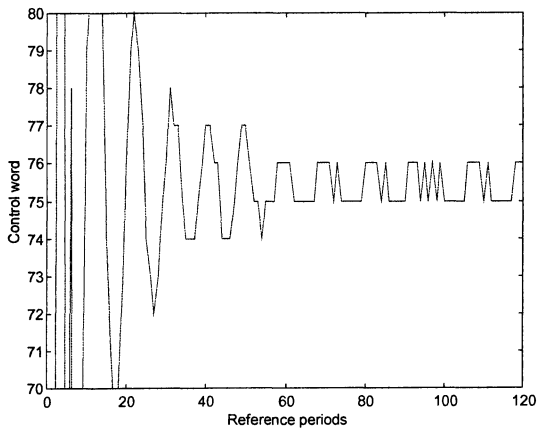


Figure 11. Phase lock.

6. CHIP LAYOUT

Figure 12 shows a chip photo of a prototype chip containing the digital PLL, which can be used as an on-chip IP-block. The actual core area for the digital PLL is limited to 0.07mm^2.

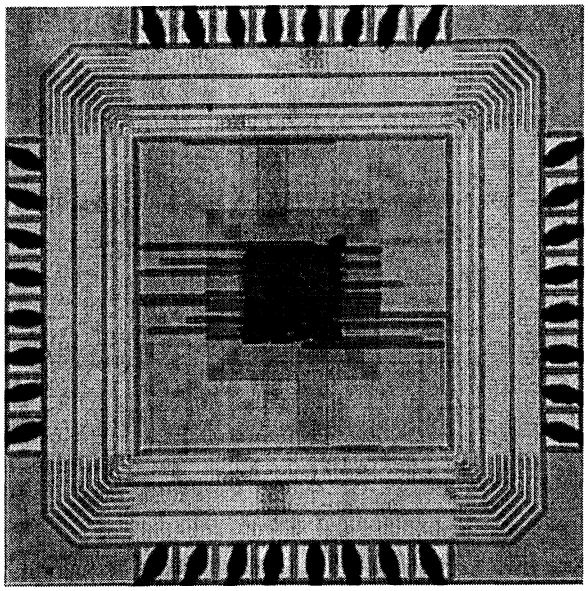


Figure 12. Chip Photo.

7. RESULTS AND CONCLUSIONS

A prototype of a standard-cell digital PLL clock multiplier is designed using a 0.35μm CMOS process. The digital PLL is designed for a frequency range of 170-360MHz and occupies about 0.07mm^2 of on-chip area. Instead of a charge pump and an analog filter, a counter and a recursive digital filter is used as a loop filter. A digitally controlled oscillator with high resolution is made from a ring oscillator with additional tri-state gates. The high resolution enables accurate frequency control and low phase noise.

The digital PLL is implemented using cells found in an ordinary standard-cell library, which makes it portable between technologies in a netlist format.

A prototype chip containing the digital PLL is fabricated and measurements are currently performed.

8. FUTURE WORK

The current design is made as a digital design in a schematic netlist editor. A more convenient method would be to have a synthesizable VHDL description of the PLL. This makes system simulation including the PLL possible in a digital VHDL simulator and it would also make the design completely process independent.

Another improvement is to make a new time to digital converter which has higher resolution without using any reference oscillator.

Using digital technique, it is rather easy to scale numbers, hence it would be possible to implement a PLL with adaptive loop gain.

The above changes together with other adjustments is currently work in progress.

9. REFERENCES

[1] P. Nilsson and M. Torkelson, "A Monolitic Digital Clock-Generator for On-Chip Clocking of Custom DSP's", *IEEE J. Solid-State Circuits*, vol.31, No. 5, pp. 700-706, May. 1996.

[2] T. Olsson , P. Nilsson, T. Meincke, A. Hemani and M. Torkelson. "A Digitally Controlled Low-Power Clock Multiplier for Globally Asynchronous Locally Synchronous Designs", In Proceedings of ISCAS'2000, Geneva, May 2000.

[3] R. E. Best, Phase-locked loops, McGraw-Hill, 1984.

[4] J. M. Rabaey. Digital Integrated Circuits: A Design perspective, Prentice hall, 1996.

[5] P. Andreani, F. Bigongiari, R. Roncella, R. Saletti and P. Terreni, "A Digitally Controlled Shunt Capacitor CMOS Delay Line", Analog Circuits and Signal Processing, Kluwer Academic Publishers, Volume 18, pp. 89-96.

Distortion in Multistage Feedback Amplifiers

Martin L. Lantz and Sven Mattisson

Abstract— **A symbolic model that predicts the linearity of multistage negative feedback amplifiers is presented. Memory-less behavior is assumed, allowing a 3rd-order power series to represent the amplifier system. This delimitation facilitates a synthesis approach.**

It is shown that, devices with high gain, even when they are associated with gross nonlinearities, are to be preferred. Furthermore, the contribution of amplifying stages to the nonlinearity is ranked by their signal-to-bias ratios, not by their position in the cascade.

The symbolic model applies to a majority of amplifier configurations. BJTs and FETs are treated in a unified way. Global and local feedback, as well as balanced and cascade configurations, are considered. Moreover, key device parameters, that predominately determine the linearity, are identified.

The high precision of the model has been confirmed by computer simulations.

I. Introduction

To minimize the power consumption of an amplifier with regard to its specified noise, linearity, and speed performance it is essential that the proper active device types, configurations, and bias currents are used [1]. Heuristics during amplifier synthesis can be avoided if a symbolic description of the relation between performance and synthesis parameters (i.e., device, configuration, and bias) is available.

A memory-less system model will represent the amplifier. Such a model is accurate within the signal frequency interval where memory effects are minute. Furthermore, a 3rd-order power series representation of the memory-less system will be suitable when the signal magnitude is small. This series rules the distortion figures of merit (e.g., intermodulation (IM) and harmonic distortion (HD)).

II. Active Devices

It is desirable to predict the general behavior of the amplifier system, with a minimum of effort. To this end, the primary effects are distinguished. The primary effects are modeled by simple functions able to describe the generic behavior of active devices. When greater detail is needed secondary effects are also considered.

A. Primary Effects

The exponential function couples a BJT input voltage to its output current,

$$i_o = g_{\mathrm{BJT}}(v_i) = I_S \exp(v_i/V_T),$$

where the constants have their usual meanings ($V_T = kT/q$). This function is approximated by the power series, found by Taylor series expansion about the nominal input voltage, v_{0i},

$$i_o = g_{\mathrm{BJT}}(\tilde{v}_i) = I_Q + \frac{I_Q}{V_T}\tilde{v}_i + \frac{I_Q}{2V_T^2}\tilde{v}_i^2 + \frac{I_Q}{6V_T^3}\tilde{v}_i^3, \qquad (1)$$

where $\tilde{v}_i$ is incremental signal, i.e., the deviation from v_{0i}. I_Q is the quiescent current (i.e., bias point). We otherwise assume, for simplicity, that signals are related linearly,

$$\frac{\tilde{i}_i}{\tilde{v}_i} = g_{i\mathrm{BJT}} = \frac{I_Q}{\beta_F V_T} \qquad \frac{\tilde{i}_o}{\tilde{v}_o} = g_{o\mathrm{BJT}} = \frac{I_Q}{V_A}, \qquad (2)$$

where β_F is the current gain factor and V_A is the forward Early voltage: we will not discriminate between incremental, nominal, or model parameter definition for the BJT current gain factor.

FETs that are biased in the saturation region with zero source bulk voltage ,$V_{SB} = 0$, are modeled by the quadratic function

$$i_o = g_{\mathrm{FET}}(v_i) = K(v_i - V_{th})^2,$$

where the constants have their usual meanings (V_{th} is the threshold voltage). As before, a power series representation is devised, this time we recognize the effective gate voltage also known as gate overdrive, $V_P = v_{0i} - V_{th}$,

$$i_o = g_{\mathrm{FET}}(\tilde{v}_i) = I_Q + \frac{2I_Q}{V_P}\tilde{v}_i + \frac{I_Q}{V_P^2}\tilde{v}_i^2. \qquad (3)$$

Again, the remaining signals are related linearly (V_E is the Early voltage),

$$\frac{\tilde{i}_i}{\tilde{v}_i} = g_{i\mathrm{FET}} = 0 \qquad \frac{\tilde{i}_o}{\tilde{v}_o} = g_{o\mathrm{FET}} = \frac{I_Q}{V_E}. \qquad (4)$$

Notice the analogous influence that V_T and V_P have on (1) and (3), respectively. These parameters are linked in yet another way. Like a BJT, a FET shows an exponential characteristic when biased at low V_P values, approaching V_T: viz, it operates in weak inversion.

B. Output-Related Notation

Nonlinearity is naturally related to the output, where signals are larger. Two output-related power series representations are needed in the sequel. They are defined here. A general description of both the power series (1) and (3) is

$$i_o = g(\tilde{v}_i) = k_0 + k_1\tilde{v}_i + k_2\tilde{v}_i^2 + k_3\tilde{v}_i^3, \qquad (5)$$

where $k_0 = I_Q$ and $k_1 = g_m$. The linear part of the output signal, $\tilde{i}_{1o} = k_{1g}\tilde{v}_i$, is distinguished, allowing the series (5) to be written

$$\tilde{i}_o = \tilde{i}_{1o} + k_{2o}\tilde{i}_{1o}^2 + k_{3o}\tilde{i}_{1o}^3, \qquad (6)$$

where the subscript 'o' identifies the new series coefficients.

The signal-to-bias ratio is the ratio of the incremental signal magnitude and the nominal (i.e., bias) signal value. The incremental and linear ratios are

$$m = \frac{\tilde{i}_o}{I_Q} \qquad \text{and} \qquad m_1 = \frac{\tilde{i}_{1o}}{I_Q},$$

respectively. Their significance is elaborated in section V. A further representation of the series (5) is now possible,

$$m = m_1 + \kappa_2 m_1^2 + \kappa_3 m_1^3, \qquad (7)$$

where the coefficients are dimensionless and identified by a prime.

The relation between the coefficients of the three power series representations are

$$\kappa_2 = k_{2o}I_Q = \frac{k_2}{k_1^2}I_Q \qquad \text{and} \qquad \kappa_3 = k_{3o}I_Q^2 = \frac{k_3}{k_1^3}I_Q^2. \qquad (8)$$

Based on these coefficients we introduce

$$\kappa_3' = \kappa_3 - 2\kappa_2^2, \qquad (9)$$

as this pattern frequently appears when the 3rd-order nonlinearity of devices in feedback configuration is determined, see section IV.

The signal-to-bias power series coefficients for BJTs and FETs are given in table I. Apparently, these coefficients are independent of the bias point as well as of any device parameters.

All signals discussed hereafter are incremental, therefore, we omit the tilde from now on.

This work was supported by the Swedish Foundation for Strategic Research (SSF) program; Personal Computing and Communication (PCC).

M. L. Lantz is with the Department of Electroscience, Lund Institute of Technology, SE-221 00 Lund, Sweden.

S. Mattisson is with Ericsson Mobile Platforms AB, Consumer Products, SE-221 83 Lund, Sweden.

TABLE I

SIGNAL-TO-BIAS-RATIO POWER SERIES COEFFICIENTS, DEFINED IN (7) AND (9), DESCRIBING PRIMARY EFFECTS OF AMPLIFYING STAGES.

Amplifying stage	κ_2	κ_3	κ_3'
common emitter (CE), BJT	1/2	1/6	−1/3
common source (CS), FET	1/4	0	−1/8
antiseries CE	0	−1/3	−1/3
antiseries CS	0	−1/8	−1/8
complementary-parallel CE	0	1/24	1/24
complementary-parallel CS	0	0	0

C. Circuit Model

A circuit model of an amplifying stage is given in Fig. 1. Its elements represent the power series (5) and the linear relation (2) or (4), BJT or FET respectively. Amplifying stages frequently appear in cascade configurations. Therefore, the model of the jth stage is surrounded by elements of the preceding $j-1$th stage as well as the succeeding $j+1$th stage.

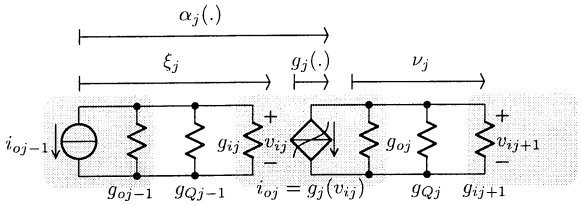


Fig. 1. Memory-less circuit model of an amplifying stage. It is distinguished by the central shaded area. The other elements originate in the preceding and succeeding amplifying stages as well as the biasing network.

The amplifier is treated on an abstraction level where the information signals and bias signals are considered orthogonal. Orthogonalization is possible if the signals have different characteristics in some domain (i.e., voltage, current, frequency, differential/common mode, or time). This view allows the synthesis of the biasing network to be handled later in the design trajectory [2].

The loading of the bias network of the jth stage is modeled by a linear conductance, g_{Qj}. The presence of a biasing network is otherwise understood.

D. Signal Flow Model

A controlled source represents a unilateral transfer. An apt signal flow representation of amplifying stages is, accordingly, based on the exclusive use of inputs and outputs of controlled sources as signal definitions. Fig. 2 illustrates such a general signal flow representation of an amplifying stage.

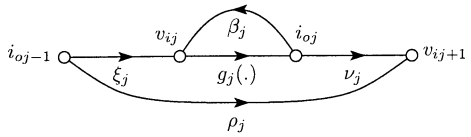


Fig. 2. Signal flow graph of an amplifying stage.

The four network parameters ξ_j, ν_j, β_j, and ρ_j in Fig. 2 correspond to the effects of input and output circuit loading, feedback, and direct feedthrough, respectively. The parameters are real constants given a memory-less system description.

The network parameters are also identified in Fig. 1. Obviously, both the feedback and direct feedthrough parameters are zero, $\beta_j = \rho_j = 0$. This is typical for a nonfeedback stage.

In the sequel, nonfeedback stages, assumed to appear in cascade, will be described by the stage current transfer, $\alpha_j(.)$. The transfer spans outputs of controlled sources,

$$i_{oj} = \alpha_j(i_{oj-1}) = g_j(\xi_j i_{oj-1}). \tag{10}$$

The stage current transfer will be represented by an output-related power series, see (6). Note that such a representation is independent of the input circuit loading effect, ξ_j. Therefore, the higher order coefficients of the stage current transfer are $\alpha_{2o} = k_{2o} = \kappa_2/I_Q$ and $\alpha_{3o} = k_{3o} = \kappa_3/I_Q^2$, which together with I_Q can be derived from table I. This is convenient since only the linear part of the stage current transfer (labeled stage current gain, $\alpha_{1j} = g_{1j}\xi_j$) then needs to be computed. By inspecting Fig. 1 one finds that the stage current gain is

$$\alpha_{1j} = \frac{-g_{1j}}{g_{oj-1} + g_{Qj-1} + g_{ij}}. \tag{11}$$

E. Key Device Parameters

The key parameters are those parameters whose values have to be known accurately as they predominately determine the system behavior. The key device parameters are found when the expression for the stage current gain, (11), is expanded by (2), (3), and (4).

Thus, the current gain, β_F, and forward Early voltage, V_A, are key BJT parameters. Given a prevailing input conductance influence, $g_{ij} \gg g_{oj-1} + g_{Qj-1}$, the stage current gain for CE stages becomes

$$\alpha_{1\mathrm{CE}j} = \frac{-g_{1j}}{g_{ij}} = -\beta_F. \tag{12}$$

The ability to manipulate the FET channel aspect ratio gives the designer additional degrees of freedom. Consequently, the key parameter set is not unique. We chose the effective gate voltage, V_P, and the Early voltage, V_E. They are both design parameters, i.e., they relate to device parameters as well as to the channel width and length.

Due to the zero input conductance of FETs, the preceding $j-1$th stage output conductance significantly contributes to the stage current gain. Assuming that the preceding stage output conductance is dominant, $g_{oj-1} \gg g_{Qj-1}$, the stage current gain for CS stages reduces to

$$\alpha_{1\mathrm{CS}j} = \frac{-g_{1j}}{g_{oj-1}} = \frac{-2I_{Qj}}{V_{Pj}} \frac{V_{Ej-1}}{I_{Qj-1}}. \tag{13}$$

The effective gate voltage decreases as the channel width is enlarged. Moreover, extending the channel length will boost the Early voltage. Evidently, both actions help improve the stage current gain. Knowing that large devices are slow one recognizes that there will be a trade-off between speed and linearity whenever FETs are involved.

F. Secondary Effects

Secondary effects (e.g., matching; parasitic series resistances, e.g., R_B and R_E; nonlinear base emitter conductance; high-level injection, I_K; and mobility reduction) can be introduced by a functional representation of the coefficients κ_2 and κ_3. Without loss of generality, we may, thus, address secondary effects implicitly in the following sections.

Whenever the linearity performance relies on either κ_2 or κ_3' being zero, one should be aware that the nonlinearity will be dominated by secondary effects.

III. Amplifying Stages

The structure given in Fig. 1 is an apt model for the basic as well as the balanced configurations. They are both unilateral. The feedback and cascade configurations do not fit this model, which is why their treatment will be postponed until sections IV and VI.

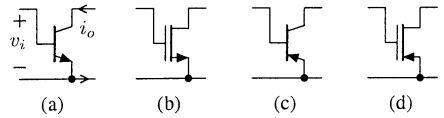


Fig. 3. Basic amplifying stages: (a) common emitter (CE), (b) common source (CS), (c) complementary CE, and (d) complementary CS.

A. Basic

The common emitter (CE) stage and the common source (CS) stage, depicted in Fig. 3, are basic, nonfeedback configurations. They are envisaged as the atom amplifying stages from which all other stages are composed.

Table I shows that the CS stage is intrinsically more than twice as linear as the CE stage. The nonzero κ_3' of CS stages suggests that FETs have a nonzero 3rd-order nonlinearity when operated in a feedback configuration, despite $\kappa_3 = 0$.

A complementary device is different in that its input and output signal has a reverse polarity, i.e., $i_o = g_{\text{comp}}(v_i) = -g(-v_i)$. This has the effect of reversing the polarity of κ_2 in table I.

B. Balanced

The balanced configuration compensates for even nonlinearities. Its operation can be traced to the addition of two suitable transfers. We treat the complementary-parallel and the antiseries configurations, exemplified in Fig. 4. While the first uses two complementary devices, the second uses identical devices.

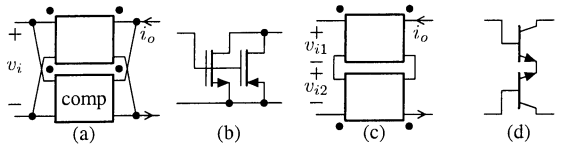


Fig. 4. Balanced amplifying stages: (a) complementary-parallel configuration; (b) complementary-parallel CS, i.e., inverter; (c) antiseries configuration; and (d) antiseries CE, i.e., differential stage.

The complementary-parallel configuration is represented by

$$i_o = g_{\text{CP}}(v_i) = g(v_i) + g_{\text{comp}}(v_i), \tag{14}$$

where the complementary devices need to match perfectly to counterbalance all even nonlinearities. Clearly, this will be difficult.

The numbers in table I suggest that the CS complementary-parallel configuration (i.e., inverter) will be perfectly linear. This is an artifact of primary effects modeling. Obviously, secondary effects need to be considered when the complementary-parallel configuration of CS stages is investigated.

The antiseries configuration (i.e., differential stages), see Fig. 4, is represented by

$$i_o = g_{\text{AS}}(v_i) = g(v_{i1}) = -g(-v_{i2}), \tag{15}$$

where i_o has an implicit relation to $v_i = v_{i1} + v_{i2}$.

This implicit relation is governed by the condition that both devices output signal currents are equal.

Naturally, the reduction of even nonlinearity will be limited by matching. Due to parameter uncertainty, a linearity improvement of no more than one order of magnitude can be expected.

IV. Memory-Less Feedback System

The feedback configuration can be applied both globally and locally. Linear global feedback, found in most high-performance amplifiers, comprises a spanning feedback two-port network connected to a cascaded active two-port network (active part), see Fig. 5.

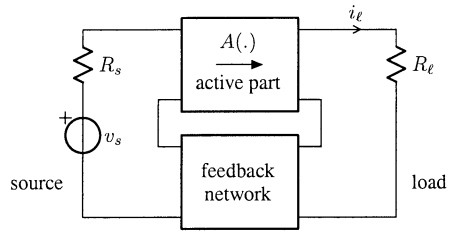


Fig. 5. A series-series feedback configuration constituting a transconductance (V→I) amplifier.

Such an amplifier fixes the relation, called transmittance A_t, between the source signal, e_s, and the load signal, e_ℓ, where e marks electrical quantity. We limit our treatment to signals represented in the voltage and current domain. Hence, solely single-feedback loop amplifiers will appear here. Furthermore, it is assumed that the desired transmittance is both linear and memory-less.

A. Signal Flow Model

All nonlinearities are contained within the active part. Linear modeling of the source and load condition (i.e., R_s and R_ℓ) is assumed to be accurate. Also, the feedback network will be linear, due to the desired linear transmittance.

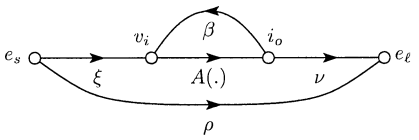


Fig. 6. Signal flow graph describing a negative feedback amplifier, where $A(.)$ comprises all nonlinearities. Other relations are assumed to be linear.

The signal flow representation of a negative feedback amplifier, illustrated in Fig. 6, is similar to that of an amplifying stage. The network parameters ξ, ν, β, and ρ are equivalent to those defined in section II-D. All network relations are linear. The nonlinearity of amplifiers can still be modeled, since the active transfer, $A(.)$, comprises all nonlinearities.

A power series representation of the system in Fig. 6 can be calculated using series reversal. These coefficients have been derived in [3] as well as in [4] where loading and direct feedthrough effects were ignored, however.

We chose a load-related notation (similar to the series (6)) of the memory-less feedback system,

$$e_\ell = e_{1\ell} + A_{\ell 2} e_{1\ell}^2 + A_{\ell 3} e_{1\ell}^3. \tag{16}$$

In practical amplifiers, the direct feedthrough effects are incidental, and can often be ignored, i.e., $|\rho| \ll |\nu\xi A_1/(1 - A_1\beta)|$. The system coefficients are now given by

$$\begin{aligned} A_{\ell 2} &= \frac{A_{2o}}{\nu(1 - A_1\beta)} \\ A_{\ell 3} &= \frac{A_{3o} + 2A_{2o}^2 A_1\beta/(1 - A_1\beta)}{\nu^2(1 - A_1\beta)}, \end{aligned} \tag{17}$$

where A_{2o} and A_{3o} are the output-related power series coefficients of the active transfer, and $A_1\beta$ is the loop gain.

Both A_{2o} and A_{3o} are present in the expression for $A_{\ell 3}$. Therefore, when operated in a feedback configuration, devices with quadratic transfers will also show 3rd-order nonlinearity. The numerator will read $A_{3o} - 2A_{2o}^2$ when the loop gain is high, $|A_1\beta| \gg 1$. This explains the importance of relation (9).

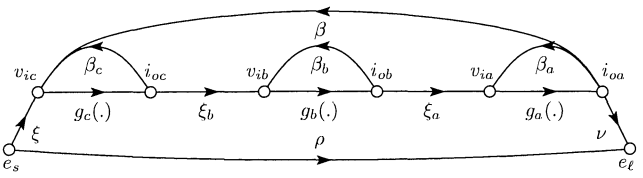


Fig. 7. Signal flow graph of a three-stage negative feedback amplifier. The direct feedthrough parameters ρ_a, ρ_b, and ρ_c have been omitted for clarity.

B. Cascade Feedback Configuration

The active part of a high-performance negative feedback amplifier comprises several amplifying stages connected in cascade, $A(.) = a(b(c(.)))$. They will be labeled with the letters a, b, and c starting with the output stage, see Fig. 7.

Implementations with more than three stages are rare, since the chances of efficient frequency compensation of such amplifiers are limited. However, a larger number of stages is feasible if local feedback is applied.

The following discussion uses a three-stage implementation for demonstrative purposes. The analysis is by no means limited in this regard, however. Any number of amplifying stages can, in principle, be investigated. Yet only cascade combinations of amplifying stages that preserve negative feedback are useful for amplifying purposes.

A signal flow description of a three-stage negative feedback amplifier is illustrated in Fig. 7. It is devised by repetitive combination of the amplifying stage model in Fig. 2 and the amplifier model in Fig. 6.

It is illustrative to first have a look at a system where stage feedback and direct feedthrough effects are minute, i.e., β_a, β_b, β_c, ρ_a, ρ_b, and ρ_c are all zero. The system coefficients are found by combining the coefficient expressions for feedback systems (17) and cascade systems. The system coefficients are then

$$
\begin{aligned}
A_{\ell 2} &= \frac{1}{F\nu}\left(\frac{\kappa_{2a}}{I_{Qa}} + \frac{\kappa_{2b}}{\alpha_{1a}I_{Qb}} + \frac{\kappa_{2c}}{\alpha_{1a}\alpha_{1b}I_{Qc}}\right) \\
A_{\ell 3} &= \frac{1}{F\nu^2}\left(\frac{\kappa'_{3a}}{I_{Qa}^2} + \frac{\kappa'_{3b}}{\alpha_{1a}^2 I_{Qb}^2} + \frac{\kappa'_{3c}}{\alpha_{1a}^2\alpha_{1b}^2 I_{Qc}^2}\right. \\
&\quad \left. + \frac{-2\kappa_{2a}\kappa_{2b}}{\alpha_{1a}I_{Qa}I_{Qb}} + \frac{-2\kappa_{2a}\kappa_{2c}}{\alpha_{1a}\alpha_{1b}I_{Qa}I_{Qc}} + \frac{-2\kappa_{2b}\kappa_{2c}}{\alpha_{1a}^2\alpha_{1b}I_{Qb}I_{Qc}}\right),
\end{aligned}
\tag{18}
$$

where the return difference $F = 1 - A_1\beta$ is defined for brevity. Also, the signal-to-bias-ratio power series coefficients in table I are identified. Furthermore, high global loop gain was presumed when $A_{\ell 3}$ in (18) was computed. From the structure of the system coefficients (18) we may draw some general conclusions:

1. Increasing the loop gain (i.e., the return difference, F) is one of the most efficient ways to improve linearity. High loop gain is therefore a characteristic of high-performance amplifiers. There is an upper limit to how much one amplifying stage can contribute to the loop gain. If loop gain in excess of this amount is desired additional amplifying stages are needed.

2. The output stage signal needs to be larger than the load signal due to the output-loading effect, ν. The linearity drops proportionally.

3. In the early days before feedback was known, increasing the bias current, I_Q, was the only means of reducing distortion in amplifiers. Unfortunately, such amplifiers suffer from low efficiency. The bias currents can be reduced while maintaining performance if the loop gain is increased appropriately. In situations where the available loop gain is insufficient (e.g., high-frequency signal processing) efficiency still has to be sacrificed.

4. The choice of technology dictates the device coefficient values, κ_2 and κ'_3. They are, in general, of the same order of magnitude. Therefore, the ability of amplifying stages to contribute to the loop gain is more important. Low-gain stages which are almost linear are therefore avoided in favor of high-gain stages with gross nonlinearity.

5. The stage current gain, α, is an important figure-of-merit for amplifying stages, since it is the upper limit to its loop gain contribution. High-stage current gain reduces the influence of preceding stages on distortion. When the output stage current gain is high, other stages can often be safely ignored. Such a condition considerably simplifies the amplifier design trajectory.

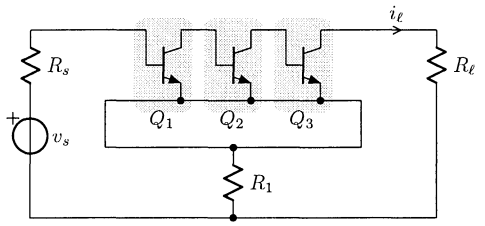


Fig. 8. Signal path of a triple CE stage transconductance (V→I) amplifier.

C. Example: A BJT Amplifier

To be able to estimate the nonlinearity of the amplifier in Fig. 8, we will compute its system coefficients, $A_{\ell 2}$ and $A_{\ell 3}$. The output conductance of the CE stages and the biasing network are ignored, since their influence on the signal path is here assumed to be insignificant compared with that of the input conductances.

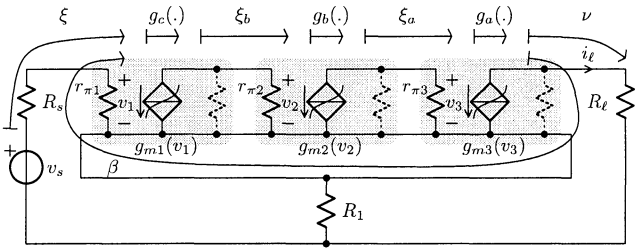


Fig. 9. Memory-less circuit model of a triple CE stage transconductance (V→I) amplifier.

We switch temporarily to a conventional element notation, e.g., $g_{m3} = g_{1a}$ and $r_{\pi 1} = 1/g_{ic}$. By inspecting Fig. 9 one finds; $\alpha_{1a} = -g_{m3}r_{\pi 3} = -\beta_F$, $\alpha_{1b} = -g_{m2}r_{\pi 2} = -\beta_F$, $A_1 = g_{m1}\beta_F^2$, $\beta = -r_{\pi 1}R_1/(r_{\pi 1}+R_s+R_1)$, and $\nu = 1$. Table I gives us: $\kappa_2 = 1/2$ and $\kappa'_3 = -1/3$. Note that only linear transfers have to be computed. Assuming $|A_1\beta| \gg 1$ the system coefficients (18) reduce to

$$
\begin{aligned}
A_{\ell 2} &= \frac{r_{\pi 1} + R_s + R_1}{R_1\beta_F^3}\left(\frac{1}{2I_{Q3}} + \frac{-1}{2\beta_F I_{Q2}} + \frac{1}{2\beta_F^2 I_{Q1}}\right) \\
A_{\ell 3} &= \frac{r_{\pi 1} + R_s + R_1}{R_1\beta_F^3}\left(\frac{-1}{3I_{Q3}^2} + \frac{-1}{3\beta_F^2 I_{Q2}^2} + \frac{-1}{3\beta_F^4 I_{Q1}^2}\right. \\
&\quad \left. + \frac{1}{2\beta_F I_{Q3}I_{Q2}} + \frac{-1}{2\beta_F^2 I_{Q3}I_{Q1}} + \frac{1}{2\beta_F^3 I_{Q2}I_{Q1}}\right).
\end{aligned}
$$

Recall that $r_{\pi 1}$ is inversely proportional to I_{Q1}, see (2). As a result, an increase in the input stage bias current, I_{Q1}, will often be advantageous, even at the expense of output stage bias current, I_{Q3}, given a limited current budget.

Judging by the form of the coefficient $A_{\ell 2}$ the 2nd-order nonlinearity will be minute when $I_{Q3} = \beta_F I_{Q2}$. This canceling effect is due to the negative transfer of common emitter stages. One must be aware, however, that only primary effects have been dealt with.

V. Signal-to-Bias Ratio

The conditions for the bias currents to achieve low distortion are twofold. First, bias currents need to be sufficiently large to avoid clipping. Clipping occurs when the signal excursion forces a device to behave like a switch. Then either the output voltage or current of the device is zero. We assume that this first condition is honored, i.e., class A biasing. Second, the total available bias current needs to be partitioned in an adequate way to minimize the system coefficients, (18).

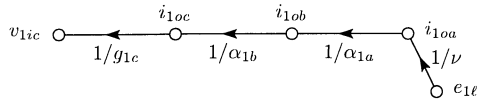


Fig. 10. Linear anticausal representation of the active signal path.

The output signal magnitude of stages are found by anticausal analysis, see Fig 10. A linear representation of the signals is sufficient. The signal-to-bias ratios (7) of each individual stage are

$$m_{1a} = \frac{e_{1\ell}}{\nu I_{Qa}} \qquad m_{1b} = \frac{e_{1\ell}}{\nu\alpha_{1a} I_{Qb}} \qquad m_{1c} = \frac{e_{1\ell}}{\nu\alpha_{1a}\alpha_{1b} I_{Qc}}. \tag{19}$$

We see that high-stage current gain helps to reduce the signal-to-bias ratios. Note, m_{1j} might be negative, whenever stages have negative current gain.

The identification of the ratios (19) in (18) results in a compact and illustrative representation of the system coefficients,

$$\begin{aligned} A_{\ell 2} e_{1\ell} &= \frac{1}{F}\left(m_{1a}\kappa_{2a} + m_{1b}\kappa_{2b} + m_{1c}\kappa_{2c}\right) \\ A_{\ell 3} e_{1\ell}^2 &= \frac{1}{F}\big(m_{1a}^2\kappa'_{3a} + m_{1b}^2\kappa'_{3b} + m_{1c}^2\kappa'_{3c} \\ &\quad -2m_{1a}m_{1b}\kappa_{2a}\kappa_{2b} - 2m_{1a}m_{1c}\kappa_{2a}\kappa_{2c} \\ &\quad -2m_{1b}m_{1c}\kappa_{2b}\kappa_{2c}\big). \end{aligned} \tag{20}$$

It is now evident that the relationship between linearity and the signal-to-bias ratios is paramount. Moreover, all stages rank equally in this regard. It is thus the stage with the largest signal-to-bias ratio that dominates the linearity performance.

The system representation (20) determines the harmonic and intermodulation distortion, i.e., $\mathrm{HD}_2 = \frac{1}{2}\left|A_{\ell 2}\hat{e}_{1\ell}\right|$, $\mathrm{HD}_3 = \frac{1}{4}\left|A_{\ell 3}\hat{e}_{1\ell}^2\right|$, $\mathrm{IM}_2 = \left|A_{\ell 2}\hat{e}_{1\ell}\right|$, and $\mathrm{IM}_3 = \frac{3}{4}\left|A_{\ell 3}\hat{e}_{1\ell}^2\right|$.

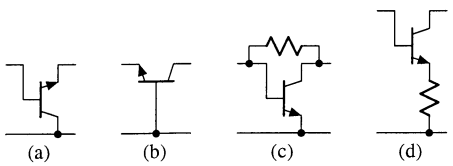


Fig. 11. Local feedback amplifying stages: (a) common collector, i.e., emitter follower; (b) common base, i.e., current follower; (c) shunt stage; and (d) series stage.

VI. Local Feedback

Local feedback, exemplified in Fig. 11, is deployed chiefly to facilitate biasing or frequency compensation. Its use marks the linearity performance of the negative feedback amplifier, however.

The available global return difference, $F_{\mathrm{avail}} = 1 - \alpha_{1a}\alpha_{1b}g_{1c}\beta$, is reduced in proportion to the local return differences, $F_j = 1 - g_{1j}\beta_j$.

The power series coefficients for local feedback stages (e.g., a_{2o} and a_{3o}) are now found by utilizing (17), while ignoring the output-loading effect. Once the stage coefficients are known, (18) is used to derive the system coefficients, so that

$$\begin{aligned} A_{\ell 2} &= \frac{1}{F_{\mathrm{avail}}\nu}\left(\frac{\kappa_{2a}F_bF_c}{I_{Qa}} + \frac{\kappa_{2b}F_a^2F_c}{\alpha_{1a}I_{Qb}} + \frac{\kappa_{2c}F_a^2F_b^2}{\alpha_{1a}\alpha_{1b}I_{Qc}}\right) \\ A_{\ell 3} &= \frac{1}{F_{\mathrm{avail}}\nu^2}\left(\frac{\kappa'_{3a}F_bF_c}{I_{Qa}^2} + \frac{\kappa'_{3b}F_a^3F_c}{\alpha_{1a}^2I_{Qb}^2} + \frac{\kappa'_{3c}F_a^3F_b^3}{\alpha_{1a}^2\alpha_{1b}^2I_{Qc}^2}\right. \\ &\quad + \frac{-2\kappa_{2a}\kappa_{2b}F_aF_c}{\alpha_{1a}I_{Qa}I_{Qb}} + \frac{-2\kappa_{2a}\kappa_{2c}F_aF_b^2}{\alpha_{1a}\alpha_{1b}I_{Qa}I_{Qc}} \\ &\quad \left. + \frac{-2\kappa_{2b}\kappa_{2c}F_a^3F_b}{\alpha_{1a}^2\alpha_{1b}I_{Qb}I_{Qc}}\right). \end{aligned} \tag{21}$$

From this we conclude that the influence of local feedback on the system coefficients is threefold. First, the global loop gain is reduced. Second, the stage coefficient value is decreased, due to local linearization. Third, signals that precede the local feedback stage have to be greater, due to the reduced stage current gain.

The linearity of negative feedback amplifiers is generally not improved by applying local feedback. This is evident in (21): the system coefficients are enlarged by an increase of the local return differences, F_j. Nevertheless, linearity improves within the narrow interval where cancellation effects occur, see section VII.

The linearity loss can be slight, if local feedback is applied to the amplifying stage that dominates the linearity performance. The least harmful place for local feedback is therefore often the output stage. Unfortunately, other stages quickly become important as local loop gain increases. Therefore, the exchange between global and local loop gain, with almost no linearity loss, will only be possible for moderate levels of local loop gain.

The contribution of output and intermediate stages to $A_{\ell 2}$ in (21) are equal in magnitude, for the local loop gain value of

$$F_{a,\mathrm{2cancel}} = \left|\alpha_{1a}\frac{\kappa_{2a}I_{Qb}}{\kappa_{2b}I_{Qa}}\right|^{\frac{1}{2}}. \tag{22}$$

Cancellation effects appear in the neighborhood of this value. When surpassed, severe linearity loss occurs.

As an application, frequency compensation is assisted by a global loop gain reduction. This reduction can be implemented without degrading the linearity if the local loop gain is kept lower than the value in (22).

A. Example: Local Series Feedback at the Output Stage

Fig. 12 shows a negative feedback amplifier with local series feedback, R_3, at the output stage. The local loop gain is essentially proportional to R_3. The system coefficients are given by (21), where F_a is nonunity.

This example substantiate the accuracy of the model. The system coefficients (21), where secondary effects have been acknowledged, are compared with computer simulations (transient analysis followed by discrete Fourier transformation) using commercial SiGe HBT device models. The agreement is excellent, see Fig. 13, where $\hat{\imath}_{\ell,\mathrm{IP2}}$ and $\hat{\imath}_{\ell,\mathrm{IP3}}$ are load referred intermodulation distortion intercept currents (the circumflex denotes signal amplitude).

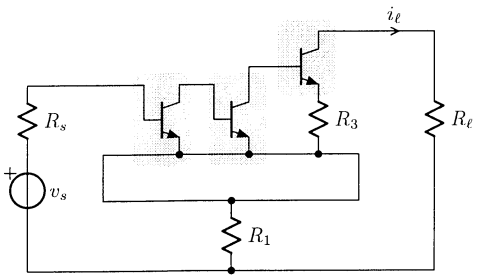


Fig. 12. Signal path of a three-stage transconductance (V→I) amplifier with local series feedback, R_3, at the output stage.

Furthermore, Fig. 13 demonstrates how no linearity performance is sacrificed at moderate levels of local feedback. In contrast, dramatic linearity degradation takes place at high levels of local loop gain. Moreover, the predicted 2nd-order cancellation effect, (22), manifests itself in the $\hat{\imath}_{\ell,\mathrm{IP2}}$ peaking. Note, the functional representation of the coefficients κ_2 and κ_3, that modeling secondary effects, shifts the cancellation instant somewhat.

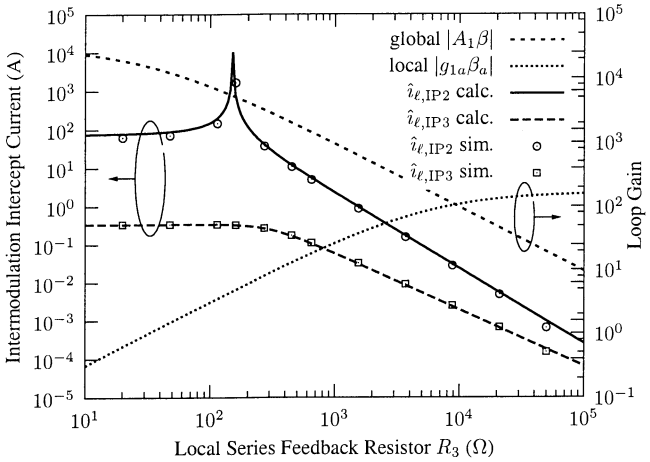


Fig. 13. Three-stage transconductance (V→I) amplifier with local series feedback, R_3, at the output stage. Calculated (including secondary effects) and computer-simulated load referred intermodulation distortion intercept currents, $\hat{\imath}_{\ell,\mathrm{IP2}}$ and $\hat{\imath}_{\ell,\mathrm{IP3}}$, are compared. $A_1\beta$ is the global loop gain and $g_{1a}\beta_a$ is the local loop gain. The signal conditions are: $R_1 = 1\mathrm{k}\Omega$, $R_s = 50\Omega$, $R_\ell = 50\Omega$, $\hat{\imath}_\ell = 20\mu\mathrm{A}$, $I_{Q1} = I_{Q2} = I_{Q3} = 1\mathrm{mA}$, $V_{CEQ} = 2\mathrm{V}$, and $g_Q = 1\mathrm{mS}$. Devices used represent typical SiGe HBT technology: $\beta_F = 220$, $I_K = 7.88\mathrm{mA}$, $V_A = 22.86\mathrm{V}$, $R_B = 52.45\Omega$, and $R_E = 2.99\Omega$. The compression point is at $\hat{\imath}_{\ell,\mathrm{CP}} = 1\mathrm{mA}$ ($-16\mathrm{dBm}$). The right axis can be read as: $\mathrm{dBm} = 44 + 20\log(\mathrm{A})$.

VII. Nonlinearity Cancellation

Nonlinearity cancellation occurs when the sum of the terms in the system coefficients (18) is zero. Designs that rely on cancellation require that parameters are both well-known and controlled. Parameter uncertainty hampers the yield, however, and effective countermeasures are expensive. Besides, it is not likely that both 2nd and 3rd-order system coefficients can be canceled simultaneously. When nonlinearity cancellation does occur, it originates from some of the effects described next.

A. Second-Order Cancellation

Some of the stage current gains, α_1, or coefficients κ_2 need to be negative, to enable cancellation of $A_{\ell 2}$ in (18). The majority of amplifying stages have a negative current gain, refer to (11). There will, consequently, be plenty of stage combinations where 2nd-order cancellation is conceivable. Another possibility is the presence of a complementary device, since its κ_2 has reverse polarity, see section III-A.

Cancellation can also occur due to local feedback. Consider, for instance, an implementation using two CE stages with equal bias currents. Cancellation subsequently occurs when the local loop gain is equal to $\sqrt{\beta_F}$, see (22).

B. Third-Order Cancellation

Third-order cancellation is less likely, since the stage current gains, α_1, are squared in the dominant terms of the coefficient $A_{\ell 3}$ in (18). However, the complementary-parallel CE stage provides an opportunity to cancel 3rd-order system coefficient, since its κ_3' has opposite polarity.

The proportion of the two terms within $A_{\ell 3}$ in (17) depends on the amount of global loop gain. Therefore, $A_{\ell 3}$ will cancel at the return difference magnitude of $F = A_{3o}/(A_{3o} - A_{2o}^2)$, which corresponds to a loop gain magnitude of $-1/2$ for a single CE stage and 0 for a single CS stage implementation. Remember, this type of cancellation does not apply to local feedback. This can be understood by realizing that the 3rd-order system coefficient depends both on the 2nd and the 3rd-order local feedback stage coefficients, so canceling only one of the stage coefficients does not suffice.

VIII. Discussion

To conclude, a symbolic model that links the linearity of multistage negative feedback amplifiers and the synthesis parameters (e.g., configuration and bias currents) has been presented. The unified treatment of devices as well as stage configurations supports the significance of the model. Only linear transfers need to be computed, facilitating a synthesis application.

It has been shown that the gain of an amplifying stage is more important than its intrinsic linearity. Due to their high gain, the use of BJTs seems beneficial when high demands are put on both speed and linearity. If FETs are preferred they should, by the same token, be biased at low effective gate voltages.

Guidelines for optimal bias current allocations have been given. It is the amplifying stage with the largest signal-to-bias ratio that determines nonlinearity. Notwithstanding, a relatively large input stage bias current is often beneficial for BJT implementations.

The causes of nonlinearity cancellation have been identified. They explain why 2nd-order cancellation is more common than 3rd-order.

Local feedback is first and foremost a way to facilitate frequency compensation. The local loop gain should be kept moderate, if severe linearity loss is to be avoided. The pole split (i.e., Miller) frequency compensation is a frequency-dependent form of local feedback and might therefore influence the linearity in an adverse way. It is recommended that frequency compensation techniques which do not interfere with the active part (e.g., phantom zero compensation [5]) are attempted first.

References

[1] J. Davidse and E. H. Nordholt, "Basic considerations concerning the application of semicustom IC techniques for the processing of analogue electronic signals," *J. Semi-Custom ICs*, vol. 5, no. 2, pp. 5–11, 1987.

[2] C. J. M. Verhoeven and A. van Staveren, "Systematic biasing of negative feedback amplifiers," in *Proc. DATE*, pp. 318–322, Mar. 1999.

[3] J. W. T. Eikenbroek, *Development of an Integrated AM Shortwave Upconversion Receiver Front-End.* PhD thesis, Delft University of Technology, 1989.

[4] S. Narayanan, "Application of Volterra series to intermodulation distortion analysis of transistor feedback amplifiers," *IEEE Trans. Circuit Theory*, vol. CT-17, pp. 518–527, Nov. 1970.

[5] M. S. Ghausi and D. O. Pederson, "A new design approach for feedback amplifiers," *IRE Trans. Circuit Theory*, vol. 8, pp. 274–284, Sept. 1961.

On-chip signaling for future SOC systems

Imed Ben Dhaou and Hannu Tenhunen
Electronics System Design Laboratory, Dept. of Electronics
KTH-Electrum, Electrum-229
Royal Institute of Technology,
SE-16440 Kista, Sweden
E-mail: imed@ele.kth.se

Abstract

In deep-submicron technology, interconnection delay and signaling power consumption for global communication started to be several orders of magnitude larger than gate power consumption and gate delay. This paper surveys various schemes for energy efficient signaling. An analogy between digital communication and on-chip signaling is described. Finally, the potential of the deployment of each signaling scheme for future SoC is discussed.

1 Introduction

In deep-submicron technology interconnect capacitances of global wires is several order of magnitude larger than intrinsic capacitances of CMOS transistors. Further, As the feature sizes decreases, the number of transistor per unit area increases accordingly, consequently the average wire-length increases and the spacing between interconnects decreases. As a result, the crosstalk capacitance between adjacent lines increases as the technology scales down. As a result, the peak crosstalk noise increases accordingly [1]. In addition, the noise margin of the CMOS circuits is decreasing due to the continuous reduction of supply and threshold voltage.

Traditionally, in submicron regime, the most efficient way for sizable energy-saving has been based on lowering the supply voltage. This results in the speed degradation of CMOS gates. In order to restore the speed, techniques like pipelining and parallelization have been widely used with the main focus on gate parameters, that is, delay and power consumption.

In deep-submicron, this technique has started to be inadequate because of the following reasons (i) the continuous scaling of the supply voltage; (ii) the increasing DSM noise due to e.g. increasing number of switching devices and the increase of the crosstalk to self capacitance ratio; (iii) the continuous decrease of the noise margin of the CMOS circuitry, which makes them highly vulnerable to DSM noise; and (iv) the constant increase of interconnection delay for medium and long wires as the technology scales down.

The wire, defined as a communication medium, is used to carry signals from one gate to the other. This means in DSM the power consumed by the gates such as AND, XOR, OR, INV, and MUX, which is referred to as *intrinsic power dissipation* in [15], started to be several orders of magnitude smaller than the power needed for signaling between two connected gates. The latter is referred to as *on-chip signaling power consumption* in [15]. These facts show that any efficient power saving scheme should address the problem associated with wires.

In order to tackle noise, delay, and power in DSM, our approach is to design a power-efficient, robust signaling techniques that allow for reliable communication between CMOS gates. This will be referred to as *energy-efficient on-chip signaling*.

This paper reviews recently published techniques for energy efficient on-chip signaling. It also reports an analogy between digital communication over band-limited channel in the presence of severe fading and on-chip signaling in the presence of DSM noise. This analogy is very helpful to understand and thereby solve some of the communication/ signaling bottlenecks such as bus encoding [9], bus sharing [14], and multi-level signaling [12].

The rest of the paper is organized as follows. In Section.2, the impact of interconnect parasitics on propagation delay is studied followed by problem statement. Efficient wire modeling techniques are described in Section.3. Section 4 surveys different signaling techniques for future SoC. Finally, the paper is concluded in Section 5.

2 Problem statement

In order to quantify the impact of interconnect on both propagation speed and power dissipation, in [15], we have studied two cascaded inverters shown in Fig.2.

Dynamic power dissipation and propagation delay of the circuit shown in Fig.2 are given by, respectively, Eq.1 and Eq.2

$$P_{driv} = \underbrace{\alpha f V_{dd}^2 dC}_{P(d,f)} + \underbrace{\alpha f V_{dd}^2 C_{intr}}_{P_{intr}}, \quad (1)$$

where α is the switching probability, P_{intr} is the intrinsic power consumption, $P(d, f)$ is the power consumed due to signaling over a wire of length d.

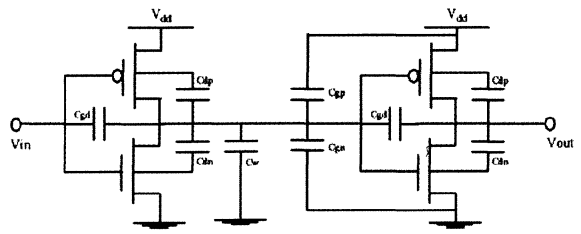


Figure 1. Cascaded inverters.

Similarly, the propagation delay is given by

$$t_p = \underbrace{\frac{\beta C_w}{2V_{dd}}}_{t(d)} + \underbrace{\frac{\beta C_{intr}}{2V_{dd}}}_{t(intr)}, \quad (2)$$

where $\beta = \frac{L_{tr}}{W_{tr}}\left(\frac{1}{k'_n} + \frac{1}{3k'_p}\right)$, k'_n and k'_p are respectively the transconductance of the PMOS and NMOS transistor, W_{tr} and L_{tr} are respectively the channel width and length of the NMOS transistor, where it is assumed that the PMOS transistor is three times wider than the NMOS transistor[2].

From Eq1 and Eq.2 we see that the power dissipation of the driving inverter is proportional to the wire-length, proportional to the square of the supply voltage and proportional to the product of switching activity and the wiring capacitance. For global communication, top metal layers are used for on-chip signaling. The top metal layers have lower resistivity but larger capacitance as compared to lower metal layers. That means signaling power consumption of top metal layers is larger than signaling power consumption of lower metal layers. However, the signal propagates faster on top metal layers than on lower ones. The described equation shows that propagation speed is linearly proportional to the wire-length.

The above equations were derived considering a simple lumped capacitance model and using long-channel transistor model. These simplifications led to the the conclusion that the propagation delay is linearly proportional to the wire-length. However, for some specific wire length ranges and for some specific wiring geometry, the wire must be modeled using either a distributed RC or RLC circuit [13]. For the case of RC wire model, the propagation delay is proportional to the square of the wire length.

2.1 Problem formulation

To address the wire bottleneck in deep-submicron technology, two problem formulation that aids the design of energy efficient signaling were described in [11] and [12]. The first problem formulation was described for voltage mode signaling where the aim was to reduce the energy dissipated while driving a global interconnect given delay and noise budgets.

The second problem formulation was generalized for an arbitrary Bit Error Rate (BER). In order to establish a closed form expression of the BER, in [12], voltage mode signaling was shown to be exactly pulse amplitude modulation, and the interconnect was accurately modeled as a low-pass filter. This formulation is very general and thus it is re-considered in this work.

Let us define by $\gamma_{i,j}$ the quantity that measures the power efficiency of the signaling scheme used to transmit data from B_i to B_j, which is given by

$$\gamma_{i,j} \triangleq \frac{P_{i,j}}{R_{i,j}}, \quad (3)$$

where $P_{i,j}$ is the power consumed by B_i to send the data to B_j over a distance $d_{i,j}$ at a given target data rate $R_{i,j}$ and target BER. The problem formulation for this generalized case is depicted in Fig. 2. In order to solve the optimization problem described in

Problem formulation
minimize $\gamma_{i,j}$ subject to
- The achievable data-rate should be equal to $R_{i,j}$
- $BER_{i,j} \leq \tau$

Figure 2. Problem formulation for multiple on-chip transmitters/receivers.

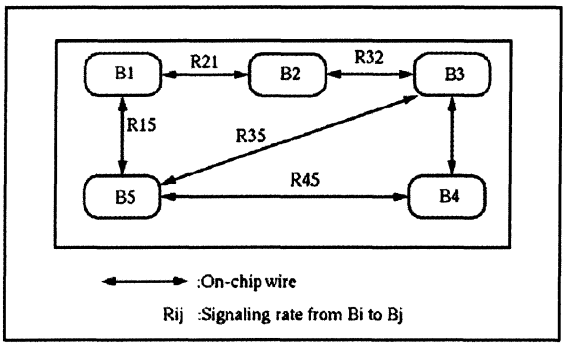


Figure 3. Power efficient signaling problem for the case of multiple on-chip transmitters/receivers.

Fig.2, it is mandatory to efficiently model the interconnect which is treated in subsequent section.

3 Efficient wire modeling

In the Gigahertz range, Maxewell's equations for the transverse electromagnetic (TEM) waves and quasi-TEM are reduced to Telegraph's equations. For higher operating frequency, a full-wave analysis must be carried out to solve Maxwell's equations. Given the validity range of quasi-TEM, the study presented in the work is confined to the TEM and quasi-TEM case.

In the quasi-TEM structure, a conductor of length d is accurately approximated using infinitesimal lumped RLC elements.

The amplitude and phase of the voltage waveform transmitted on the on-chip wire is the solution of the differential equation given in (4).

$$\frac{\partial^2 v(x,t)}{\partial x^2} = RC\frac{\partial v(x,t)}{\partial t} + LC\frac{\partial^2 v(x,t)}{\partial t^2}, \quad (4)$$

where R, L, and C are the per-unit value of the resistance, inductance and capacitance. Often the transmission line is coupled to one or many transmission lines. Thus, in order to elaborate an accurate closed-form expression of the propagation delay, the crosstalk noise and the power dissipation, the DFE given by Eq.4 should be generalized to account for the coupling effect among parallel wires [13].

The signal propagated on-chip can be either current of voltage waveform. In CMOS and other silicon technologies the signal is ramp in shape. The clock signal is a special case where the ramp signal is repeated periodically with periodicity equals inverse of the clock frequency. It follows from the approximation theory that any periodic signal can be represented using a Fourier series. Let $s(t)$ be the clock signal of period T, that is $s(t) = s(t + kT)$, where k is a positive integer. Fourier series of $s(t)$ is given by Eq.5.

$$s(t) \approx a_0 + \sum_{n=1}^{+\infty}\left[a_n \cos(\frac{2n\pi t}{T}) + b_n \sin(\frac{2n\pi t}{T})\right], \quad (5)$$

where $a_n = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}\left[s(t)\cos(\frac{2\pi nt}{T})dt\right]$, $b_n = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}\left[s(t)\sin(\frac{2\pi nt}{T})dt\right]$, and $j = \sqrt{-1}$. This means that at any instant of time, the clock signal is accurately approximated using an infinite summation of sinusoidal signals. This suggest that the signal $s(t)$ has a finite bandwidth. In the context of high-speed signaling, the widely used expression of the signal bandwidth of a ramp signal is

$$f_{max} = \frac{0.35}{T_r}, \quad (6)$$

where T_r is the rise time. Physically, the signal bandwidth defines the maximum frequency, f_{max}, which corresponds to cut-off frequencies at the -3 dB points [7]. The maximum frequency as well as physical parameters of the medium (length, wire geometry), determine the appropriate interconnect model.

Traditionally in micron technology the maximum clock frequency was few MHz and the number of metal layers was at maximum two, thus, an interconnect of length d was accurately approximated using one lumped L, T or Π RC model. In [8], it was found that Π_3 and T_3 accurately approximate a distributed RC wire as compared to the L lattice model. In very high speed design the inductance must be used to model a long wire. Given the angular frequency, $w = \frac{0.7\pi}{T_r}$, the wire is modeled as distributed RLC if $R < Lw$ or $T_r \leq l\sqrt{LC}$. Again, for some specific frequency range, the wire parasitics parameters i.e. R, L, and C are independent of the frequency. However, as the frequency increases beyond the skin effect frequency, then the RLC wire must be modeled using R(f)L(f) and C(f) model.

Similarly to the distributed RC wire, the problem faced to simulate and to elaborate a closed form expression of output voltage and delay expression is the estimation of the number of ladder Π, T, or L cells to appropriate model a wire of length d given the driver rise time. Lots of works have done in this area. In [5] a closed form expression that computes the number of lumped elements, given by Eq.7, is reported. In Star-Hspice, the number of lumped elements used to approximate a metal wire using Eq.8 [6]. In [7] an iterative algorithm that computes the appropriate number of cells is reported.

$$N_\pi = \frac{10d\sqrt{LC}}{T_r}, \quad (7)$$

where T_r is the signal rise time.

$$N_\pi = 1 + \frac{20d\sqrt{CL}}{T_r}. \quad (8)$$

The widely used technique to approximate a given wire of length d is to segment it into N equally sized sections. Each section can be approximated using either Π or T model. In [13], it was found

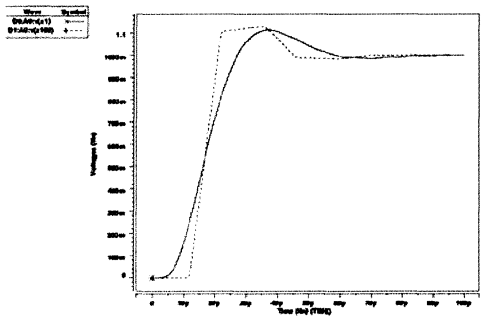


Figure 4. Transient simulation of an RLC wire of length 1mm that has the following per-unit values, $R = 11\omega/mm, C = 86.178fF/mm$, and $L = 1.5528nH/mm$. The wire is terminated by a capacitive load $C_L = 3fF$ and driven by a driver that has an output resistance of value $R_0 = 100\Omega$. The dashed line is the plot of the transient simulation of the wire modeled using 100 lumped elements. The solid line is the same plot if the wire is modeled using one lumped element.

that T and Π models are indifferent for larger N. This undeniably means that for an appropriate number of lumped element, the delay formula obtained using a wire that is approximated using N lumped T model leads to the same results if it is modeled using N lumped Π model. However, the simulation of an interconnect that has larger number of N is costly in both time and memory requirements. There are have been lots of research work that aimed at reducing the order of interconnect based on model order reduction such as Padè approximation. This issue is however beyond the scope of this work.

4 Efficient signaling techniques

Before we proceed into the description of several energy efficient signaling, it is worthwhile to review basic concepts used to (i) characterize different signaling schemes and (ii) optimize the singling scheme. These concepts are delay, power dissipation, and noise immunity. The expression of interconnection delay and power dissipation of voltage mode signaling for a wide range of interconnect models are described in [13]. It was found that the crosstalk capacitance can drastically increase both propagation delay and dynamic power dissipation. For two coupled wires, this worst case scenario occurs when both victim and aggressor switch in opposed direction.

4.1 Bus encoding for low-power

From Eq.1, we know that, in the absence of crosstalk noise, the power dissipation is proportional to the switching activity and the wire length. This suggest that the signaling power dissipation can be reduced by reducing the switching activity. This can be accomplished using an appropriate bus encoding scheme. In the last decade, lots of research works have been carried out, see e.g. [4] and [9].

In the presence of crosstalk noise, the energy dissipation depends on the switching of neighboring wires. In this case, the encoding scheme should minimize inter-wire transition which can be accomplished using for example Transition Pattern Coding (TPC) as proposed in [18].

Another popular technique to optimize the power dissipated by busses is to reduce its number. This can be accomplished by sharing busses between multiple resources which can be done using multiple access techniques such as DS-CDMA as proposed in [14].

4.2 Voltage mode signaling

The interconnection delay of an RC wire is proportional to the square of the wire length. In order to propagate the signal at higher speed, the wire must be partitioned so that the delay becomes linear on the wire length. This approach is commonly known as buffer insertion for optimum delay [3]. However, this scheme cannot guarantee a noise free functionality, or *perfect communication*[11]. To alleviate this problem, the wire must be buffered to guarantee both minimum delay and noise free operation.

Buffer insertion for noise and/or delay optimization comes at the expense of increased power dissipation [11]. To address this issue, the approach proposed in [11] and [17] is based on low-swing signaling combined with buffer resizing. The problem with this voltage mode signaling is its vulnerability to power supply noise. As the technology scales down, the impact of power supply noise on circuit reliability is quickly becoming intolerable. This calls for efficient method to deal with power supply noises. One of the most popular technique is current mode signaling.

4.3 Multi-level current mode signaling

It is well known that current mode signaling is faster, more immune and more energy efficient than voltage mode signaling. In [16], the propagation delay of both current mode and voltage mode were derived. The interconnect was modeled as ladder RC circuit. The wire is terminated at a load resistance R_L, and the driver is modeled as current source in parallel with a resistance of value R_B. The expressions of interconnection delay for current and voltage mode are given, respectively by Eq.9 and Eq.10.

$$\Delta_v = \frac{d^2RC}{2}\left(1 + 2\frac{R_B}{dR}\right). \tag{9}$$

$$\Delta_i = \frac{d^2RC}{2}\left(\frac{R_B + \frac{dR}{3}}{R_B + dR}\right). \tag{10}$$

Lemma 1 *Current mode signaling propagates faster than voltage mode signaling if and only if the wire length satisfies the following relation.*

Proof: *Suppose that voltage mode signaling propagates faster than voltage mode signaling if the wire parameters belong to some specific range. Then, the following relation must hold*

$$1 + 2\frac{R_B}{dR} \leq \frac{R_B + \frac{dR}{3}}{R_B + dR}. \tag{11}$$

The right hand side of the equation can be written as

$$\frac{R_B + \frac{dR}{3}}{R_B + dR} = \frac{R_B + dR}{R_B + dR} - \frac{2\frac{dR}{3}}{R_B + dR}. \tag{12}$$

If we combine Eq.11 and Eq.12, we obtain

$$2\frac{R_B}{dR} \leq -\frac{2\frac{dR}{3}}{R_B + dR}. \tag{13}$$

This relation is impossible since the right hand side term in the above equation is always negative. This concludes the proof. □.

In [12] it was found that quaternary current mode signaling is a promising candidate for signaling over long on-chip wires. A more elaborated formula for encoder-decoder pair insertion is described in [10].

4.4 Differential signaling

In differential signaling, the signal and its complement are transmitted. The receiver regenerates the original signal by comparing the two received differential signals. Mathematically, the signal x_k is encoded into two differential signals y_{2k} and y_{2k+1} as given by

$$x_k = y_{2k} - y_{2k+1}. \tag{14}$$

For example, if the signal to be transmitted is 1V than one possible implementation would be to have $y_{2k} = -y_{2k+1} = 0.5V$. Low Voltage Differential Signaling (LVDS) was found to be more robust and more energy efficient than low-swing voltage mode signaling combined with buffer insertion [19]. It was also found that differential signaling greatly reduces inductive coupling. However, the problem with this scheme is the increased number of signal paths which can have adverse effects such increasing both crosstalk noise and chip area. Additionally, this scheme is based on voltage mode signaling. Thus, it is highly vulnerable to DSM noise.

5 Conclusions

This paper described three popular signaling schemes for future SoC. The impact of the wires on both propagation delay and dynamic power dissipation is reviewed. It was pointed out that in the future technologies, the impact of crosstalk noise on both interconnection delay and dynamic power dissipation is becoming more severe. In addition, the power supply noise is increasing as the technology scales down. These issues suggest that multi-level current mode signaling is a very effective way for signaling over long wires. To efficiently address the increasing number of on-chip busses, multiple access techniques or packet routing might be the ultimate solution to this problem. For signaling over on-chip busses, the impact of switching activity of neighboring wires on energy dissipation is getting higher than the impact of temporally switching of single wire. This motivates the needed for encoding schemes that aim at minimizing the inter-wire switching activity. This scheme is referred to as TPC and is also shown to be outstanding in reducing interconnection delay. In this line of defense, a very popular signaling scheme which was not explored in this work is to deploy RF to transmit timing critical global signals, such as clock signal [20]. To this end, we strongly believe that the potential of differential signaling is very tiny.

Acknowledgments

The authors would like to thank Jimson Mathew for carefully reading and commenting on earlier versions of this paper. This work was supported by the strategic research program Personal Computing and Communication (PCC).

References

[1] J. M. Wang, Q.Yu and E.S. Kuh, " Coupled noise estimation for distributed RC interconnect model", in Proc. DATE, pp.664 -668, 1999.

[2] J.M. Rabaey, " Digital Integrated Circuits: A Design Perspective", Prentice-Hall 1996, ISBN 0-13-178609-1.

[3] C. Alpert and A. Devgan, " Wire Segmenting for Improved Buffer Insertion", Proc. Design Automation Conf. June 1997.

[4] S. Ramprasad, N. Shanbhag, I. Hajj, "A Coding Framework for Low-Power Address and Data Busses", IEEE Trans. On VLSI, No. 2, pp. 212-221, June 1999 .

[5] W-K. Chen, " The VLSI handbook", The CRC Press and the IEEE Press, 1999.

[6] Avant! Corporation, "Star-HSPICE Manual", Release 99.2, June 1999.

[7] T. Dhaene, D. Zutter, "Selection of Lumped Element Models for Coupled Lossy Transmission Lines", IEEE Trans. On CAD, vol. 11, issue 7 , pp. 805 -815, July 1992.

[8] T. Sakurai, " Approximation of Wiring Delay in MOSFET LSI", IEEE J. Of Solid-State Circuits, Vol.SC-18, No.4, pp.418-426, Aug. 1983.

[9] M. Stan, W. Burleson, "Low-power encodings for global communication in CMOS VLSI", IEEE Trans. on VSLI Systems, Issue 5, no. 4, 1997, pp. 49-58.

[10] I. Ben Dhaou, E. Dubrova and H. Tenhunen, "Power Efficient Inter-Module Communication for Digit-Serial DSP Architectures in Deep-Submicron Technology", in Proc. ISMVL, pp.61 -66, May 2001.

[11] I. Ben Dhaou, V.Sundararajan, H. Tenhunen and K. K. Parhi, "Energy Efficient Signaling in Deep Submicron CMOS Technology", in Proc. ISQED, pp. 319 -324, March 2001.

[12] I. Ben Dhaou, E. Dubrova, H. Tenhunen, and M. Ismail, " Current mode Low Power On-Chip Signaling in Deep-Submicron CMOS Technology", IEEE Trans. On Circuits and Systems I, Fundamental Theory and Applications, TCAS, 2001 (Submitted)

[13] I. Ben Dhaou, "Low Power Design Techniques for Deep Submicron Technology with Application to Wireless Transceiver Design ", PhD Thesis, Royal Institute of Technology, draft version.

[14] R. Yoshimura, T.B.Keat, T.Ogawa, S.Hatanaka, T.Matsuoka, K.Taniguchi, "DS-CDMA Wired Bus With Simple Interconnection Topology for Parallel Processing System LSIs", in Proc. ISSC, pp. 370-371, Feb. 2000.

[15] I. Ben Dhaou, V.Sundararajan, H. Tenhunen and K. K. Parhi, "Energy Efficient Signaling in Deep Submicron CMOS Technology", in Proc. ISCAS, pp.319 -324, 2001.

[16] E. Seevinck, P.J. van Beers, and H. Ontrop, "Current-mode Techniques for High-Speed VLSI Circuits with Application to Current Sense Amplifier for CMOS SRAM's", JSSCC, Vol.26, No. 4, pp.525-535, April 1991.

[17] I. Ben Dhaou, Tenhunen and K.K. Parhi, " Energy efficient signaling in Deep-Submicron Technology", Special Issue on Timing Analysis and Optimization for Deep Sub-Micron (DSM) IC, VLSI design journal, 2002.

[18] P.P. Sotiriads, A. Chandrakasan, "Bus energy Minimization by Transition Pattern Coding (TPC) in Deep Sub-Micron Technologies ", Proc. ESSIRC, pp.322-233, 2000.

[19] Y. Massoud, J. Kawa, D. MacMillen, J. White, "Modeling and Analysis of Differential Signaling for Minimizing Inductive Crosstalk", Proc. DAC, pp. 804-809, 2001.

[20] F. M. Chang, V. P. Roychowdhury, L. Zhang, H. Shin, Y. Qian, "RF/Wireless Interconnect for Inter- and Intra-chip Communications", Proc. of the IEEE, vol.89, issue 4, pp. 456 -466, April 2001.

NEW ASTRONOMY WITH THE ODIN SATELLITE

Åke Hjalmarson

Centre for Astrophysics and Space Science at Chalmers University of Technology,
Onsala Space Observatory, SE-439 92 Onsala, Sweden. E-mail: hjalmar@oso.chalmers.se

Abstract

On 20 February 2001 the Odin satellite - a new observatory for sub-millimetre wave spectroscopy – was launched from Svobodny in far-eastern Russia. The Odin project is a Swedish-lead, shared (50/50%) astronomy/aeronomy mission, supported by space agencies and scientists in Canada, Finland, France and Sweden. The satellite has been developed and is operated by the Swedish Space Corporation. The astronomy part of the mission is focussed on observations of spectral line emission from interstellar H_2O and O_2.

Odin houses a high-precision 1.1m diameter offset Gregorian telescope equipped with a very flexible cryogenic submm/mm receiver package. The frequency range 541-581 GHz is covered by three tuneable Schottky mixers and a fourth Schottky mixer covers the band 486-504 GHz. A 119 GHz fix-tuned HEMT preamplifier has been installed to allow very sensitive searches for interstellar O_2. All receivers are operated in single-sideband mode. Any combination of four, three, or two receivers (depending upon the available power and mode of operation) can be used in combination with two auto-correlation spectrometers (bandwidth range: 100-800 MHz; resolution range: 0.125 -1 MHz) and an acousto-optical spectrometer, AOS (bandwidth: 1000 MHz; resolution 1 MHz). The Odin antenna beam-widths are about 2' and 9' at 557 and 119 GHz.

This talk presents the in-orbit characteristics of the Odin satellite observatory and highlights some new astronomy results from the first months of regular observations.

Information on the Odin project can be found on the Odin web-page of the Swedish National Space Board (http://www.snsb.se/Odin/Odin.html).

Radio studies of physics in space: New ideas and new facilities

Bo Thidé

Swedish Institute of Space Physics, Physics in Space, Box 537 SE-751 21 Uppsala, Sweden. E-mail: bt@irfu.se

Abstract

With the huge LOFAR radio telescope, proposed to be built in the Netherlands and Germany, and the radar sub-facility LOIS, proposed in southern Scandinavia and to be operated in the 10–240 MHz frequency range, a number of ideas of new radio studies of physics in space have been put forwars. We describe some of these new ideas as well as the facilities themselves.

1 Introduction

Since the pioneering radio experiments of the early 20th century, many active radio and radar techniques have been developed for the study of the near-Earth space. The literature describing this development and the employment of the various techniques is plentiful (*e.g.*, *Davies* [1], *Hunsucker* [2], and *Kohl et al.* [3]).

The advent of efficient digital techniques for system control and data handling allows for the flexible, cost-effective design of large, long-wavelength (HF/VHF) transmitting and receiving antenna arrays. Thus, we now have the ability to build long-wavelength radio and radar facilities which can be used for studies with a sensitivity and sophistication that has not been previously possible, thereby opening a new frequency window in active deep-space radio research.

An example of such a facility is the proposed Low Frequency Array (LOFAR) radio telescope[1] which will operate in the 10-240 MHz frequency range. Its 13,000 dipole antennas will be clustered in roughly 100 stations spread over a region 400 km across. Connected through fibre optical links, this interferometric array will be capable of digitally forming multiple simultaneous beams that can be electronically steered, will possess extreme frequency agility, and will provide a flexible system for distributed control, signal processing, monitoring, and remote operation.

Scandinavian space, radio, and IT scientists have proposed that the receive-only LOFAR radio telescope be augmented by a ground-based digital transmitter/radar facility called LOIS.[2] Fully configurable and operated in software, the LOFAR-LOIS combination will enable unprecedented deep-space HF/VHF radar applications.

[1] http://www.lofar.org

[2] www.lois-space.org

Space radar techniques can be uniquely powerful methods of obtaining information not only on near-Earth space but also on planetary objects including planets, moons, asteroids, and comets; the interplanetary medium; and the Sun. Measurements of the distribution of echo power versus direction angle (image), time delay (range), Doppler frequency (radial velocity), spectral width (turbulence), and radio wave polarisation (magnetic field) constitute multi-dimensional observations of potentially high resolution, given high enough signal-to-noise ratios.

A key factor in space radar investigations is human control of the transmitted signal used to illuminate the target. While virtually every other astronomical technique relies on passive measurement of reflected sunlight or naturally emitted radiation, radar uses coherent illumination whose time/frequency/phase structure and polarisation state are designed by the scientist. The general stratagem of a radar observation is to transmit a signal with well-defined characteristics and then, by comparing the echo to the transmission, deduce the target's properties. Hence, the observer is intimately involved in an active observation and, in a very real sense, performs a controlled experiment on the target.

2 Remote Sensing of Geospace and the Inner Heliosphere

The inner heliosphere is usually defined as the part of the circumsolar region where the solar magnetic field deviates 45 degrees or less from the radial direction. This corresponds roughly to radial distances out to one astronomical unit ($1\,\mathrm{AU} = 1.496 \times 10^{11}$ m) and encompasses the sunward part of geospace.

While processes in the inner heliosphere and geospace are of decisive importance for Earth's interplay with its space surrounding and for life on Earth itself, many of these processes are still not well understood.

In situ measurements of the interplanetary medium in the near-solar environment at about 0.3 AU$= 62R_\odot$ ($1R_\odot = 6.96 \times 10^8$ m $= 109$ Earth radii) were performed in the *Helios* spacecraft missions [4] but measurements of the regions much closer to the Sun have to be made with remote sensing techniques. The LOFAR-LOIS combination will allow for such remote sensing from the ground.

2.1 Solar radar

Kindled by the discoveries by *Hey* and *Southworth* in the early 1940s [5, 6] that the Sun emits radio waves at metre and centimetre wavelengths, systematic studies of the solar environment with radio methods have by now been carried out for five 11-year solar cycles [7]. In 1952 *Kerr* [8] proposed that radar studies of the Sun could be performed. Further considerations of the solar radar concept were made a few years later by *Bass and Braude* [9] and *Kerr* [10].

However, the solar atmosphere absorbs some of the radar signal intensity according to the well-known optical depth formula $I(\omega) = I_0 \times \exp[-\int \kappa(\omega) dz]$, where κ is the absorption index and $\kappa(\omega)$ is principally determined by the solar atmospheric density and temperature profiles and is found to exhibit a ω^2 dependence. This limits the frequency range for a solar radar to below 100 MHz or so.

The main constraint on the lowest usable frequency is set by the ionosphere. It is transparent for frequencies of the order 5–10 MHz and higher, depending on the site, time of the day and time of the year. But even for frequencies above this cut-off, corrections for the ionospheric distortions on both the transmitted and received radar signals must be made with self-calibration techniques or similar.

Direct probing of the solar atmosphere with radar techniques have been performed by a number of workers:

1. In 1959, *Eshleman et al.*, in California, detected the corona at 25.6 MHz, at a time when the coronal radar cross section was greatly enhanced over the average level [11]. The point of reflection was determined at 1.7 photospheric radii from the centre of the Sun.

2. From 1961 to 1969 *James* made a single daily observation of the corona at local noon, using a much more powerful 38.25-MHz radar system at El Campo in Texas, [12, 13, 14, 15]. The echoes were typically blue-shifted.

3. In the mid-1960s, *Campbell and Parrish* detected the corona using a 40-MHz transmitter at Arecibo with much lower effective radiated power (ERP) than the El Campo radar [16, 17]. They confirmed *James'* results but were severely limited by the low sensitivity of the 40-MHz Arecibo system.

4. In 1977 and 1978 *Benz and Fitze* attempted to observe scatter from Langmuir waves in the corona using a 2.6 GHz, 500 MW CW transmitter at Arecibo [18, 19, 20] but were not successful, most likely because of the very high frequency used.

5. From 1996 to 1998 solar radar experiments were carried out at 9 MHz using the Russian Sura high-power radio transmitter together with the Ukrainian UTR-2 radio telescope in a bistatic mode [21, 22], and there was a likely marginal detection of the corona with a Doppler bandwidth of about 40 kHz.

The El Campo radar, built by MIT's Lincoln Laboratory detected 38.25-MHz radar echoes from the Sun for a period of nine years in the 1960s [23, 24, 25]. Huge, rapidly-moving targets were occasionally observed, but this was before the space-borne coronagraph discovery of coronal mass ejections (CMEs), and the physical nature of these radar targets was a mystery [26]. It has been proposed that radar echoes from CMEs were being detected [27].

Despite its tremendous potential, solar radar is still an essentially unexplored diagnostic technique. Factors which were not available to *James* and which are a strong motivation for reopening the field of fundamental solar radar research at this point in time are

- Correlative observations from existing or upcoming spacecraft, as well as modern ground-based solar observatories.
- Modern polarimetric techniques.
- Multi-frequency, multi-polarisation radar technologies.
- An improved general knowledge of the corona.
- Dramatically improved modelling and computing ability.

Reliable detection and monitoring of CMEs is of great practical importance. CMEs that impact the Earth's magnetosphere can result in hundreds of millions of dollars in damage to spacecraft, communications, and electrical power systems [28]. A system of powerful, flexible low-frequency transmitters operated in software, coupled with a high-angular-resolution receiving array, would form a cost-effective system to detect and track CMEs. The rich information inherent in this measurement could open an entirely new window on CME studies, yielding their angular distribution, ranges, and line-of-sight velocities. Here we present a study of the feasibility of performing such radar studies, starting with what we know about the solar atmosphere and CMEs.

2.2 High-Altitude Radar

The solar wind kinetic energy drives several processes in the magnetosphere of the Earth, *e.g.*, convection, current systems and the visual auroral displays. Coronal mass ejections (CMEs) are believed to cause major magnetospheric storms [29] corresponding to increased solar wind energy input [30]. The increase of energy input rate during CMEs or during other solar wind disturbances will most likely increase the turbulence level at the magnetopause [31].

With a deep-space radar, we hope to establish the signatures of the bow shock and the magnetopause during a

variety of solar wind conditions, including CME encounters, and subsequently to investigate the processes responsible for transferring energy and momentum to the magnetosphere. A primary task would be to establish a relation between the characteristics of the magnetopause as measured by the radar and the energy input rate into the magnetosphere.

We will search for radar signatures of the auroral acceleration regimes and study their relationship to the energy input rate. Combining the new data with other information we will address issues such as substorm triggering loading-unloading, and cause-and-effect relationships.

2.3 Low-Altitude Radar

Space-borne experiments have shown that strong ion-acoustic turbulence can exist on auroral zone magnetic filed lines in the magnetosphere at altitudes between 4000 and 13000 kilometres [32, 33, 34]. Radar signals in the upper HF and lower VHF ranges may scatter coherently off this long-electron-Debye-length (λ_{De}) plasma turbulence so that they can be observed by LOFAR and other high-sensitivity radio telescopes with favourable observing geometries and optimum wavelengths λ_{radar}. Theoretical arguments by *Ginzburg and Rukhadze* [35] suggest that the optimum scattering occurs when $\lambda_{radar} \approx 5\lambda_{De}$, which means that a radar frequency of 85 MHz would yield maximum echoes from altitudes around 3000 km; lower frequencies will scatter off magnetospheric ion-acoustic turbulence at higher altitudes.

The first tentative observation of coherent HF radar scattering off magnetospheric ion-acoustic turbulence was made at the Russian Sura facility in 1991 by *Gurevich et al.* [36]. Later, in 1995, scattering off magnetospheric turbulence at about 6000 km was observed in experiments performed at Tromsø where the EISCAT HF facility was operated in a radar mode [37]. The echoes were typically a factor 10^{-20} weaker than the transmitted pulses.

The use of LOFAR along with LOIS as a deep-space radar is expected to contribute further to these studies of transient enhancements in backscattered power from the lower magnetosphere. Incoherent scatter radars have single, point beams, so while they may be used with great accuracy to locate scattering sources in relation to auroral filaments, they will not give a bigger picture—morphology, dynamics, and relation to magnetospheric phenomena—without being operated in (mechanical) scanning mode, which introduces time/space ambiguities into the interpretation of the data.

The extreme bandwidth and sensitvity of LOFAR will make it an excellent instrument for the observations of radio pulses associated with cosmic particles hitting the moon regolith and extended air showers in the Earth's atmosphere. The extended air shower ionisation trails in the atmosphere may be studied with radar techniques similar to meteor trail diagnostics and has the potential of yielding very important clues to the nature of the cosmic particles.

References

[1] Kenneth Davies. *Ionospheric Radio*. Number 31 in IEE Electromagnetic Waves Series. Peter Peregrinus Ltd., Stevenage, Herts., UK, 1990.

[2] Robert D. Hunsucker. *Radio Techniques for Probing the Terrestrial Ionosphere*. Number 22 in Physics and Chemistry in Space. Springer-Verlag, New York, 1991. ISBN 0-387-53406-7.

[3] Harry Kohl, R. Rüster, and K. Schlegel, editors. *Modern Ionospheric Science, A Collection of Articles published on the Occasion of the Anniversary: "50 years of Ionospheric Research in Lindau"*. European Geophysical Society, Katlenburg-Lindau, 1996. ISBN 3-9804862-1-4.

[4] Rainer Schwenn and Eckart Marsch. *Physics of the Inner Heliosphere*, volume 1, Large-Scale Phenomena. Springer-Verlag, Berlin, Heidelberg, New York, 1990. ISBN 3-540-52081-3.

[5] J. S. Hey. Solar radiation in the 4–6 metre radio wave-lenght band. *Nature*, 157:47–48, 1946.

[6] G. J. Southworth. Microwave radiaiton from the Sun. *J. Franklin. Inst*, 239:285–297, 1945.

[7] Mukul R. Kundu. *Solar Radio Astronomy*, volume 52, chapter 17. Interscience Publishers, New York, 1965.

[8] Frank J. Kerr. On the possibility of obtaining radar echoes from the Sun and the planets. *Proc. IRE*, 40:660–666, 1952.

[9] F. G. Bass and S. Ia. Braude. On the question of reflecting radar signals from the Sun. *Ukrain. J. Phys.*, 2:149–163, 1957.

[10] Frank J. Kerr. Radio echoes from the Sun, Moon, and the planets. In *Handbuch der Physik*, volume LII, pages 449–464. Springer-Verlag, Berlin, 1959.

[11] V. R. Eshleman, R. C. Barthle, and P. B. Gallagher. Radar echoes from the Sun: Man's first contact with the sun opens new approaches for the study of solar events. *Science*, 131:329–332, 1960.

[12] J. C. James. Radar echoes from the sun. *IEEE Trans. Mil. Elect.*, MIL-8:210–225, July–October 1964. Identical to *IEEE Trans. Antennas Propagat., AP-12*, 876-891.

[13] J. C. James. Radar studies of the sun at 38 Mc/s. *Astrophys. J.*, 146:356–366, 1966.

[14] J. C. James. Radar studies of the sun. In J. V. Evans and Tor Hagfors, editors, *Radar Astronomy*. McGraw-Hill Book Co., New York, 1968. Library of Congress Catalog Card Number 66-29751.

[15] J. C. James. Some observed characteristics of solar radar echoes and their implications. *Solar Phys.*, 12:143–151, 1970.

[16] Alan Parrish. Solar radar experiments, 1967. Technical Report 300, Center for Radiophysics and Space Research, Cornell University, Ithaca, New York, April 1968.

[17] Don Campbell. Private communication, 1998.

[18] Arnold O. Benz and H. R. Fitze. Microwave radar observations of the Sun. *Astron. Astrophys.*, 76:354–355, 1979.

[19] A. O. Benz and H. R. Fitze. First solar radar observations in microwaves. In Mukul R Kundu and T E Gergely, editors, *Radio Physics of the Sun*, pages 247–250. International Astronomical Union, Paris, 1980.

[20] H. R. Fitze and A. O. Benz. The microwave solar radar experiment, I, Observations. *Astrophys. J.*, 250:782–790, 1981.

[21] Paul Rodriguez, A. Konovalenko, O. Ulyanov, Lev M. Erukhimov, Yu. Belov, Yu. Tokarev, Anatoliy Karashtin, and K. van't Klooster. Detection of HF radar reflection from the solar corona. In *Transactions of the AGU, EOS Supplemental Issue*, volume 78, page F538. American Geophysical Union, 1997. (Paper SH21A-15, AGU Fall Meeting, San Fransisco, CA, December 8–12, 1997).

[22] P. Rodriguez. Radar studies of the solar corona: A review of experiments using HF wavelengths. In K. S. Balasubramaniam, S. L. Kiel, and R. N. Smartt, editors, *Radio Astronomy at Long Wavelengths*, volume Geophysical Monograph 119, pages 155–165. American Geophysical Union, 2000.

[23] W. G. Abel, J. H. Chisholm, P. L. Fleck, and J. C. James. Radar reflections from the Sun at very high frequencies. *J. Geophys. Res.*, 66(12):4303–4307, 1961.

[24] W. G. Abel, J. H. Chisholm, P. L. Fleck, and J. C. James. A VHF solar radar system. In *IRE Int. Conv. Record, vol. 10, pt. 5*, pages 58–66, 1962.

[25] J. C. James. A 38 Mcps solar radar system. *NEREM Record*, 7:186–187, 1965. [Northeast Electronics Research and Engineering Meeting, Session: Radar Astronomy, FAM-1, Fri, 5 Nov 1965].

[26] J. H. Chisholm and J. C. James. Radar evidence of solar wind and coronal mass motions. *Astrophys. J.*, 140:377–379, 1964.

[27] P Rodriguez. High frequency radar detection of coronal mass ejections. In K S Balasubramaniam, S L Kiel, and R N Smartt, editors, *Solar Drivers of Interplanetary and Terrestrial Disturbances*, volume Conference Series 95, pages 180–188. Astronomical Society of the Pacific (ASP), 1996.

[28] C. Dybas. Scientists to present results of "solar storm" research—massive events disrupt power, cause aurora displays. National Science Foundation Press Release 96-20, 8 May 1996.

[29] Zdeněk Švetska. Varieties of coronal mass ejections and their relation to flares. *Space Sci. Rev.*, 95/1-2:135–146, 2001.

[30] S.-I. Akasofu. Explosive magnetic reconnection: Puzzle to be solved as the energy supply process for magnetospheric substorms. *EOS*, 66:9, 1985.

[31] L. Rezeau and G. Belmont. Magnetic turbulence at the magnetopause, a key problem for understanding solar wind/magnetosphere exchanges. *Space Sci. Rev.*, 95/1-2:427–441, 2001.

[32] M. Temerin, K. Cerny, W. Lotko, and F. S. Mozer. Observations of double layers and solitary waves in the auroral plasma. *Phys. Rev. Lett*, 48:1175–1179, 1982.

[33] Rolf Boström, Georg Gustafsson, Bengt Holback, Hannu Koskinen, and Paul Kintner. Characteristics of solitary waves and weak double layers in the magnetosphere plasma. *Phys. Rev. Lett.*, 61:82–85, 1988.

[34] Gunnar Holmgren and Paul M. Kintner. Experimental evidence of widespread regions of small-scale plasma irregularities in the magnetosphere. *J. Geophys. Res*, 95:6015–6023, 1990.

[35] V. L. Ginzburg and A. A. Rukhadze. *Waves in magnetized plasma*. Nauka, Moscow, 1975.

[36] Alex V. Gurevich, Andrey M. Babichenko, Anatoliy N Karashtin, and Victor O. Rapoport. HF sounding of the auroral magnetosphere. *J. Geophys. Res.*, 97(A6):8693–8696, 1992. (In Russian: Pis'ma ZhETF, *53*, (1991), pp. 139–143).

[37] Bo Thidé, Tobia Carozzi, and Michael Rietveld. Unpublished results, 1995.

A Novel SIS Mixer for Applications in Radio Astronomy

Vessen Vassilev[1], Christophe Risacher[2] and Victor Belitsky[3]

[1 2 3] Onsala Space Observatory - Chalmers University of Technology
[1] E-mail: vevas@oso.chalmers.se
[2] E-mail: risacher@oso.chalmers.se
[3] E-mail: belitsky@oso.chalmers.se

Abstract

This paper presents a prototype of a sideband separation super heterodyne receiver for 85-115 GHz based on SIS (Superconductor Insulator Superconductor) junctions as a frequency down converter (mixer).

Sideband separation is achieved by using a quadrature scheme with RF power distributed between the two identical mixer junctions pumped by a local oscillator (LO) with 90° phase difference.

To divide the input RF signal and to couple it to the mixers fabricated on the quartz substrate we designed a new structure - a waveguide-to-microstrip double-probe transition. The RF power is thus coupled from the input waveguide to the probe structure on the substrate and divided between the two mixer junctions.

At the conference we present details on the design of some of the mixer components and measurements of the waveguide-to-microstrip double probe transition and cryogenic IF amplifiers. The mixer is currently under fabrication and first mixer test results are expected in April-May 2002.

1 Introduction

The Atacama Large Millimetre Array (ALMA) is an international project in radio astronomy instrumentation: with its 64 sub-millimetre 12-meter diameter antennas, ALMA is planned to be the largest and the most sensitive instrument for millimetre and sub millimetre radio astronomical observations. Sited in the Chajnantor region of the Chilean Andes at the altitude of 5000 m, ALMA heterodyne instrument will cover frequencies from 30 to 960 GHz.

With SIS mixer noise temperature approaching level of 2-5 times the quantum limit, the noise performance of a double-side band (DSB) super-heterodyne receiver often is limited by the atmospheric noise fed into the system via the image band. Thus, single sideband (SSB) operation is required to further increase the system sensitivity.

Sideband separation can be achieved using a quadrature scheme as shown in Figure 1, the RF power is divided between the two identical DSB mixers and the local oscillator (LO) is applied to the mixers with 90° phase difference.

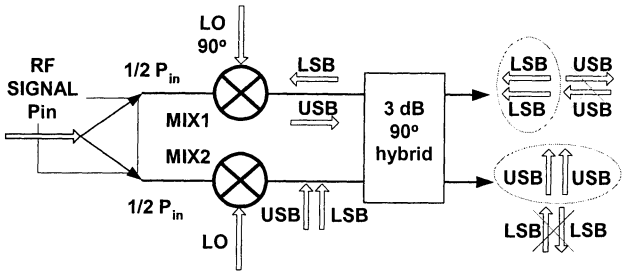


Figure 1 Block diagram of the sideband separation mixer. The crossed out vectors (upper and low side band, USB & LSB) at the hybrid outputs are the rejected sidebands (180° phase difference).

The mixers outputs at intermediate frequency (IF) are connected to a quadrature 3-dB hybrid. Since one of the sidebands at each output of the hybrid is combined in-phase and the other out of phase (180°), sideband cancellation occurs and the both sidebands appear separated.

The quadrature scheme does not use any RF filter components and has been demonstrated at mm-wavelengths [1, 2]. The degree of sideband suppression is directly related to the magnitude and phase balance of the divided RF and LO power applied to the mixers [3] as well as the rest of the circuitry components symmetry.

2 Mixer Description

The main components of the mixer can be identified in the mixer-block layout shown in Figure 2. The RF input is a corrugated horn feed divided into two sections to facilitate the machining. The mixer block used split-block technique and RF power is coupled to the mixer substrate in the middle of the waveguide section.

The LO power is applied through a waveguide at the opposite side of the mixer block employing a waveguide quadrature 90° 3-dB hybrid.

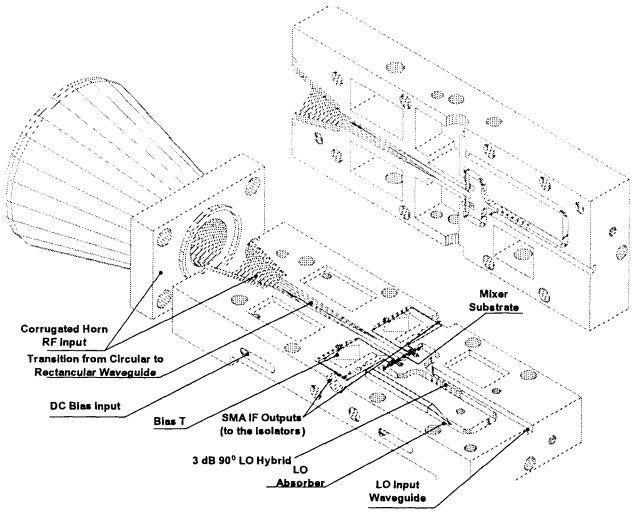


Figure 2 Mechanical layout of the sideband separation mixer. A conventional split-block technique is used to fabricate the machining of the mixer mount. The waveguide dimensions are 1.2x2.4 mm.

The mixer block consists of two identical parts dividing symmetrically all waveguide structures. The bottom part of the mixer block accommodates the mixer substrate, bias T for SIS junction DC bias and the termination of the idle port of LO hybrid. Unused volumes in the block are removed in order to reduce thermal capacity of the structure operating at 4 K ambient temperature.

The IF outputs of the SIS mixers are connected to cryogenic isolators and low noise IF amplifiers (gain and phase balanced) and then combined in a 90^0 IF hybrid, where the sideband cancellation occurs, as shown in Figure 3.

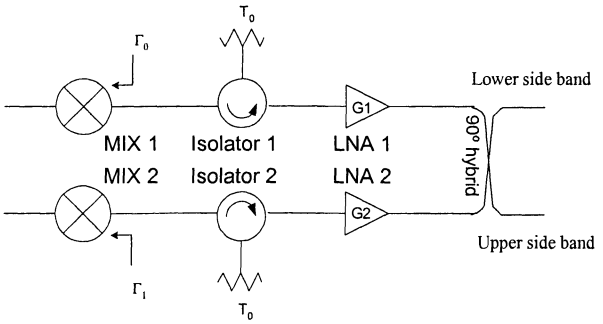


Figure 3 Detailed block diagram of the IF part of the sideband separation mixer.

Isolators are needed since the SIS resistance at IF can become negative and thus cause instabilities in the amplifiers.

Placing the isolators and amplifiers before the hybrid gives lower system noise temperature compare to the scheme with the hybrid placed between the mixers and the isolators. The drawback of this configuration is that additional phase and gain imbalance are added to each channel by these components and therefore limiting the sideband rejection.

Figure 4 shows the amplitude and phase imbalance relation in order to get a 10dB and a 20 dB sideband rejection. The Imbalance margin is quite large and, for example, in order to get 10 dB rejection, a 30° phase and a 3-dB gain imbalance can be tolerated, which should be achievable. The gain and phase asymmetry of the different elements in our scheme were estimated and the total gain and phase imbalance gives values of ±1.5 dB and ±15° correspondingly; therefore rejection should safely be better than -10 dB.

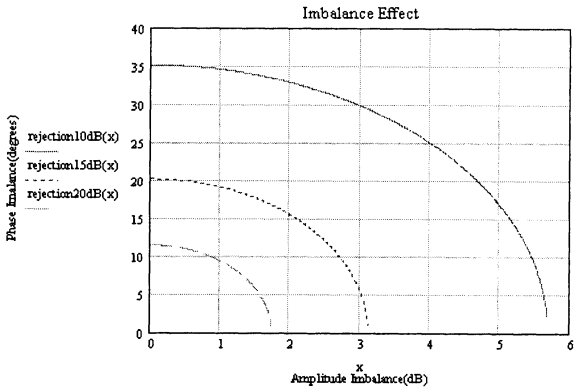


Figure 4 Plot showing relation phase and gain imbalance limits in order to achieve 10 dB and 20 dB sideband rejection.

3 Mixer Components

A closer view of the mixer substrate is given in Figure 5. To ensure a high degree of symmetry in the SIS junction performance and the geometry of the mixer circuitry, all mixer components are integrated on the same compact substrate. The LO power is coupled to the ends of the substrate through waveguide to microstrip transitions.

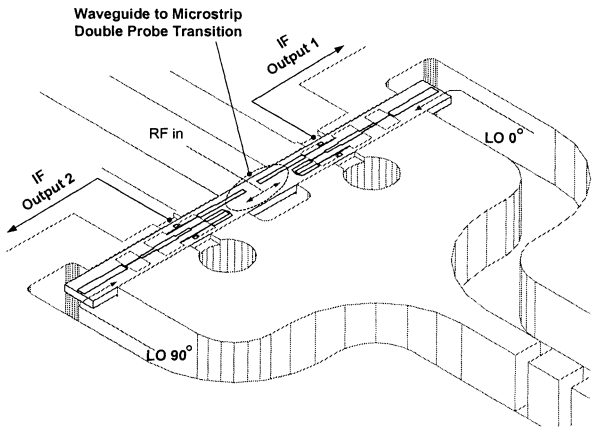


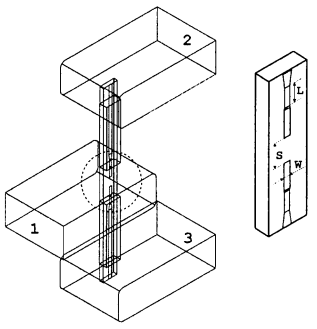
Figure 5 The mixer substrate coupled to the waveguides. The divided LO power is coupled at the ends of the substrate. The RF power is coupled to the substrate in the middle and divided between the two mixer junctions by the waveguide to microstrip double probe transition. The mixer substrate is a Z cut crystal quartz with dimensions 0.7/8.74/0.15mm (W/L/H). The substrate size is chosen such that it does not allow waveguide modes inside the substrate channel.

3.1 Waveguide to microstrip double probe transition

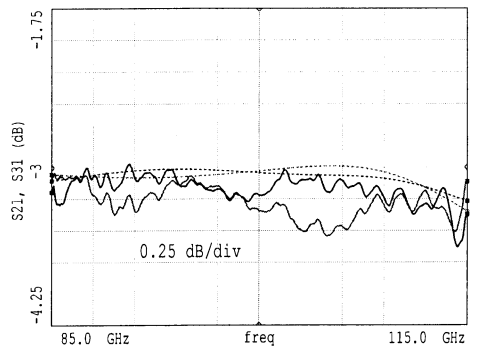
To achieve a simple design combining low-loss, wide-band symmetrical power division and the

waveguide-to-microstrip transition, we employed a new device where a double-probe structure is coupled to the E-field in a waveguide 1, as shown in Figure 6a. The transition has a simple geometry and does not require any lumped termination load, which is difficult to fabricate with satisfactory accuracy using thin film technology.

Scalar measurements of the waveguide to microstrip double probe transition were performed at 85-115 GHz [4]. To facilitate the measurement, additional microstrip to waveguide transitions were designed forming a waveguide-to-waveguide 3-dB power divider as shown in Figure 6a. The transmission S21 and S31 is plotted in Figure 6b, the typical total loss is ≈0.3dB over 30% of the bandwidth and it includes the losses from the single waveguide to microstrip transitions.



a)



b)

Figure 6 a. Waveguide to microstrip power divider as it was measured at 100GHz. The microstrip ports were replaced with microstrip to waveguide transitions. The waveguide is WG-10, the structure dimensions are S=340 μm, W=95 μm, L=270 μm and the substrate cross section is 610x150 μm. **b.** Measured transmission S21 and S31 of the double probe transition (solid lines) and the simulated transmission (dashed lines).

To ease the phase measurement a scale model was built at 8.5-12.5 GHz and measured with a vector network analyser. The measurements performed at 10 GHz and 100 GHz showed that the waveguide-to-microstrip double probe transition gives very good magnitude balance and excellent phase symmetry of the divided RF signal over bandwidth covering all frequencies specific for the waveguide dominant mode and providing a possibility of achieving very good degree of the sideband separation.

3.2 LO injection directional coupler

A directional coupler shown in Figure 7 follows the waveguide-to-microstrip double probe transition and is used to inject LO power to the junctions. To provide low RF signal loss, the LO coupler is designed for –15 dB coupling using a single section of λ/4 long microstrip coupled lines.

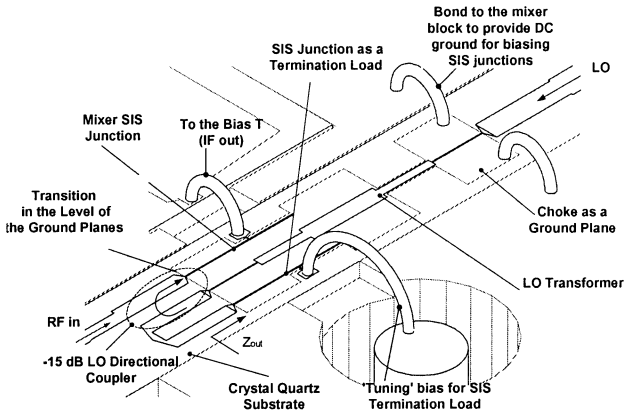


Figure 7. A closer view of the mixer components: the LO injection directional coupler uses the crystal quartz substrate as a dielectric and the bottom substrate channel as a ground plane, the SIS tuning circuitry is based on thin film SiO_2 dielectric and choke structure serves as a ground plane. A second SIS junction and its tuning circuitry is providing real impedance Z_{out} to terminate the idle port of the LO coupler.

In order to simplify fabrication of the SIS mixer chip, the LO directional coupler should use the same dielectric utilized by the SIS junction and RF tuning circuitry (250 nm of SiO_2). However, we found that the required spacing between the lines, which is a critical dimension, should be less than 1 μm and is difficult to produce with the required accuracy by optical lithography. Therefore the LO directional coupler together with the waveguide to microstrip double probe transition is build employing the crystal quartz substrate as a dielectric and the bottom of the substrate channel as a ground plane [5, 6]. Thus the width of the lines and the spacing between them is increased substantially. The impedance of the coupled lines should be as low as possible for easy match with the impedance of the SIS junction (≈ 5 Ω). To keep the output impedance of the coupler low, the lines should be kept as wide as possible. In this case the width of the lines is limited by the width of the substrate, which should be kept sub critical to prevent propagation of waveguide modes inside the substrate channel.

The idle port of the LO coupler must be terminated with matched impedance. Initially a thin film absorption line was considered as a termination but the required length was found to be impractically long to fit the available area on the substrate. Therefore, an SIS junction identical to one used for mixing was found to be suitable to provide required termination [6].

3.3 RF circuitry

The principal design approach is to provide as good as possible RF coupling between the signal source and the SIS mixer junction. The real part of the SIS junction impedance R_{RF} can be derived from the junction current-voltage (IV) characteristic in the presence of LO pumping signal. The complex impedance is obtained by adding SIS junction geometrical capacitance, Cj, specific for the particular junction area in parallel to the R_{RF}. The equivalent circuit of SIS junction at RF is shown in Figure 8. To match the SIS complex impedance to the real impedance of the signal source, Cj should be resonated out at RF. We used a short microstrip line, equivalent to inductance, L_t, at RF, connected in parallel. A $\lambda/4$ open stub provides ground for L_t. R_{RF} is then matched to the RF source impedance by a l/4 transformer section.

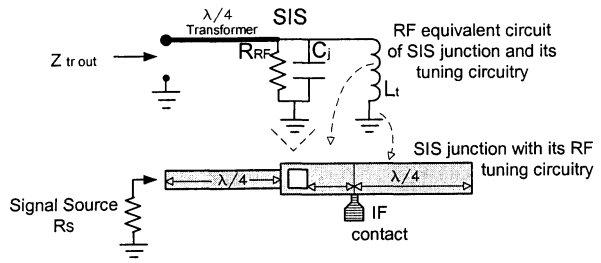


Figure 8 Equivalent circuit of SIS junction and its tuning microstrip-based circuitry at RF. The impedance of the pumped SIS junction, R_{RF}, connected in parallel with its geometrical capacitance C_j. A piece of line acting as inductance, L_t, grounded via $\lambda/4$ open stub, resonates C_j at RF centre frequency. The impedance R_{RF} is matched to the RF source impedance R_S via a $\lambda/4$ transformer.

The RF tuning circuitry shown above has the advantage that IF can be extracted at a point which is virtual ground at RF, and therefore no filtering is needed to prevent RF leakage towards the IF output. SIS impedance in presence of LO power is derived with the help of a program which calculates R_{RF} vs. frequency and LO pumping power based on the measured or modelled IV characteristic of the junction [7]. The impedance of the signal source is considered as the output impedance of the waveguide-to-microstrip double probe transition and has been derived using HFSS [8].

3.4 Shift in the level of the ground plane

Since the LO directional coupler uses the crystal quartz substrate as a dielectric and the bottom of the substrate channel as a ground plane, whereas the rest of the mixer circuitry uses a 250 nm thin film SiO_2 as a dielectric required to fabricate the SIS junctions with the ground plane placed on the top of the substrate, a special measure is required to achieve effective transfer in the levels of the ground planes

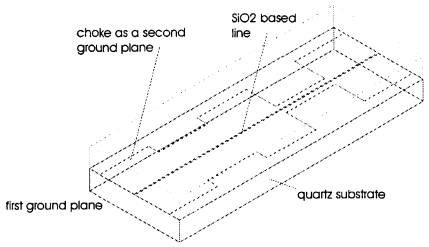


Figure 9 Choke structure on the quartz substrate provides coupling of the first and the second ground planes (separated by the quartz substrate).

The solution to this problem is to shape the upper ground plane as a low pass filter (choke) for the RF, as shown in Figure 9. The choke consists of a series of quarter-wavelength sections open circuited at the ends resulting in zero impedance observed at the ends of the choke with respect to the first ground plane (the bottom of the substrate channel). Hence at the points of interface the choke is short circuited to the first ground plane for the RF. The choke serves as a ground for SIS tuning circuitry as well as the guiding circuitry bringing LO power to the LO directional coupler (Figure 7). The choke is also bonded to the mixer mount to provide ground to the SIS junction DC bias.

3.5 Low noise IF amplifiers

The cryogenic low noise 2-stage amplifiers were developed for frequency band 3.4-4.6 GHz, using commercial Mitsubishi MGFC4419G GaAs HEMT transistors. The amplifier design was done using Agilent ADS and special attention was paid to model the passive components, the matching circuitry and to use accurate cold S-parameters. The amplifier gain of 28 dB and the noise temperature of 3.0 K at ambient temperature of 12 K over the entire 3.4-4.6 GHz band were measured, which is the state-of-the-art performance at this frequency using GaAs transistors [9]. The measured performance of seven fabricated amplifiers are completely consistent.

Two amplifiers with equal gain and phase delay were chosen for the sideband separation mixer and their measured performance is shown in Figure 12. The gain and phase similarity between them is significant with only ± 0.2 dB gain imbalance and less than 5 degree phase imbalance, which should not degrade the rejection between the sidebands noticeably.

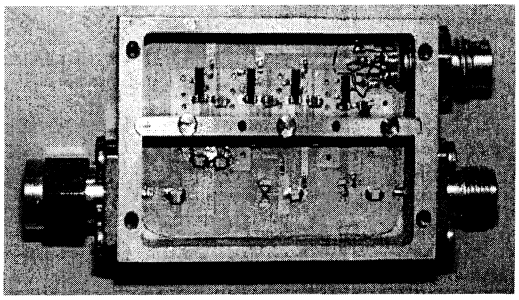


Figure 11 Photograph of a 3.4-4.6 GHz low noise amplifier, using 2 stage GaAs HEMT chip transistors. Bias circuitry is on the top of the picture, separated by a wall from the RF lines on the bottom part.

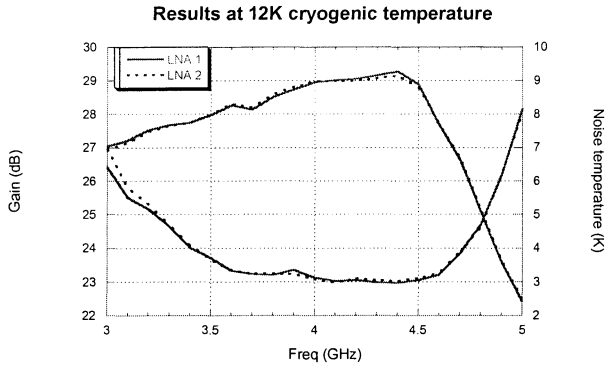


Figure 12 The amplifiers measured performance at 12 K ambient temperature. Bottom curves depict the noise temperature and upper curves shoes the amplifiers' the gain.

4 Conclusion

We present the design of a prototype sideband separation mixer for 85-115 GHz. The mixer is based on a new device - waveguide-to-microstrip double probe transition, which makes possible integration of all mixer components on the same compact substrate and thus ensures a high degree of similarity in the SIS junction performance and the geometry of all the mixer elements. Figure 10 shows a picture of the mixer block, the mixer substrate with the junctions is under fabrication and first measurement results are expected at the end of May 2002.

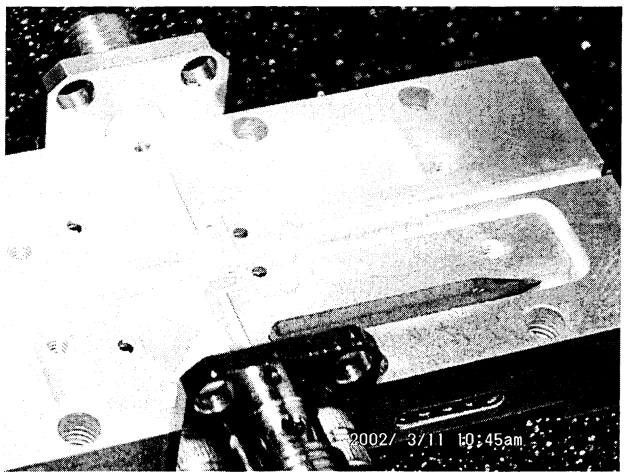


Figure 10 The mixer block under assembling. The picture shows the LO quadrature hybrid with installed absorber at the idle port with the 'bias T' filters and IF SMA connectors installed.

Cryogenic IF amplifiers have been designed and measured for 3.4-4.6 GHz frequency band. The typical measured gain and noise temperature are 28dB and 3.0 K respectively at an ambient temperature of 12 K. Two amplifiers having gain imbalance of only ± 0.2 dB and phase imbalance less than 5 degrees were chosen to be used in the sideband separation receiver.

Acknowledgements

Dr. I. Angelov is greatly acknowledged for providing precise cold S-parameters for the transistors used in the amplifiers. Professor R. Booth was always supporting and encouraging us for successful completion of this work.

References

[1] A. R. Kerr, S.-K. Pan and H. G. LeDuc, "An integrated sideband separating SIS mixer for 200-280 GHz", *Proc. of the Ninth Space Terahertz Technology Symposium, Pasadena*, USA, March, 1998.

[2] R. L. Akeson, J. E Carlstrom, D. P. Woody, J. Kawamura, A. R. Kerr, S. -K. Pan and K. Wan, "Development of a Sideband Separation Receiver at 100GHz", *Proc of Fourth International Symposium on Space Terahertz Technology*, pp.12-17, March, 1993.

[3] A. R. Kerr and S.-K. Pan, "Design of Planar Image Separating and Balanced SIS Mixers," *Proc. of Seventh International Symposium on Space Terahertz Technology*, March 12-14, 1996.

[4] V. Vassilev, V. Belitsky, D. Urbain, S. Kovtonyuk, "A New 3 dB Power Divider for MM-Wavelengths", *IEEE Microwave and Wireless Components Letters*, page 30-32, vol.11, Jan. 2001.

[5] V. Vassilev and V. Belitsky, "Design of Sideband Separation SIS Mixer for 3 mm Band", *12th International Symposium on Space Terahertz Technology*, Feb. 2001.

[6] V. Vassilev, "Development of Components for MM-Wave Radio Astronomy Instrumentation", *Technical report No.416L Department of Radio and Space Science, Chalmers University of Technology*, Dec. 2001.

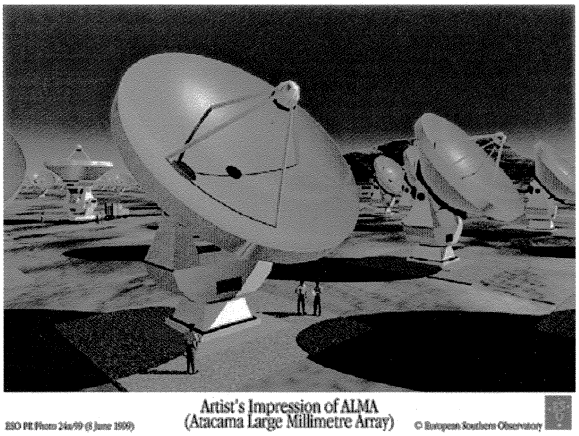
[7] John R. Tucker, Marc J. Feldman, "Quantum Detection at Millimeter Wavelengths", *Review of Modern Physics*, Vol. 57, No. 4 October 1985.

[8] HP HFSS, Hewlett-Packard Company Test and Measurement Organization 5301 Stevens Creek Blvd., Bldg. 51L-SC Santa Clara, CA 95052-8059.

[9] C. Risacher, M. Dahlgren, V. Belitsky, "A Low Noise 3.4-4.6 GHz Cryogenic Amplifier", *Proceedings of "GHz 2001 Symposium", November 2001, Lund, Sweden.*

ALMA, THE ATACAMA LARGE MILLIMETRE ARRAY

R.S. BOOTH
Onsala Space Observatory,
Chalmers University of Technology,
S-43992 Onsala, Sweden
E-mail: roy@oso.chalmers.se



Abstract

The Atacama Large Millimetre Array is a joint European – US project (with possible Japanese involvement) to build a milllimetre/submillimetre array on a high (5000 m) plateau in Northern Chile. It is a revolutionary instrument bringing large-scale aperture synthesis techniques to millimetre/sub-millimetre astronomy to facilitate precision imaging on milli-arcsecond scales. The array will consist of 64 x 12m diameter antennae, giving a collecting area of 7000 squ .m., seven times the biggest existing millimetre telescope and, with its excellent site and state of the art receivers, an observational speed up time of nearly 10,000 for continuum observations. It will be a major force, overall, in astrophysics; a mm/submm equivalent of the Next Generation Space Telescope facilitating studies of star-forming galaxies across almost the entire Universe and star-forming regions across the Galaxy. ALMA is one of the highest priority projects in astronomy today and it promises to revolutionise our subject, opening up new horizons in science. ALMA is specified to operate in the wavelength range 3.5 mm to 350 micron and although much of the technology is in place, it presents enormous technical challenges in terms of mass production, and maintenance of the specifications under (remote) operational conditions at high altitude.

1. Introduction

The millimetre/submillimetre (mm/submm) wavebands are among the richest in astronomy. It is at these wavelengths that we find spectral lines from most interstellar molecules and the peak of the thermal spectrum of cold dust in space. At submm wavelengths we also find important atomic transitions. The mm/submm bands are the only parts of the electromagnetic spectrum where we can detect cold dust and molecules from distant, young, high-redshift galaxies, making them of paramount importance for the study of the early universe. These bands too, are the only parts of the spectrum where we can determine fine scale kinematics of young stellar discs that may form planets, and of the ejected envelopes of old stars which are forming dust grains and enriching the interstellar medium with heavy elements, carbon, oxygen and nitrogen.

Present day mm/submm wavelength instruments are barely scratching the surface in these of these fascinating fields but we are able to image the few nearby proto-stellar discs and, aided by gravitational focussing, have detected a few distant galaxies to a red-shift, z=4.69. Clearly, the provision of an imaging telescope with an order of magnitude improvement in both sensitivity and resolution is badly needed. Here we describe that instrument, the Atacama Large Millimetre Array, ALMA, which will facilitate studies of stellar evolution across the Galaxy and of star forming galaxies across the Universe. But it will go far beyond these main scientific drivers and have a major impact on all of astronomy; it will be a millimetre

equivalent of the ESO Very Large Telescope (VLT) or the Hubble Space Telescope (HST).

2. The ALMA project

Such is the expected scientific impact of a large millimetre array that, until quite recently, three separate project studies were being conducted: in Europe, Japan and the United States [see 1]. However, in mid 1997, the North American and European groups decided to explore the possibility of merging their ideas and their concepts. The result is the current Phase 1 study of ALMA, the Atacama Large Millimetre Array. In addition to the Euro/US collaboration, it is likely that the Japan will join the ALMA project, when their funding climate improves.

The ALMA concept then, represents a fusion of ideas from two earlier projects: the Millimetre Array [2], a US project which has been studied since the 1980s and driven scientifically by the requirement to image (large) molecular clouds, and the Large Southern Array [3,4], a European project driven by relatively recent discoveries of emission from carbon monoxide and dust in galaxies at high red shift ($z > 2$). This is oversimplified because there was significant scientific overlap among the projects. However, the US preferred small high frequency (submm) antennae in relatively compact configurations while the LSA concept called for a much larger collecting area (10,000 m^2), higher resolution (longer baselines) and a more limited spectral coverage. This was based on the fact that the cosmological red shift causes spectral lines and even the peak of the dust spectrum to be shifted long-ward in wavelength so, it was argued, dispensing with the need for higher frequencies.

The overlapping scientific goals of the two projects made it fairly easy to agree the specifications of the joint array. ALMA will consist of 64 antennae, each 12m in diameter giving a total collecting area of 7000 m^2. The array will be built on the site chosen for the MMA, a 5000 m high plateau called Llano de Chajnantor in the Chilean Andes. The wavelength coverage is specified to be 3.5 mm to 350 micron, with the possibility of an added channel centred at 35 GHz. Baselines will extend to 10 km, or more, although the array will be re-configurable and transporters will be used to move the telescopes. The full specification is given in Table 1. Receiver frequencies are given in Table 2

Table 1. Summary of the proposed ALMA specification

Aperture Synthesis Array		
	Number of telescopes	64
	Number of baselines 2016	
	Total collecting area	7000 m^2
	Maximum baseline	12 km
	Maximum angular resolution	10 mas (at a wavelength of 1 mm)
Antenna specification		
	Diameter	12 m
	Surface accuracy	25 micron rms
	Pointing accuracy (relative)	0.6 arcsec
Initial Correlator (lag architecture)		
	Cross - correlations	
2016 /frequency band/polarisation		

Bandwidth	16 GHz
Maximum resolution	30 kHz (0.04 km/s at 230 GHz)

Table 2: Receivers and frequency bands (see MMA memo 213)

Band no.	Frequency range (GHz)	Receiver type
1	30 – 40	HEMT amplifier
2	67 – 90	HEMT
3*	89 – 116	SIS mixer
4	125 – 163	SIS mixer
5	163 – 211	SIS mixer
6*	211 – 275	SIS mixer
7*	275 – 370	SIS mixer
8	385 – 500	SIS mixer
9*	602 – 720	SIS mixer
10	787 – 950	SIS mixer

*Receiver bands with highest priority

The site chosen for the ALMA is close to the ultimate available on the earth and it is very dry; the measured annual median opacity at 230 GHz is 0.05. [4]. However, because it is not particularly comfortable to work at an altitude of 5 km, it is proposed to operate the array remotely from the small village of San Pedro de Atacama, 60 km away, or at a lower altitude site, somewhere in between. The site is less than 10 km from a newly paved trunk road across the Andes and the driving time to San Pedro is about one hour.

3. The ALMA Site

The Atacama desert in Northern Chile is one of the driest places on Earth so it is not surprising that all three millimetre arrays had chosen sites in this area. At mm/submm wavelengths the atmosphere is only partially transparent so the best sites for observations should be high and dry to minimise the attenuation from oxygen and water vapour. For interferometry their atmosphere should also be stable to minimise the fluctuations in relative path length from the source to the individual telescopes (good radio seeing), thereby simplifying phase calibration schemes. Since the pointing accuracy of the individual antennae is critical, the wind conditions on the site should be as light and gust-free as possible and the wind conditions must be taken into account in the antenna design.

Sites with submm quality were chosen by the US and Japanese project teams. Both are extensions of the same high plateau, which extends some 20-30 km in a north easterly direction between the peaks of the mountains Chajnantor and Chascon. Both are close to the newly paved main trunk road between Chile and Argentina. The ALMA project has chosen to

study the higher, more southerly site of Llano Chajnantor and several sets of site monitoring equipment have been installed. Site testing has been going since 1995 as a collaboration among the three groups involved. The site co-ordinates have been refined through GPS measurements by the Onsala group and are shown below. (Johansson and Gradinarsky, private communication):

Geodetic Co-ordinates of Llano Chajnantor:
Latitude: 23deg 01min 21.8sec S
Longitude: 67deg 45min 15.7sec W
Altitude: 5074.8161 m

4. Measurements of site quality

Among the site monitoring equipment on Llano Chajnantor, both the US and Europeans have set up wind speed indicators, 225 GHz radiometers used to monitor the atmospheric opacity through measurement of the sky brightness at this frequency and phase measuring interferometers. The US campaign has been discussed by Radford and Holdaway [5] and we site their data here.

Typical optical depth measurements derived from the 225 Ghz radiometers seem to correlate well with direct atmospheric transparency measurements at 350 micron, particularly at low opacities. The cumulative zenith optical depth data between early 1995 and late 1999 are shown in figure 2. Apart from the summer excursions and the 1997 el niño event, the optical depth remains below 0.05 for something like 50% of the time. In fact the atmospheric transparency in this region is exceptional. Fourier transform spectroscopy, conducted by the Japanese team in June, 1998 on Pampa La Bola, has shown that there is 20% atmospheric transmission at 1.5 THz (Ishiguro, private communication)!

The atmospheric seeing is measured by a pair of 1.8m diameter antennae connected as an interferometer with a baseline of 300m, which monitors the 11.2 GHz signal from a geo-stationary satellite [6]. The phase difference between the signals received by the two antennae is measured continuously. Since the atmosphere is non-dispersive (away from line centres), it is possible to scale the low frequency data to at least 350 GHz. The cumulative results of the path length fluctuations are shown in figure 2. For a wavelength of 1mm. the median phase is better than a radian about one third of the time. But this is not the full story since the phase fluctuations show a diurnal variation, the phase stability decreasing as the wind increases in the afternoons.

For good image reconstruction with the array, it is important that atmospheric phase fluctuations are less than 20 degrees; 10 degrees for the very highest dynamic range which should be at least 10,000. Thus, it is imperative that we devise good phase calibration schemes. Often in radio interferometry, calibration is performed by switching between the source of interest and a nearby known point source. At the highest frequencies, however, there are fewer known point sources and atmospheric fluctuations tend to be rapid, thus demanding fast slewing of the antennae, placing costly demands on their mechanical construction. Another scheme is being tested in Chile.

Since the variable atmospheric water vapour also dominates the sky emissivity fluctuations, a method, suggested by Richard Hills, for correcting the interferometer phase by measuring the strength of the 183 GHz water line along the line of sight to each antenna, is being tested on

the ALMA site. Alongside the 11 GHz interferometer antennas we have mounted 183 GHz radiometers – radiometers which monitor the full power from the atmospheric water line and give a measure of the total water vapour along the line of sight [7,8] from which the excess path can be determined. The radiometer data will be compared with the interferometer measurements of the relative path length fluctuations.

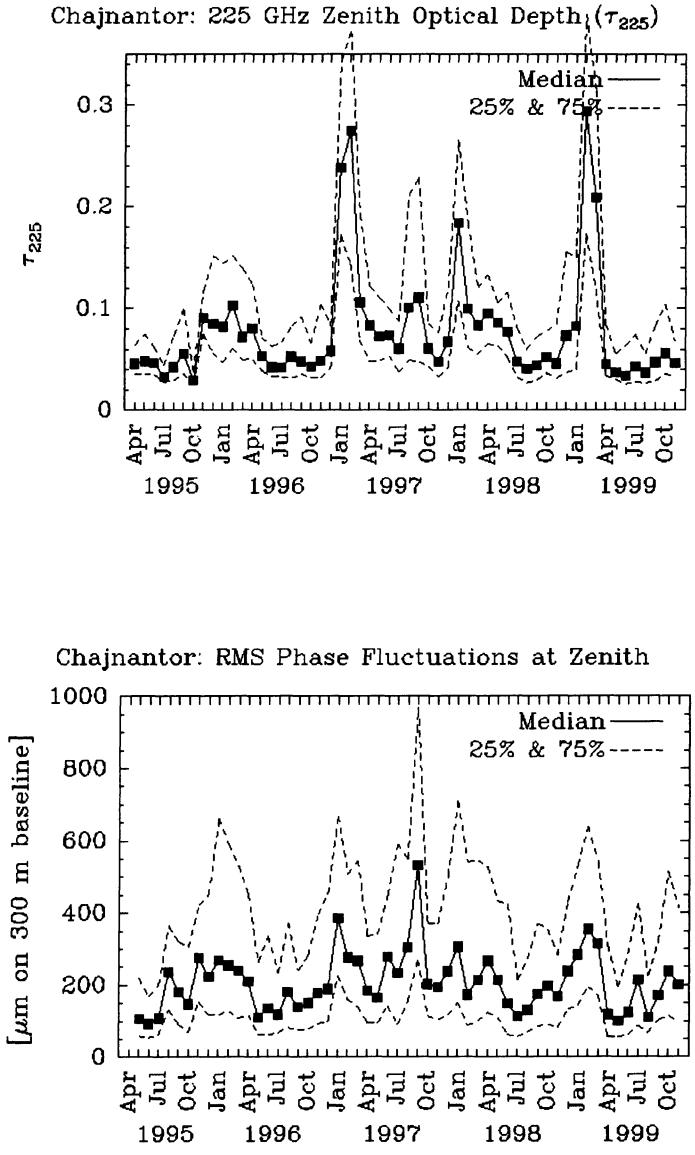


Figure 1. *Cumulative measurements of atmospheric transparency and phase on the ALMA site, Llano Chajnantor. (see text).*

It is only on such excellent sites as Chajnantor that the (strong) 183 GHz line will not be saturated; previous phase correction has been done using 225 GHz radiometers, i.e. monitoring the power in the wing of the 183 GHz line, or by monitoring the 22 GHz water line, which is optically thin under most conditions [see 9]. If 183 GHz radiometry proves to be a reasonably successful phase correction technique, each antenna of ALMA will be fitted

with such a radiometer. Even if the technique is not perfect, we hope that it will be accurate enough to permit a longer than otherwise switching cycle between the source and phase calibrator.

5. The scientific potential of ALMA

The new array will have a collecting area of 7000 m^2 and 2016 simultaneous baselines. On the longest baselines, the synthesised beam will be 0.03" at 230 GHz. The highest resolution at the highest frequency will be finer than 10 mas. Overall, given the excellent site, better instrumentation and the large number of baselines, ALMA will give a vastly increased sensitivity and image quality over what is available today. For example, with the bandwidth of 8 GHz and adding both polarisation channels, the 5-sigma point source sensitivity of the array at a frequency of 230 GHz is 60 micro-janskys in one hour. This means that the array will be much faster to reach a given sensitivity than today's arrays. The speed-up factor over the current arrays is nearly 10,000 for continuum observations: 50 on account of the collecting area, about 16 for the increase in bandwidth and the dual polarisation and a factor of around 10 for the for the reduced atmospheric attenuation (better site) and possible small receiver improvement. For lines the speed up factor is around 500. ALMA will be so sensitive that many thousands of stars will be detectable. A planet like Jupiter at the distance of Alpha Centauri will be detected in about 150 hours!

A good review of ALMA science has been given by Downes in the LSA/MMA Feasibility report [10]; see also [11]. The main scientific drivers for ALMA are the origins of galaxies and stars and even their planets and the array is designed to have appropriate point source sensitivity and resolution. Thus we expect that the array will be used to study:

- the formation and evolution of galaxies and quasars in the early universe
- the formation of evolution of stars and their planets, both in the Galaxy and nearby galaxies.

But it will go far beyond these drivers and have a major impact on all branches of astrophysics, and even in planetary science. (At a distance of 1 AU, the linear resolution of ALMA at 230 GHz will be 25 km!). It will make millimetre astronomy accessible to all astronomers and could well have a user community as large as the VLT! Let us look briefly at just a few areas of importance.

5.1. Galaxy formation in the early Universe

The detection of CO and dust at red-shifts approaching 5, or at an epoch when the Universe was less that 10% of its present age [see12,13], show that some galaxies had formed and the first star bursts had enriched their interstellar medium even at this early epoch – much earlier than had previously been assumed from optical observations alone! Submillimetre continuum observations with the SCUBA bolometer array on JCMT [14] are detecting what are probably dusty proto-galaxies which have no optical counterparts and Ivison [15] has detected prolific amounts of dust in a radio galaxy at z = 4.25. The inferred enormous gas mass suggests that star formation may still be in progress. The fact that star-bursts probably produce vast quantities of dust which probably obscure their optical light probably makes ALMA the best instrument with which to study the epoch of galaxy formation.

5.2. Stellar evolution

ALMA will be make an important impact over all areas of stellar evolution and not only stellar birth and death, the more usual realms of millimetre astronomy. Many types of star will be detected by the in the continuum. It is to be expected that ALMA will produce a quantum leap in the study of radio stars - as large as that produced by the centimetric capabilities of the VLA [16]. About 100 stars have been detected at millimetre wavelengths at flux limits of 10 – 30 mJy; they cover the entire HR diagram and include O and B stars, W-R stars, hot stars with shells, pre-main sequence stars, giants and supergiants. According to the MMA 'Sun and Stars Working Group' (see 'Science' on the ALMA www site) 100,000 stars would have been detectable with the MMA with one hour integration each. The ALMA collecting area is 3.5 times the MMA, so this time will be reduced to less than 6 minutes with ALMA.

The significance of millimetre interferometry in the study of star formation has been shown by the remarkable detection of Keplarian rotation in the disc surrounding the binary star GG Tau. IRAM interferometer images in the ^{13}CO line by Dutrey et al [17] show a probable protoplanetary disc. Future high resolution, high sensitivity observations will not only enhance this work but enable astronomers to extend such observations to many more systems. e.g. An interferometer with a resolution of 0.05 arc seconds at 230 GHz would resolve a 5 AU disc at a distance of 100 pc.

Studies of the chemical properties of circumstellar material surrounding evolved stars are highlighted by another IRAM interferometer study of the carbon star IRC 10216 [18]. More than 50 different molecules have been detected in the dusty envelope of this star. The different molecules have different distributions, e.g. sodium chloride, NaCl, a highly refractory molecule seems to be produced exclusively in the stellar atmosphere while C_2H, a product of ion-molecule chemistry, is seen only in the outer envelope. Such studies require good linear resolutions, which are reachable in IRC 10216 which is only 200 pc away; to extend such studies to more distant objects requires both finer resolution and higher sensitivity.

5.3. Solar System

Millimetre-wave spectroscopy of *comets* has resulted in the detection of more than 20 molecules in their atmospheres. The ALMA with its high sensitivity and resolution will allow maps of the molecular distribution to be made on relatively short time scales, facilitating studies of changing structures in the coma. The 0.1 arcsec beam of ALMA represents a linear resolution as fine as 70 km at a distance of 1 AU.

Asteroids will also be studied with ALMA. Continuum observations will provide details of their size and albedo and comparative observations at submillimetre, millimetre and centimetre wavelengths will probe their sub-surface temperatures, giving important clues to the outer structure of these bodies [19].

Marten [20] has discussed observations of *planetary atmospheres* with a millimetre array. Combined with studies from spacecraft, ALMA observations will enhance both studies of planets but also their satellites. Millimetre spectroscopy is an important tool in the study of planetary atmospheres. Examples of important goals for ALMA are the detection of trace species like sulphur bearing molecules in the atmosphere of Venus, since sulphur is a key element governing its complex atmospheric chemistry, and organic compounds in the Martian atmosphere.

ALMA will facilitate profound advances in the study of planetary satellites. An important advance came with the discovery of SO_2 in the Jupiter satellite, Io; with ALMA one should be able to detect other minor compounds like SO and H_2S from Io's volcanic plumes.

6. *Project status*

The ALMA phase 1 study has been underway for two years. The project has been organised into study groups concerned with different parts of the system, prototype receivers are being developed and the prototype antennae are soon due for delivery

6.1. The telescopes

With 64 antennae, trouble free operation and ease of maintenance are extremely important so the telescopes should be simple and robust. The specifications of 25 micron surface accuracy and pointing to better than 0.6 arcsec have never been achieved for open air millimetre telescopes but a US/European working group has considered that the specifications are feasible and several design studies have been published [21,22]. Because of the tight specification independent US and European calls for tender have been issued and two prototype antennae will be purchase and compared. A first series of tests on the prototypes, including interferometric tests, will be performed at the US VLA site in N.Mexico before placing the final antenna contract. In this way it is hoped to minimise the antenna test work on the high site.

6.2. Configuration and reconfiguration of the array

Applications of ALMA call for all configurations from a tightly spaced array for mapping Galactic clouds, to a system with the full resolution for early universe work. Cost considerations call for minimisation of the number of pads and the number of moves. With these factors in mind, Conway [23] among others, has discussed configuration geometries which will reduce the number of antenna stations (pads) and even enable the array to be continuously reconfigured within the shortest 3 km baselines, requiring a major reconfiguration only for the large 12 km arrangement, which would be a 12 km diameter circle.

Conway's idea, which is now accepted as the basis for the 3 km configuration, is for a 'zoom' array based on a three armed logarithmic spiral geometry. At intermediate resolutions, the spiral has a constant pitch angle, which is chosen so that the azimuth changes by 1.5 turns for a ratio of 8 in radius. At these intermediate scales, the pattern is self-similar. A reduction of scale gives rise to exactly the same baseline coverage (except for a rotation, unimportant because of the near circularly symmetry). The array is 'continuously' reconfigured by moving three antennae twice per week, and in this way it is possible to have two full cycles per year through the 3km configuration.

6.3. Receivers

It is foreseen that SIS (superconductor-insulator-superconductor) mixer technology will be used for all receiver bands except the lowest frequency band near 40 GHz. The SIS receivers will each be installed in a 'cartridge', and the cartridges will be grouped into a single large dewar. In the event of a fault, a whole new receiver assembly will be installed.

Side-band separation SIS mixers have been discussed [24] because of their potential to improve the sensitivity for line measurements by reducing the atmospheric noise in the image side-band. At NRAO an experimental 200 – 300 GHz system has been developed in which all the RF components are on a single chip; another interesting design is under investigation by the Onsala receiver group [25].

It is important to realise that between 30 and 900 GHz, there are of order ten operating frequency bands in the atmospheric windows. If we choose to observe with dual polarisation in all of these bands, the LMA calls for about 1500 receivers (with spares). Such a large number of SIS receivers (all must be cooled to 4 K) has never been contemplated before and some serious production arrangements must be set up for their manufacture.

6.4. Local oscillator

A crucial part of the receiver is the local oscillator system since a heterodyne receiver requires a high purity signal to mix with the incoming astronomical signal. This local oscillator must have a power level of several microwatts, be highly coherent across the array and locked in phase to a microwave reference source. Most millimetre LO signals are derived by multiplication of a lower frequency reference signal but an interesting innovation which might be simpler and cheaper, and lends itself to distribution across the array (over distances of 10 km or more) has been suggested by Payne et al [26]. This is a 'photonic local oscillator' in which two laser signals, separated by the required local oscillator frequency, are distributed around the array in optical fibres and mixed at the telescopes. It seems that the limitation to the realisation of such a system today is the availability of photomixers with sufficient power conversion efficiency but work is continuing.

6.5. Correlator

The ALMA correlator will be big, there being 2016 baselines per polarisation, per frequency interval (1024 typically). Thus conventional correlator architectures may need to be re-considered. Developments in computer technology continue apace and it has been decided therefore to adopt a two-stage approach to the correlator construction: NRAO will design and build an interim correlator, based on conventional ideas, for the first operational (sub)array of say 15 – 20 antennae and only at that stage will the design of the full correlator be decided.

6. Timescale

Because of funding delays, the phase 1 study will continue well into 2002. The prototype antennae are already out of phase, there being a delay in delivery of the European prototype. This probably means that the antenna testing will not be complete until the end of 2003, thus delaying the commencement of construction on the site until at least 2005. The array should be completed by the end of 2011.

7. Acknowledgements

In this paper I have called freely on material in the ALMA proposals. Much technical material has been assembled in the former MMA and now ALMA series of memos that can be located on either the ESO or NRAO ALMA web sites.

8. References

1. Booth, R.S., 1997. Proc ESA Symp. On The Far Infrared and Submillimetre Universe, ESA publ. SP-401, p. 235.
2. R. Brown, R., 1990. 'The Millimetre Array' proposal to the US National Science Foundation,
3. Booth, R.S., 1994. IAU Colloquium 140, Astronomy with Millimetre and Submillimetre Wave Interferometry, eds M. Ishiguro and Wm. J. Welch (Astron Soc. Pacific Conf. Series)
4. Downes, D., 1994 in Frontiers of Space and Ground-based Astronomy. Eds W. Wamsteker, M.S. Longair and K. Kondo. (Kluwer)
5. Radford, S. and Holdaway, M., Proc. SPIE, vol 3357, 254, 1998, ed T. G, Phillips.
6. Holdaway, M. A., Radford, S. J. E., Owen, F. N., & Foster, S. M., 1995, Millimeter Array Memo 129
7. Lay, O.P., Proc. SPIE, vol 3357, 254, 1998
8. Delgado, G. *et al*, 1999. ALMA Memo 271
9. Welch, W.M., 1999 in Reviews of Radio Science 1996-1999. Oxford. Ed. W Ross Stone.
10. Downes, D. 1998. 'LSA/MMA Feasibility Study', ESO.
11. Shaver, P. & Booth, R.S., 1998. ESO Messenger, no 91. p. 26.
12. Ohta, K. *et al.*, 1996. Nature, 382, 426.
13. Omont, A. *et al.* 1996. Nature, 382, 428.
14. Smail,I., Ivison, R.J., Blain, A.W., 1997. ApJ, 490, L5.
15. Ivison, R.J. *et al*, 1998. Ap.J., 494, 211.
16. Pallavicini, P. & White, S.M., 1995, in Science with Large Millimetre Arrays. Springer. Ed P. Shaver. p.268
17. Dutrey, A., Guilloteau, S., Simon, M., 1994. Astron. Astrophys., 280, 149.
18. Guélin, M., Lucas, R., Neri, R. 1995, in Science with Large Millimetre Arrays. Springer. Ed P. Shaver. p.276
19. Crovisier, J., 1995, ibid, p. 316
20. Marten, A., 1995, ibid, p. 308
21. Plattner, D., 1998. Proc. SPIE, 3357, 755.
22. Andersen, T., Koch, F., Shaver, P. 1998. ibid, p.770.
23. Conway, J., 1998, MMA Memo 216; 1999, ALMA Memo 260; 2000, ALMA Memo 283.
24. Webber, J.C. *et al*, 1998. Proc. SPIE, 3357, 122.
25. Vassilev, V. & Belitsky, V., 2001. Proc 12th International Symposium on Space Terahertz Technology.
26. J. Payne et al, Proc. SPIE, 3357, 143, 1998

Global Interplanetary Dust Distribution Measurements with the EISCAT and Arecibo HPLA radars

Asta Pellinen-Wannberg[1], Gudmund Wannberg[2], Assar Westman[3], Ingemar Häggström[4], John D. Mathews[5], Diego Janches[6], and David D. Meisel[7]

[1]Swedish Institute of Space Physics, Box 812, 981 28 Kiruna, Sweden. E-mail: asta@irf.se.

[2]EISCAT Scientific Association, Box 164, 981 23 Kiruna, Sweden. E-mail: ugw@eiscat.com.

[3]EISCAT Scientific Association, Box 164, 981 23 Kiruna, Sweden. E-mail: assar@eiscat.com.

[4]EISCAT Scientific Association, Box 164, 981 23 Kiruna, Sweden. E-mail: ingemar@eiscat.com.

[5]CSSL, Penn State University, University Park, PA 16802, USA. E-mail: JDMathews@psu.edu.

[6]Arecibo Observatory, Box 995, Arecibo, Puerto Rico 00613. E-mail: djanches@naic.edu.

[7]SUNY Geneseo, Geneseo, NY 14454, USA. E-mail: meisel@geneseo.edu

Abstract

High Power Large Aperture (HPLA) radars can observe the plasma clouds formed around micrometeoroids or dust particles of 1-1000 μm size as they impinge on the E-region. Observations performed with the EISCAT and Arecibo radars have yielded many dust parameters, such as instantaneous velocities, vector velocities, decelerations, scattering cross-sections, particle sizes, masses and fluxes. The results have established the HPLA facilities as important instruments for micrometeor studies in an otherwise almost inaccessible size range. The technique will be used for a systematic study of the spatial distribution of interplanetary dust in a collaborative project using the Arecibo radar to observe the near-ecliptic component and EISCAT for the northern hemisphere dust component. The observations will be performed at different times of year to sample the ecliptic plane all around the sun. The results are expected to significantly improve our knowledge of the interplanetary dust distribution at 1 AU ecliptic crossing.

1 Introduction

The characteristics of the interplanetary dust and particle distributions are important for understanding the near-Earth space environment in terms of impact risks and effects on low Earth orbit satellites. High power large aperture radars have recently been shown to be useful in a wide range of applications concerning small particles entering Earth's atmosphere from space. These radars can observe the impacting meteoroids with very high resolution enabling origin determinations through their orbits. Simultaneously they can monitor the effects on the background ionosphere through the standard incoherent scatter process. A review of the method is given in [1].

It is proposed to use the Arecibo and EISCAT radars to make coordinated observations of the interplanetary dust particle populations at 1 AU. So far most dust detectors have been carried on board interplanetary missions such as Ulysses, Galileo and Cassini. By combining the results from Arecibo near-ecliptic and EISCAT northern hemisphere observations, statistically significant knowledge of the global dust population dynamical distribution at 1 AU will be established.

2 Radar facilities

There are only about ten HPLA radar systems in the world. All the incoherent scatter radar facilities and some military radars can be included in this class. Meteor observations have been performed at Jicamarca Observatory in Peru [2], EISCAT Scientific Association in Northern Scandianvia [3], Arecibo Observatory in Puerto Rico [4], Millstone Hill Radar in Massachusetts, USA [5], Sondrestrom Radar on Greenland, MU radar in Japan [6] and the ALTAIR Radar on the Marshall Islands [7]. Although all these facilities have much higher power and larger aperture than traditional meteor radars, they are also different in many respects and can be expected to both confirm and complement each others' observations.

2.1 EISCAT

EISCAT, the European Incoherent SCATter facility, is located in Northern Scandinavia and on Svalbard. The EISCAT systems that have been used for meteor observations are the tristatic UHF at 930 MHz and

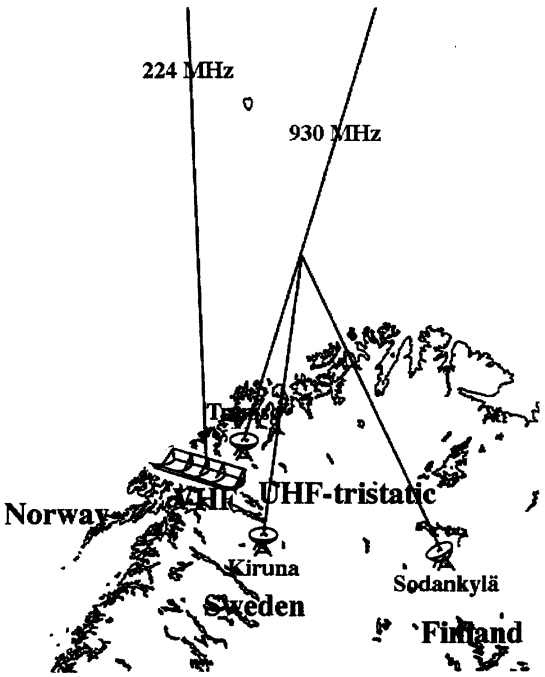


Figure 1: The EISCAT mainland radar sites in Northern Norway, Sweden and Finland. The tristatic UHF facility comprises three 32 meter parabolic antennas. The transmitter/receiver is located in Tromsø and the remote receivers in Kiruna and Sodankylä. The monostatic VHF transmitter/receiver is collocated with the UHF radar in Tromsø. The aperture and type of this radar antenna is a 40 m x 120 m parabolic trough.

the monostatic VHF at 224 MHz [3], [8], [9], [10], [11], [12] [13]. Figure 1 shows the locations of the EISCAT UHF and VHF radar sites. The geographical coordinates are 69° 35' N and 19° 14' E for Tromsø, 67° 52' N and 20° 26' E for Kiruna and 67° 22' N and 26° 38' E for Sodankylä site. In the tristatic mode EISCAT can be used to observe vector velocities and thus meteoroid trajectories. Orbits up to ± 60° from the local vertical can be defined. The upper limit here is probably set by the longer path through the atmosphere rather than due to the angle at which the meteoroid can technically be observed.

2.2 Arecibo

The Arecibo Observatory (AO) 430 MHz radar uses the world's largest antenna, the 305 m diameter radio telescope located in Puerto Rico. Many basic and application studies on meteors and dust have been performed with this facility [4], [14], [15], [16], [17], [18],[19], [20]. The coordinates of the site are 18° 21' N and 66° 45' W. At Arecibo, orbits for meteoroids impacting at up to ± 15° from the local vertical can be determined. Arecibo looks at the ecliptical plane during the nighttime in winter and during the daytime in summer. Due to the combined effect of the latitude and the earth's tilting on its orbit the vertical for Arecibo deviates about 40° at most from the ecliptic plane during a 24 h period.

3 The Measurements

The HPLA radars principally observe head echoes characterised by Doppler shifts corresponding to typical meteor velocities, while the classical meteor radars observe the targets through the formation of an ionisation trail along their path. The direct observation of the moving target with simultaneous high spatial and time resolution provides a series of instantaneous values of many meteor parameters, such as velocity, deceleration, size etc.

3.1 Techniques

The various HPLA radar facilities were originally designed for different purposes, with different technical solutions regarding siting, antenna types and sizes, choice of frequency, transmitter characteristics etc. To achieve optimal performance in dust measurements, the measurement technique must therefore be adapted to the individual radar system.

Meteors and dust can only be detected when they produce echoes significantly above the receiver noise floor. As the signal-to-noise ratio of a unit cross-section point target at a given range is directly proportional to antenna aperture squared, it is clear that the Arecibo observatory with its 305-meter dish enjoys a big advantage. Arecibo can observe particles down to submicron size and achieve meteor event rates of up to 1000/hour. Their measurement scheme uses 40 μs uncoded pulses together with a sampling rate of 1 μs^{-1}. As the targets of interest are transient point scatterers and the background is dominated by the relatively slowly varying incoherent echo from the equatorial E region plasma, this simple method is sufficient for many purposes.

The tristatic EISCAT UHF system comprises three 32-meter dish antennas located in the auroral region. Due to the relatively small antenna apertures, the higher operating frequency (930 MHz vs. 430 MHz) and the fairly restrictive signal processing possibilities, the best event rates observed so far are only about 10/hour [9]. Since the instantaneous signal-to-noise ratios observed at EISCAT are small and the signal from the background E region can be strong and highly dynamic in e.g. auroral events, pulse compression by binary phase modulation and coherent decoding is applied to improve the effective SNR. 13x3 μs and 13x2 μs Barker codes have been used in the

earlier experiments. A 32x2 μs binary code was introduced recently and will be used in future coordinated HPLA dust observations.

A complication peculiar to the early EISCAT measurements was that the very small amount of memory available in the digital correlator forced high time resolution measurements to be done by accumulating data from 32 series of ten consecutive pulses on top of each other. While retaining a basic 6-ms time resolution, this technique also raises the baseline variance by a factor of 5.7, thus potentially masking many weak echoes. After a major upgrade of the EISCAT facilities in 2000/2001, this complication is no longer a factor. In the newest version of the EISCAT meteor mode, (16+16) bits of complex amplitude data from the receiver are recorded to disk and/or DLT at 1.67 MHz sampling rate. Event search and analysis is done off-line on a pulse-by-pulse basis, thus achieving the lowest possible system noise floor.

3.2 Strategies

A systematic observation series of four annual measurements of the interplanetary dust distribution will be performed. The operations will be scheduled close to the vernal equinox, the summer solstice, the autumnal equinox and the winter solstice to cover extreme points on the earth's orbit at about one astronomical unit distance from the sun. Figure 2 shows the experiment topology in space.

Every measurement will last for 24 hours to cover the full Earth revolution. Since the earth moves in its orbit at about 30 km/s velocity, it will encounter many meteors on the forward side around 0600 local time. 12 hours later at the same geographical point the meteors have to catch up with the Earth. Thus the event rates as well as the velocities are expected to be lower.

A first preliminary 24 hour measurement session was run on the EISCAT tristatic UHF radar starting on March 19, 2002 at 1100 UT. As described above, a new modulation was used as well as new signal prosessing and data recording facilities. In earlier tristatic measurements the Sodankylä site has been the limitation for the observations due to its location far away from the transmitter. By moving the tristatic point closer to Sodankylä the common scattering volume increases and the polarisation topology will be better, thus increasing the probability of observing meteors at this site. Monthly observations at Arecibo have also been performed since January 2002, and will continue until January 2003.

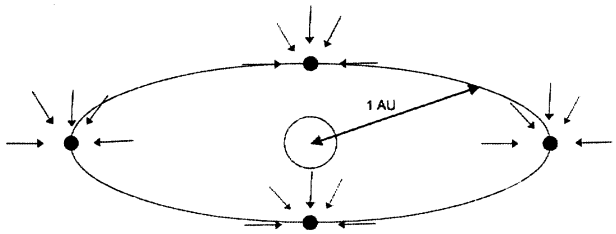


Figure 2: The measurement strategy for the solar system dust distribution observations at about 1 astronomical unit distance from the sun in the middle of the figure. The measurements are performed in four points along the Earth's orbit - during vernal and autumnal equinox and summer and winter solstice. During 24 hour measurements the Arecibo radar samples a full rotation of the earth the close to ecliptical flux and the EISCAT radar the Northern Hemisphere flux.

4 Summary

The use of HPLA radars for a global survey of the interplanetary dust distribution at 1 AU is just in its infancy. About ten years of meteor/dust observations at EISCAT and Arecibo, and later even with other facilities, have established the method as important for studies of otherwise almost inaccessible particle size ranges. Systematic observations along the earth's trajectory will be performed both at EISCAT and Arecibo during the coming few years. By combining the results from the near ecliptic and the Northern Hemisphere observations, statistically significant knowledge of the global dust population dynamical distribution at 1 AU will be established.

Acknowledgments

The EISCAT Scientific Association is supported by the *CNRS* of France, the *MPG* of Germany, the *PPARC* of the United Kingdom, the *NFR* of Norway, the *VR* of Sweden, the *SA* of Finland, and the *NIPR* of Japan. This work is supported by the European Office of Aerospace Research and Development, Air Force Office of Scientific Research, Air Force Research Laboratory, under Contract No. F61775-01-WE084. This work is also part of the NSF/CEDAR Postdoctoral Fellowship (ATM-0004284) awarded to Diego Janches.

References

[1] Pellinen-Wannberg, A., The high power large aperture radar method for meteor observations, Proceedings of the Meteoroids 2001 Conference,

Swedish Institute of Space Physics, ESA SP-495, 443-450, 2001.

[2] Chapin, E., and E. Kudeki, Radar interferometric studies of long duration meteor echoes observed at Jicamarca,*J. Geophys. Res.*, 99, 8937-8949, 1994.

[3] Pellinen-Wannberg, A., and G. Wannberg, Meteor observations with the European incoherent scatter UHF radar, *J. Geophys. Res.*, 99, 11 379-11 390, 1994.

[4] Zhou, Q., C. A. Tepley, and M. P. Sulzer, Meteor observations by the Arecibo 430 MHz incoherent scatter radar, I. Results from time-integrated observations, *J. Atmos. Terr. Phys.*, 57, 421-431, 1995.

[5] Erickson, P. J., F. D. Lind, S. M. Wendelken, and M. A. Faubert, Meteor head echo observations using the Millstone Hill UHF incoherent scatter radar system, Proceedings of the Meteoroids 2001 Conference, Swedish Institute of Space Physics, ESA SP-495, 457-463, 2001.

[6] Zhou, Q. H., J. D. Mathews, and T. Nakamura, Implications of meteor observations by the MU radar, *Geophys. Res. Lett.*, 28, 7, 2001.

[7] Close, S., S. M. Hunt, M. J. Minardi, and F. M. McKeen, Analysis of Perseid meteor head echo data collected using the Advanced Research Projects Agency Long-Range Tracking and Instrumentation Radar (ALTAIR), *Radio Sci.*, 25, 1233-1240, 2000.

[8] Pellinen-Wannberg, A., and G. Wannberg, Enhanced ion-acoustic echoes from meteor trails, *J. Atmos. Terr. Phys.*, 58, 1-4, 495-506, 1996.

[9] Wannberg, G., A. Pellinen-Wannberg, and A. Westman, An ambiguity-function-based method for analysis of Doppler decompressed radar signals applied to EISCAT measurements of oblique UHF-VHF meteor echoes, *Radio Sci.*, 31, 497-518, 1996.

[10] Westman, A., A. Pellinen-Wannberg, and G. Wannberg, Meteor head echo height effect observed by the European Incoherent Scatter UHF and VHF radars, in Assar Westman's Ph.D. thesis, IRF Scientific Report 246, 1997.

[11] Pellinen-Wannberg, A., A. Westman, G. Wannberg, and K. Kaila, Meteor fluxes and visual magnitudes from EISCAT radar event rates: a comparison with cross-section based magnitude estimates and optical data, *Ann. Geophysicae*, 16, 1475-1485, 1998.

[12] Pellinen-Wannberg, A., A. Westman, and G. Wannberg, A three-dimensional meteor head echo Doppler shift method for the EISCAT UHF radar, Meteoroids 1998 proceedings, Astron. Inst., Slovak Acad. Sci., 83-86, 1999.

[13] Janches, D., A. Pellinen-Wannberg, G. Wannberg, A. Westman, I. Häggström, and D. D. Meisel, Tristatic observation of meteors using the 930 MHz EISCAT radar system, submitted to *J. Geophys. Res.*, 2002.

[14] Mathews, J. D., K. P. Hunter, D. D. Meisel, V. S. Getman, and Q. Zhou, Very High Resolution Studies of Micrometeors using the Arecibo 430 MHz Radar, *Icarus*, 126, 157, 1997.

[15] Zhou, Q. H., J. D. Mathews, and Q. N. Zhou, Incoherent scatter radar study of impact of meteoric influx on the nocturnal E-region ionization, *Geophys. Res. Lett.*, 1833-1836, 1999.

[16] Janches, D., J. D. Mathews, D.. D. Meisel, V. Getman, and Q. Zhou, Doppler Studies of Near-Antapex UHF Radar Micrometeors, *Icarus*, 143, 2, 347-353, 2000a.

[17] Janches, D., J. D. Mathews, D.. D. Meisel, and Q. Zhou, Micrometeor Observations using the Arecibo 430 MHz Radar: I. Determination of the Ballistic Parameters from Measured Doppler Velocity and Deceleration Results, *Icarus*, 145, 2, 2000b.

[18] Mathews, J. D., D. Janches, D. D. Meisel, and Q. H. Zhou, The Micrometeoroid Mass Flux into the Upper Atmosphere: Arecibo Results and a Comparison with Prior Estimates, *Geophys. Res. Lett.*, 28, 1929, 2001.

[19] Janches, D., J. D. Mathews, D.. D. Meisel, and Q. Zhou, Orbital Properties of the Arecibo Micrometeors at Earth Intersection, *Icarus*, 150, 206-218, 2001.

[20] Meisel, D. D., D. Janches, and J. D. Mathews, Extrasolar Micrometeors Radiating from the Vicinity of the Local Interstellar Bubble, *ApJ*, 567, 323-341, 2002.

Incoherent scatter spectra for non-Maxwellian plasmas

Mikael Hedin[1] and Ingemar Häggström[2]

[1]Swedish Institute of Space Physics, Box 812, 981 28 Kiruna, Sweden. E-mail: mikael.hedin@irf.se.
[2]EISCAT Scientific Association, Box 164, 981 23 Kiruna, Sweden. E-mail: ingemar@eiscat.com.

Abstract

We present an extension of the standard incoherent scatter spectral analysis. The standard assumption of the analysis is that there is only thermal electrons present in the ionosphere. This is often not the case. The major deviation from the standard spectrum can bee seen in the electron line, a weak part of the spectra between the prominent ion line and the plasma line.

This extension should be important in analysis of incoherent scatter spectra observed during conditions when there are particularly large deviations from the Maxwellian electron distribution such as during auroral precipitation, for high temperatures as obtained during ionospheric radiowave heating (HF-pumping) and at altitudes lower than 200 km all the time.

1 Introduction

The possibility to measure the scatter of radio waves from free electrons in the ionosphere using powerful transmitters and sensitive receivers was predicted in 1958 [1] and the first observation was reported the same year [2]. Since then, the technique has developed into a standard instrument for space exploration, and today there are several system around the earth in regular operation, *e.g.,* Arecibo in Puerto Rico, Millstone Hill in Massachusetts, USA and EISCAT in northern Scandinavia.

In [1] the scatter from free thermal electrons was predicted to give a weak scattered signal, Doppler broadened by the corresponding thermal motion. The first observations, however, detected a significantly stronger and more narrow signal, and it was soon realized that the full spectral shape of the return signal contained the information. In the present work we will broaden the view and again study the originally considered, wider part of the spectrum, to learn more about non-thermal conditions.

As the ionosphere is a plasma, albeit weakly ionized, the electrons do not move freely, but are coupled to the ions by the electric field. This make the equations for the kinetic properties of electrons and ions coupled, with complex mutual dependencies. The frequencies of the scattered return signal is modulated by these motions, *e.g.,* Doppler broadening and Doppler shift. Analyzing the scattered signal shape it is thus possible to determine macroscopic properties of the plasma, such as electron density, ion and electron temperature and plasma drift.

2 Interpreting the incoherent scatter signal

The traditional analysis of the incoherent scatter spectra, used in routine operation of the production systems, uses the (relatively) high power part of the spectrum. The *ion line* is the central part, typically extending over several kHz around the radar transmitter frequency.[1] An example is given in figure 1.

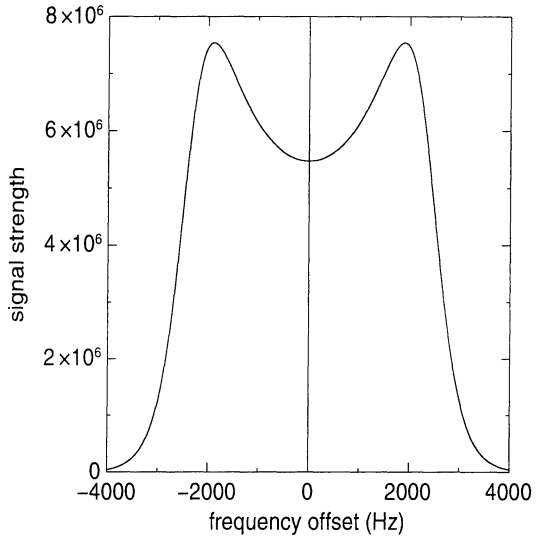


Figure 1: Signal strength for the ion line (in arbitrary units) The plasma properties are determined from the shape of the ion line, roughly as follows: the electron density is the spectral height, the first moment is the plasma drift velocity, the width depends on the ion temperature and the depression in the middle depends on electron temperature and collision frequency.

[1]The width of the scatter lines all depends on the transmitter frequency, and is given here for the EISCAT VHF system at 224 MHz

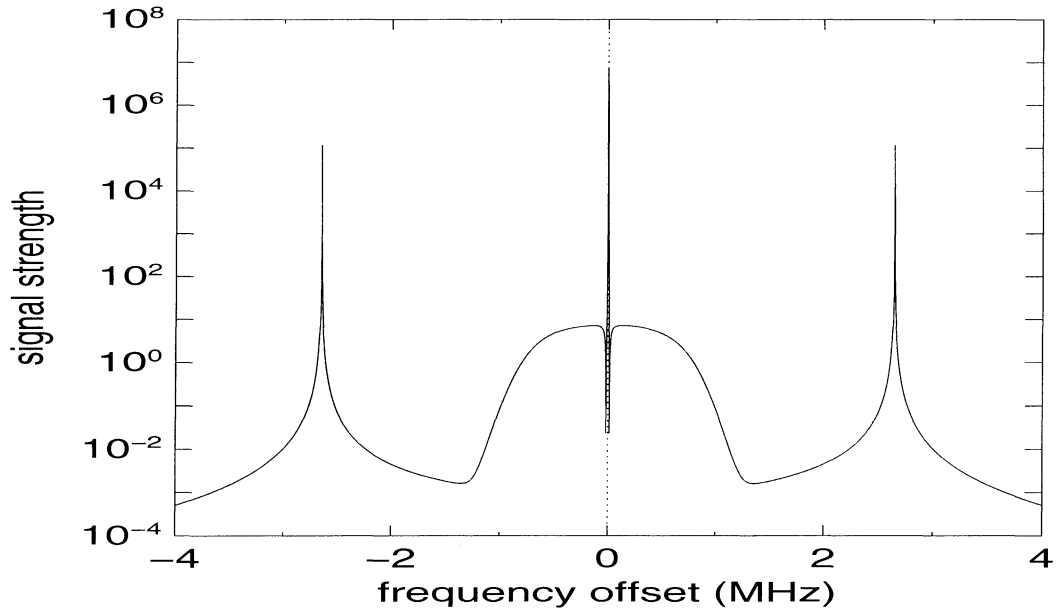


Figure 2: The incoherent scatter spectra for thermal conditions. From the center and out, we see the narrow ion line, the weak electron line and the two plasma lines.

There are also often run experiments to use the *plasma line*, under some conditions a strong signal, narrow in width and located at offsets from the transmitter frequency close to the plasma frequency. A typical offset value is ± 3 MHz, as in the example in figure 2.

The Doppler broadened scatter from the free electrons at frequency offsets ± 1 MHz is called the *electron line*.[2] This part has not been used in standard operations, since the signal is difficult to detect above the measurement noise, but with recent improvements of instrument hardware, such experiments have become possible. As we will show, this part is modulated and by the presence of supra-thermal electrons with energies $\approx 0.1 - 10$ eV, and strong enhancements occur for such conditions.

3 Calculating the spectra using the Nyquist theorem

Now we turn to the task of actually determining the spectra of the incoherent scatter signal. To do this, we make use of the Nyquist noise theorem. It was originated for electrical systems in [3] and generalized to arbitrary systems and dimensions in [4, 5] (not to be confused with the Nyquist theorem on frequencies).

The application of the Nyquist theorem to incoherent scatter was pioneered in a series of articles [6, 7, 8, 9, 10] for various ionospheric condition. We use these results as a basis for the current work.

[2]This value is also for the EISCAT VHF system

3.1 Summary of the theorem

The Nyquist noise theorem assumes that a generalized harmonic force F is applied to a system, and the system response is X, such that the instantaneous rate of work done by the force is $F_a X_a$ (both assumed to vary with time as $e^{\ i\omega t}$ where ω is the angular frequency). If the response is linear, it can be expressed as

$$X_a = Y_{ab}(\omega) F_b, \tag{1}$$

and then, in the absence of external forces and in thermal equilibrium, there is a spontaneous response in the system, determined by the *admittance* Y of the system as

$$\langle |X_a X_b| \rangle \mathrm{d}\omega = \frac{KT}{2\pi} \mathrm{Re}[Y_{ab}(\omega) + Y_{ba}(\omega)] \mathrm{d}\omega, \tag{2}$$

where K is the Boltzmann constant and T the temperature of the system. Equivalently this can be regarded as a spontaneous action using (1). This is the thermal noise, present in all kinds of system, originating from the quantum nature of matter.

3.2 Application to the ionosphere

We regard the ionospheric plasma as a collection of separate particle populations, electrons and several species of ions, which each make up a separate system which we apply the Nyquist theorem to. Here we use vector notation, as we use three dimensions, and we are only interested of the response in the direction of the applied force so no tensor is needed[10]. We use

the definition of the admittance and the continuity equation for each population:

$$N\vec{u} = Y(\vec{F} + q\vec{E}), \tag{3a}$$

$$\Delta N = kNu/\omega, \tag{3b}$$

and the poison equation for the electric field

$$-ik\varepsilon_0\vec{E} = \sum q\Delta N. \tag{4}$$

where the summation is over all species.

The incoherent scatter cross section is

$$\sigma\,\mathrm{d}\omega = r_e^2\langle|\Delta N_e|^2\rangle\mathrm{d}\omega, \tag{5}$$

where r_e is the classical electron radius. Together with (3) and (4) this evaluates to

$$\sigma = \frac{N_e r_e}{\pi\omega}\left\{\left|\sum_j \mu_j y_j + ik^2{\lambda_D}^2\right|^2 \mathrm{Re}\{y_e\} + |y_e|^2\sum_j \eta_j\,\mathrm{Re}\{y_i\}\right\} \cdot \left|y_e + \sum_j \mu_j y_j + ik^2{\lambda_D}^2\right|^{-2} \tag{6}$$

where

$$\eta_j = N_j {q_j}^2/N_e e^2 \tag{7}$$

and

$$\mu_j = \eta_j T_e/T_j \tag{8}$$

and the index j stand for the different ion species. Here we use the normalized admittance y, defined as

$$Y_k = \frac{N_k\omega y_k}{k^2 K T_k}. \tag{9}$$

In this form we have neglected the bulk drift of the different species, which will change ω to $\omega - \vec{k}\cdot\vec{v}$ in (3b), and hence (6) and (9) will have to be modified appropriately.

The admittances are calculated from the Boltzmann equation, and forms for different assumptions can be found in the references in section 3, *e.g.,* the appendix in [7].

3.3 Extending to multiple electron distributions

In the ionosphere, the relaxation time for the plasma is long, *i.e.,* the particles interaction in a thermal sense is very weak, and the equilibrium condition is not very strict. As the plasma is normally stationary on a time scale of seconds, a long time compared to the radar probings, we can allow the populations to have different thermal properties, *e.g.,* temperature (T) and bulk drift velocity ($\vec{v}$). As we assume thermal equilibrium for the individual populations, their particle distribution will be Maxwellian:

$$f_0 = N_0\sqrt{m/2\pi KT}\,e^{-m\vec{v}^2/2KT}. \tag{10}$$

This also applies to the electrons if we have a quasi-stable distribution which is not strictly thermal, *i.e.,* not Maxwellian. Then we can regard it as several "species" of electron, which together constitute the present distribution. We then extend (3) and (4) to apply for also the different electron populations.

The resulting cross section is then

$$\sigma = |R|^{-2}\sum_e\left\{\left|R - \frac{{q_e}^2 N_e y_e}{T_e}\right|^2 \frac{N_e y_e^R}{\omega} + \left|\frac{q_e N_e y_e}{T_e}\right|^2\left(S - \frac{{q_e}^2 N_e y_e^R}{\omega}\right)\right\} \tag{11}$$

where superscripted y^R denotes the real part of y and we use the shorthands

$$R = \sum_e \frac{N_e y_e {q_e}^2}{T_e} + \sum_i \frac{N_i y_i {q_i}^2}{T_i} + i\varepsilon_0 k^2 K \tag{12}$$

and

$$S = \sum_e \frac{N_e y_e^R {q_e}^2}{\omega} + \sum_i \frac{N_i y_i^R {q_i}^2}{\omega}. \tag{13}$$

4 Comparison of the spectra

To compare these new spectra with the traditional results, we use data from Monte Carlo simulations of the electron distribution during a HF-pump situation at auroral latitudes, *e.g.,* relevant for EISCAT experiments[11]. The modification of the electron distribution is caused by inelastic collisions of electrons of energy $2 - 4\,\mathrm{eV}$, mainly with N_2, to such high degree that the results a significant deviation from the Maxwell distribution in the range 1.5 to 5–7 eV, as shown in figure 3.

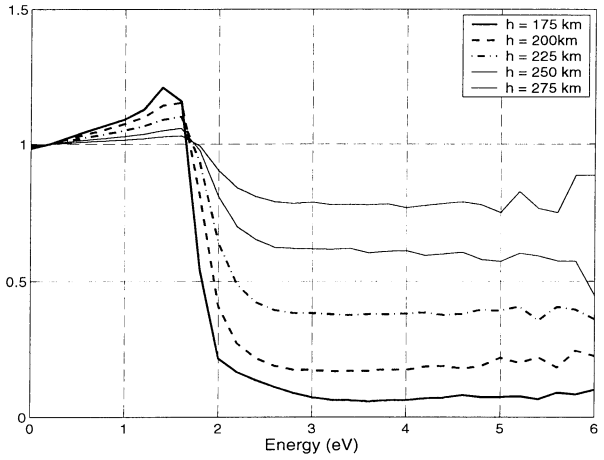


Figure 3: Quotient of the distribution of a pure Maxwellian distribution and the modeling results for several heights.

In the spectra in figure 4 we see that the ion and plasma line are not affected much, but the electron line is enhanced by 4 orders of magnitude at it's peak. This is such a significant difference that it should be possible to distinguish the signal from the background noise, that would hide the non-enhanced electron line.

5 Experimental evidence

To test these theoretical predictions, an special purpose program has been written for the EISCAT VHF system in Tromsø, named ELIN. This experiment was run during two periods in 2001, together with the Heater and the UHF system in a regular mode. The experiment uses a 2.5 MHz wide receiving band, with the sender frequency 0.25 MHz from the lower edge, to make sure the high half part of the electron line falls completely in the receiving frequency band.

The data from a one minute integration in figure 5 shows that there probably is an electron line present. However, we cannot be sure that the heating really happened in the measuring range at this particular time, and if increased temperature was produced of not. To get a quantitative measurement, the temperature increase from the heater must be larger, and at the right height. It might also be possible to make a longer integration, but that requires the ionospheric conditions not to change as to destroy or alter the electron line signal.

Acknowledgement

The authors thanks Dr. Björn Gustavsson, NIPR, Japan for valuable discussions and comments on this work.

References

[1] W. E. GORDON. Incoherent scattering of radio waves by free electrons with applications to space exploration by radar. *Proc. IRE*, 46:1824, 1958.

[2] K. L. Bowles. Observation of vertical-incidence scatter from the ionosphere at 41 Mc/sec. *Phys. Rev. Lett.*, 1(12):454–455, December 1958.

[3] H. Nyquist. Thermal agitation of electric charge in conductors. *Phys. Rev.*, 32:110–113, July 1928.

[4] Herbert B. Callen and Richard F. Greene. On a theoreme of irreversible thermodynamics. *Phys. Rev.*, 86(5):702–710, June 1952.

[5] Richard F. Greene and Herbert B. Callen. On a theoreme of irreversible thermodynamics. II. *Phys. Rev.*, 88(6):1387–1391, December 1952.

[6] J. P. Dougherty and D. T. Farley. A theory of incoherent scattering of radio waves by a plasma. *Proc. Roy. Soc. London*, A, 259:79–99, 1960.

[7] D. T. Farley, J. P. Dougherty, and D. W. Barron. A theory of incoherent scattering of radio waves by a plasma: II. Scattering in a magnetic field. *Proc. Roy. Soc. London*, A, 263:238–258, 1961.

[8] J. P. Dougherty and D. T. Farley. A theory of incoherent scattering of radio waves by a plasma: 3. Scattering in a partly ionized gas. *J. Geophys. Res.*, 68(19):5473–5486, 1963.

[9] D. T. Farley. A theory of incoherent scattering of radio waves by a plasma: 4. The effect of unequal ion and electron temperatures. *J. Geophys. Res.*, 71(17):4091–4098, 1966.

[10] W. E. Swartz and D. T. Farley. A theory of incoherent scattering of radio waves by a plasma: 5. The use of the Nyquist theorem in general quasi-equilibrium situations. *J. Geophys. Res.*, 84(A5):1930–1932, 1979.

[11] B. Gustavsson, T. Sergienko, I. Häggström, and F. Honary. Simulation of high energy tail of electron distribution function. Submitted to Adv. Space Res., 2001.

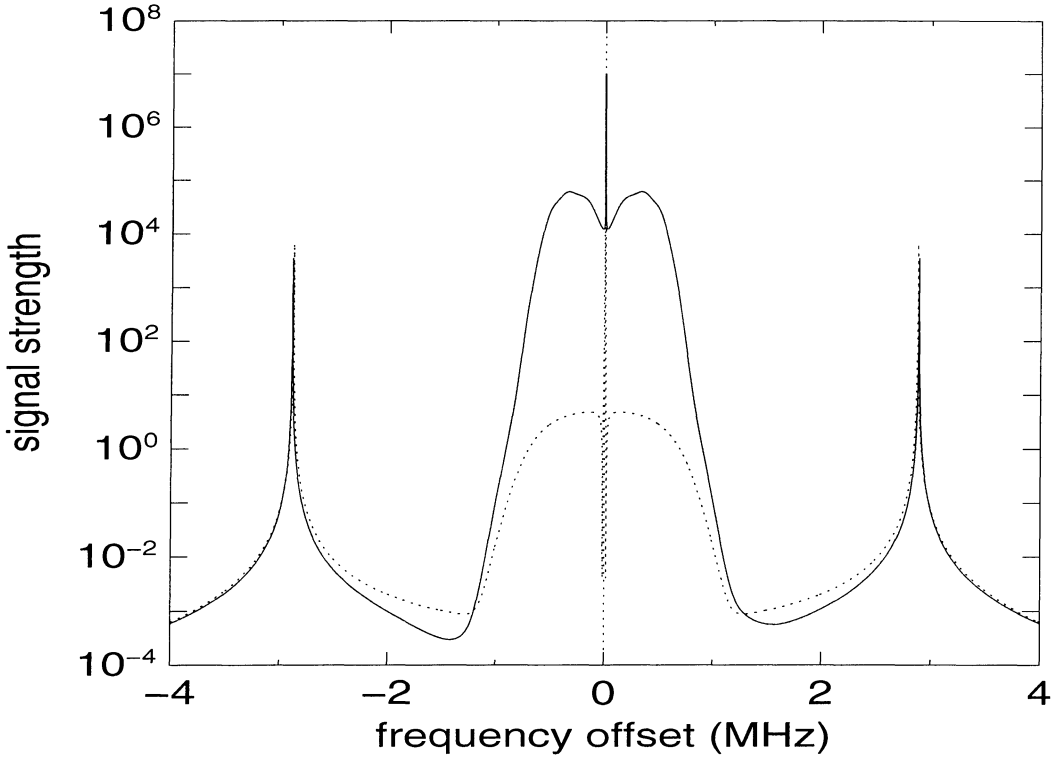


Figure 4: The incoherent scatter spectra with enhanced electron line (full line) superimposed on an standard spectra for the same conditions (dashed line).

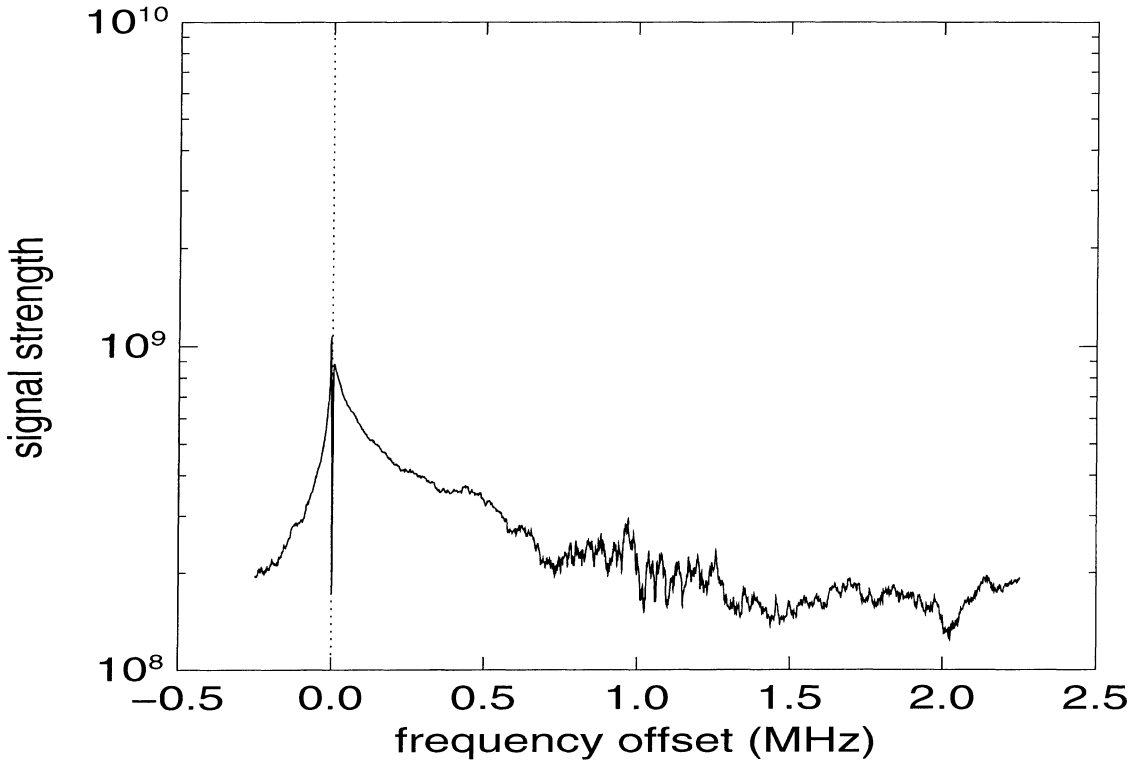


Figure 5: Data plot from the ELIN experiment ELIN@VHF 2001.10.14 18:04, integrated for 1 minute.

Plasma and electron line measurements with the EISCAT radars

Ingemar Häggström[1], Mikael Hedin[2] and Björn Gustavsson[3]

[1]EISCAT Scientific Association, Box 164, SE-98123 Kiruna, Sweden. E-mail: ingemar.haggstrom@eiscat.com
[2]Swedish Institute of Space Physics, Box 812, SE-98128 Kiruna, Sweden. E-mail: mikael.hedin@irf.se
[3]National Institute of Polar Research, 1-9-10 Itabashi, Tokyo 173-8515, Japan. E-mail: bjorn@nipr.ac.jp

Abstract

The normal incoherent scatter spectrum consists of the ion line, the electron line and the two plasma lines. These lines depends on different parts of the electron velocity distribution: the ion line is sensitive to electron energies in the sub-eV range and the plasma line to energies of around 10 eV and the electron line for energies in between.

The plasma lines, offsetted from the transmitted frequency by the frequency of the Langmuir waves of the plasma, are naturally enhanced by photo- and auroral electrons. From the relative magnitudes and position one can deduce line-of-sight currents and amount of suprathermal electrons. Measurements, using the EISCAT UHF radar, in the active auroral E-region are presented.

The electron line has a large spectral width and several orders of magnitude lower power than the other lines and has so far never been measured at EISCAT. Theoretical calculations of the expected electron distribution shows features depending on the neutral gas properties and the results are far from Maxwellian, leading to enhancements in the electron line. The recent upgrades at EISCAT makes it possible to try to measure the electron line properties and in March 2001 an experiment was run using the VHF radar and the Heater to control the electron temperature. Results from the experiment and comparisons to theory are presented.

1 Introduction

The normal incoherent scatter spectrum consists of the ion line, the electron line and the two plasma lines. These lines depends on different parts of the electron velocity distribution: the ion line is sensitive to electron energies in the sub-eV range and the plasma line to energies of around 10 eV and the electron line for energies in between.

The ion line is the main line in the spectrum and is reletively easy to measure as it is rather strong and narrow. This line is used in almost all of the incoherent scatter radar measurements and enables extraction of many of the thermal parameters of the plasma constituents, such as density, temperature and bulk velocity.

The plasma lines, offsetted from the transmitted frequency by the frequency of the Langmuir waves of the plasma, are naturally enhanced by photo- and auroral electrons. During the enhancements it is possible to detect the lines, with power levels sometimes exceeding the ion line. The offset from the transmitter frequency is the Langmuir frequency, which depends on the plasma parameters. Because of this, the line is not easy to detect, especially during active conditions when the plasma undergoes large variations in fractions of seconds. In practice, most of the succesful plasma line measurements have been carried near the F-region peak where the height gradient of the Langmuir frequency is minumum, and in HF heating situations where the Langmuir frequency matches the heater frequency.

The electron line has a considerable larger spectral width and several orders of magnitude lower power than the other lines. Due to that, there have been almost no measurements done and till now at EISCAT none.

Recent developments in the incoherent scatter theory, as well as technical progress, have now raised the interest in the plasma and electron lines. In this paper we present the first fits of a multi-Maxwellian electron distribution using plasma line measurements in the vicinity of auroral electrons. We will also use predictions of deviations from Mawwellian in the lower energy range for electrons, causing enhancements of the electron line. In the first EISCAT electron line experiment we show that these predictions are in the right direction, but need refinements, especially in the higher energy range.

2 Theory

The full incoherent scatter spectrum from a Maxwellian plasma, typical for the lower auroral F-region, is shown in Fig. 1. One can easily distuingish four major features: the two plasmalines at large frequency offsets, the central wide electron line and the central narrow ion line.

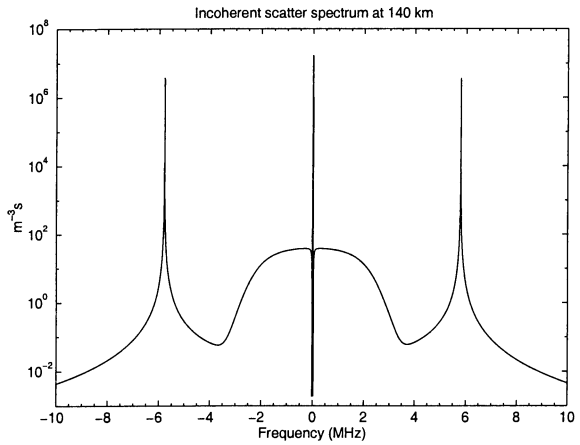


Figure 1: The wide frequency incoherent scatter spectrum for the EISCAT UHF radar.

By separating the different major type of particles into electrons and ions it is possible to extend the theory for multimaxwellian distributions of both electrons and ions. Häggström et al. (2000) has shown that even a moderate addition, less than 1 percent, of low energy electrons will enhance the plasma line considerable, with power levels exceeding the ion line.

Almost any distribution can be studied in a similar way, since it is possible to construct it from a large number of maxwellians. Gustavsson et al. (2001) made in-depth theoretical calculations of the expected electron distribution in the F region showing features depending on the neutral gas properties and the results shows distributions that sometimes differs considerable from Maxwellian.

3 Measurements

3.1 Plasma lines

The Swedish-Japanese EISCAT campaign in February 1999 were devoted to measurements close to and inside auroral arcs together with the Swedish auroral imaging system ALIS. For this, an optimised experiment were developed for spectral measurements of the ion line in the E-region together with four narrow bands for plasma line detection. In Häggström et al. (2000) the experiment is fully described and the large number of plasma line enhancements are discussed.

Of the about 200 enhancements, only one had good signatures for spectral investigations. There are a number of reasons for this. The experiment was not fully optimised for spectral plasma lines, due to compromises between data size and height resolution. Whenever the gradients of the Langmuir frequency of the plasma, both in time and height, are larger than the receiver band width, the line will be smeared out allover the receiver band and no distinct line, suited for analysis will be detected. In the active auroral E region this happens most of the time.

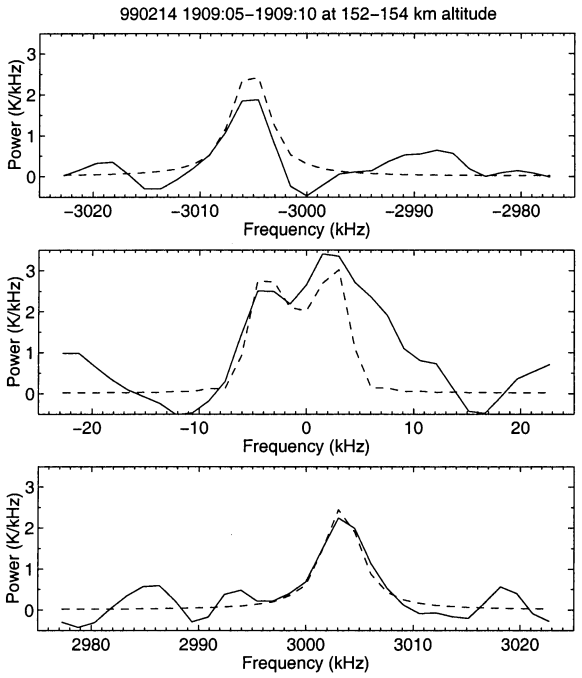


Figure 2: Measured spectra of the both plasma lines and the ion line and the best 7-parameter fit of the theoretical spectrum. The parameters fitted were $N_e = 9.7 \cdot 10^{10}$ m ^{3}, $T_e = 600$ K, $T_i = 228$ K, $v_i = 100$ ms ^{1}, $v_e = 483$ ms ^{1} and a suprathermal distribution with $n_e = 2.2 \cdot 10^{8}$ m ^{3} and $T = 7$ eV.

Anyhow, there were one, and this have now been analysed in detail. Fig. 2 shows the three lines together with the best fit using two superposed electron distributions, one representing the normal thermal plasma and one representing a secondary distribution caused by energetic auroral precipitation. One can derive an electrical current from the parameters the fit provides from $j = q_e n_e (v_i \quad v_e)$, and in this cases that amounts to 6 μAm ^{2}, which is maybe slightly high, but not unreasonable.

3.2 Electron lines

The electron lines have sofar not been possible to detect at EISCAT, but theoretical calculations [1], shows that the VHF band of EISCAT, 224 MHz, would be the most feasible frequency. However, in the last year considerable efforts have been done to upgrade the EISCAT radars, with receiving capabilities of extreme flexability and content. By utilising the new features an experiment was run in March 2001 on the VHF radar to try to detect the electron line. In order to modulate the F region electron temperature and the width of the electron line, also the EISCAT HF heater were run in a 10 min on — 5 min off cycle.

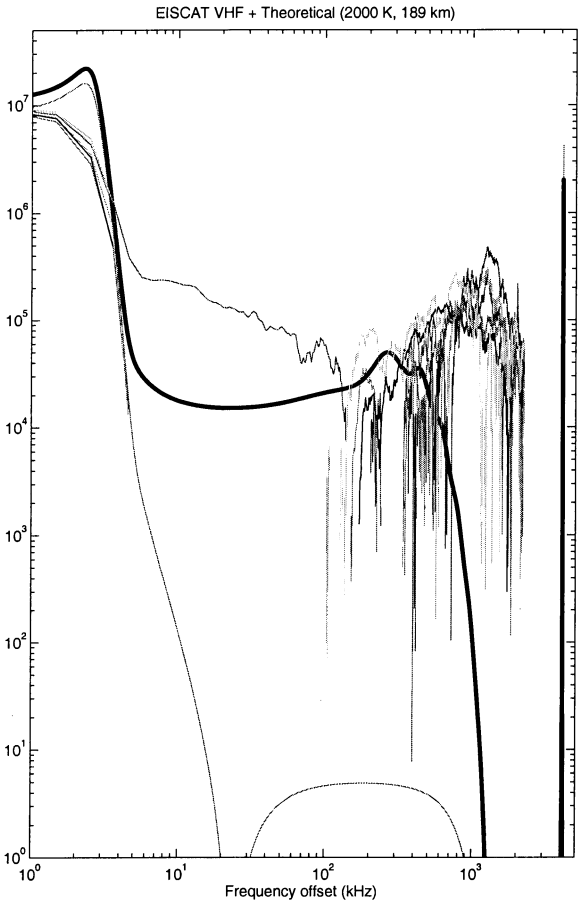


Figure 3: Measured wide band VHF incoherent scatter spectra at different times in the lower F region. The thick line represents a theoretical spectrum based on a expected calculated electron distribution. The thin theoretical line in the bottom is a theoretical spetrum of a Maxwellian electron distribution.

Sofar only 30 minutes of data have been analysed, but the results of this are very promising. Fig. 3 shows the measured spectrum at a number of heights, with and without the heater. During this time, however, the plasma frequencies of the E region were higher than the heater frequency, preventing the heater wave to reach F region heights. So, no heating modulation could be expected. Included in the figure are an expected spectrum, based on the calculations by Gustavsson et al (2001). The measured spectra are close to the expected up to 0.5 MHz offset, where the theoretical drops off and the measured continues to grow. The measured spectra peak between 1 and 1.5 MHz. This discrepancy should certainly be smaller if the theoretical calcutions of electron distribution went to finer structures, especially at larger energies.

4 Summary

The paper uses the recent progress in theoretical calculations of incoherent scatter and, together with experimental data, peer into new fields of research in auroral ionospheric physics. By using incoherent scatter radars it is now possible to investigate the electron distribution in much greater detail at energies up to about 10 eV. This is especially important when trying to understand plasma processes in active areas.

Acknowledgements

We are indebted to the Director and staff of EISCAT for operating the facility. EISCAT is an International Association supported by Finland (SA), France (CNRS), the Federal Republic of Germany (MPG), Japan (NIPR), Norway (NFR), Sweden (VR) and the United Kingdom (PPARC).

References

[1] Häggström I., M. Hedin, T. Aso, A. Pellinien-Wannberg and A. Westman, Auroral field-aligned currents by incoherent scatter plasma line observations, Adv. Polar Upper Atmos. Res., 14, 103-121, 2000.

[2] Gustavsson B., T. Sergienko, I. Häggström, F. Honary, Simulation of high energy tail of electron distribution function, Submitted to Adv. Space Res, 2001.

Analysis of HF enhanced 6300 Å airglow

Björn Gustavsson[1], Tima Sergienko[2], Urban Brändström[2], Thoma Leyser[3] and Åke Steen4[4]

[1]National Institute of Polar Research, Tokyo 173-8515, Japan. E-mail: bjorn@irf.se.

[2]Institutet för Rymdfysik, Box 812, 981 28 Kiruna, Sweden. E-mail: urban@irf.se, tima@irf.se

[3]Institutet för Rymdfysik, 755 91 Uppsala, Sweden. E-mail: thomas.leyser@irfu.se

[4]Remspace, Linköping, Sweden. E-mail: ake.steen@remspace.com

Abstract

This report presents the first estimates of the three-dimensional volume emission rate of enhanced O(1D) 630 nm airglow caused by HF radio wave pumping in the ionosphere. A region of enhanced airglow was imaged by three stations in the Auroral Large Imaging System (ALIS) in northern Scandinavia. These images allowed for a tomography-like inversion of the volume emission of the airglow. The altitude of maximum emission was found to be around 235 ± 5 km with typical scale sizes of 20 km. The altitude of maximum emission is found to be approximately 10 km below the altitude of the enhanced ion line and 15 km above the altitude of maximum electron temperature. Comparisons of the measured altitude and temporal variations of the 630 nm emission with modeled emission caused by $O(^1D)$ excitation from the high energy tail of a Maxwellian electron distribution show significant deviations.

1 Introduction

A powerful high-frequency (HF) radio wave transmitted into the ionosphere may enhance airglow, which result from inelastic collisions between energized electrons and atmospheric constituents. By the dependence on energization of electrons the enhanced airglow can be used to study the dissipation of pump-driven plasma turbulence. In addition, by the impact excitation of states in neutral atoms and molecules the airglow can be used to determine a range of parameters in the ionosphere and thermosphere (e.g. [1]).

The airglow at 6300 Å from the lowest excited states of oxygen $O(^1D)$ and $O(^1S)$ has been studied in repeated experiments since the early 1970s at low latitude (Arecibo) [2, 3, 6], at midlatitudes (Platteville Colorado, Moscow, Sura) [4, 5, 6] and at auroral latitudes [7, 8]. Apart from being one of the few fields in space physics where repeatable experiments can be performed it is also (to our knowledge) here the only quantitative theory in space physics can be found: The "thermal theory" put forward by [9] where it is suggested that "Large 6300-Å airglow intensity enhancements observed in ionosphere heating experiments are excited by thermal electrons". According to this theory are:

1. The electron distribution Maxwellian during HF-pumping, at all altitudes and times.
2. The variation in electron, ion and neutral temperatures (T_e, T_i, T_n respectively) can be determined from the heat equations.
3. The $O(^1D)$ excitation is obtained from the thermal excitation rate (thermally averaged cross section for electron impact excitation)
4. The 6300 Å emission is then calculated from the $O(^1D)$ continuity equation.

With simultaneous observations of the enhanced 6300 Å airglow, T_e and electron concentration n_e it is possible to quantitatively test this theory.

2 Experiment and Data

In February 1999 we were able to test this theory in an experiment with simultaneous optical observations with ALIS [10] imaging the enhanced airglow [8] simultaneously with EISCAT UHF measurements of n_e, T_e and T_i during a period from 1630 to 1830 UT on Feb 16th when the EISCAT Heating facility transmitted about 125 MW effective radiative power in O-mode heating the night time F region ionosphere with a 2 min and 2 min off cycle until 1720 UT and a 4 min on and 4 min off cycle from 1724 UT. In this study we will concentrate on the period from 1740 to 1800 UT. For the period 1740 to 1756 UT the 6300 Å airglow was measured by the ALIS stations in Silkimuotka (68.03 N, 21.69 E), Abisko (68.36 N, 18.82 E), and Nikkaluokta (67.85 N, 19.00 E) as shown in Figure 1. Exposures were made synchronously every 10 s, and the exposure time was 5 s. For this period it was possible to apply a tomography like inversion to retrieve the spatial distribution of volume emission and thus determine the altitude variation of the enhanced 6300

Å emission at every time step [11] - the result showed typical scale sizes of 20 km. For this period we ob-

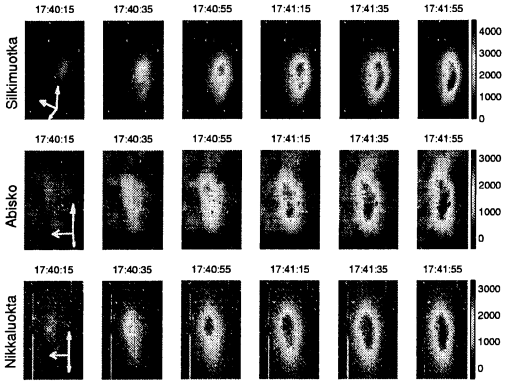


Figure 1: The initial growth period of the pulse between 1740 and 1744 UT as seen from Abisko, enhanced intensity is shown in counts. The light blue structures in the images from Abisko are due to clouds in the lower atmosphere. The arrows are clockwise from vertical the projection of north, zenith, and east directions.

served T_e of up to 3500 K, while T_i not increased by more than a few 100 K and never exceeded 1300 K as, is show in figure 2. There it is also seen that n_e has a slow and stable decrease and the ion flows are low.

3 Data analysis

In this section we will test whether the thermal hypothesis can explain both the optical and radar observations. In order to do so we do a detailed modeling of the 6300 Å airglow we use a time-dependent one-dimensional ionospheric model. In this model we solve the coupled time-dependent continuity equations for the minor neutral species, $O(^1D)$, $O(^1S)$, $N(^4S)$, $N(^2P)$, $N(^2D)$, $N(^2P)$, NO and the ions O_2^+, NO^+, N_2^+, O^+ the equation of motion for ions and electrons as well as the electron and ion energy equations. Input parameters of the model are the neutral composition and temperature, MSIS-90 atmospheric model [12], the ultraviolet solar flux and precipitating particle flux. A more exhaustive description of the source and loss processes of the other minor species included in this model as well as the numerical methods used in solving the equations is outside the scope of this report.

To obtain the initial conditions (most important the ion composition of the lower F region ionosphere) the equations are solved from 1200 UT with input of the MSIS neutral atmosphere for the prevailing geophysical conditions.

The main ionospheric sources of $O(^1D)$ that are

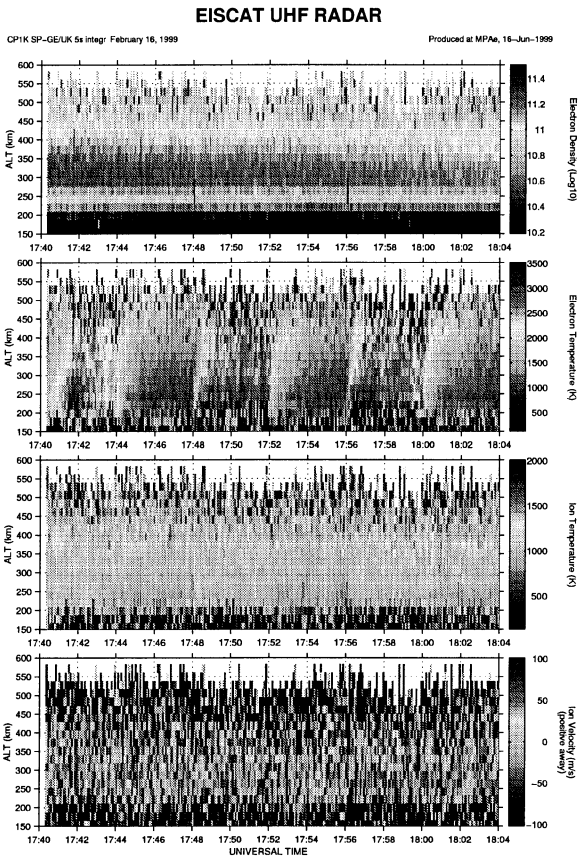


Figure 2:

affected by the HF pumping are dissociative recombination

$$O_2^+ + e_{th} \rightarrow O(^3P, ^1D, ^1S) \quad (1)$$

with a reaction rate [13]

$$\beta = 1.9 \times 10^{-7}(T_e/300)^{-0.5} \quad (2)$$

with a $O(^1D)$ yield of 1.2 atoms per reaction, and direct electron impact

$$O(^3P) + e(E > 1.96 \text{ eV}) \rightarrow O(^1D) + e \quad (3)$$

with a reaction rate [14]

$$\alpha(T_e) = 0.502\sqrt{T_e}\frac{(9329 + T_e)}{(51813 + T_e)^3}\exp(-\frac{22756}{T_e}) \quad (4)$$

for excitation from a thermal electron distribution. The increase in electron temperature caused by the HF-pump wave will decrease the reaction rate of dissociative recombination (equation (1)) according to equation (2), and provided that the electron distribution remains Maxwellian over the energy range 2–10 eV, increase the rate coefficient of equation (3). The HF-pump wave induced variation in 6300 Å airglow will then be determined by the relative importance of the two reactions. To model the 6300 Å airglow response to the HF pump, we need accurate calculations of the ionosphere-thermosphere composition and temperature.

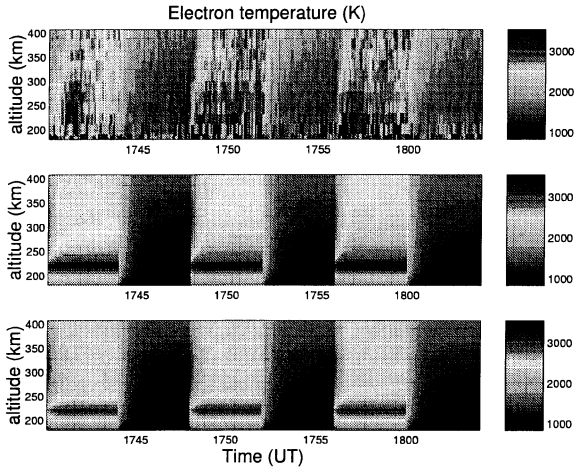


Figure 3: (top) EISCAT UHF measurements of electron temperature, (middle) modeled electron temperature with best fit of temperature, and (bottom) modeled electron temperature giving best fit with optical data.

To test the thermal hypothesis, where the electron distribution is assumed to remain Maxwellian during the interaction with the HF-pump wave [2, 14, 15] and the excitation of $O(^1D)$ is caused by a thermal tail above 1.96 eV, we search for a heat source that reproduces the EISCAT UHF electron temperature measurements between 1740 and 1804 UT, both concerning the time variation and altitude distribution. The best fit with the EISCAT UHF data is obtained with a 5 km wide heat source of 6×10^4 eV/m^3/s at an altitude of 220 km for which both the altitude and the temporal variation of the electron temperature agree well with what is observed, as shown in the Figure 3 where the measured electron temperature for the period 1740 to 1804 UT is plotted in the upper panel and the modeled electron temperature, henceforth denoted as the "hot temperature," from this source, is plotted in the middle panel. The $O(^1D)$ emission caused by excitation of the high-energy tail of the hot temperature electron distribution corresponds to an increase of $\sim$ 200% above the background compared to the observed increase that is only between 50 and 70%. Here it should also be noted that the EISCAT measurements are made along the magnetic field line at approximately 13°south while the airglow enhancement are centered approximately 20 km north of the EISCAT UHF line-of-sight which should decrease the observed enhancement by another factor of 2 to between 25 and 35 %. Furthermore, neither the

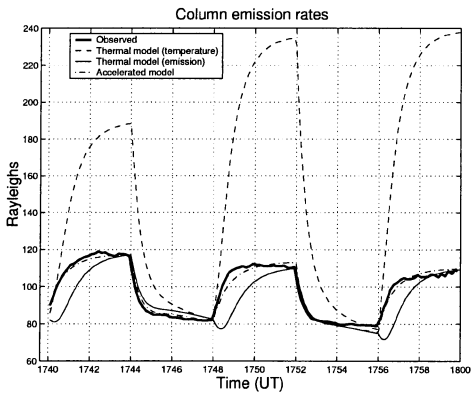


Figure 4: Measured (thick solid) and modeled column emission. The dashed curve is airglow emission calculated from the "hot temperature" profiles, the thin solid curve the airglow emission calculated from the "warm temperature" profiles, and the dot-dashed curve the airglow emission calculated with an electron distribution that corresponds to the hot temperature profile up to 2 eV and a constant isotropic electron source from 2–10 eV.

time variation, where the modeled intensity growth does not level off as markedly as the observed (as is seen in Figure 4) nor the altitude distribution plotted in Figure 5, agrees very well with what is observed. In particular, the altitude of maximum emission is at too low an altitude by 18 km. The opposite ap-

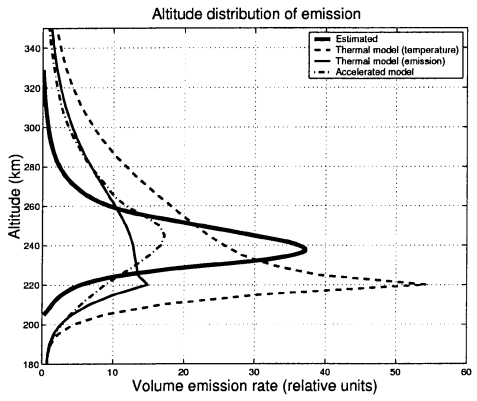


Figure 5: Comparison between estimated and modeled altitude distribution of the 6300 Å emission. The thick solid curve is the estimate (from tomography) altitude distribution of 6300 Å emission. The dashed curve is the airglow modeled with excitation from the high-energy tail of a Maxwellian electron distribution with the "hot electron temperature". The thin solid curve is the airglow distribution as above but with the 'warm electron temperature'. The dot-dashed curve is the airglow distribution from an isotropic source at the altitude of the enhanced ion line.

proach, to search for a heat source that gives a good fit with the optical data, results in an estimated heat source of 4×10^4 eV/m^3/s. This also leads to incompatibilities. There is an initial lag time in the O(^1D) 6300 Å airglow due to an initial decrease in the source from dissociative recombination, and the calculated electron temperature (henceforth denoted "warm temperature") is systematically lower than the electron temperatures observed by the EISCAT UHF radar (as can be seen in the bottom panel of Figure 3) for all times and all altitudes, especially the peak temperature that is now only $\sim$ 3070 K compared to the measured 3500 K.

The other main hypothesis is that the enhanced excitation of O(^1D) is due to a nonthermal electron distribution in the energy range 2–10 eV. It has been suggested that the source of this distribution is acceleration of electrons by Langmuir turbulence [16, 17]. To model the O(^1D) 6300 Å airglow response to this source, we assume an electron source at the same altitude as the altitude of the HF-pump wave enhanced ion line altitude with isotropic flux and a uniform energy distribution in the range 2 – 10 eV. The excitation rate of O(^1D) from this source is calculated with a Monte Carlo model of the electron transport in the Earth's atmosphere [18]. The O(^1D) excitation from dissociative recombination is taken into account, while the contribution from direct impact excitation of the thermal electrons is absent in this case. An electron source with an energy flux of 1.8×10^9eVcm^{-2}s^{-1} gives a 6300 Å emission that has the best agreement with the observed emissions.

The direct O(^1D) excitation rate α (equation (4)) can be written as

$$\alpha \propto \int_{1.96}^{\infty} Ef(E)\sigma(E)dE, \qquad (5)$$

where E is the electron energy in eV, $f(E)$ is the electron distribution function, and $\sigma(E)$ is the integral excitation cross section for direct electron excitation of O(^1D). Since the modeled enhancement of the column emission from the hot temperature is approximately a factor of 3 larger than observed we can conclude that

$$\int_{1.96}^{\infty} Ef(E)\sigma(E)dE \approx 0.3 \int_{1.96}^{\infty} Ef_{th}(T_e, E)\sigma(E)dE, \quad (6)$$

where $f_{th}(E, T_e)$ is a Maxwellian electron distribution with electron temperature T_e. With only measurements of 6300 Å airglow it is not possible to determine the shape of the high-energy tail of the electron distribution. A consequence of the above relation is, however, that the main characteristic of the electron distribution from 2–10 eV is a depletion, as shown in Figure 6, i.e., the electron distribution function is found to be sub-thermal during ionospheric HF pumping.

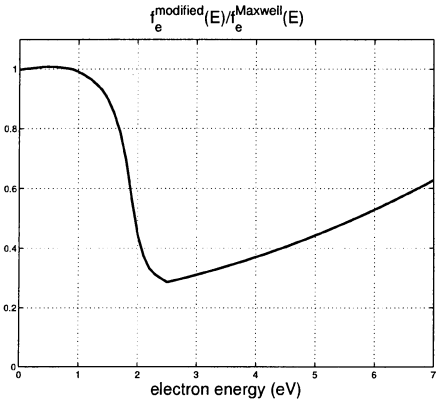


Figure 6: Estimated correction factor, that should be produced by inelastic electron–nitrogen collisions.

One likely reason for this is the energy loss of 2–3 eV electrons to the vibrational levels of N_2. It has long been known that this electron excitation of the vibrational states of N_2 is strong and and several authors [19] have reported it to peak between 1×10^{-16} and 3×10^{-16}cm^2 for 2–3 eV electrons. The effect of electron excitation of the N_2 vibrational states has been calculated to produce significant dips in the electron spectra both during precipitation of high-energy

electrons [19] even at 250 km altitude; the same process should cause a decrease in the electron energy distribution relative to the purely thermal case also during HF-pump experiments.

References

[1] P. A. Bernhardt, M. Wong, J. D. Huba, B. G. Fejer, L. S. Wagner, J. A. Goldstein, C. A. Selcher, V. L. Frolov, and E. N. Sergeev. Optical remote sensing of the thermosphere with HF pumped artificial airglow. *J. Geophys. Res.*, 105(A5):10,657–10,671, MAY 2000.

[2] A. A. Biondi, D. P. Sipler, and R. D. Hake. Optical ($\lambda - 6300$) detection of radio frequency heating of electrons in the F region. *J. Geophys. Res.*, 75:6421–6424, 1970.

[3] H. C. Carlson, V. B. Wickwar, and G. P. Mantas. Observations of suprathermal electrons accelerated by HF-excited instabilities. *J. Atmos. Terr. Phys.*, 44:1089–1100, 1982.

[4] J. C. Haslett and L. R. Megill. A model of the enhanced airglow excited by rf-radiation. *Radio Sci.*, 9(11):1005–1019, November 1974.

[5] T. G. Adeishvili, A. V. Gurevich, S. B. Lyakhov, G. G. Managadze, G. M. Milikh, and I. S. Shluyger. Ionospheric emisson caused by an intense radio wave. *Sov. J. Plasma Phys.*, 4:721, 1978.

[6] P. A. Bernhardt, C. A. Tepley, and L. M. Duncan. Airglow enhancements associated with plasma cavities formed during ionospheric heating experiments. *J. Geophys. Res.*, 94(A7):9071–9092, July 1989.

[7] P. Stubbe, H. Kopka, H. Lauche, M. T. Rietveld, A. Brekke, O. Holt, T. B. Jones, T. Robinson, A. Hedberg, B. Thidé, B. Crochet, and H. J. Lotz. Ionospheric modification experiments in northern Scandinavia. *J. Atmos. Terr. Phys.*, 44(12):1025–1041, 1982.

[8] B. Urban E. Brändström, Thomas B. Leyser, Åke Steen, Michael T. Rietveld, Björn Gustavsson, Takehiko Aso, and M. Ejiri. Unambiguous evidence of HF pump-enhanced airglow at auroral latitudes. *Geophys. Res. Lett.*, 26(23):3561–3564, Dec 1999.

[9] G. P. Mantas. Large 6300-Å airglow intensity enhancements observed in ionosphere heating experiments are exited by thermal electrons. *J. Geophys. Res.*, 99(A5):8993–9002, 1994.

[10] Åke Steen and Urban Brändström. A multistation ground-based imaging system at high latitudes. *STEP Int. Newsl.*, 3:11–14, 1993.

[11] B. Gustavsson, T. , Sergienko, M. T. Rietveld, F. Honary, Å. Steen, B. U. E. Brändström, T. B. Leyser, Aruliah A. L., T. Aso, and M. Ejiri. First tomographic estimate of volume distribution of enhanced airglow emission caused by hf pumping. *J. Geophys. Res.*, 106(A12):29105–29123, December 2001.

[12] A.E. Hedin. Extension of the MSIS thermospheric model into the middle and lower atmosphere. *J. Geophys. Res.*, 96(A2):1159–1172, February 1991.

[13] Stan C. Solomon, P. B. Hays, and V. J. Abreu. The auroral 6300 Å emission: Observation and modeling. *J. Geophys. Res.*, 93(A9):9867–9882, 1988.

[14] G. P. Mantas. Large 6300-Å airglow intensity enhancements observed in ionosphere heating experiments are excited by thermal electrons. *J. Geophys. Res.*, 99:8993, 1994.

[15] George P. Mantas and Herbert C. Carlson. Reinterpretation of the 6300-Å airglow enhancements observed in ionosphere heating experiments based on analysis of Platteville, Colorado, data. *J. Geophys. Res.*, 101(A1):195–209, January 1996.

[16] G. Meltz and F. W. Perkins. Ionospheric modification theory: Past, present and future. *Radio Sci.*, 9:885–888, 1974.

[17] A. V. Gurevich, Ya. S. Dimant, G. M. Milikh, and V. V. Va'skov. Multiple acceleration of electrons in the regions high power radio-wave reflection in the ionosphere. *J. Atmos. Terr. Phys.*, 47:1057–1070, 1985.

[18] V. E. Ivanov and T. I. Sergienko. *Interaction of the Auroral Electrons With the Atmospheric Gases (The Monte-Carlo modeling).* Nauka, St. Petersburg, 1992.

[19] Manfred H. Rees. *Physics and Chemistry of the Upper Atmosphere.* Cambridge Univ. Press, New York, 1989.

Fundamental Potential and Limits of Fiberoptic Systems and Cellular Microwave Radio Systems in relation to Local Area Networks

Olle Nilsson
Chalmers University of Technology, Goteborg, Sweden
Phone: (46) 300 541475, Fax: (46) 300 541519
Email: nilsson@ep.chalmers.se

Abstract:
We first consider the fundamental physical limits on fiberoptic communication systems and demonstrate that the long term potential is huge indeed. We include a discussion of switching capacity in future electronic systems which also promises to be very large.

Noting that there is strong evidence that although the presently technology driving application is the trunk networks the main costs will be associated with the local area networks and the subscriber terminals we try to identify more fundamental obstacles to realise low cost optical LANs and subscriber terminals with very high capacity.

We discuss some attractive applications requiring such capacities which could motivate the necessary investments.

We then consider fundamental physical limits on cellular wireless radio systems and try to make a long term comparison between fiberoptic and wireless LANs and their combination.

In many places we try to raise provoking questions rather than providing definite answers.

Which are the fundamental physical limits on fiberoptic communication systems?
In a fiber to the home scenario more than 1 Gbps equivalent circuit switched capacity may well be feasible in the future. The capacity limit of a fiber as determined from Shannons channel capacity theorem by foreseeable available fiber bandwidth (presently about 30 nm) and a power level, about 1 mW, low enough not to cause too much non-linearities is probably at least 100 Tbps.
Accounting for the noise generated in the fiber amplifiers Lars Thylén, Stockholm, has estimated that with present fibers one could ultimately reach about 20 Tbps over long spans and about 40 Tbps over a few km.

Using two polarisations and wavelength multiplexing 160 channels of each 40 Gbps a team from NEC has experimentally reached 6.4 Tbps over 186 km.
On the other hand **single channel electronics and electro-optical converters will probably be limited to considerably less than 1 Tbps**, perhaps to a few hundred Gbps (today about 40 Gbps). However, this will still allow about hundred subscribers using 1 Gbps each to share the same electronic and electro-optical circuit.

Switching

To get some idea of the problem one may, as a thought experiment, consider the total network as one large switch and apply classical switching theory to calculate the necessary number of switch elements, i.e. gates, in different known configurations. Thus with the number of subscribers $N = 10^9$ one finds that a non-blocking switch of straightforward crossbar type requires about 10^9 gates per subscriber. This may not seem prohibitive but one must remember that the power consumption per gate at 100 Mbps is about 10μW with present day technology leading to a power consumption of 10 kW per subscriber. Even if future technology may well reduce the power consumption by a factor of hundred or more such a switching architecture does not seem realistic from e. g. practical topological reasons. Employing well-known architectures for very low blocking probability, e. g. modified Banyan structures, on the other hand, the number of gates required per subscriber is proportional to $\log_2 N$ rather than N. Then both the number of gates and the power consumption per subscriber become insignificant even if the proportionality constant is taken as large as 1000. Thus with $N = 10^9$ the number of gates per subscriber becomes of order 30 000 and the power consumption of order 0.3 W using present day technology. Such a switch architecture seems also well suited to be implemented in a geographically distributed structure.

In conclusion simple estimates indicate that although the presently technology driving application is the trunk networks **the main costs will be associated with the local area networks and the subscriber terminals** where the costs are not shared by a large number of subscribers.
It seems therefore that **very high capacity, self-routing, low cost, virtually non-blocking networks** could be built. They would simply provide fast and reliable point-to-point and multi-point connections and they would not require any operator in the traditional sense. Independent service providers could be connected as subscribers and engaged when needed by other subscribers. Large such networks could evolve by scaling small local networks already operating according to such principles. Then an interesting question is:

Which are the fundamental obstacles to realize low cost fiber optical LANs and subscriber terminals with very high capacity?

So far we have been able to identify only two. One is the cost for the physical installation of the fiber. Not the costs of the fibers and opto-electronic components themselves that by mass-production can be reduced drastically. Consider the very advanced opto-electronics in

standard cheap CD-player! The other, perhaps in a shorter perspective more important obstacle, is that before there is mass-production and low costs there is no big market. This leads to the next question:

Which applications requiring high capacities could motivate the necessary investments to install fiber and to set up mass-production?

There are at least two rather different applications that in time can become important:

- Hi-fi video with resolution and dynamics approaching that of the human eye (requiring about 100 Mbps equivalent circuit switched capacity).
- Rapid downloading huge amount of data for processing in future very high capacity user friendly personal computers with some form of artificial intelligence.

Key factors for the realization of very high capacity LANs may therefore be our ability to develop reasonably cheap and practical very high quality displays and user friendly very high capacity computers with fast large capacity memories. This will probably be possible but it may well require a considerable time.

In the mean time, however, new, as yet unforeseen, applications may of course become apparent.

Neither must we forget that there is a large variety of less cost-sensitive professional and semi-professional applications yet to be explored, aiming at increasing the work-force efficiency in nearly all sectors of the society.

Long term comparison between fiber optic and wireless LANs and their combination.

Simply because of the much lower carrier frequencies it seems safe to assume that microwave WLANs with truly mobile subscriber terminals can never reach the capacities of a fiber based LAN.

Another well-known difference between a fiber optical network and a cellular microwave network is that the latter due to the use of freely propagating waves is subject to interference problems and requires far reaching regulations concerning frequency allocations.

However, the present day and even the so called fourth generation cellular systems have a fairly large margin to fundamental limits. Several tens of Mbps should be possible and that would be enough for a lot of new applications.

This conclusion is based on the fact that there is a simple relation between the subscriber effective antenna area, the area covered by each beam from the base station and the free transisson loss and that the number of subscribers sharing the same beam from the base station limits the bandwidth per subscriber. Applying Shannon's capacity theorem and limiting the subscriber transmitter power to of order 1 mW one finds that with a reasonable density of base stations an equivalent circuit switched capacity of some 30 Mbps should be within reach although requiring very sofisticated microelectronics.

An operational low cost extremely high capacity fiber optical network would greatly facilitate the technical realization of such a high capacity cellular network. It could provide connections

between base stations and to the trunk network. Fiber optical distribution of microwave signals and perhaps even power are also important options.

It should finally be pointed out that **free space, i.e. wireless, optical or microwave links** that can largely overcome the interference problems and in principle have capacities at least of order 10 Gbps will be important options in circumstances where fibers cannot easily be installed.

A 4.0 GHz Differential Frequency Synthesizer For HiperLAN/2 in 0.18 μm CMOS

Håkan Magnusson and Andreas Kämpe

Radio Electronics, LECS/IMIT KTH-Electrum 229, 164 40 Kista, Sweden. E-mail: {hmag, aka-lps} @imit.kth.se

Abstract

This paper presents the design of a 4.0 GHz frequency synthesizer targeting HiperLAN/2. The frequency synthesizer utilizes a frequency doubler to reduce the frequency pulling. The reference frequency is set to 125 MHz. This shifts the reference spurs outside the HiperLAN/2 band. All circuits are fully differential and realized in a 0.18μm CMOS process. The complete PLL consumes 86 mW, having a simulated phase noise of -133 dBc/Hz at 2 MHz offset from a 2.0 GHz center frequency.

1 Introduction

The recent growth in Wireless communication and especially WLAN has increased the demand for low cost solutions. It would be very attractive to integrate the complete transceiver in a CMOS process. This would offer a possibility to integrate all components, mixing analog and digital circuits on the same chip and thereby lowering the cost. Another advantage with integration is that the components do not have to go off chip to interface each other; this reduces stray capacitance and stray inductance and is especially important for Radio Frequency (RF) blocks in a transceiver, such as the Local Oscillator (LO).

The target for this frequency synthesizer is HiperLAN/2, the transceiver utilizes a heterodyne architecture, which makes it necessary to have two local oscillators. This configuration reduces the problem with DC offset but it introduces a problem with the mirror frequency. The HiperLAN/2 specification requires an Image Rejection Ratio (IRR) of -97 dB for an intermediate frequency (IF) larger than 1.325 GHz. This is very difficult to achieve with the techniques that are available today. Therefore a lower IF of 1.2 GHz is chosen where the IRR requirement is only -86.5 dB. The first local oscillator will then operate at 4.0 GHz and the second at 1.2 GHz. To reduce the problem of frequency pulling, a frequency doubler is used after the VCO. Figure 1 shows a block diagram of the designed frequency synthesizer.

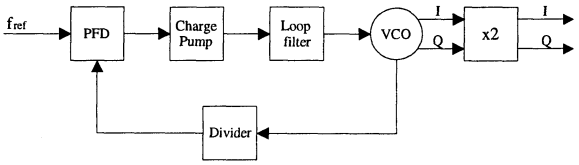


Figure 1: *Block diagram of the frequency synthesizer*

2 Synthesizer architecture

The frequency synthesizer is implemented with a phase/frequency detector, charge pump, loop filter, VCO and a divider.

The reference frequency for the PLL was chosen as high as 125 MHz. This results in reference spurs 2x125 (frequency doubler) MHz away from the center frequency, thus outside the RF bandwidth for HiperLAN/2 (5.15 – 5.35 GHz). The PLL design can thereby be relaxed. Another advantage is that the bandwidth of the PLL can be increased; this can reduce the phase-noise of the VCO. The division ratio of the divider is also smaller with a high reference frequency. A reference frequency of 125 MHz gives a division ration of 16.

Designed circuits use differential structures to reduce the sensitivity to power supply variations and substrate noise.

2.1 Phase/Frequency Detector

The phase/frequency detector (PFD) used here is a modified pre-charged type [2]. It consists of 16 transistors and 18 inverters. The inverters are used to give the output pulse a certain pulse width by creating a delay in the signal path. This pulse width has to be larger than 400 ps, otherwise the input stage to the charge pump will not work properly. The pulse is then to fast for the transistors in the charge pump. All transistors in the PFD use the smallest possible width to limit the current in the circuit.

2.2 Charge pump

Most of the charge-pumps today are single-ended. This is due to design simplicity and low power consumption, but with the new submicron processes, differential solution has become more interesting. Designing a differential charge-pump is not at trivial matter, especially if it has to operate at a high frequency. The limitations in the current-mirrors and charge injections are some of the problems that occur. A circuit that can handle these problems is presented in Figure 2. It is a modified version of an earlier charge-pump [3].

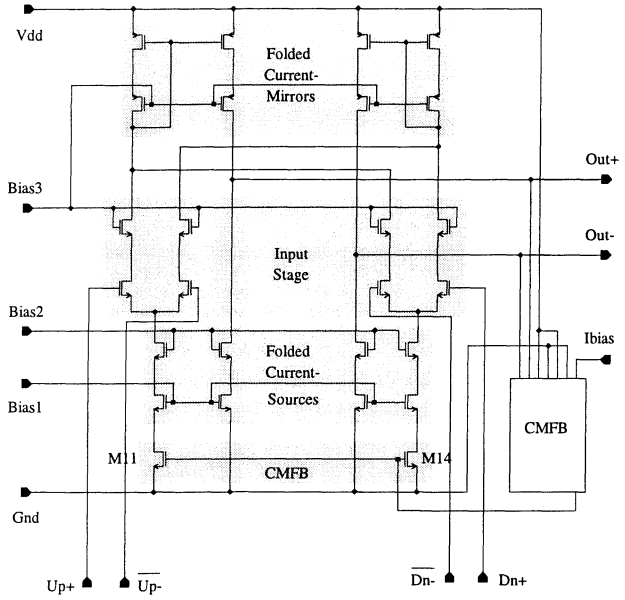


Figure 2: *Differential charge pump*

The current mirrors in Figure 2 never turn off, which is the biggest advantage with this circuit. Current mirrors in switching mode can only handle pulses with at least some nanoseconds in pulse width. The reason is that it takes time to open a channel in the transistor. By leaving the channel open, this time can be avoided. The current mirror can operate with pulses wider then 400 ps. This is necessary since the reference frequency is set to 125 MHz (8ns period time).

The fast switching input transistors introduce a problem. When the transistors turns off, the charge that is stored at the gate, is dissipated. This causes an increased current on the output and hence a change in the output voltage. This phenomenon is referred to as charge injection, causing a ripple at the outputs. Since all circuits are differential, this problem would normally be neglected. Unfortunately the VCO has an inferior common-mode rejection, resulting in reference spurs at the VCO output. These reference spurs occur 125 MHz away from the center frequency and therefore outside the intended frequency band, as discussed in the previous chapter. The reference spurs should be as small as possible, to avoid intermodulation products in the VCO. The use of small switching transistors can reduce this problem.

The current-mirror and current-sources need to be biased so all the transistors are in the active region. The bias circuit achieves this, but it requires a minimum voltage drop over the transistors of 0.4 V. This limits the output voltage range to 0.4-1.4 V. To maximize this voltage range and to ensure that the transistors are in active region, a common-mode feedback circuit (CMFB) has been added. This circuit controls the bottom current-mirror, M11 and M14, of the cascade, thereby controlling the common-mode voltage.

2.3 Loop filter

The loop-filter is an essential part of the PLL since it determines the stability of the loop.
The loop-filter that was used is of second order, i.e. the PLL is of third order. The resistance, R, was divided in two R/2 resistances each connected to one terminal of C_2 as to create a symmetrical topology and layout for the differential low-pass filter, shown in Figure 3. The optimum values for the components with a charge pump current of 30 μA was found to be $C_2 = 2.2$ pf, $C_2 = 25$ pf and $R = 29$ kΩ.

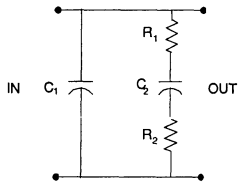


Figure 3: *Loop filter*

2.4 Divider

The frequency divider is one of the digital parts of the PLL. The divider consists of four divide by two circuits in series, thereby creating a division ratio of 16. Each divide-by-two circuit consists of two pseudo-NMOS D-flip-flop whose outputs are connected back to its input, as seen in Figure 4.

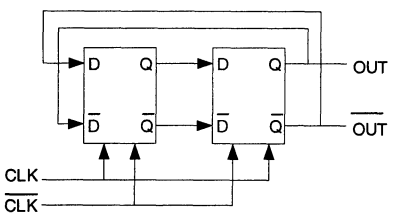


Figure 4: *Divide-by-two circuit*

3 VCO

The VCO is based on two differential LC-oscillators coupled in a ring structure, shown in Figure 4.

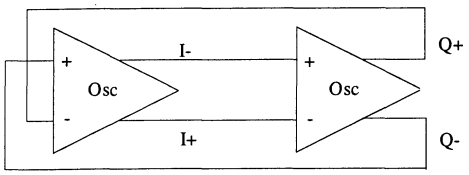


Figure 4: *Oscillators coupled in ring structure*

This structure offers quadrature outputs, in-phase (I) and quadrature (Q), without the need for a poly phase filter.

A buffer with a rail-to-rail swing is connected to the outputs of the VCO. This buffer has two purposes, to increase the swing for the divider and also to provide isolation from the mixer. The output of the buffer will be connected to a frequency doubler to gain extra isolation between the VCO and the mixer. This frequency doubler also relaxes the VCO constraints by reducing the oscillation frequency of the VCO from 4 GHz to 2 GHz.

3.1 Cross-coupled LC oscillator

Each oscillator consists of complementary cross-coupled transistors, equipped with an LC-tank, shown in Figure 5. This type of oscillator was chosen due to its differential operation, large output swing and low phase-noise for a given current. The low phase-noise is mostly due to the LC-tank, but also the complementary structure improves the performance [4]. The connection of the two oscillators in a ring configuration requires the inputs to be separated from the outputs; this was possible using a driver-stage, consisting of an NMOS and a PMOS transistor on top of each other, similar to an inverter.

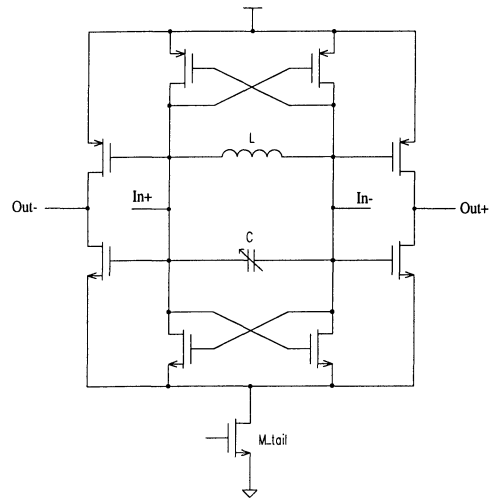


Figure 5: *One oscillator including two driver stages*

3.2 LC tank

The frequency is controlled by the LC-tank, having the well-known equation

$$\omega_0 = \frac{1}{\sqrt{LC}} \qquad (1)$$

The capacitance is tunable, thereby controlling the frequency of the oscillation. The tunable capacitor, called a varactor, consists of two N-MOS and two P-MOS transistors, shown in Figure 6.

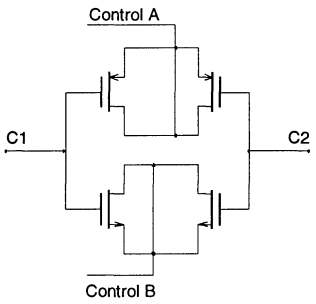


Figure 6: *MOS varactors*

The varactors capacitance is tuned by controlling the drain/source voltage, the use of both P-MOS and N-MOS transistors as varactors enables a differential tuning.

As the oscillator has a very large voltage swing, the capacitance of the varactors will be averaged over a period. This results in distortion but the high Q of the LC-tank will effectively attenuate these harmonics.

For maximum varactor Q the shortest gate length should be used. This is due to the losses in the resistive channel of the MOS-transistor. However as the gate length is reduced, the minimum achievable capacitance will increase [5], thereby reducing the tuning range.

3.3 Output Buffer

The buffer consists of two identical circuits; each one has a differential input and a single-ended output. With the use of two identical circuits and by shifting the polarity of the other input pair, a differential output is achieved, as shown in Figure 7.

With this buffer it is possible to achieve a rail-to-rail voltage swing. This is due to the output stage, a push-pull configuration consisting of M2 and M4. Transistor M1 controls the current through the current mirror (M3 M4) and thus transistor M4 follows the input signal, In+.

The advantage of this output stage compared to an ordinary push-pull circuit found in an inverter, is that the current glitches will be reduced. In an ordinary push-pull circuit, both the gates are connected to each other. When the gate-voltage shortly reaches half the supply voltage, both the PMOS and the NMOS transistors will have a conducting channel, causing a current glitch.

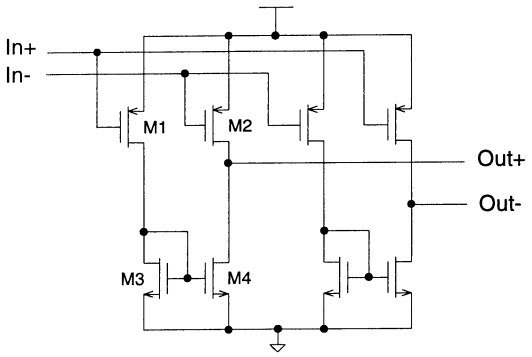


Figure 7: *Output buffers*

To reduce the power consumption and to maximize the tuning range, the loading capacitance had to be reduced. This was achieved by using the shortest gate lengths on all switching MOS-transistors, 0.18µm.

The widths of the MOS-transistors in the cross coupling were designed to obtain maximum amplitude and symmetry for minimum current. Due to different mobility between NMOS and PMOS transistors, the PMOS-transistors have to be 3 times wider than the NMOS-transistors to ensure equal rise and fall times. In this case, the high Q of the inductor resulted in equal rise and fall times, without the need for a 3 to 1 ratio in width. This symmetry is very important, since it affects the Impulse Sensitivity Function (ISF) of the oscillator [6]. If the rise and fall times are made equal then the DC component of the ISF will be minimal, thus reducing the phase noise and lowering the $1/f^3$ noise corner.

4 Results

The circuit has been simulated in Cadence design environment. All simulations were done using transient analysis due the large signal amplitudes. A summary of the simulation results can be seen in Table 1.

The phase noise was found to be –133 dBc/Hz at 2 MHz offset, using a Periodic Steady State (PSS) simulation. Figure 8 shows the phase noise at different offset frequencies.

A wide VCO tuning range is a desirable feature to compensate for process variation. The tuning range for the designed VCO was found to be 1.6 – 2.22 GHz, shown in Figure 9

The reference spurs were simulated using a FFT on the PLL output. The resulting frequency spectrum is shown in Figure 10. From the figure it can be seen that the reference spurs appears 125 MHz away from the center frequency and 50 dB below the carrier. The locking behavior of the PLL can be seen in Figure 11.

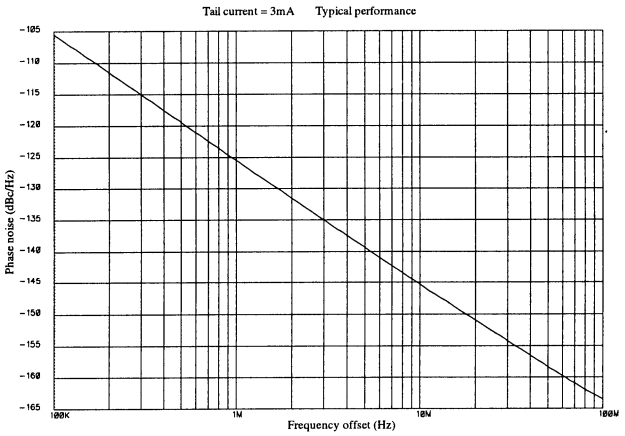


Figure 8*: Phase noise for the VCO*

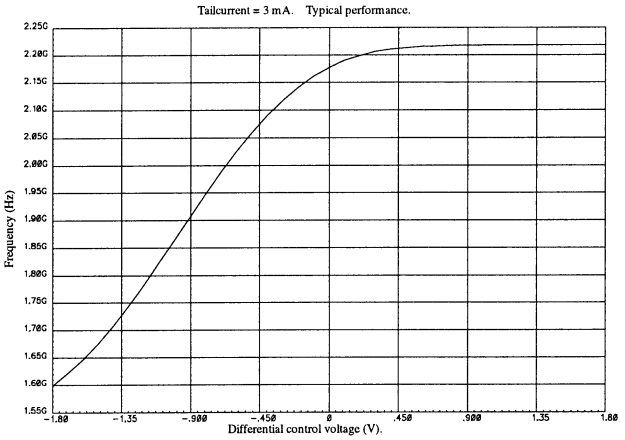


Figure 9*: Tuning range of the VCO*

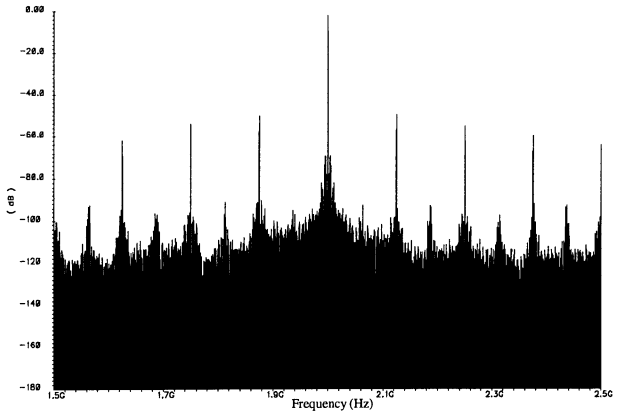


Figure 10*: Output spectrum.*

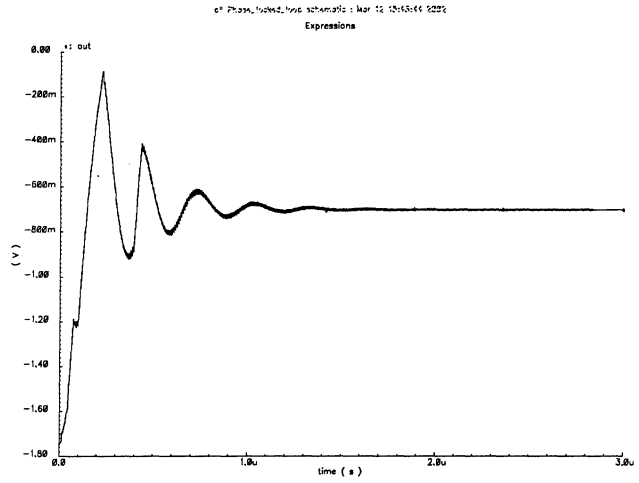


Figure 11: *PLL locking behavior*

Table 1: *Summary of technical data*

Parameter	Technical data
Reference frequency	125 MHz
Reference spurs	-50 dBc
VCO *Tailcurrent = 3 mA* *Center frequency = 2GHz*	
Phase noise @ 600 kHz	-121 dBc/Hz
Phase noise @ 1.875 kHz	-132 dBc/Hz
Phase noise @ 20 MHz	-151 dBc/Hz
Tuning range	1.60 – 2.22 GHz
Power consumption	
PFD	3.20 mW
Charge-pump	21.5 mW
Bias circuit	3.58 mW
Divider	35.1 mW
VCO incl buffers	22.5 mW
VCO excl buffers	11 mW
Total incl buffers	**85.9 mW**

5 Conclusion

The design and simulation of a fully differential frequency synthesizer hase been presented. The complete PLL consumes 86 mW, with a simulated phase noise of –133 dBc at 2 MHz offset from a 2.0 GHz center frequency.

6 References

[1] J. Khun-Jush, P. Schramm, U. Wachsmann, F. Wenger. Structure and Performance of the HIPERLAN/2 Physical layer, http://WWW.ericsson.com/Wlan/te-h2-publ.asp

[2] G.B Lee, P.K. Chan and L.Siek. A CMOS phase frequency detector for charge pump phase-locked loop, Circuits and Systems, 2000. 42nd Midwest Symposium on , Volume: 2 , 2000 Page(s): 601 - 604 vol. 2

[3] Woogeun Rhee. Design of High-Performance CMOS charge pumps in Phase-Locked Loops, Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium on , Volume: 2 , 1999 Page(s): 545 - 548 vol.2

[4] Ali Hajimiri and Thomas H. Lee. Design issues in CMOS differential LC oscillators, IEEE J. Solid-State Circuits 34, pp. 717-724, 5 May 1999.

[5] Alain-Serge Porret, Thierry Melly, Christian C. Enz and Eric A. Vittoz. Design of high-Q varactors for low-power wireless applications using a standard CMOS process, IEEE J. Solid-State Circuits 35, pp. 337- 345, 3 March 2000

[6] Ali Hajimiri and Thomas H. Lee. A General Theory of Phase Noise in Electrical Oscillators," IEEE J. Solid-State Circuits 33, pp. 179-193, 2 Feb 1998.

A Highly Integrated Single-Stage RF Down-Converter For Future Adaptive X-band Radar Systems

Robert Malmqvist[1], Andreas Gustafsson[2], Mattias Alfredsson[3] and Aziz Ouacha[4]

[1]FOI Swedish Defence Research Agency, P.O. Box 1165, 581 11 Linköping, Sweden. E-mail: rma@foi..se.
[2]FOI Swedish Defence Research Agency, P.O. Box 1165, 581 11 Linköping, Sweden. E-mail: andreasg@foi.se.
[3]FOI Swedish Defence Research Agency, P.O. Box 1165, 581 11 Linköping, Sweden. E-mail: matal@foi.se.
[4]FOI Swedish Defence Research Agency, P.O. Box 1165, 581 11 Linköping, Sweden. E-mail: aziz@foi.se.

Abstract

An 8-10GHz two-chip single-stage RF down-converter intended for the microwave receiver modules of a digital beamforming X-band radar array antenna is presented. The use of such an agile and compact receiver front-end could potentially enable a significant reduction of the receiver size and complexity in future adaptive X-band radar systems. Simulated results together with estimations show a promising RF front-end performance.

1 Introduction

A low vulnerability to jamming signals due to electronic warfare or electromagnetic interference, for example, is of prime importance in modern radar systems. A reduction of size and cost of the microwave receiver modules in future adaptive radar array antennas is also becoming increasingly important [1]. Despite this demand, such receiver modules are traditionally being fabricated using hybrid circuit technologies that are neither compact nor very suitable for low-cost production. Thus, in order to be able to realize much more compact and cost-effective multi-channel adaptive radar systems in the future, low-cost integration methods that can be used to reduce the receiver module size and complexity have to be developed. As a consequence of this, increased interest has recently been focused on the possibility of using frequency tunable high gain and low noise active monolithic microwave integrated circuit (MMIC) filters to reduce the vulnerable bandwidth of frequency hopping radar systems [2]. If such a MMIC filter is combined with an MMIC image rejection mixer a radar frequency (RF) front-end that combines the attractive feature of small size at low cost with a good ability to reject unwanted signals could potentially be realized. Such a front-end could possibly also reduce the number of down-converting stages required in an agile radar receiver by allowing a greater down-conversion step to be made. Rejection at the image frequency ($f_{image}=f_{RF}-2f_{IF}$, where f_{IF} denotes the down-converted intermediate frequency) should however be high enough to avoid the effect of interfering signals that may be present at this frequency. That is, the image rejection of the front-end should be in the same order as the required spurious-free dynamic range (*SFDR*) [1]. A block diagram of a single-stage RF front-end of the type described above is shown in Fig. 1.

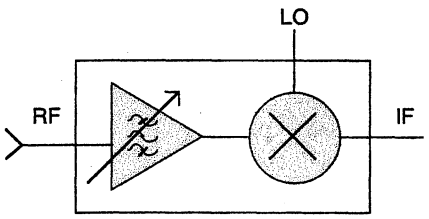


Fig. 1. Block diagram of a single-stage RF down-converter composed of a frequency tunable high gain and low noise active MMIC filter and an image rejection MMIC mixer.

Typical RF front-end requirements in future adaptive radar systems are summarized in Table I. The specification is based on our experience in the area and on assumed radar system requirements. Note that an *SFDR* of 109dB/Hz$^{2/3}$ corresponds to an *SFDR* of 60dB when a noise bandwidth (B) of 20 MHz is assumed.

TABLE I
TYPICAL RF FRONT-END REQUIREMENTS IN FUTURE ADAPTIVE RADAR SYSTEMS

Agile frequency range	≥20-40%
Conversion gain (G)	≥10dB
Noise figure (*NF*)	≤5 dB
Input third order intercept point (IIP_3)	≥-6 dBm
Spurious-free dynamic range (*SFDR*)	≥109dB/Hz$^{2/3}$

To our knowledge, relatively few papers have been published on the realization of agile on-chip receiver front-ends at X-band (8-12GHz) or above. Zelley [3] presented a 12-18GHz single-chip transceiver with 20dB of image rejection. An 8.7-9.8GHz two-chip (tunable active filter and mixer) front-end with an IF of 360MHz was presented in [4]. The maximum possible image rejection and conversion gain of the filter/mixer-combination were estimated at 46dB and –3dB respectively. One purpose of the presented work in this

paper is to investigate if it is possible to improve these results by replacing the filter used in [4] with a recently developed 7.9-9.7GHz tunable active MMIC filter [5]. The filter and the mixer and the combined two-chip filter/mixer front-end are shortly described in section 2. In section 3 simulated results of the front-end are compared with estimations of the front-end performance based on corresponding measured filter and mixer data. Finally, possibilities and difficulties of realizing an agile single-stage X-band radar receiver front-end with adequate performance are shortly discussed in section 4.

2 Single-stage RF down-converter

In this paper, we propose a two-chip single-stage RF down-converter intended for the microwave receiver modules of future adaptive X-band radar array antennas. The down-converter consists of two previously developed X-band MMIC's: a tunable active filter [5] and a 360MHz-IF image rejection mixer [4]. The two MMIC's have been fabricated using a 0.2μm GaAs/AlGaAs PHEMT based MMIC process.

2.1 Tunable active filter

Recursive active MMIC filters are promising for narrow band and low noise applications since high-Q filters of this type can be designed with high gain in combination with a noise figure approaching that of the low noise amplifier (LNA) used in the filters [6]. A frequency tunable active MMIC filter with relatively high gain and out-of-band rejection together with noise and large signal properties that are close to adequate for the application in mind is presented in [5]. This filter implementation consists of two second order recursive active filters placed between two quadrature couplers in a balanced configuration. The filter center frequency tuning is implemented using self-switched three-bit time shifters that enable eight discrete tuning states [7]. A block diagram of the filter topology studied and a photo of the MMIC filter are shown in Figs. 2 and 3 respectively.

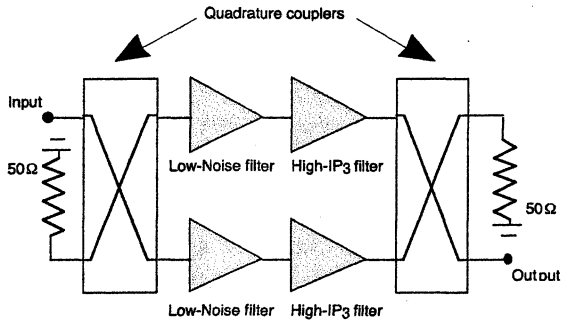


Fig. 2. A balanced filter topology composed of two second order filters placed between two quadrature couplers. The two second order filters consist in turn of two cascaded tunable recursive active filters (a *Low-Noise filter* in and a *High-IP3 filter*).

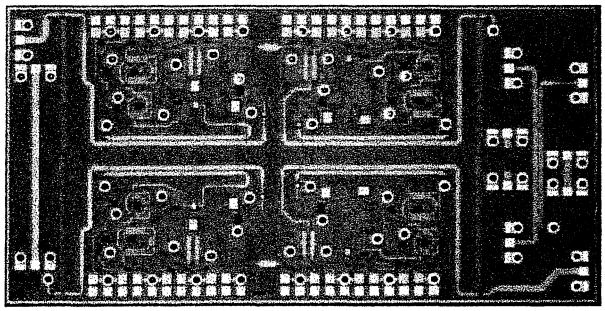


Fig. 3. Photo of the tunable active X-band MMIC filter and some break out circuits (chip dimensions: 3mm x 6mm).

2.2 Image rejection mixer

In the proposed single-stage front-end solution a mixer with a relatively high internal image rejection is typically needed. The mixer should also have high isolation between the RF and LO ports. To meet these demands a double balanced mixer where the mixing stages are implemented using dual-gate field effect transistors (FET's) could be used. To achieve high values of image rejection implies however tough requirements on phase and amplitude balance of such a mixer, something that can be rather difficult to maintain over a large bandwidth. Therefore this type of mixer normally has two ways of operating, fixed or tuned DC bias. With fixed bias we mean that the mixer DC biasing conditions are kept constant for all frequencies. Tuned bias means on the other hand that the DC voltages applied to the dual-gate FET mixing stages are varied in order to obtain the best possible phase and amplitude balance at each frequency of interest. The MMIC image rejection mixer used in our two-chip front-end is the double balanced 360MHz-IF mixer design presented in [4]. A photo of this MMIC mixer is shown in Fig. 4.

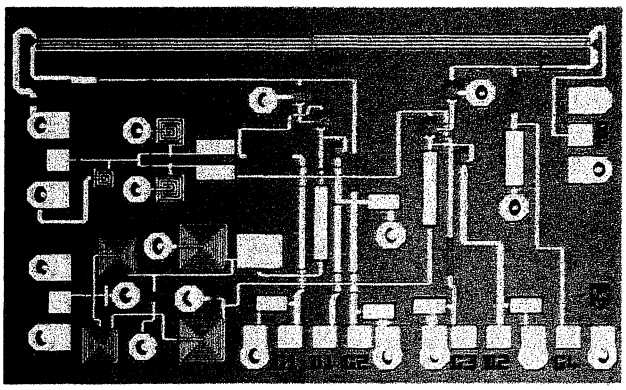


Fig. 4. Photo of the double balanced X-band MMIC image rejection mixer (chip dimensions: 1.6mm x 2.5mm).

2.3 Two-chip filter/mixer front-end

With reference to the filter and the mixer described above, we have realized a two-chip 360MHz-IF X-band front-end by simply cascading these two MMIC's using

a bonding-wire. In the two circuit designs all circuitry needed for impedance matching and transistor biasing is included on-chip. However, in order to be able to experimentally evaluate the front-end it has to be mounted in a test fixture. All DC bias pads are therefore also connected using bonding-wires. Figure 5 shows the two-chip filter/mixer front-end mounted in a test fixture.

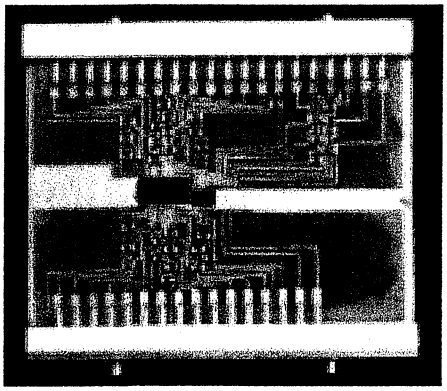


Fig. 5. Photo of the combined two-chip filter/mixer single-stage X-band front-end mounted in a test fixture.

3 Results

So far, we have not been able to experimentally evaluate the two-chip filter/mixer front-end. Typical simulated data of the front-end over the agile frequency band are shown in Table II. All simulated results have been obtained using *HP EEsof* CAD-tool Libra v.6.0 together with foundry-provided circuit component libraries. As a comparison, the corresponding estimated front-end performance based on measured data obtained from break out circuits that contain the filter and the mixer used in the front-end are also presented in this table. Measured results (see Table II) show that the filter is discretely tunable between 7.9-9.7GHz (corresponding to a relative agile bandwidth of 20%). However, according to simulations the active filter is tunable between 7.8-9.5GHz. For this reason, we have chosen to simulate the front-end over the same frequency band. Figure 6 shows typical simulated values of front-end image rejection (which in this case means the relative suppression of signals at 2IF below RF compared with signals at RF) in the case of fixed mixer bias at all eight possible filter center frequency tuning states. The filter bias has been tuned to achieve maximum possible filter gain and out-of-band rejection. The filter and mixer DC currents are in this case 230mA and 40mA from 3V and 6V DC supply, respectively. These values correspond to a front-end DC power dissipation of 930mW.

As shown in Table II the expected value of front-end image rejection (based on measured filter and mixer data) is 47-49dB. Figure 6 shows, on the other hand, that the corresponding simulated values of front-end image rejection in the case of fixed mixer bias typically varies between 27-45dB over the agile bandwidth. The discrepancy between simulated and estimated values of front-end image rejection can partly be explained by the fact that the amount of impedance mismatch between the filter and the mixer varies with frequency. This can however be compensated for by re-tuning the mixer DC bias. In principle, it should be possible to achieve the highest possible value of front-end image rejection over the whole agile bandwidth if we allow the mixer DC bias conditions to be re-tuned for each filter tuning state. It also should be noted that in an experimental measurement set-up an extra degree of freedom exists since the DC bias can be varied in a continuous way by simply turning the knobs of the different voltage sources that are connected to the device under test. This is on the other hand not the case when simulating the front-end in a microwave CAD simulator that rather requires a single set of bias points to be used for each circuit simulation. Simulating the front-end can thus be relatively time-consuming and therefore becomes impractical especially when considering several discrete filter center frequency tuning states over a large agile bandwidth. Furthermore, re-tuning the mixer bias also can be rather difficult and impractical when using the front-end in a realistic system. This is however an area of on-going research. The use of an analog self-calibration technique based on automatic phase and amplitude adjustments of image rejection mixers was recently proposed in order to improve the maximum obtainable values of image rejection for a 900MHz RF front-end [8].

The simulated values of front-end conversion gain in the case of fixed mixer bias varies between 5-14dB over the agile frequency band (see Fig. 6.) Compared with estimations based on measured filter and mixer data these values are about 7-11dB higher (see Table II). This discrepancy can however in turn to a large extent be explained by a corresponding deviation between the measured and simulated values of filter gain [5].

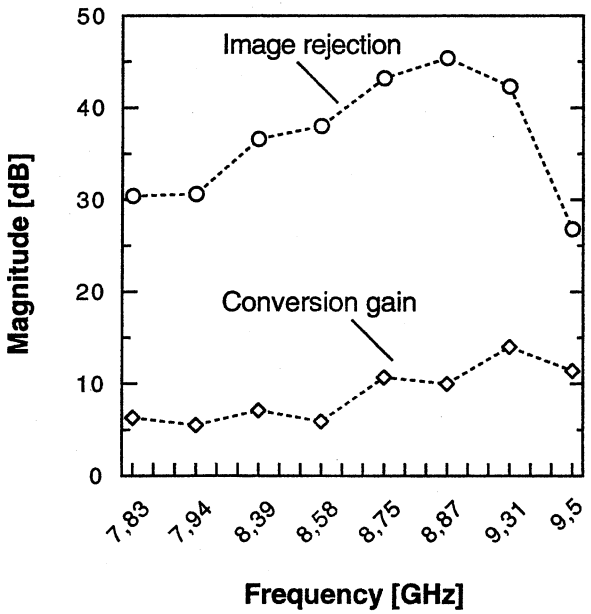


Fig. 6. Simulated image rejection and conversion gain for the front-end in the case of fixed mixer bias at each filter center frequency tuning state.

TABLE II
SUMMARY OF RESULTS

Results	Tuning range	Image Rejection [dB]	Conversion Gain [dB]	NF [dBm]	IIP_3 [dBm]	SFDR* [dB] *B=20MHz
Filter [5] (measured)	7.9-9.7GHz	7-9	11-16	6	≈0	≈63
Mixer (measured)	8-12GHz	40	-13	13	+7	63
Front-end (estimated†)	7.9-9.7GHz	47-49	-2 to 3	7(†)	-10 to -5(†)	56-59(†)
Front-end (simulated)	7.8-9.5GHz	27-45	5-14	Not Available	Not Available	Not Available

(†) *Estimated results based on the measured filter and mixer data.*

At this point, we have not been able to successfully simulate the noise and large signal properties of the front-end. However, based on measured data of the filter and of the mixer the noise figure of the filter/mixer front-end is estimated at 7dB (see Table II). Table II further reveals that the estimated value of front-end IIP_3 varies between −10 to −5dBm over the agile bandwidth. These estimated results correspond to 56-59dB of front-end *SFDR* when a noise bandwidth of 20MHz is assumed.

4 On-chip single-stage RF front-end for future adaptive radar systems

In this section we will shortly discuss possibilities and difficulties of realizing an agile single-stage X-band image rejection receiver front-end compared with if we use a front-end that is based on two or more down-converting stages. An image rejection receiver enables cancellation of signals that are present at the image frequency (which lies 2IF away from the RF). For a given value of the RF, a certain amount of image rejection is more easily achieved using a higher value of the IF. Several down-converting stages are normally required in the front-end when a high IF is used. On the other hand, choosing a lower value of the IF could enable a reduction of the number of down-converting stages needed in the front-end.

A 7.9-9.7GHz single-chip filter/mixer front-end with an IF of 1GHz was previously developed [9]. In this front-end the tunable active filter also used in our presented 360MHz-IF front-end was combined on the same die as a 1GHz-IF image rejection mixer. The measured maximum possible values of front-end image rejection and conversion gain over the agile frequency band were 60dB and 0dB respectively. In the receivers of a digital beamforming antenna AD-converters with 10-14 bits are normally required [2]. The relatively high IF of 1GHz for the front-end in [9] implies that a second down-converting stage will be needed since an IF of that order can be considered too high for today's standard ADC's when such a high number of bits are required.

Compared with a two-stage RF down-converter solution a more simplified front-end architecture could be to realize an on-chip single-stage down-converter composed of a tunable filter and a mixer with an IF that is sufficiently low for 10-14 bits standard ADC's. It is believed than an IF in the order of 300-400MHz could be considered low enough for today's under-sampling ADC's with 10-14 bits. A comparison of the estimated front-end performance presented in Table II with the typical requirements given in Table I indicates that the noise and large signal properties of the evaluated 360MHz-IF single-stage front-end is not very far from being adequate for the application in mind. In principle, it should be possible to fulfil the noise and large signal requirements given in Table I by careful re-optimization of the filter and mixer designs. Compared with the 1GHz-IF single-chip front-end in [9], however, the 360MHz-IF front-end presented in this paper is expected to have somewhat lower values of image rejection (below 50dB). As mentioned previously, the total image rejection of a radar receiver should be in the same order as the required value of *SFDR* (~60dB when B=20MHz). To be able to achieve an image rejection of that order (at an IF of 360MHz) over a large agile bandwidth probably requires the use of a mixer image rejection self-calibration technique as described in e.g. [8]. If this method also can be adopted for wide-band X-band MMIC image rejection mixers with sufficiently low values of the IF (i.e. in the order of 300-400MHz) single-stage on-chip X-band radar front-ends with adequate performance could potentially be realized. Compared with conventional microwave receivers that are fabricated using hybrid circuit components, such front-ends could enable a significant reduction of the microwave receiver size and complexity in adaptive X-band radar systems.

5 Conclusion

A 7.9-9.7GHz two-chip single-stage receiver front-end composed of a tunable active MMIC filter and a 360MHz-IF image rejection MMIC mixer has been

presented. Simulated results show that this front-end can achieve almost 50dB of image rejection. To achieve even higher values of image rejection over the whole agile receiver bandwidth, however, some kind of automatic mixer image rejection self-calibration technique probably has to be implemented. Estimations based on measured data of the filter and of the mixer imply that the evaluated front-end can achieve a noise and large signal performance that is close to what should be adequate for the microwave receiver modules of adaptive X-band radar systems. The results presented seem promising for the possibility that agile on-chip single-stage RF front-ends with sufficiently low values of the IF potentially could be used to enable a significant reduction of the microwave receiver size and complexity in such future advanced X-band radar applications, for example.

Acknowledgement

This work was financially supported by *The Swedish Armed Forces (FM)* and *The Swedish Foundation for Strategic Research (SSF)*.

References

[1] L. Pettersson, M. Danestig, and U. Sjöström, "An experimental S-band digital beanforming antenna," *IEEE AES Syst. Mag.*, vol. 12, no. 11, pp. 19-26, Nov. 1997.

[2] A. Gustafsson, M. Danestig, R. Malmqvist, and A. Ouacha, "Tuneable S-band filter for on-chip receiver," *Proc. of 1998 Asia-Pacific Microw. Conf.*, pp. 781-784, Dec. 1998.

[3] C. A. Zelley, P. A. Gould, P. D. Mundy, and R. W. Ashcroft, "A J-band transceiver MMIC with image rejection," *Proc. of 30th Europ. Microw. Conf.*, pp. 417-420, Oct. 2000.

[4] A. Gustafsson, M. Alfredsson, M. Danestig, R. Malmqvist, and A. Ouacha, "A fully integrated radar receiver front-end including an active tunable band pass filter and image rejection mixer," *Proc. of 2000 Asia-Pacific Microw. Conf.*, pp. 99-102, Dec. 2000.

[5] R. Malmqvist, A. Gustafsson, M. Alfredsson, and A. Ouacha, "A tunable active MMIC filter for on-chip X-band radar receiver front-ends," to be published in *2002 IEEE MTT-S Int. Microw. Symp. Dig.*, June 2002.

[6] M. Danestig, H. Johansson, A. Ouacha, and S. Rudner, "Low-noise active recursive MMIC filters," *1997 IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 705-708, June 1997.

[7] M. Danestig, A. Ouacha, C. L. Kien Tsoen, T. Tieman, and S. Rudner, "Recursive filters employing transmission type phase shifters and novel self-switched time shifters for frequency tuning, " *Proc. of 28th Europ. Microw. Conf.*, pp. 352-357, Oct. 1998.

[8] R. Montemayor, and B. Razavi, "A self-calibrating 900-MHz CMOS image-reject receiver," *Proc. of 26th European Solid-State Circuits Conference*, Sept., 2000.

[9] R. Malmqvist, M. Alfredsson, A. Gustafsson, and A. Ouacha, "A 7.9-9.7GHz on-chip radar receiver front-end for future adaptive X-band smart skin array antennas," to be published in *2002 IEEE MTT-S Int. Microw. Symp. Dig.*, June 2002.

A Modular Asynchronous Wrapper

Jonas Carlsson[1], Weidong Li[1], Tormod Njølstad[2], Kent Palmkvist[1], Lars Wanhammar[1], and Shengxian Zhuang[1]
[1]Department of Electrical Engineering, Linköping University
SE-581 83 Linköping, Sweden
{jonasc, weidongl, kentp, larsw, zhuangsx}@isy.liu.se
[2]Norwegian University of Science and Technology (NTNU), Trondheim, Norway

ABSTRACT

In this paper we discuss some aspects of designing low power asynchronous wrappers for use in GALS based communication systems. We propose a general method for designing asynchronous wrappers for use in asynchronous multiport-to-multiport communication channels with low swing busses and variable power supply voltages.

1. INTRODUCTION

Design of communication systems-on-Chip (*SoC*) using deep submicron CMOS processes [1][2][3][4][5] poses some new problems besides the increased system complexity. One major problem comes from the fact that future chips will be larger and operate at very high clock frequencies. Hence it is difficult to use a global clock. Problems due to large power consumption and clock skew will also become significant [6]. Another problem is the increasing interconnection RC delay due to the scaling of interconnections.

1.1 COMMUNICATION SYSTEMS

Most communication and signal processing systems have a data-flow structure that easily can be mapped to a corresponding hardware structure. We advocate that this structure should be communication centered with data-driven, fixed- or semi-fixed function units. For example a video encoder for CCITT H.261 standard as shown in Figure 1 has a typical data-flow structure.

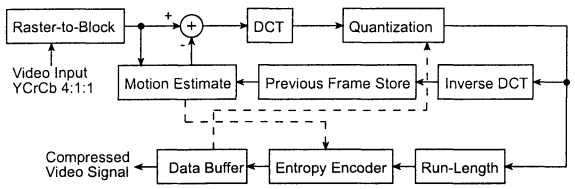


Figure 1. Functional view of CCITT H.261 video encoder

2. GALS

The communication-based design approach not only simplifies the system design phase, but also simplifies the mapping to a heterogeneous software/hardware (S/H) structure that may contain both general processors as well as optimized fixed-function and weakly programmable hardware units (modules). We advocate that the communication between the units should be data-driven (asynchronous handshaking protocol) while the units use local, synchronous clocking schemes internally, i.e., a GALS (Global Asynchronous Local Synchronous) based approach. The GALS based approach matches the data-driven system architecture well since it consists of local, data-driven functional units that operate with a synchronous clock and communicate using an asynchronous communication protocol.

The GALS based approach for building large systems-on-chip using deep submicron CMOS technology is getting increased attention due to several factors. It partitions the system into independent parts, which allow a short time to market and encourages the designer to reuse old designs and to use IP blocks.

The use of a robust asynchronous handshaking protocol to communicate between the units has many potential advantages besides the obvious reduction in power consumption in the clock network and idling of non-busy units, such as encapsulating different clock domains and subsystems and isolate the internal details of the units.

The use of old units may introduce problems as they have not been designed to communicate, i.e., problems involving different clock rates, various timing aspects, and control signals. The current design paradigm is using a single clock domain or a few clock domains. Whenever there is a need to communicate between different clock domains the designer must be very careful and make a thorough test of the interface. Similar problems also arise when design use gated clocks to reduce the power consumption. With the increasing system complexity, validation and testing will become a major part of the design. A more generalized form of communication should therefore be introduced early in the design process, that does not rely on the use of a common clock to indicate valid timing signals.

Since the synchronous modules are isolated from all other modules they can use their own clock and power supply voltage. This also enables the possibility to vary clock frequency and power supply at runtime on one synchronous module without affecting any other synchronous module.

2.1 COMMUNICATION CHANNELS

Communication between the synchronous clocked units is handled by the asynchronous wrapper that encapsulate the internal properties of a unit and provide one or several asynchronous communication channels.

In the next section we discuss the design of a module-based wrapper that can be use in a large variety of cases.

2.2 ASYNCHRONOUS WRAPPERS

The asynchronous wrapper can be realized in many different ways. Liljeberg et al. [5] use a synchronous/asynchronous interface unit (SAInt unit) containing dual-port RAM as a FIFO queue. This structure is targeted at ULSI but can contain significant overhead for smaller applications. Muttersbach et al. [2] uses hazard-free two-level AND-OR circuit for handshake, which also was implemented on chip. Bormann and Cheung [4] uses, as Muttersbach [2], a modular wrapper but only with different ports as modules. Njølstad et al. [1] describes a wrapper that also supports locally rate-adapted dynamic voltage scaling for the module.

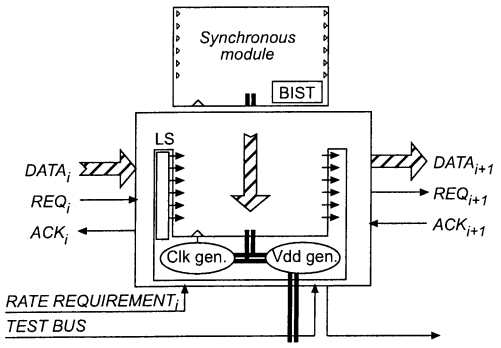


Figure 2. Asynchronous wrapper and synchronous module.

Figure 2 shows an asynchronous wrapper and a synchronous module. Through $RATE\ REQUIREMENT_i$ the local power supply voltage can be scaled down so that the synchronous module process data at the rate data arrive. By scaling the power supply voltage the speed will also be scaled, i.e., the speed of the circuit $f(V_{DD})$ depends on power supply voltage. The synchronous module can be generated by a state-of-the-art synthesis tool and inserted into the wrapper and the wrapper will then convert from/to the voltage outside the wrapper to/from the right voltage for the synchronous module.

From the designers point of view, a simple and structured way to introduce the asynchronous wrapper into the system architecture is needed. The asynchronous interface should be modular to allow for simple control of its features, modularity also allows the designer to choose what modules to include in the wrapper.

3. POINT-TO-POINT COMMUNICATION

3.1 HANDSHAKE PROTOCOLS

Sending data from one module to another module in the GALS approach will not use a common clock to define when data is stable. Instead are handshake signals used to define when data is available and when data has arrived. This is also known as bundled data. The two most common asynchronous handshake protocols are the two- and four-phase protocols. Both protocols uses a request and an acknowledge signal as shown in Figure 3.

In the case of a two-phase protocol the transmitting module is indicating that data is available by a transition on the request signal. The receiver then uses the data and when finished sends a transition on the acknowledge signal. This completes the transmission, and the next data exchange will start with a new transition on the request signal. This protocol is also known as transition or non-return-to-zero (*NRZ*) signaling.

In the case of a four-phase protocol the transmitting module is indicating the start of a data transfer by a transition on the request signal. The receiving module acknowledges this by a transition on the acknowledge signal. The transmitting module then performs another transition on the request signal restoring the value to its initial value. The receiving module restores the acknowledge signal after the request signal is restored. This is also known as level or return-to-zero (*RTZ*) signaling. The data valid period can be divided into three schemes [7] and they look different depending on which side that is transferring the data. In Figure 3 the port that is asserting *Req* is also sending data.

Apparently the four-phase protocol is less efficient in terms of signal transitions on the bus control signals. It is however easier to implement the control circuitry, as the request or acknowledge signal can be used immediately to control input latches in the receiver.

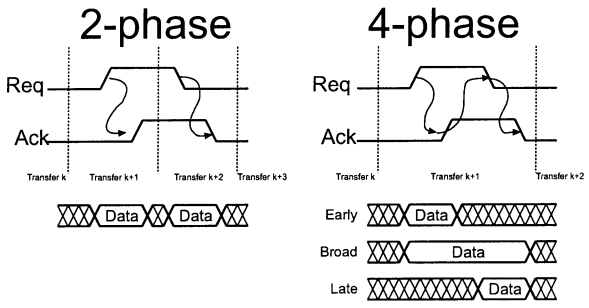


Figure 3. Two-phase and four-phase protocols.

From Figure 3 we notice that a combination of the data schemes early and late will generate the same data speed as two-phase. A drawback would then be that two data are always transferred even if only one data is in the queue, but then again this could be used to half the bus size and send one data in two packets.

3.2 HANDSHAKE PORTS

Point-to-point communication can be divided into several types. When looking at how the local clock is handled there is two possible types, demand or poll, for the handshake the port could be either active or passive and for the data the port is output or input. Demand stops the internal clock as soon as a the synchronous part tries to send/ receive data, poll port only stops the internal clock under the handshake process if necessary. An active port starts the handshake (i.e., drives the request signal) and the passive part ends the handshake.

The signal transfer graph (STG) in Figure 4 describes a demand active port with *En* and *Ack* as input and *Req* and *Str* as output, our realization is shown in Figure 5. A similar STG can be draw for a demand passive port which realization is shown in Figure 6. (*Reset* signal is added to set the port to a known starting state)

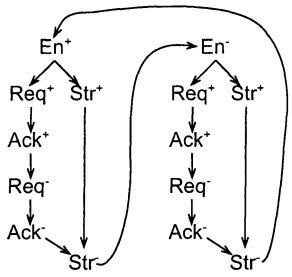


Figure 4. STG for a demand active port.

The realization in Figure 5 and Figure 6 do not wait for the clock to go low as in [2], thus the handshake can start earlier.

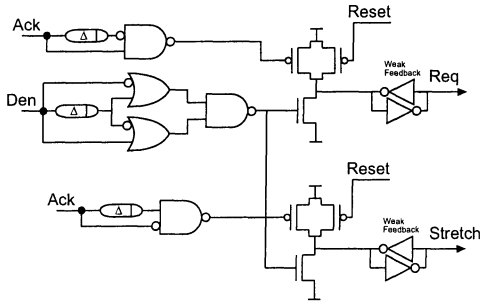


Figure 5. Demand active port.

This gives the port the advantage to start the handshake early and use the whole clock cycle instead of only 50% of a clock cycle. Section 7.1 describes a stretchable clock controller.

Figure 5 and 6 looks almost the same so the following description is for Figure 6. The handshake starts by a transition (i.e., either a positive or negative transition) on *Den* which will generate a positive puls on *Xa*. This pulse sets *Stretch* to high and when *Xa* goes low again the feedback will hold *Stretch* high as log as necessary. When *Stretch* is high, the AND gate allows *Ack* to follow *Req*. The handshaking are now enabled and when the handshake ends, i.e., *Ack* goes low, it will generate an pulse on *Xb* which will reset *Stretch* back to zero. *Ack* will stay low until a new handshake is enabled by a transition on *Den*.

The handshake ports above are implemented with 0.35µm standard CMOS technology and simulated with Spectre™ in Cadence™. The simulation is based on schematic.

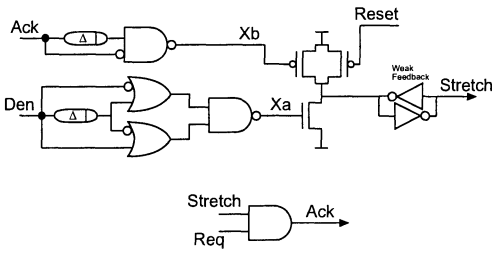


Figure 6. Demand passive port.

Table 1. Simulation results for one handshake.

Signals	Time
$DenA^+$ -> $StrA^-$	1.41 ns
$DenP^+$ -> $StrP^-$	1.42 ns
Req^+ -> Ack^-	692 ps
Energy	8.6 pJ

Signals in Table 1 that is ending with either A or P, refers to Active respectively Passive port (e.g., *DenA* refers to *Den* in demand active port). The simulation results shows that the maximum clock frequency is 700 MHz before the *Stretch* signal will start to delay the clock.

4. MULTIPORT-TO-MULTIPORT COMMUNICATION

Many of the modules needs to communicate with more than one other unit. It is therefore necessary to create multiport transmission lines that allows a module to communicate with more than one module. Although the signaling, protocol and its corresponding control circuits for point-to-point communication have been well studied, few research efforts have been given to the design of communication methodology between multi-port modules due to its complexity and application-dependent variations [3][8]

In terms of interfacing methods that are most frequently used in GALS systems, there could be two forms of multiport communication. Either is a single output driving multiple inputs, or multiple outports driving the same input. In other words, we can call them either a multiple-output unit interfacing several single-input units or several single-output units interfacing with a multiple-input unit. A combination of these two forms is not con-

sidered here.

In both cases, the single-input or single-output units can be specified in the same way discussed in section 3.2 using two- or four-phase protocol. Thus the main consideration should be taken on the multiple-output or input units in the design of multiport to multiport communication. One possible multiple-output setup is shown in Figure 7, where $A1$ is active output port and $B1$ and $B2$ is passive input ports.

For the design of multiple-output units, which is shown in Figure 7, the data can be sent to both $B1$ and $B2$ or just to one of them. When data is transferred to both $B1$ and $B2$, $Req1$ and $Req2$ can be either separate or combined to one request signal. The schedule for the request signal to $B1$ and $B2$ could be different for different application, e.g., $Req2^+$ could be delayed to after $Ack1^-$ all depending on the application.

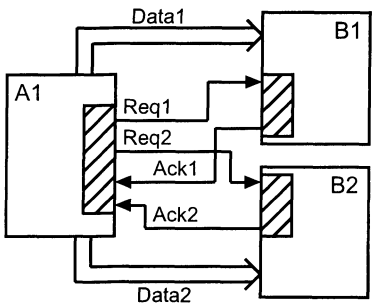


Figure 7. Multiple-output unit interface.

With the multiple-input units, as shown in Figure 8, we have a similar consideration as with multiple-output units.

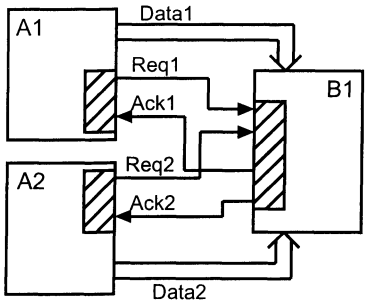


Figure 8. Multiport-input unit interface.

There are also several approaches to implement the asynchronous interface circuits embedded in multiple-input module which depends on scheduling of data communication between multiple modules.

Assuming that computation of the multiple-input unit can not be started until it received all the data transferred from units in front and all the communication channels share the same data path but each has its own control signals. The data communication between modules are mutually exclusive. That is to say, if module $A1$ is communicating with module $B1$, the communication between module $A2$ and module $B1$ has to be delayed until the first communication between $A1$ and $B1$ has ended. If $A1$ and $A2$ shares the same data path, i.e., tristate bus, the Early or Broad data schemes can not be used because both $A1$ and $A2$ would then try to drive the bus in the same time. A solution could be to either use a delayed version of the Broad scheme, i.e., delay until Ack^+, or use the Late scheme.

5. LOW POWER BUSSES

The busses are unique communication channels for different functional units or modules. The busses consume a significant part of power due to the high switching activities and large loads. The increasing level of system integration on chip requires more busses. The higher system performance demands also higher speed requirement for the busses, which further increases the power consumption of busses. Low power busses are therefore important for the systems.

For the CMOS circuits, the dynamic power consumption can be defined by $P_{dyn} = \alpha C_{load} V_{swing} V_{dd} f$. In order to reduce the power consumption, various techniques have been proposed. To reduce the switch activity factor α, bus encoding scheme is proposed [9]. To reduce the capacitance load for the busses, the busses use often thinner metal wires, which has less area capacitance to the ground. Low swing bus is used for reduce the V_{swing} term and discussed in [10].

6. MULTIPLE V_{DD}

Some modules may be much faster than required. That would mean that they are often idle. One way to exchange this into lower power consumption would be to use a lower power supply voltage for these modules. This would of course require that all signals going in and out of the module should have level shifters.

Some modules may have an utilization which is varying. One efficient solution would be to vary the power supply voltage depending on the computation requirement. The variation could either be accomplished by selecting power supply voltage between a few predefined values, or a local DC/DC converter could generate the correct power supply voltage. The simplest form of a DC/DC converter would be some type of voltage doubler.

More general DC/DC converts can be used to provide finer regulation of the supply voltage than the two methods above, but require large external components for the low-pass filter, so they are not suited for several locally controlled on-chip voltage supply generators.

7. LOCAL CLOCKS

The use of asynchronous communications between the modules removes the need for a common synchronous clock to be used by communicating modules. It is thus possible to use a local clock within the module, and to have it work at a rate which is only determined by the requirements posed by the requirement on throughput in the local module. It is even possible to stop the clock completely once the module is idle.

This allows for substantial power savings as the main part of the power consumption is related to switching.

Some functions in DSP have a bursty communication pattern. For example, the FFT (depending on the implementation) will generate bursts of data. The modules feeding and getting fed by the FFT may be using a fixed sample rate. The requirements on the FFT may then be relaxed if a buffer is added between the fixed rate input module and the FFT.

The buffered communication link between two modules should use a FIFO. This unit can be located at either end of the link, or be distributed into it. A distributed version would use a micropipeline structure, and the local registers would then double as storage elements and line drivers. In the case of a FIFO located at the receiver or transmitter end, the FIFO should be located at the end which has the shortest sample rate (most bursty in nature). This will allow the communication link use a less bursty data transfer.

7.1 STRETCHABLE CLOCK CONTROLLER

A stretchable clock controller is a key component to GALS systems. By stretching the low phase of clock and adjusting the number of inverters we can provide the synchronous module a highly flexible clock to avoid synchronization failure. The pausable clock controller (PCC) is the first scheme applied to GALS systems that address correct arbitration [3]. Its main drawback is that a metastability could happen when the request and the rising edges of the clock arrive simultaneously, because it takes a strategy to "toss a coin" to determine which passes the mutual exclusion circuit. Another stretchable clock [4] has a simple structure using only two basic gates, but it could be unreliable because the output can generate glitches. Our stretchable clock controller shown in Figure 9 has a similar architecture to [4].

The generated local clock signal is feed back together with the *Stretch* signal to the input terminals of the Müller *C* element. According to the properties of *C* element, if *Stretch* is asserted to Low, the output of *C* element will follow the inputs and a clock signal is generated. If *Stretch* is asserted to high, the input *Xa* is set to low, the output of *C* element could be either at low or high. However, the output will eventually be maintained at low level. Thus the next rising edge is postponed by the $Stretch^{+}$. For the multiple requests of stretching the local clock, it can be obtained by connecting all the $Stretch_{1\text{-}i}$ to an OR gate. As soon as one stretch signal goes high, the clock will be stretched.

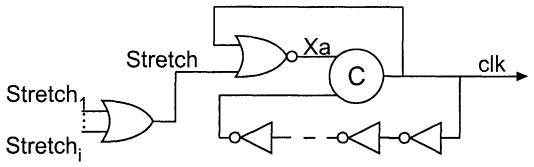


Figure 9. Stretchable clock generation.

8. CONCLUSION

This paper we discuss the design and implementations of communication systems based on the GALS approach and the design of asynchronous communication channels based on a modular wrapper.

REFERENCES

[1] Njølstad T. et al., "A socket interface for GALS using locally dynamic voltage scaling for rate-adaptive energy saving," Proc. 14th Annual IEEE Intern. ASIC/SOC Conf., Rochester, pp. 110-116, 2001

[2] J. Muttersbach, T. Villiger and W. Fichtner, "Practical Design of Globally-Asynchronous Locally-Synchronous Systems," Proc. Int. Symp. on Advanced Research in Asynchronous Circuits and Systems, 2000, Eilat, Israel, pp. 52-59, March 2000.

[3] K.Y. Yun and R.P. Donohue, "Pausible clocking: a first step toward heterogeneous systems," Proc. of Int. Conf. Computer Design (ICCD), Austin, Texas, USA, pp. 118-123, 1996.

[4] D.S. Bormann and P.Y.K. Cheung, "Asynchronous Wrapper for Heterogeneous Systems," Proc. Int. Conf. on Computer Design (ICCD), Austin, Texas, USA, pp. 307-314, 1997.

[5] P. Liljeberg, J. Plosila and J. Isoaho, "Asynchronous interface for locally clocked modules in ULSI systems," Proc. Int. Conf. on Circuits and Systems, ISCAS 2001, Sydney, Australia, pp. 170-173 vol. 4, 2001.

[6] P. Gronowski et al., "High-performance microprocessor design", IEEE J. Solid-State Circuits, Vol. 33 Issue 5, pp. 676 -686, May 1998.

[7] A.M. Peeters, "Single-Rail Handshake Circuits," Ph.D. Thesis, Eindhoven University of Technology, June 1996.

[8] T.H.-Y. Meng, R.W. Brodersen and D.G. Messerschmitt, "Automatic synthesis of asynchronous circuits from high-level specifications," IEEE on Trans on Computer-Aided Design, Vol. 8, No. 11, pp. 1185-1205, Nov., 1989

[9] M. Stan and W. Burleson, "Bus-Invert Coding for Low Power I/O," IEEE Trans. on VLSI Systems, Vol. 3, No. 1, pp. 49-58, March 1995.

[10] H. Zhang and J. Rabaey, "Low-Swing Interconnect Interface Circuits," Proc. Int. Symp. on Low Power Electronics and Design, Monterey, California, USA, pp. 161-166, Aug. 1998.

[11] D.M. Chapiro, "Globally-Asynchronous Locally-Synchronous Systems," Ph.D. Thesis, Stanford University, Oct. 1984.

A tuned, inductorless, recursive filter LNA in CMOS

Christer Svensson[1], Peter Caputa[2], and Stefan Andersson[3]

[1]Electronic Devices, Department of EE, Linköping University 581 83 Linköping. E-mail: christer@isy.liu.se.
[2]Electronic Devices, Department of EE, Linköping University 581 83 Linköping. E-mail: petca@isy.liu.se.
[3]Electronic Devices, Department of EE, Linköping University 581 83 Linköping. E-mail: stean@isy.liu.se.

Abstract

An active recursive filter approach is proposed for the implementation of an inductorless, tuned LNA in CMOS. Such an LNA was designed and fabricated in a 0.8 μm CMOS process. In simulation, the feasibility of this type of LNA was demonstrated, and reasonably good performance was obtained. The fabricated device shows a center frequency tuning range from 250 MHz to 975 MHz. Gain and Q-value are tunable in a wide range. The LNA exhibits an input referred 1 dB compression point of -31 dBm and a noise figure of approximately 3 dB measured at 900 MHz center frequency.

1 Introduction

The frontend of most wireless receivers needs a bandpass filter and a low noise amplifier. Normally, this is accomplished by the combination of an external LC-filter and a bandpass LNA utilizing integrated inductors. It would be very attractive if the inductors could be avoided, as they are quite hard to integrate and use large chip area. It would furthermore be very attractive if the bandpass filter could be electronically tunable in both frequency and bandwidth. The present work is an attempt to design an LNA with these attractive properties. Being a first attempt, the goal is to demonstrate the feasibility of a new approach, not to demonstrate a final solution. In the following, we will first discuss the new approach. We will then describe an experimental design followed by results from both circuit simulation and measurements on the actual device. Finally, we will give a discussion and draw some conclusions.

2 Active recursive filter approach to the LNA

Active recursive bandpass filters have recently become interesting for microwave applications [1, 2, 3]. In these applications, a bandpass filter is accomplished according to the model in Fig. 1. This model is based on positive feedback through a delay.

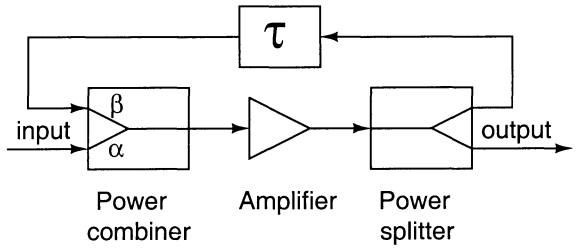


Figure 1: Microwave recursive filter.

In the microwave case, bandpass filters are aimed for replacing the combination of LNAs and resonator-based filters, leading to smaller area, and considerably increased tunability. Let us now see how the principle of active recursive filters can be applied to a standard digital CMOS process without using inductors, resonators, or transmission lines.

We propose a scheme according to Fig. 2. where the input stage power combiner is replaced

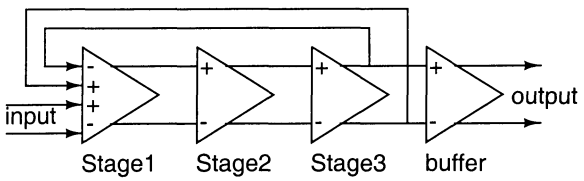


Figure 2: CMOS recursive filter.

by a two-input amplifier. The necessary delay is replaced by delays through the three amplifier stages. Frequency tunability is achieved through variation of the amplifier delay, which can be studied through the amplifier time constant, τ. The delay through the amplifier chain corresponds to a phase shift ϕ_{delay} according to

$$\phi_{delay} = \pi + 2\pi n = \omega_c \tau = 2\pi f_c \tau \qquad (1)$$

Furthermore, by connecting the positive/negative output to the negative/positive input, 180 degrees

low-frequency phase shift is provided. Together with the delay through the amplifier chain, the necessary phase shift of 360 degrees (and its multiples) is obtained. This implies $f_c=k/\tau$, where k is a positive integer. The proposed scheme in Fig. 2 can be viewed as a "nonoscillating ring oscillator" utilizing the same frequency tuning mechanism as used in VCOs based on ring oscillators [4, 5]. The delay of each amplifier is proportional to its unity-gain frequency. In other words

$$\tau \propto \frac{C_L}{g_m} \tag{2}$$

This means the time constant τ can be controlled by the bias current through each amplifier.

The Q-value can be tunable through variation of the loop gain. Low gain broadens the amplification peak which results in a lower Q-value. Finally, we have chosen a fully differential design, for its robustness and low sensitivity to supply noise [4].

A schematic for the recursive filter in Fig. 2 is shown in Fig. 3. The first stage should have two differential inputs, which is achieved by having two differential pairs connected in parallel, sharing the load transistors. The differential input signal is fed to a common-gate stage having an input impedance of $1/g_m$, chosen to achieve 50 Ω. Malmqvist found that minimum noise figure was obtained by having maximum coupling coefficient for the input signal (α close to 1 in Fig. 1) [3]. We choose g_m of the feedback input in the first stage to 1/4 of the g_m of the input transistor. Stage 2 and 3 in Fig. 2 are based on simple differential amplifiers. The general current level is controlled by the bias voltage Vb1, and is used for loop delay control and hence for frequency tuning. Furthermore, the gain in stage 1 and 2 is controlled by PMOS transistors connected across the differential outputs. The gain can thus be tuned by the bias voltage Vb2 connected to these PMOS devices. This facilitates a loop gain control, which is independent of the delay control. A simple bias circuit is used to generate bias to the PMOS loads and to the input transistor gates. Output buffers are not shown in Fig. 3.

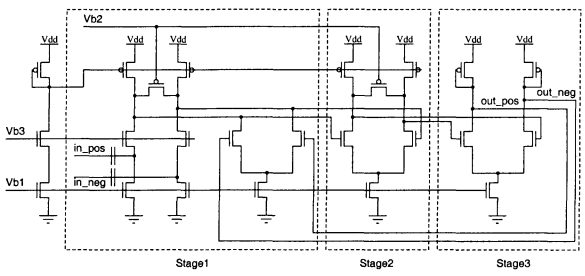


Figure 3: Filter LNA schematic.

3 Measured Performance

The LNA previously described was fabricated in a 0.8 μm CMOS process having a supply voltage of 5 V. A chip microphotograph is shown in Fig. 9 where the total circuit area (including pads) is 0.78 mm^2. In Fig. 10 we see the die bonded directly onto a printed circuit board in order avoid large bonding wire inductances. The behavior of the measured and simulated circuit agreed quite well. Table 1 shows both simulated and measured results for a case where the LNA was tuned to an input frequency of 900 MHz.

Parameter	Simulated	Measured
Q-value (tunable)	20	36
Gain (tunable) [dB]	30	27
Noise Figure [dB]	3.4	~3
Input impedance [Ω]	100	-
Power consumption [mW]	41	104

Table 1: Simulated and measured parameters of the filter LNA at 900 MHz center frequency.

Simulation of the circuit showed a center frequency tuning range of at least 590-1200 MHz. The center frequency span of the fabricated device was found to be 250-975 MHz. Fig. 4 shows the measured transfer characteristics when the LNA was tuned to a center frequency of 450 MHz, 675 MHz, and 900 MHz, respectively. In all three cases, the center frequency gain was set to 15 dB. The corresponding total power consumption (including bias circuitry and output buffers) was measured to 22.7 mW, 50 mW, and 104 mW, respectively.

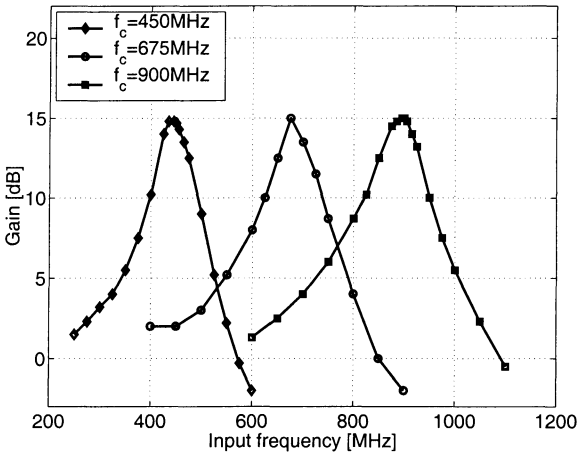


Figure 4: LNA transfer characteristics for a center frequency of 450 MHz, 675 MHz, and 900 MHz.

By raising the control voltage Vb2, the gain and Q-value of the device is increased while leaving the

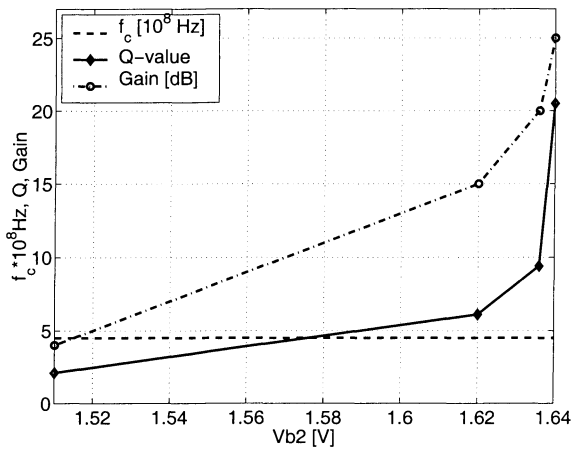


Figure 5: Q-value and gain vs. control voltage Vb2 at 450 MHz.

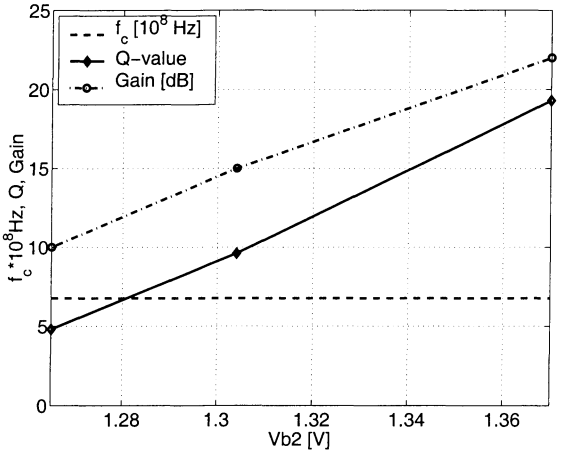


Figure 6: Q-value and gain vs. control voltage Vb2 at 675 MHz.

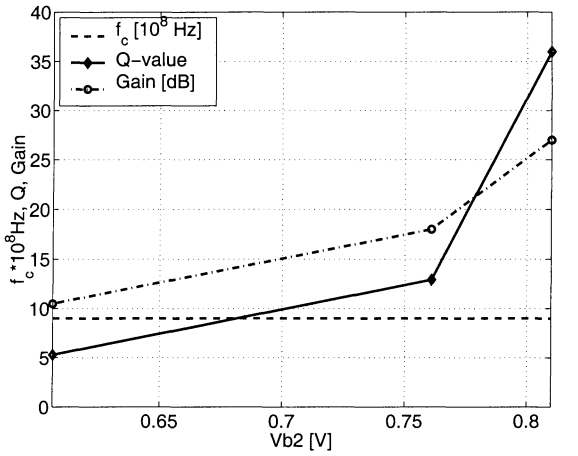


Figure 7: Q-value and gain vs. control voltage Vb2 at 900 MHz.

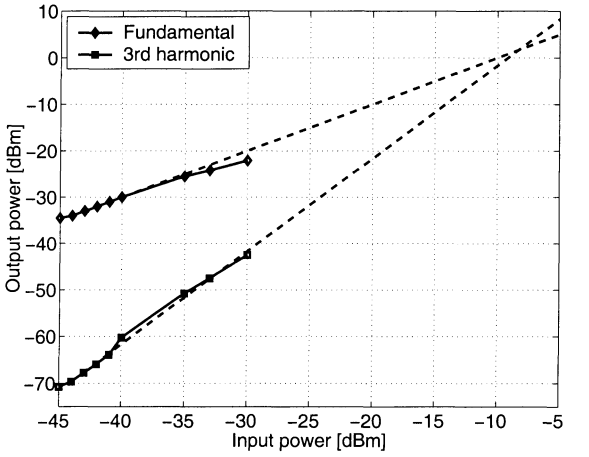


Figure 8: Fundamental and third-order harmonic.

center frequency essentially unaffected. Fig. 5-7 display examples of the measured gain and Q-value tunability vs. Vb2 when the LNA was tuned to a center frequency of 450 MHz, 675 MHz, and 900 MHz, respectively.

The input referred 1 dB compression point was recorded to -31 dBm. This rather low value is to some extent due to reduced gain (less than one) and nonlinearity of the output buffers. Fig. 8 shows the fundamental and third-order harmonic in a two-tone measurement where the input frequencies are 900 MHz and 905 MHz. To estimate IP3 we need a slope of three for the third-order harmonic, which is not the case in Fig. 8 where the slope equals two. Therefore we cannot estimate IP3. Instead we note a SFDR equal to 30 dB for an input power of -40 dBm.

A noise figure of around 3 dB was obtained by directly using a spectrum analyzer according to a method described in [6]. This method employs the equation

$$NF = N_o - G - 10\log BW + 174dB + 1.7dB \quad (3)$$

where N_o is the noise level measured on the spectrum analyzer, G is the LNA gain, and BW the is the IF bandwidth setting of the analyzer. The 174 dB factor is the constant kT which is the Boltzmann constant times temperature. The 1.7 dB is a correction factor which is the sum of bandwidth, log amplifier, and detector correction. This is a fairly good number to use for just about any analyzer.

4 Discussion

A new form of bandpass LNA, based on the principle of recursive filters, was proposed and implemented in a 0.8 μm CMOS process. Measurements have shown that this approach is feasible, at least for applications

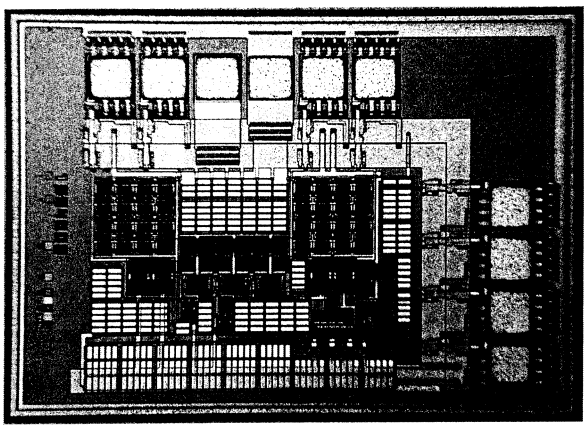


Figure 9: Chip microphotograph.

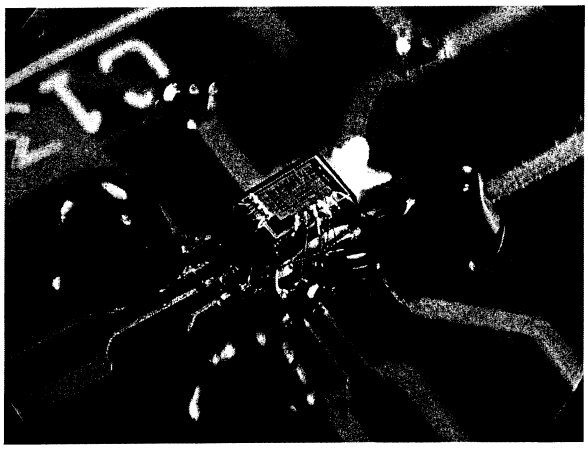


Figure 10: Direct bonding of the chip to the printed circuit board.

with not too high performance demands. The tuning range in frequency is very wide and extends from 250 MHz up to 975 MHz. Moreover, the Q-value of the device is quite high and can, together with the gain, be tunable in a broad range.

For high Q-values this design will be very sensitive to parameter and bias variations [2]. This is natural as we are utilizing a circuit close to oscillating. The circuit must therefore be controlled by a control system, for example based on replica, just oscillating, oscillator [7].

At a center frequency of 900 MHz, the noise figure of the amplifier is around 3 dB. This corresponds quite well to the expected value of 3.4 dB obtained during circuit simulation. The nonlinearity of the circuit is reflected by a 1 dB compression point of -31 dBm (input referred) and SFDR of -30 dB at an input power of -40 dBm.

In conclusion, we have demonstrated the feasibility and benefits of a low power, inductorless, recursive filter LNA in CMOS, useful at least in less demanding applications.

5 Acknowledgements

Some of this work was performed during Prof. Christer Svensson's sabbatical leave at Circuit Research Lab., Intel Corp., OR. He wants to acknowledge Shekhar Borkar and Krisnamurthy Soumyanath for their hospitality.

The authors would also like to thank M.Sc. Johan Månsson for participation in the physical design work and Ph.D student Daniel Wiklund for designing the printed circuit board.

References

[1] M. Delmond, L. Billonet, B. Jarry and P. Guillon, *Microwave Tunable Active Filter Design in MMIC technology using recursive concepts*, IEEE 1995 Microwave and Millimeter-Wave Monolithic Circuits Symposium, Orlando, FL., May 1995, pp. 105-108.

[2] R. Malmqvist, M. Danestig, S. Rudner and C. Svensson, *Theoretical Analysis of Sensitivity and Q-Value for Recursive Active Microwave Integrated Filters*, IEE Proceedings of Antennas and Propagation, vol. 146, pp. 247-252, August 1999.

[3] R. Malmqvist, M. Danestig, S. Rudner and C. Svensson, *Some Limiting Factors for the Noise Optimization of Recursive Active Microwave Integrated Filters*, Microwave and Optical Technology Letters, vol. 22, pp. 151-157, August 1999.

[4] A. Hajimiri, S. Limotyrakis and T. Lee, *Jitter and Phase Noise in Ring Oscillators*, IEEE Journal of Solid State Circuits, vol. 34, pp. 790-804, June 1999.

[5] D.A. Johns and K. Martin, *Analog Integrated Circuit Design*, Wiley 1997, ISBN 0-471-14448-7

[6] T.S. Laverghetta, *Handbook of Microwave Testing*, Artech House 1981, ISBN 0890060703

[7] H. Huang and E.K.F. Lee, *A 1V Active RC Filter with On-Chip Frequency and Q Tuning*, Proc. of the 25^{th} European Solid-State Circuits Conf., pp. 370-373, Sept. 1999.

Integrated Self-Calibrating Image-Rejection Technique for a 5.2 GHz HIPERLAN/2 Receiver

Mikael Nylund

Radio Electronics LECS/IMIT KTH-Electrum 229, 164 40 Kista, Sweden. E-mail: rel-mnd@imit.kth.se

Abstract

A self-calibrating image-rejection technique for a 5.2 GHz HIPERLAN/2 receiver is examined and designed. The calibration technique is applicable to receivers using *time division duplexing* (TDD). The receiver uses a Weaver architecture with phase calibration circuits. We obtain an image-rejection of 64 dB, without using any discrete components. A chip is designed in a 0.18 µm CMOS-process.

1 Introduction

Wireless local area networks (WLANs) for 2.4 GHz have increased dramatically during the last years and this has led to a need for more frequency space. The frequencies that are available are located in the 5 GHz band. A challenge for the designers is to make both cheap and high performance transceivers for this band.

HIPERLAN/2 is a standard for WLANs, using TDD and *time division multiple access* (TDMA) in the bands 5.15-5.35 GHz and 5.47-5.725 GHz. Each channel is 20 MHz wide with 52 sub-carriers and the modulation scheme is *orthogonal frequency division multiplexing* (OFDM). The data bitrates ranges from 6 Mbit/s to 54 Mbit/s.

2 Image-rejection

From specifications [1], a full receiver would require 86.5 dB image-rejection ratio (IRR). One way of dealing with the image-rejection problem is to put a SAW-filter in front of the *low noise amplifier* (LNA) in the receiver chain. The filter has to be realized off-chip and is hence not attractive. Uncalibrated we might only reach 30 dB IRR. It is also hard to make integrated image-rejection filters before the downconversion, because of the low Q-value that can be obtained using integrated techniques. What is left is to use some image-reject receiver architecture, such as Weaver, Hartley or to use polyphase filters after the downconversion.

The most common way of dealing with image-rejection in integrated receivers is to use polyphase filters, even though there are many drawbacks. It is possible to realize passive polyphase filters that gives image-rejection over 60 dB over the entire channel-bandwidth [2]. These filters then need to be of order 5 or more and that leads to an attenuation of the wanted signal by 10-20 dB. The realization of a 5th order polyphase filter will furthermore consume a lot of chip area. Another method is to design an active polyphase filter [3]. The advantages are less consumed area and less attenuation of the wanted signal. However, there is a big drawback, the power consumption increases with the channel bandwidth and HIPERLAN/2 has quite wide channels.

Only the Weaver is attractive among the image-reject receiver architectures. This is due to the shift-by-90°-network that is incorporated in a Hartley architecture. The shift-by-90°-network is often realized as an RC-CR combination and the image-rejection is only good at one single frequency.

3 The Weaver architecture with calibration.

A Weaver architecture is chosen to obtain large image-rejection, see Figure 1. Phase and gain error between the I- and Q-paths lead to the need for calibration. The IRR that can be obtained with phase and gain error can be shown to be given by Equation (1) [4].

$$\mathrm{IRR} = \frac{1+(1+\varepsilon)^2+2(1+\varepsilon)\cos\alpha}{1+(1+\varepsilon)^2-2(1+\varepsilon)\cos\alpha} \qquad (1)$$

Where α is the phase error and ε is the gain error.

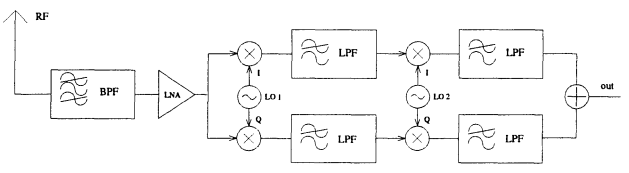


Figure 1: The Weaver architecture

A 60 dB IRR leads to hard but reachable requirements on the phase and gain errors, see Figure 2. The full 86.5 dB IRR can be reached with an external band pass filter.

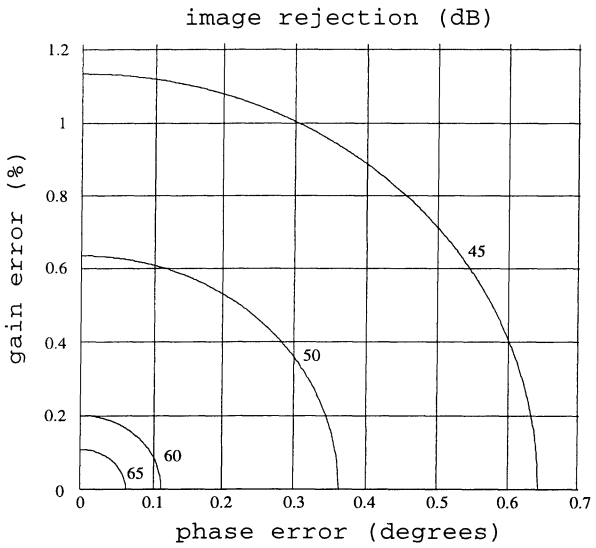


Figure 2: IRR as a function of phase and gain error.

The method with calibration is interesting and it has been tested for GSM earlier [5]. The Weaver architecture is more sensitive to phase error than gain error and therefore we only implement the phase calibration here.

3.1 Phase calibration

The method uses a known reference signal, that is located in vicinity of the image frequency and that is fed into the receiver during calibration. Since the receiver uses TDD, the calibration is done when the receiver is in transmit mode. During this time only LNA_2 is on, see Figure 3. The reference signal is downconverted with the I-component of the first *local oscillator* (LO) and the Q-component of the second LO. This signal is then multiplied with the rest of the image. After the multiplication two frequency components are present, one at DC and one upconverted frequency. The DC-component contains information about both the magnitude and the sign of the phase error and is called the error voltage, V_α. It can be shown that it is proportional to V_{ref}'s amplitude, the gain in the system, A, and the phase error, α, see Equation (2).

$$V_\alpha = A \cdot V_{ref} \sin(\frac{\alpha}{2}) \approx A \cdot V_{ref} (\frac{\alpha}{2}) \qquad (2)$$

The DC-component is then filtered out using a *low pass filter* (LPF) and fed back to two variable delay cells. These delay cells control the output of the second LO and are operated in anti-phase.

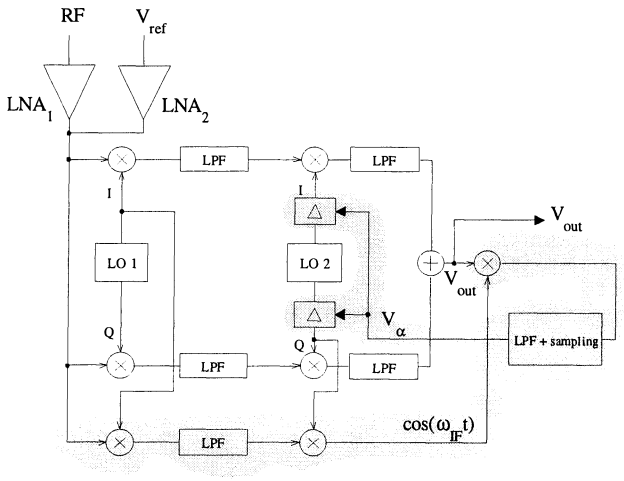


Figure 3: The Weaver architecture with phase calibration (grey).

Another very important issue of the LPF is to introduce a dominating pole, which guarantees stability in the feedback. The control system uses only proportional control, this means that the phase error can never be forced to zero, it can however be made arbitrarily small [6]. It can be shown that the error is proportional to the gain in the system, se Equation (3). A larger gain on the other hand requires a lower cut-off frequency in the LPF, leading to a slower system.

$$\lim_{t\to\infty} \alpha * \frac{y(t)}{x(t)} = \lim_{s\to 0} s \frac{\alpha}{s} \frac{Y(s)}{X(s)} = \frac{\alpha}{G} \qquad (3)$$

After the control system has settled, the error voltage, V_α, is sampled on capacitors and the receiver is ready for reception.

3.2 Gain calibration

The gain calibration can be done in a similar way as the phase calibration, however there are some differences. The reference signal has to be downconverted with the I-component in both downconversions and then multiplied with the rest of the image. The error voltage, V_ε, is obtained after the LPF and it contains information about both the sign and the magnitude of the gain error, see Equation (4). The error voltage is also independent of the phase error.

$$V_\varepsilon = V_{out} \cdot \sin(\omega_{IF} t) = \frac{\varepsilon}{2} \cdot \cos(\frac{\alpha}{2}) \approx \frac{\varepsilon}{2} \qquad (4)$$

The error voltage can finally be fed back to *variable gain amplifiers (VGAs)*, see Figure 4.

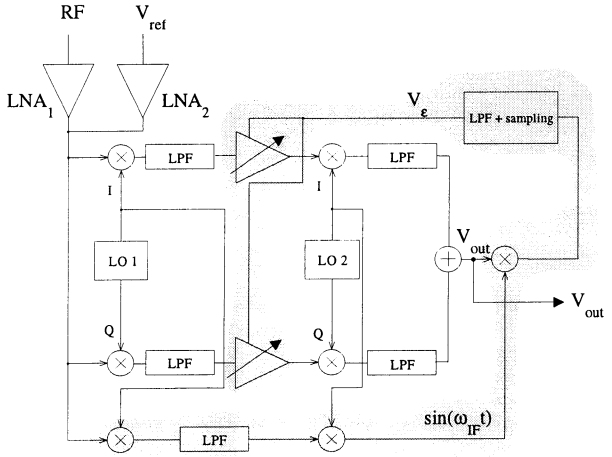


Figure 4: The Weaver architecture with gain calibration (grey).

4 Designed circuits

The chip is designed in a 0.18 µm CMOS-process, with 1.8 V supply voltage. The circuits are made fully differential to reduce substrate interference. The circuits that are realized are the detection mixer, the filter, the sampler and the delay block.

The detection mixer is based on a classic double balanced Gilbert-mixer. The mixer has then been extended with an active cascode load in order to obtain large gain.

An additional circuit for offset-cancellation has been added to the cascode load. The offset originates from selfmixing of the LO signal in the detection mixer. This is especially troublesome when the interesting signal is at DC. This offset-cancellation circuit, senses the difference at the output from the load and feeds back a corrective current to the input, see Figure 5. This forces the DC-offset towards zero.

A common-mode feedback system is also applied to control the bias of the load. This stabilizes the output common-mode voltage and it also protects the circuit against variations in the manufacturing process.

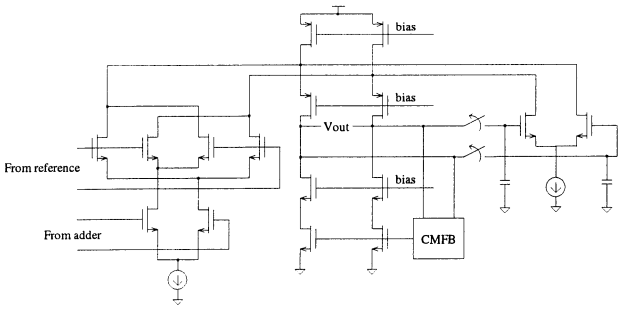


Figure 5: The detection mixer with load and offset cancellation.

The purpose of the filter is to suppress the upconverted component from the mixer. Furthermore the filter has to introduce adequate phase margin in order to guarantee stability. The signals entering the detection mixer have the same frequency, around 10 MHz, resulting in an upconverted component around 20 MHz. As the frequency in the baseband is quite low, this filter could not be implemented as a simple RC-network. This is because of the large chip area it would have consumed. The filter should also show first order behavior, in order not to introduce too much phase shift. A high impedance is obtained by using a transistor biased in the sub-threshold region. Together with a capacitor, this filter acts as a first order filter for low frequencies. When the frequency is increased, the filter stops working as a filter, probably due to capacitive coupling. This problem can be solved by introducing a second pole at a higher frequency, resulting in a first order behavior overall. This filter is over 20 times smaller than it's corresponding RC counterpart.

The bias voltage of the transistors is very sensitive to parameter variations. Designed circuits incorporate protection against process variations and temperature variations. Figure 6 shows the principle for the filter.

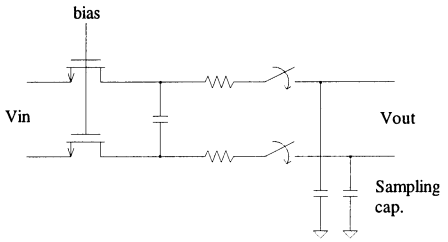


Figure 6: The filter, sampling switch and sampling capacitors.

The delay block is placed after the second LO instead of in the signal path. Then nonlinear switching circuits could be used to obtain the delay. The intermediate frequency is 1.2 GHz, and hence this is the operation frequency of the delay block. The delay is obtained by interpolation between two signal paths, one fast and one slow. The slow path uses two differential pairs and an RC-network to make the signal slow. The delay can be made slower or faster by applying positive or negative differential control voltage on the control input. The schematic is depicted in Figure 7.

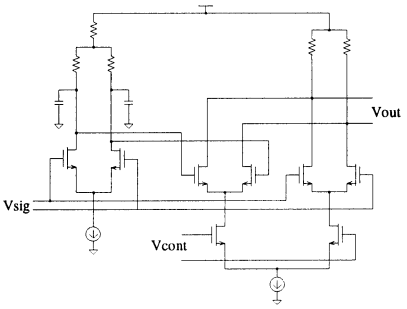


Figure 7: The delay block

6 Simulations

The simulations have been performed in Cadence using SpectreRF. With an initial phase error of 5°, this method is able to improve the image-rejection from 28 dB to over 64 dB.

The power consumption for the calibration circuits including output buffer is 36 mW, during calibration.

The settling time for the calibration system is short, see Figure 8. Nominal settling time is less than 500 ns. Due to temperature and process variations, the settling time can be enlarged to a maximum of 3 µs.

The function of the circuit is verified by simulations for all process corners and temperature variations around room temperature, i.e. 27 ± 10 °C.

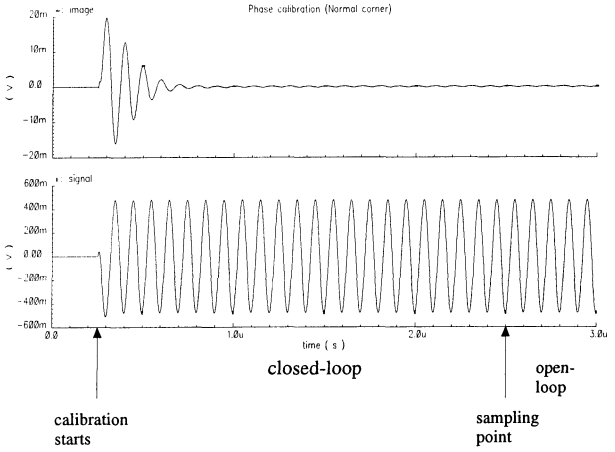


Figure 8: The image and the signal during calibration.

7 Summary

We have developed and analyzed an analog calibration technique of an image-reject receiver. It uses a reference signal for calibration purposes. The technique is aimed for TDD in TDMA systems, such as HIPERLAN/2, for the 5 GHz-band.

The simulated IRR, of our designed circuits, when only performing phase calibration is better than 64 dB. This contributes an improvement by 30-40 dB over un-calibrated image-reject receivers.

The power consumption of the system is 36 mW.

8 References

[1] Adiseno. An Integrated HIPERLAN/2 Receiver: Design specifications v.1.0. Radio Electronics Group, LECS, IMIT, KTH 2002.

[2] F. Behbahani, Y. Kishigami, J. Leete and A. A. Abidi. CMOS 10 MHz-IF Downconverter with On-Chip Broadband Circuit for Large Image-Supression. Symposium on VLSI Circuits Digest of Technical Papers, 1999, pp. 83-86.

[3] Jan Crols and Michiel Steyaert. An AnalogIntegreraded Polyphase Filter for a High Perfomance Low-IF Receiver. Symposium on VLSI Circuits Digest of Technical Papers, 1995, pp. 87-88.

[4] José M. Páez-Borrallo and Francisco J. Casajús Quirós. Self Adjusting Digital Image Rejection Receivers For Mobile Communications. IEEE Vehicle Tech. Conf. 2, 1997, pp. 686-690.

[5] Raymond Montemayor and Behzad Razavi. A Self-Calibrating 900-MHz CMOS Image-Reject Receiver. Proc. 2000 Eur. Solid-State Circuits Conf., 2000, pp. 292-295.

[6] Torkel Glad and Lennart Ljung. Reglerteknik Grundläggande teori. Studentlitteratur, Lund 1989.

Mobile Wireless Access to Fixed Networks

Johan M Karlsson[1] and Ove Edfors[2]

[1]Dept. of Telecommunication, Lund University, Box 118, 221 00 Lund, Sweden

[2]Dept. of Electroscience, Lund University, Box 118, 221 00 Lund, Sweden

Abstract

This PCC project deals with mobile wireless access to fixed networks. In particular, the project addesses the quality-of-service problems that occur when low-quality wireless links are used as extensions of high-quality wire-line networks.

1 Problem Area

The core of the problem area addressed in this project is the use of both wired and wireless physical channels as access to a network. The large difference in quality of the connections can cause several problems. As an example, the discrepancy in the physical layer implies serious problems for wireless ATM (wATM) where the strict ATM demands on QoS cannot be met in most wireless situations. Fixed high speed networks are usually built on the assumption that the error rate is very low. ATM has been optimized in the sense that neither error nor flow control is implemented due to the assumed very low error rate (designed for optic fibers). However, when we enter the wireless ATM concept, we meet new challenges. Even though we use ATM as an example here, essentially the same arguments are valid for other technologies initially developed for fixed networks. A thorough understanding of the time varying physical layer is essential for making the correct changes for an enhanced performance.

2 Approach

We strongly believe that the situation described above can be greatly improved by proper QoS negotiations between the different communication layers. We therefore combine efforts by researchers from different research areas. Competence in wireless technology is used to make sure that relevant information is available to higher layers, while competence in networking is used to make the most out of the current conditions in the wireless channel. Researchers in both fields are forced to not only assume a certain behaviour of the other layer, but understand, negotiate and discuss the physical and pratical limitations endorsed by their respective knowledge areas. This interaction is used to make appropriate trade-offs in their respective areas to meet the overall performance goals and strategies. Several levels of communications will be necessary, as shown in Figure 1. For instance, the slowest changes in the physical layer may be communicated all the way to the application layer (APP) while the most rapid changes will have to be handled very close to, or within, the physical layer (PHY).

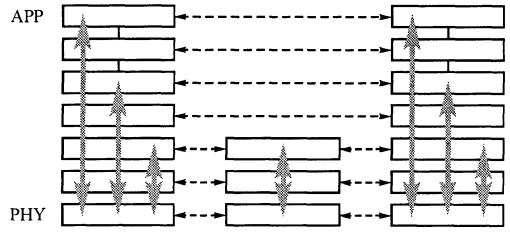


Figure 1: *Communication of quality information between network layers.*

From the user's point of view, and probably also the network management's, the performance measurements that are important are the quality and quantity of the transmission between the two end-points in the connection. Our aim is to improve this by using our different skills in a unified way. The best way of getting a better understanding of this entire chain and to be able to give an end-to-end performance measure, is a better understanding of the interaction between the radio and network layer.

The specific demands from different services used by the mobile users are forced to be communicated between the network and the radio layer to be able to handle them efficiently from the systems point of view. Without the knowledge of the services, the network QoS and what the physical layer can deliver, the suboptimizations that are taking place today within these areas will continue. We strongly feel that the strategy of searching suboptimal solutions, only to preserve the isolation between layers in the ISO/OSI model, can be greatly improved upon.

Networks with wireless connections can be of several different types, as outlined in Figure 2 – all the way from fixed networks with dedicated wireless access points (e.g. HIPERLAN) to ad-hoc networks

(e.g. BlueTooth). The focus of this project is on fixed networks with wireless access.

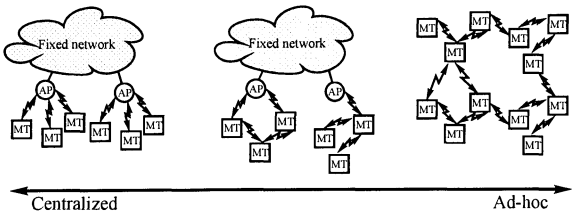


Figure 2: *Different levels of centralization of the wireless access. The focus of the project is on the two left types.*

3 Objective

The objective of this project is to develop a unified approach to *"mobile wireless access to fixed networks"*. By this we mean a global optimization of the entire network, including both fixed and wireless connections. In more detail, the objective can be expressed in three parts:

- Improving and tracking the varying quality of wireless connections,
- providing adaptive protocols that can take these variations into account, and
- optimizing the entire network, both fixed and wireless connections, as one unit by allowing efficient inter-layer communication and thereby obtaining, *e.g.*, more efficient scheduling.

A Procedure for Combining Positioning and History–Based Measurements in Dynamic Guard Channel Schemes

Roland Zander and Johan M Karlsson

Department of Communication Systems, Lund Institute of Technology, Lund University, Box 118, SE–221 00 Lund, Sweden. E–mail: {rolandz,johan}@telecom.lth.se

Abstract

Insertion of guard channels in a cellular network is a way of giving priority to on–going calls by having capacity exclusively reserved for handover purposes. In advanced dynamic guard channel schemes, positioning or history–based measurements are usually used to determine the need for guard channels. Herein, a scheme called *Predictive and Adaptive Resource Reservation* (PARR) combining these two types is introduced. This is very efficient since history–based measurements give an initial idea of the call movements in a cell, while positioning can be used to adjust the movement predictions of a specific call with time according to its actual movements so far. To improve the history–based predictions, the calls are divided into groups depending upon mobility and latest visited cell, where separate measurements are performed for every single group. Simulations show that PARR outperforms a pure history–based scheme.

1 Introduction

In order for an operator to be successful in the field of wireless communication, it is crucial to keep the subscribers pleased, which can be achieved by offering high quality services to a relatively small cost. Since it is much more frustrating for a subscriber to have an on–going call dropped than a new call blocked, it is a good idea for the operator to give priority to on–going calls. One way of obtaining this in a cellular network is to have some channels in the cells, called guard channels, exclusively reserved for handover purposes. For a new call to be accepted, all guard channels plus one "ordinary" channel must be unoccupied at the time of the call arrival, while a handover call only needs one available channel to be accepted. Since the number of handovers per call will increase in the future due to cell–size reduction, it will become increasingly important to handle handovers in a satisfactory way.

In the fixed guard channel scheme, a fixed number of guard channels is at all times allocated in a cell [1]. Thus, the number of allocated guard channels is not time varying, but may very well differ between different cells. The fixed guard channel scheme is a simple way of giving priority to on–going calls but unfortunately it is not especially efficient due to its poor flexibility. This flexibility problem will get progressively worse with time because the variance of the handover arrival rate will increase with the introduction of services that demand high bandwidth, e.g. streaming video.

The dynamic guard channel scheme was introduced to counter the above–mentioned flexibility problem. Here, the number of allocated guard channels is dependent upon the momentary network conditions. For example, a guard channel may be allocated in a cell for every 10^{th} on–going call in the adjacent cells [2]. A well–designed dynamic guard channel scheme provides a higher quality of service than its fixed counterpart. Furthermore, the high flexibility makes it feasible to provide a guaranteed average quality of service to on–going calls, such as a handover failure rate of 0.1%, which cannot be achieved in the fixed guard channel scheme. Unfortunately, the accompanying complexity increase causes more processing and more communication between base stations.

There exist two main types of advanced dynamic guard channel schemes, namely those involving history–based measurements and positioning respectively. History–based schemes use movement patterns of previous calls when determining the number of guard channels to allocate in a cell [3]. From movement and mobility measurements, the number of on–going calls in a cell expected to make a handover attempt to a specific adjacent cell can be probabilistically determined.

Global Positioning System (GPS) and received signal strength are two potential techniques for determining the position of a mobile device in a cellular network. One way of using positioning in dynamic guard channel schemes is to allocate a guard channel in a cell if the distance between the cell border and an active subscriber in an adjacent cell is estimated to be below a predetermined threshold [4]. A more gradual approach where the allocated capacity is increased with decreased distance between subscriber and cell border may also be applied. Previous positioning measurements can also be taken into account in the allocation procedure. Then, the direction and speed of an on–going call can be estimated and the future movement may be extrapolated with the use of linear regression [5,6]. The extrapolation line is used to predict the target cell of the potential handover attempt, while the estimated time for the handover attempt may be used to determine the size of the reserved capacity.

Much can be gained if a combination of positioning and history–based measurements is applied in dynamic guard channel schemes. History–based measurements give an initial idea of the call movements in a cell, while positioning can be used to adjust the movement predictions of a specific call with time according to its actual movements. A scheme combining these two types was introduced in [7], but has limitations since it assumes that the base stations have exact geographical knowledge about the road network within its coverage area.

In this paper, a combination scheme called *Predictive and Adaptive Resource Reservation* (PARR) working under more general conditions is introduced. The history–based part of the scheme measures the probability for a call in a particular cell to make a handover attempt (handover probability) and the percentage of these attempts made to a certain target cell (handover direction probability). In order to improve the efficiency of the scheme, the calls are divided into groups depending upon mobility and most recent source cell (latest visited cell), where separate measurements are performed for every single group [8,9]. Further, the cells are divided into zones according to the principle of closest adjacent cell and positioning is used to determine the cell zone of every single on–going call. The original measurement groups are divided into subgroups where every subgroup considers the calls that belong to the measurement group and have visited a certain cell zone.

The remaining of this paper is organized as follows. In section 2, the history–based part of PARR is described in detail, while the description of the positioning part can be found in section 3. The simulation model is presented in section 4 and numerical results from the simulations are shown and explained in section 5. Concluding remarks are given in section 6.

2 History–Based Measurements

PARR requires the base stations to continuously measure the cell–specific handover and handover direction probabilities. When an on–going call departs from a cell, either through a call termination or a handover, the event is registered at the covering base station. The handover probability, P_h, and the handover direction probability to adjacent cell y, $P_{hd}(y)$, are derived through formulas (1) and (2) respectively. In the formulas, D is the number of calls having departed from a cell with ζ adjacent cells and $H(y)$ is the number of handover attempts made to adjacent cell y.

$$P_h = \frac{\sum_{k=1}^{\zeta} H(k)}{D} \quad (1)$$

$$P_{hd}(y) = \frac{H(y)}{\sum_{k=1}^{\zeta} H(k)} \quad (2)$$

Since the subscriber behavior may change with time, there should be a threshold regarding how old events that are to be considered in the measurements. To ensure high flexibility, there may also be a limitation regarding how many events in total that are considered. If the number of events inside the time threshold exceeds the size limit, the events with the oldest time stamps are discarded. It is recommended to further increase the influence of the more recent events by weighing the event data vector according to a function with a negative exponential behavior.

In order to improve the accuracy of the movement predictions and thereby increasing the efficiency of the scheme, the calls are divided into groups depending upon most recent source cell, where separate measurements are performed for every single group. Calls with no previous handover are placed in a separate group. This means that a cell with ζ adjacent cells has $\zeta+1$ source cell groups. The introduction of source cell groups is very effective since calls arriving from different adjacent cells often have completely different movement and mobility patterns. This is especially obvious for a cell covering a highway where almost all subscribers follow the same path, namely the road. Obviously, the handover direction probabilities to some/all adjacent cells for subscribers traveling in opposite directions on the road differ completely from one another.

Further, the source cell groups are split into two different groups depending upon mobility. The channel holding time in the latest visited cell is used to determine if a call shall be placed in a low or high mobility group. Obviously, this procedure can only take place for calls having previously performed a handover. Thus, a total of $2\zeta+1$ separate measurements are performed for every cell. Mobility groups are mainly used because of the differences in handover probabilities between calls from different groups. A slow moving subscriber is often less likely to make a handover attempt than a fast moving subscriber.

From the measured probabilities and the number of on–going calls belonging to the different measurement groups, $C(k)$, the number of calls in a cell expected to make a handover attempt to adjacent cell y is derived. These handover expectancy numbers, $G(y)$, are continuously signaled to the base stations covering the adjacent cells where all received numbers are summed up and a total handover expectancy number indicating the current handover arrival rate is obtained.

$$G(y) = \sum_{k=1}^{2\zeta+1} C(k) P_h(k) P_{hd}(k,y) \quad (3)$$

Since all handover calls will not arrive simultaneously, the number of needed guard channels in cell y, $N_g(y)$, will be significantly smaller than the total handover expectancy number. Fractional guard channels are applied, meaning that the number of guard channels in a cell does not have to be an integer. $N_g(y)$ is derived through multiplication of the total handover expectancy number and a parameter denoted ψ (ψ<1). ψ is variable in order to meet the quality of service demands, such as a handover dropping probability of 0.1%. If the measured average handover dropping probability is higher than the requested value, ψ is increased by a predetermined fraction. To prevent the call blocking probability from rising to an unacceptable level, the parameter is decreased by a predetermined fraction if the measured average handover dropping probability is lower than the requested value. In formula (4), $G_i(y)$ is the handover expectancy number received from adjacent cell i.

$$N_g(y) = \psi \sum_{i=1}^{\zeta} G_i(y) \qquad (4)$$

3 Positioning

Even though the subscribers of a cellular network are allowed to move around freely, most subscribers remain stationary whilst using their mobile devices. Since stationary calls will not make handover attempts, no capacity has to be reserved in the adjacent cells for these calls. Consequently, significant gain can be obtained if the stationary calls can be sorted out from the moving calls. With the use of positioning, the call movements in a cell can be monitored. If a call with no previous performed handover remains stationary throughout its first t seconds, the scheme places the call in a measurement group for non–moving calls where it stays until a call movement is detected.

Further, the cells are divided into zones according to the principle of closest adjacent cell. In Figure 1, the cell zones for a rectangular cell with four adjacent cells are shown. The cell zones are used to divide the calls in the original measurement groups into subgroups. When a call departs from a cell, the event is registered in the subgroups corresponding to all the zones in a cell that were visited by the call. This means that the departure of a call having visited several cell zones is registered in more than one subgroup. Every base station keeps track of the initial zone (in the current cell), current zone and visited cell zones (in the current cell) of all on–going calls in its coverage area. When calculating the handover expectancy numbers to the adjacent cells, the current cell zone of the calls is used for determining subgroup. For the call whose movements in a cell are shown in Figure 1, the initial cell zone is A, the current cell zone is C and visited cell zones are A,B and C. Hence, the measurements of subgroup C should be used when predicting the future movements of the call.

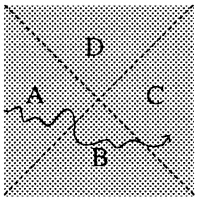
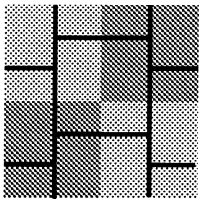


Figure 1. *Cell zones*

Unfortunately, the introduction of measurement subgroups increases the scheme complexity, leading to more processing and increased signaling levels. If only a small number of events are considered, a large number of separate measurements might give problems with the accuracy of the movement predictions. This can be countered by merging groups/subgroups with similar handover and handover direction probabilities. Rapid alterations in the number of on–going calls residing in a certain cell zone might give stability problems if the handover and handover direction probabilities differ considerably between different measurement subgroups. This problem can be reduced if a more conservative approach is applied. By introducing a neutral area in the middle of the cell, considered as a part of all cell zones, the calls will move to a new cell zone less often. Another solution is to apply a more gradual approach in the allocation procedure by using a mixture of the calculated probabilities from the initial subgroup (determined by the initial cell zone of the call) and the current subgroup (determined by the current cell zone of the call) for the on–going calls. The weights used are determined by the distance to the closest adjacent cell where a smaller distance gives the probabilities from the current subgroup a larger weight and vice versa.

4 Simulation Model

The simulation model consisted of 100 rectangular–shaped cells, with a base station located in each cell center, placed in a 10*10 ring topology with wrapped round edges. In every cell there was a simple road network with two crossings as shown in Figure 2. It was assumed that all subscribers traveled along the roads. When a subscriber arrived at a road crossing, a randomization procedure was used to determine the next direction/road segment of the subscriber. Herein, a road segment is defined as the road between two adjacent crossings or between a cell border and the first encountered crossing. In Figure 2, shades of gray is used to distinguish different cells from one another, while black is used to display the road network.


Figure 2. *Road network*

All cells had identical traffic parameters chosen to obtain realistic settings and reasonable traffic load. The distance between consecutive new call arrivals and the call lengths were exponentially distributed with mean values of 0.29 and 7 minutes respectively. The time it takes for a subscriber to travel through a road segment was either rectangular or exponentially distributed depending upon simulation scenario. A mean value of 0.5 minutes was used for the rectangular distribution, while the mean value for the exponential distribution was adjusted to hold the number of handovers per call at the same level for all scenarios. Due to simulation simplification reasons, we chose not to randomize the exact location and traveling direction of new calls arriving to a cell. Instead of tracking these calls through the different road segments, a rectangular or exponential distribution with a mean value corresponding to that of two road segments was used to describe the entire channel holding time. If a handover attempt was made by a new call, the target cell was chosen among the adjacent cells with an equal probability, namely ¼.

The efficiency of PARR (scheme IV) was evaluated through a comparison with a pure history–based guard channel scheme (scheme I) identical to the one used in the history–based part of PARR. In order to investigate the impact of the positioning–based features more thoroughly, schemes with subgroups for either the handover or handover direction probabilities were also included (schemes II and III). In Table 1, the investigated schemes are listed and briefly explained.

Scheme	*Description*
I	No subgroups
II	Subgroups only for handover probabilities
III	Subgroups only for handover direction probabilities
IV	Full subgroups (PARR)

Table 1. *Investigated guard channel schemes*

Simulations were performed for three different scenarios described in Table 2. Only the probabilities for making a left/right turn at the crossings, P_1 and P_2, separated scenario II from scenario I. In these scenarios, a rectangular distribution was used to describe the time it takes for a subscriber to travel through a road segment, while an exponential distribution was used in scenario III.

Scenario	*Distribution*	*Crossing 1*	*Crossing 2*
I	Rectangular	P_1=0.5, P_2=0.5	P_1=0.5, P_2=0.5
II	Rectangular	P_1=0.5, P_2=0.5	P_1=1, P_2=0
III	Exponential	P_1=0.5, P_2=0.5	P_1=0.5, P_2=0.5

Table 2. *Simulation scenarios*

We chose to give all subscribers identical mobility parameters that were randomized for every road segment. This means that the speed of a subscriber starting to travel along a certain road segment was not dependent upon the previous mobility of the subscriber. Consequently, no gain could be obtained from the introduction of mobility groups, making them unnecessary and the feature was therefore disabled. In order to shorten the simulation length, the guaranteed average quality of service feature was also disabled and the handover dropping probabilities were set to $3*10^{-4}$ by manual calibration of ψ. Then, the call blocking probabilities of the evaluated schemes could be compared to find the best scheme for the scenario analyzed.

5 Numerical Results

In Table 3, the simulation results of the different scenarios are shown. 95% confidence interval for the call blocking and handover dropping probabilities are given. In addition, the obtained gain from the introduction of measurement subgroups is presented. The lower limit of the obtained gain column is calculated by dividing the upper limit of the call blocking probability confidence interval with the lower limit of the corresponding number in the pure history–based scheme. The upper limit of the obtained gain column is calculated by dividing the lower limit of the call blocking probability confidence interval with the upper limit of the pure history–based scheme.

Scenario I			
Scheme	*Blocking (10^{-2})*	*Dropping (10^{-4})*	*Gain (%)*
I	3,189–3,194	2,991–3,010	0
II	3,173–3,176	2,994–3,007	0,4–0,7
III	3,001–3,005	2,991–3,013	5,8–6,0
IV	2,975–2,980	2,991–3,012	6,6–6,9
Scenario II			
Scheme	*Blocking (10^{-2})*	*Dropping (10^{-4})*	*Gain (%)*
I	3,066–3,070	2,991–3,009	0
II	3,044–3,048	2,989–3,008	0,6–0,8
III	2,872–2,876	2,992–3,011	6,2–6,4
IV	2,847–2,851	2,990–3,010	7,0–7,3
Scenario III			
Scheme	*Blocking (10^{-2})*	*Dropping (10^{-4})*	*Gain (%)*
I	4,048–4,055	2,990–3,010	0
II	4,030–4,034	2,989–3,011	0,3–0,6
III	3,887–3,892	2,990–3,010	3,9–4,1
IV	3,859–3,862	2,993–3,006	4,6–4,8

Table 3. *Simulation results*

From the simulation results presented in Table 3, we concluded that significant improvements can be obtained when combining history–based measurements with positioning. We also found that the obtained gain for scheme III is larger than the corresponding number for scheme II in all scenarios. This means that especially the target cell predictions, which are derived from the handover direction probabilities, appeared to benefit from the combination procedure. The results can be explained by larger differences for the handover direction probabilities of different measurement subgroups in comparison with the corresponding

handover probabilities. In simulation scenario II, the probabilities for a left/right turn at the second encountered crossing were altered which benefited the proposed scheme due to increased differences between the handover direction probabilities of different measurement subgroups.

When the standard deviation for the channel holding time is increased, which was the case for scenario III, it becomes less important to make correct movement predictions because of inactive guard channels [9]. A guard channel is considered to be inactive as long as the capacity that is supposed to be allocated is occupied by on–going calls. For a distribution with a large standard deviation, the probability of getting a really small value is usually higher than for a distribution with a smaller standard deviation. Really short values for the channel holding time increase the probability for a handover being performed before the allocated capacity in the target cell (next cell) is available, in which case it does not matter if the movement prediction was accurate or not.

The simulation settings were chosen to be quite advantageous towards PARR. Still, if properly calibrated, the scheme will give improvements under any network conditions. Even though significant improvements were achieved with PARR, it is important to state that the division into measurement subgroups increases the overhead in form of more processing and more communication between base stations. It also leads to smaller measurement groups, which might have a negative impact on the accuracy of the predictions.

The obtained gain from predictive oriented features such as measurement groups increases with a larger variance in the need for guard channels. This means that the importance of such features increases with the introduction of services with high bandwidth demands. High subscriber mobility and small cells lead to an increasing number of handovers, which also increase the need for good movement predictions.

6 Conclusions

A dynamic guard channel scheme called *Predictive and Adaptive Resource Reservation* (PARR), combining history–based measurements with positioning when determining the number of guard channels to allocate in the cells, has been introduced. History–based measurements give an initial idea of the call movements in a cell, while positioning can be used to adjust the movement predictions of a specific call with time according to its actual movements. The scheme continuously measures the probability for a call in a certain cell to make a handover attempt and the percentage of these attempts made to a certain target cell. To improve these probabilistic predictions, separate measurements can be performed for calls that are likely to have different movement or mobility patterns. In the scheme, the calls are divided into groups depending upon mobility and latest visited cell. Further, positioning is used to divide calls belonging to the same measurement group into subgroups according to cell zone location.

The efficiency of PARR was evaluated through simulations where it was compared to a pure history–based scheme identical to the one used in the history–based part of PARR. We found that PARR outperformed (lower failure rates) the reference scheme in every single simulation scenario. Furthermore, the high adaptability of PARR makes it feasible to provide a certain average quality of service to on–going calls. Unfortunately, the high complexity of the scheme requires more processing at the base stations as well as more communication between base stations, which has not been taken into account.

References

[1] D. Hong and S. S. Rappaport. *Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone System with Prioritized and Nonprioritized Handoff Procedures.* IEEE Transactions on Vehicular Technology, vol.35, no. 3, pp. 77–92, 1986.

[2] A–L. Beylot, S. Boumerdassi and G. Pujolle. *A New Prioritized Handoff Strategy Using Channel Reservation in Wireless PCN.* IEEE GLOBECOM '98, vol. 3, pp. 1390–1395, 1998.

[3] S. Choi and K. G. Shin. *Predictive and Adaptive Bandwidth Reservation for Handoffs in QoS–Sensitive Cellular Networks.* ACM SIGCOMM'98, pp. 155–166, 1998.

[4] C. H. Choi, M. I. Kim, T. J. Kim and S. J. Kim. *Adaptive Bandwidth Reservation Mechanism Using Mobility Probability in Mobile Multimedia Computing Environment.* IEEE Local Computer Networks 2000, pp. 76–85, 2000.

[5] M–H. Chiu and M. A. Bassiouni. *Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning.* IEEE Journal on Selected Areas in Communications, vol. 18, pp. 510–522, 2000.

[6] W–S. Soh and H. S. Kim. Dynamic Bandwidth *Reservation in Hierarchical Wireless ATM networks Using GPS–Based Prediction.* Vehicular Technology Conference 1999, vol.1, pp. 528–532, 1999.

[7] S. Choi and K. G. Shin. *Exploiting Path/Location Information for Connection Admission Control in Cellular Networks.* Wireless Communication and Networking Conference 2000, vol.3, pp. 935–940, 2000.

[8] S. Lu and V. Bharghavan. *Adaptive Resource Management Algorithms for Indoor Mobile Computing Environments.* Computer Communication Review, vol. 26, pp. 231–242, 1996.

[9] R. Zander and J. M. Karlsson. *An Adaptive Algorithm for Allocation of Dynamic Guard Channels – Impact of the Channel Holding Time Distribution.* Proceedings of Wireless 2001, pp. 300–308, 2001.

A Structure for Work-Conserving Downlink Scheduling in Wireless FDD/TDM Based Systems

Torgny Holmberg[1] and Johan M Karlsson[2]

Department of Communication Systems,Lund Institute of Technology,Lund University,
P.O. Box 118, SE-221 00 LUND, Sweden. E-mail: {torgnyh[1], johan[2]}@telecom.lth.se.

Abstract

In this paper we present a structure for the downlink scheduling performed by the base station in a Local Multipoint Distribution Systems (LMDS), e.g. HiperAccess. The scheduling mechanism is capable of multiplexing traffic from several service classes with different demands for quality of service (QoS) and it can provide limited delay bounds as well as fairness for the best-effort services.

1 Introduction

The current packet radio networks (PRN) provide, to some extent, functionality for both real-time and best-effort traffic ranging from voice to e-mail. As new forms for personal communication are constantly created, the demand on the future PRNs will most likely increase. This will in turn require an even more resource efficient scheduling than today since the capacity of the scheduled resource, the radio channel, is limited by regulations.

In this paper we study a number of schedulers used in a wireless communication system based on the Frequency Division Duplex/Time Division Multiplex (FDD/TDM) scheme. For such systems the scheduler is a statistical multiplexer and its core functionality is to provide a sharing scheme for allowing multiple traffic streams to share a common resource, i.e. the radio channel. For real-time services, the greatest challenge is to meet delay- and jitter-demands. Many schedulers, especially those that are work-conserving, do not decouple the bandwidth need from shorter delays, i.e. by admitting more capacity the packet delays are shortened. For best-effort services on the other hand, the challenge lies in providing fairness. Fairness is a quality that states that all backlogged best-effort services may transmit, in proportion to their weight, an equal amount of data during a certain time. The schedulers based on the fairness notion in the literature often attempts to approximate the Generalized Processor Sharing (GPS) discipline as means for providing fair allocation.

The schedulers investigated in this paper are work-conserving, this means that packets are scheduled if there exists packets available for transmission. This discipline results in better utilization of the capacity but has some difficulties in providing the slack necessary for eliminating jitter. Further, the important objectives of isolation and sharing are presented and how the investigated schedulers fulfill these objectives. The goal of isolation is to prevent misbehaving flows from disrupting the well-behaving flows, i.e. those transmitting in accordance to the traffic contract. The notion of sharing is simply the mixing of the different flows.

The outline of the paper is as follows. In section 2 we present the background and describe the problem. In section 3 a number of work-conserving schedulers are presented and their respective pros and cons are discussed. In section 4 the call admission and the necessity of its existence is described, followed by conclusions in section 5.

2 Background

Traffic streams of different priorities and service requirements may need to coexist and share a common radio channel. In this paper we use three types of service classes to characterize the traffic streams. These service classes are in the literature often referred to as guaranteed, predicted and best-effort. The names reflect the service commitments taken by the communication system, i.e. the QoS that the system has agreed to deliver. The first two are real-time services. Guaranteed services, e.g. (emulated) circuit switching and hard real-time applications, must under the circumstances negotiated in the QoS contract be protected from other resource demanding or misbehaving traffic sources and the packets should (author's opinion) be served upon arrival at the scheduler. A hard real-time application is an application for which the absence of on-time delivery of the packets may have severe consequences for the system. For communication systems like the one studied, such strict requirements usually do not exist and hard delay bounds rather refers to

bounded delays. Predicted services, e.g. audio and video of variable bit rate, have statistical bounds on the QoS and the system will with a certain probability provide the requested service, e.g. 90% of the packets will be delivered within the required time delay. Best-effort services, e.g. web-browsing, share the remaining capacity left by the two services of higher service requirements.

The system for which the investigated schedulers are intended is centralized and based on the FDD/ TDM principle. Centralized in the sense that the base station determines which Access Terminal (AT) that is allowed to transmit as well as the quantity. What is actually transmitted on the uplink is up to the AT to decide. An AT may multiplex several traffic streams of different service requirements and it is the AT that has the best knowledge of the preferred mixture that should be transmitted using the capacity provided by the base station. In this sense the system is distributed.

The slots in the down- and uplink phases may, for generality, be of different length and can carry several packets. On the downlink the base station transmits bursts of packets to the terminals. The number of packets in each burst and to which terminal the burst is intended is a decision taken by the scheduler alone, i.e. terminals may be excluded even though there exists packets destined to a terminal. On the uplink the backlogged terminals transmit in subslots. When a terminal may transmit and how much it is allowed to transmit is decided by the base station. An example of a system adopting such a discipline is HiperAccess.

3 Work-Conserving Scheduling

In the work-conserving disciplines the packets are scheduled whenever there are packets to schedule, i.e. the channel is never idle if there are packets in the buffers. This discipline of always trying to keep the link busy can result in traffic being transmitted with a different statistical behavior than the original traffic. This traffic distortion appears as rate- and delay-jitter. Rate-jitter is used to describe how bursty the traffic is and delay-jitter is the difference in packet delays experienced by any two packets on the same connection.

Examples of work-conserving schedulers will be explained and elaborated upon in the following subsections.

3.1 FIFO

In this discipline the packets are served in the same order as they are received and is hence a special case of the Earliest Due-Date (EDD), which is presented below, with a deadline being a fix increment to the reception time. FIFO does not provide isolation [2] and the guaranteed service class is in the mercy of the best-effort service class. Best-effort traffic is often bursty by nature and this introduces uncontrolled delay-jitter to all other service classes.

The drawbacks of the simple FIFO queue can be avoided by shaping the traffic using a leaky bucket. The output rate and bucket size of the different flows limit the burst size and inter arrival times of the packets at the scheduling buffer. Such a scheme decreases the delay-jitter. Further, the output rates of the leaky buckets can set the priority (bandwidth allocation) hence differentiating the mean queuing delay.

3.2 HOL-NP

The head-of-line non-preemptive, sometimes referred to as priority scheduling, uses separate queues for traffic streams with different priorities. The incoming packets from a flow are given a static priority and put in the queue assigned for that priority. When a packet has been transmitted the next packet is taken from the queue with highest priority. A packet of lower priority is never transmitted as long as there exist packets of higher priority. A packet of a certain priority is only affected by a packet of lower priority in the event of the higher priority packet must wait for the low priority packet that is in service upon arrival due to non-preemption. This scheme provides shorter waiting times for higher priority packets, but does not provide a guaranteed delay unless the queue length is limited or the traffic is shaped by leaky buckets [5]. This scheme is an example of how (excess) bandwidth yields lower transmission delays.

3.3 HOL-NP+

The HOL-NP strategy presented above does not consider how urgent it is to transmit a packet. By giving the packets a suitable time stamp the scheduler may allow some deviation from the static priority order as long as it does not risk violating the delay guarantees of higher priority packets. Even if a traffic stream has hard real-time requirements, it is not necessarily the case that the maximum allowed transmission delay is short. If it is possible to postpone transmission of the packet of higher priority yet keeping the delay guarantee in order to improve the chance of also transmitting the packet of lower priority on time, the bandwidth is in some sense more fairly distributed among the traffic flows. Having a too large margin of safety is essentially only a waste of bandwidth unless retransmissions are significant for the service quality.

3.4 Earliest Due-Date (EDD)

In the simplest form of EDD, also referred to as Earliest Deadline First (EDF), each packet is assigned a

deadline and the packets are served in order of earliest deadline, i.e. the deadline dynamically sets the priority [6]. However, in [6] the authors allow preemption, i.e. the algorithm was intentionally intended for a process control unit, e.g. a CPU executing processes in a real-time operating system. In this paper preemption is, as for most communication systems, not considered. In cases where the service time of the packets is constant, the EDD policy is equivalent to the Least Laxity First (LLF) policy, whose eligible time is the deadline minus the transmission time [8], hence only the EDD policy is considered in this paper.

The EDD algorithm is well covered in the literature in the context of process scheduling in real-time systems. The task invocation is referred to as *periodic*, *aperiodic* or *sporadic*. The formal definition of these seems to be author-specific. The process scheduling principle differs from the packet transmission scheduling principle due to the fact that in a communication network the packets are transmitted according to the store-and-forward principle. The packets are buffered at the input port of the switch/base station because *i*) the switching fabric is a shared resource, and *ii*) traffic shaping may be implemented using buffers. Process scheduling do not consider buffering and hence many of the known process scheduling theories do not apply.

3.5 Scheduling Best-Effort Traffic

The schedulers based on the fairness notion, as stated above, often attempts to approximate the Generalized Processor Sharing (GPS) discipline as means for providing fair allocation. GPS is an ideal (cannot be realized) fluid flow scheduling discipline where every backlogged flow is served an infinitesimal amount of data at least once during a finite time interval. This scheme provides max-min fairness allocation. To also provide the max-min *weighted* fair share, the flows are instead served an amount of data in proportion to their weights. In the literature, a scheme is stated to be fair if the scheme can provide max-min weighted fair share.

The widely accepted definition of GPS is [4]:

If a backlogged flow i with the weight $\phi_i > 0$ is served by a GPS server, then the GPS server serves $S_i(t_1, t_2)$ amount of data from flow i during the interval $[t_1, t_2)$, and for any other flow j the relationship

$$\frac{S_i(t_1,t_2)}{S_j(t_1,t_2)} \geq \frac{\phi_i}{\phi_j} \quad , \quad \forall j$$

or equivalently

$$\frac{S_i(t_1,t_2)}{\phi_i} = \frac{S_j(t_1,t_2)}{\phi_j} \quad , \quad \forall j$$

is fulfilled.

This definition assumes that the flows can be served an infinitesimal amount of data. This is, however, impossible in real networks. Data is usually served packet-by-packet and for wireless networks the channel errors are location dependent, hence a backlogged flow may be prevented from transmitting data and the fairness criterion is, at least in the short timescale, violated. Hereafter we assume without loss of generality that $\sum_j \phi_j = 1$.

There exists several packet scheduling algorithms that approximates the GPS scheduler. These Packet Fair Queuing (PFQ) algorithms are usually implemented using a virtual time $V(t)$ which determines when the packet would be served if it had been served by the GPS scheduler. How this virtual time is implemented and how it is used varies with the different PFQ algorithms. Generally, the closer the algorithm emulates the GPS scheduler, the better the fairness is. An (extremely) good approximation of the GPS scheduler may be motivated in wired networks, but not in the wireless environment, especially not in the mobile systems. The radio communication channel with its location dependent errors is significantly worse, in terms of packet errors, than the fixed networks. The strive for achieving the best possible fairness possible by choosing a complex PFQ algorithm is difficult to motivate. This insight has lead to lead-lag strategies where the unfavored flow is compensated for lost time due to channel errors. The compensation is on behalf of the flows that are ahead of its fair share [7, 9].

The best known PFQ algorithm is Weighted Fair Queuing (WFQ) [4], also known as Packetized Generalized Processor Sharing (PGPS) [10]. However, it has been proven that WFQ does not provide the fairness index initially reported [1].

One of the PFQ algorithms is Start-time Fair Queuing (SFQ) and it is described in [3]. In the article the authors show that SFQ

- provides both short average delays and short maximum delays,
- provides fairness regardless of the variation in server rate,
- has a fairness measure as good as any other fair queuing algorithm, and
- has low complexity.

The SFQ algorithm uses a start tag and a finish tag which are associated with each packet and the packets are transmitted in increasing order of the start tags. Upon arrival of the k^{th} packet in flow f, p_f^k, it is stamped with a start tag, $S(p_f^k)$, defined by

$$S(p_f^k) = \max(V(A(p_f^k)), F(p_f^{k-1})) \qquad (1)$$

where $V(t)$ is the virtual time at real time t and is equal to start tag of that packet that is in service at time t, $A(p_f^k)$ is the arrival time of the k^{th} packet in flow f and $F(p_f^{k-1})$ is the finish tag of the previous packet in flow f. The finish tag is given by

$$F(p_f^k) = S(p_f^k) + \frac{L_f^k}{\phi_f C} \quad (2)$$

where $F(p_f^0) = 0$, L_f^k is the packet length in bits and C is the capacity of the channel. In a TDM system the packet length can be considered being equal to the constant slot length, $L_f^k = L \; \forall k, f$, and hence (2) can be rewritten as

$$\begin{aligned} F(p_f^k) &= S(p_f^k) + \frac{L}{\phi_f C} = \\ &= S(p_f^k) + \frac{1}{\phi_f R_{max}} = \\ &= S(p_f^k) + \frac{1}{r_f} \end{aligned} \quad (3)$$

where R_{max} is the maximum packet rate and r_f is the mean packet rate for flow f. Since the packets are transmitted in order of start time, SFQ essentially works according to the EDD principle where the start tags determines the deadline and execution order.

4 Call Admission

Prior to connection establishment a traffic contract is negotiated between user and network. The user provides traffic parameters that characterizes the traffic behavior together with demands on the wanted service, i.e. the QoS parameters. The admission control and resource manager then takes decision upon whether the network can provide the desired quality based on the current traffic status of the network and the expected behavior of the new traffic.

Commonly used traffic parameters, for example by the Integrated Services (IntServ) working group, are

- peak rate, p,
- mean rate, r, and
- maximum burst size, b.

Using the ATM notation and depending on service class these parameters may have different meaning. For Constant Bit Rate (CBR) services the peak rate parameter (ATM:Peak Cell Rate,PCR) defines the the highest rate at which cells can be generated and transmitted and the source can generate cells at PCR for any period of time. For Variable Bit Rate (VBR) services the PCR parameter is used together with the maximum burst size (ATM: Maximum Burst Size, MBS) parameter which defines the maximum number of bits that can be transmitted at PCR. The mean rate parameter (ATM: Sustainable Cell Rate, SCR) defines the upper bound on the average cell rate and over a long period of time the average number of cells transmitted per second should not exceed SCR. ATM uses more parameters than the three presented above.

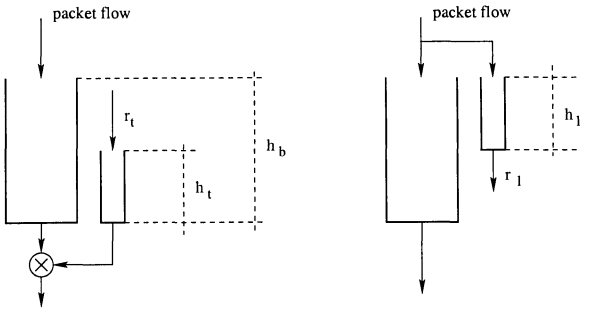


Figure 1: *Token bucket with limited queue length (left) and leaky bucket (right).*

4.1 Leaky Bucket Traffic Shaping and Conformance Tests

The packet flow is often filtered through a leaky bucket structure containing one or two leaky buckets to *i*) find those packets that are transmitted in conformance to the traffic contract and discard/mark those that do not, or *ii*) shape the traffic in accordance to the traffic parameters. The literature is often a little careless when using the term leaky bucket. Several types of leaky buckets exist where for example the so called token bucket is one, but it is sometimes used as synonymous to leaky bucket.

A token bucket is a leaky bucket with the following properties. It has a token height, h_t, which defines the maximum number of tokens allowed in the structure. A packet can pass if there exists any tokens and the token counter is decremented by one when the packet is passed. The token counter is incremented at a constant (deterministic inter arrival times) rate, r_t, but is never allowed a higher value than h_t, hence, the token bucket is a credit-based scheme. Given these parameters, the maximum number of packets that can be served during a time t is $h_t + r_t \cdot t$. With this definition of token bucket, we will hereafter use the term leaky bucket when describing a FIFO queue with queue length (bucket size) h_l and draining rate r_l. This structure is equivalent to a token bucket with parameters $(h_t, r_t) = (1, r_l)$. Further, the height of the queue length in a token bucket may also be limited, see Figure 1.

The conformance test in ATM is performed by a leaky bucket structure referred to as the Generic Cell Rate Algorithm (GCRA). The GCRA is presented using two parameters, GCRA(I, L). It may be described as using a leaky bucket of height L where each cell ar-

rival increments the bucket content with I units and the bucket is constantly drained at a rate of 1 unit per time unit. If the bucket can hold another I units upon arrival of a new cell the cell is said to be conforming to the traffic contract and it is kept for transmission. Note that the cell itself is not stored in the bucket. The GCRA is often used for flows with constant packet size (such as ATM) and it can be realized using a leaky bucket with parameters

$$h_l = (1 + L/I)l \quad , \quad r_l = l/I \tag{4}$$

where l is the packet size in units of data [5].

Using a token bucket, the token height is used to limit the burst size of packets. Given the peak rate, p, the token height is given by

$$h_t \quad = \quad (1 - r/p)b \tag{5}$$

which is easily seen by the fact that the maximum number of tokens generated during a time t, as shown before, must equal the number of packets generated during that time, $p \cdot t$, which yields $h_t = (p - r)t$. But t is the time defined by the maximum burst size, i.e. $t = b/p$, which gives (5). The same result holds for a leaky bucket.

However, instability may appear if the token bucket becomes empty. Any packets generated at a higher rate than r_t must wait for a token which sets the stability criterion for the token bucket

$$r \quad \leq \quad r_t \tag{6}$$

where there can only be equality if both the arrival and the service process are deterministic. Instability results in unbounded queue lengths and hence unbounded waiting times unless the token bucket is limited in size. This unstability is of course an unwanted phenomenon for scheduling algorithms such as EDD.

5 Conclusions

In this paper we have presented a survey of work-conserving schedulers intended for LMDS systems such as HiperAccess. Due to the fact that traffic from a wide variety of service classes and different demands on QoS should be able to coexist, it is crucial to implement traffic shaping and conformance tests in order to make the traffic somewhat more predictable. The lack of such police function would result in a system with possibly infinite delays and an incorrect call admission control. Such rate-limiting structures based on the leaky bucket and token bucket principles are also discussed in this paper.

References

[1] J. C. R Bennett and H. Zhang. Why WFQ is not good enough for integrated services networks. In *Proceedings of the 6th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV), Shonan Village International Conference Center, Zushi, Japan*, April 1996.

[2] David D. Clark, Scott Shenker, and Lixia Zhang. Supporting real-time applications in an integrated services packet network: Architecture and mechanism. In *SIGCOMM*, pages 14–26, 1992.

[3] P. Goyal, H. M. Vin, and H. Cheng. Start-time fair queuing: A scheduling algorithm for integrated servicespacket switching networks. In *ACM SIGCOMM*, volume 26(4), pages 157–168. ACM, ACM, October 1996.

[4] S. Keshav. *An Engineering Approach to Computer Networking.* ISBN 0–201–63442–2. Addison-Wesley, 1997.

[5] J.-Y. Le Boudec and P. Thiran. *Network Calculus – A Theory of Deterministic Queuing Systems for the Internet.* ISBN 3–540–42184–X. Springer Verlag, 2001.

[6] C. L. Liu and James W. Layland. Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the Association for Computing Machinery*, 20(1):46–61, January 1973.

[7] Songwu Lu, Vaduvur Bharghavan, and Rayadurgam Srikant. Fair scheduling in wireless packet networks. *IEEE/ACM Transactions on Networking*, 7(4):473–489, 1999.

[8] A. K. Mok and M. L. Dertouzos. Multiprocessor scheduling in a hard real-time environment. In *Proc. 7th IEEE Texas Conf. Comput. Symp.*, pages 5–1–5–12, 1978.

[9] T. S. E. Ng, I. Stoica, and H. Zhang. Packet fair queueing algorithms for wireless networks with location-dependent errors. In *Proceedings IEEE INFOCOM '98*, volume 3, pages 1103–1111. IEEE, IEEE, 1998.

[10] Abhay K. Parekh and Robert G. Gallagher. A generalized processor sharing approach to flow control in integrated services networks: the single node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, 1993.

Evaluating a Performance Analysis of Slow FH Systems by Simulations

André Stranne[1,2], Fredrik Florén[1,2], and Ove Edfors[1]

[1]Dept. of Electroscience, Lund University, Box 118, SE-221 00 Lund, Sweden. E-mail: andre.stranne@es.lth.se

[2]Telia Research AB, Box 94, SE-201 20, Malmö, Sweden

Abstract

In this paper the approximate and simulated throughput of Bluetooth networks, which employ a packet-based frequency-hopping scheme, are compared. The approximate throughput is given as a function of the number of adjacent networks, the packet types available, and the probability of networks selecting a certain packet type for transmission. Furthermore, the approximation was derived under the assumption that collisions result in a total loss of the data in the colliding packets. In the simulations, the throughput of a reference network in the presence of interfering networks was measured. Free-space path loss was assumed and the interfering networks were randomly placed on a circular area with the reference network at its center. It was concluded that the approximation gives reasonable values on the throughput, with increasing accuracy the closer the interferers are to the reference network units.

1 Introduction

It is expected that the number of electronic consumer products with the ability to interconnect using short-range radio interfaces will increase dramatically in the near future. One technique that will be used in these interfaces is slow frequency hopping, where frequency hops are performed on a packet basis. It enables uncoordinated use of several networks in the same frequency band.

However, interference is unavoidable and will increase with the number of transmitting units, with decreased throughput as a result. Naturally, an expression that quantifies the impact of the interference on the throughput is desired. Therefore, we have in a previous paper [2] derived an approximate expression for the throughput as a function of the number of collocated units. The derivation was performed for the case where units can transmit packets of varying lengths. Furthermore, it was assumed that the slightest overlap in time and frequency, i.e., a collision, between packets results in a total loss of the information contained in the colliding packets.

In order to evaluate the validity of the system model introduced in [2], or more specifically the assumption that collisions result in lost packets, we in this paper compare the approximate throughput of a system of Bluetooth networks with the throughput obtained by more realistic simulations. The results give indications for which scenarios the system model is valid. In the simulations, Bluetooth networks are randomly placed over a circular area of a given radius and the signal attenuation is given by the free-space path loss.

In all, this paper indicates when the approximate throughput agrees well with more realistic system models, and a deeper understanding of frequency-hopping systems that transmit packets of varying lengths is given. Herein lies the novelty of our work since previous analyses of slow frequency-hopping systems have been limited to one packet size.

The outline of the paper is as follows. The system model is given in Section 2 and the throughput analysis is summarized in Section 3. The application example along with the simulations are presented in Section 4 and, finally, some concluding remarks are given in Section 6.

2 System Model

Our system is defined to consist of N networks. We further define a network to consist of an arbitrary number of units that communicate without interference. Specifically, there is exactly one ongoing transmission within a network at a time. This implies that units within a network are coordinated in some manner.

How the resources are divided between the network units or the resulting throughput of a specific unit is not of interest here; the performance of the networks and of the system is considered as a whole. The term network transmission is used for a transmission by any unit within a network.

Networks transmit packets that consist of the following components: a header of length h_i, a payload of length l_i, and a guard interval of length d_i, where i refers to the ith packet type. The components of a type i packet can be seen in Figure 1. The inter-

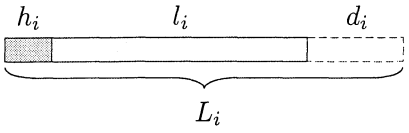


Figure 1: Definition of a packet and its components: header, payload, and guard interval. Their lengths are h_i, l_i, and d_i, respectively, and the sum of the lengths equals L_i.

val $h_i + l_i$ during which the transmitter is active, i.e., transmitting, is referred to as the *active interval.* In contrast, the transmitter is idle during the guard interval. The lengths l_i, h_i, and d_i are here specified in seconds, and $L_i = h_i + l_i + d_i$ is used for the entire packet length.

The probability of selecting a packet of type i for transmission is denoted r_i. The packet types and their corresponding probabilities are henceforth referred to as a *packet type distribution.* Further, it is assumed that networks always have packets to transmit and they are only idle during the guard interval.

When a new packet is to be sent, a packet type is selected from the set of available packet types and a new frequency channel is also selected. The new channel is selected from a set of q channels and all channels are selected with the same probability. In other words, a packet-based *slow* frequency-hopping scheme is employed. The scheme is termed slow since hops are not performed after each symbol or parts of a symbol, but after each packet. We further assume that all networks select their channels independently of all other networks.

Transmissions are not synchronized between networks, i.e., there is no coordination between networks of the start of packet transmissions.

In the analytical throughput analysis we assume that should there occur a collision, all packets involved in the collision will be destroyed and the information contained in the colliding packets is lost. On the other hand, if no collisions occur packets are assumed to be correctly received. In the simulations we use more realistic assumptions and take into account the amount of interfering energy when determining if a packet is successfully received. These two approaches are further detailed in the coming two sections.

3 Throughput Analysis

The method employed for obtaining an expression for the throughput of FH systems is the introduction of a variable-length reference packet, for which an approximation of the probability of successful transmission is found.

Consider a system of N networks. One of these networks will act as a reference network and the other $N-1$ networks will act as interferers. It is assumed that the interferers all use the same packet type distribution although extending the analysis to a case without this restriction poses no major difficulties.

Here follows the results of the derivation of the approximate throughput as presented in [2].

Let T be the length of the active interval of the reference packet and $P(\mathrm{S};T)$ be the probability of successful transmission of reference packet of length T. This probability can be expressed as

$$P(\mathrm{S};T) = \sum_{n=0}^{\infty} p_{\mathrm{tot}}(n;T)\left(1-\frac{1}{q}\right)^n. \qquad (1)$$

Here, $p_{\mathrm{tot}}(n;T)$ PDF for the total number of overlaps generated by the interfering networks during the transmission of the reference packet and $(1-1/q)^n$ is the probability of successful transmission given n overlaps.

Although the exact probability function of the sum of the number of overlaps $p_{\mathrm{tot}}(n;T)$ is possible to evaluate, it it quite tedious and the result is difficult to analyze. Therefore, $p_{\mathrm{tot}}(n;T)$ is approximated with a Dirac pulse located at the mean of the number of overlaps, i.e., at $\bar{n}_{\mathrm{tot}} = \sum_{n=0}^{\infty} n p_{\mathrm{tot}}(n;T)$.

By applying the approximation, (1) reduces to

$$P(\mathrm{S};T) = \sum_{n=0}^{\infty} p_{\mathrm{tot}}(n;T)\left(1-\frac{1}{q}\right)^n \approx \left(1-\frac{1}{q}\right)^{\bar{n}_{\mathrm{tot}}}, \qquad (2)$$

and the approximation of $P(\mathrm{S};T)$ is denoted $\tilde{P}(\mathrm{S};T)$, i.e.,

$$\tilde{P}(\mathrm{S};T) = \left(1-\frac{1}{q}\right)^{\bar{n}_{\mathrm{tot}}}. \qquad (3)$$

As derived in [2], the average total number of overlapping packets $\bar{n}_{\mathrm{tot}}$ generated by the $N-1$ interfering networks is given by

$$\bar{n}_{\mathrm{tot}} = (N-1)\frac{T + \sum_{i=1}^{M} r_i (L_i - d_i)}{\sum_{k=1}^{M} r_k L_k}, \qquad (4)$$

where M is the number of packet types.

Substitution of (4) into (2) yields a closed form approximation for the probability of successful transmission of a reference packet whose active interval length is T

$$\tilde{P}(\mathrm{S};T) = \left(1-\frac{1}{q}\right)^{(N-1)\frac{T+\sum_{i=1}^{M} r_i(L_i-d_i)}{\sum_{k=1}^{M} r_k L_k}}. \qquad (5)$$

The throughput, which is denoted by R, is the mean fraction of the channel time that is used for successful transmission of data in the payloads. Hence, throughput is given by the ratio between the mean length of

Parameter	Value
Number of channels	$q = 79$
Header length	$h = 160\,\mu$s
Guard interval	$d = 220\,\mu$s
Payload lengths	$l_1 = 250\,\mu$s
	$l_2 = 1500\,\mu$s
	$l_3 = 3000\,\mu$s
Packet type probabilities	$r_1 = r_2 = r_3 = \frac{1}{3}$

Table 1: The Bluetooth parameters.

the successfully transmitted payloads and the mean packet length, including the guard interval. If the reference network uses the same packet type distribution as the interfering networks, we have

$$R = \frac{\sum_{n=1}^{M} r_n l_n P(\mathrm{S}; L_n - d_n)}{\sum_{k=1}^{M} r_k L_k}. \quad (6)$$

Finally, by using the approximation (5) in (6), the approximate throughput, $\tilde{R}$, for the reference network becomes

$$\tilde{R} = \frac{\sum_{n=1}^{M} r_n l_n \left(1 - \frac{1}{q}\right) \left(\frac{L_n - d_n + \sum_{i=1}^{M} r_i (L_i - d_i)}{\sum_{k=1}^{M} r_k L_k}\right)^{(N-1)}}{\sum_{k=1}^{M} r_k L_k}. \quad (7)$$

4 Simulations

In order to investigate the validity of the system model presented in Section 2, a Bluetooth scenario is examined. The throughput of a Bluetooth reference network in the presence of other interfering Bluetooth networks is both calculated and simulated. It is assumed that all networks, including the reference network, use the same packet type distribution.

The parameters roughly describing Bluetooth networks, can be found in Table 1 and these are used in (7) to obtain the throughput for $N - 1$ interfering networks. Note that all packet types in the packet type distribution have the same header length, h, and guard interval, d.

The simulation setup is as follows. At the center of a circle with a radius of ρ meters a reference network is placed. The reference network, as well as the interfering networks, consist two communicating units and the distance between the two units of the reference network is y meters. The interfering units are randomly placed according to a uniform distribution over the surface of the circle. This is illustrated in Figure 2.

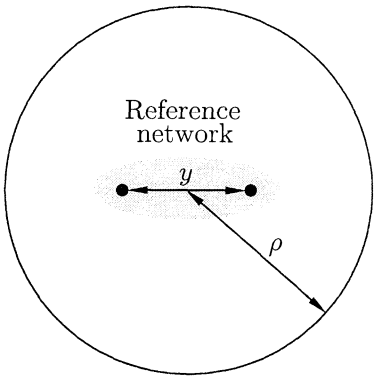


Figure 2: Illustration of the simulation setup where the reference network is placed at the center of a circle of radius ρ. The distance between the two units of the reference network is y.

Parameter	Value
Transmit power	P_{t} =1 mW
Minimum received power	$C_{\min}$ =-70 dBm
Minimum SIR	$\mathrm{SIR}_{\min}$ =11 dB
Center frequency	f_c =2.4 GHz

Table 2: The Bluetooth radio parameters used in the simulations.

For simplicity we assume that the transmission of a packet is successful if the requirements for both the minimum received power and the minimum signal-to-interference ratio (SIR) are satisfied. This is not entirely realistic either, but it is sufficient for the comparison in this paper. The parameter values used in the simulations, including transmit power and center frequency, are given in Table 2. Further details about the radio parameters can be found in [1].

To determine if a packet has been successfully received in the simulations, both the received useful energy in the packet, E_C, and the interfering energy generated by the interferers, E_I, must be calculated. We assume a static environment and the path loss, Λ, between two units is assumed to be given by the free-space path loss between isotropic antennas. Hence,

$$\Lambda = \left(\frac{4\pi x f_c}{c}\right)^2, \quad (8)$$

where x is the distance between the units, c is the speed of light, and f_c is the center frequency, which we have set to 2.4 GHz in the simulations.

The received useful power, C, at the receiver of a reference packet is

$$C = \frac{P_{\mathrm{t}}}{\Lambda_{\mathrm{ref}}}, \quad (9)$$

where Λ_{ref} is the path-loss between the two units in the reference network and P_{t} is the transmitted power.

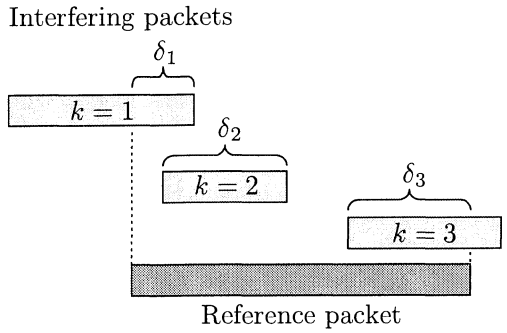


Figure 3: Three interfering packets overlapping a reference packet in both time and frequency. The length δ_k of the overlap from each packet k is shown.

Then, if the payload length of the reference packet is l_{ref}, the received useful signal energy is

$$E_C = C(h + l_{\text{ref}}). \tag{10}$$

The interfering power generated by the kth packet overlapping the reference packet in time and frequency, I_k, is given by

$$I_k = \frac{P_{\text{t}}}{\Lambda_k}, \tag{11}$$

where Λ_k is the is the path-loss between the receiver of the reference packet and the unit transmitting the kth interfering packet. Then, if there are K interfering packets overlapping the reference packet in time and frequency, the total interfering energy received is

$$E_I = \sum_{k=1}^{K} I_k \delta_k, \tag{12}$$

where δ_k is the length of the overlap by the kth interfering packet, as illustrated in Figure 3.

For each packet received by a unit in the reference network, the SIR is calculated as

$$\text{SIR} = \frac{E_C}{E_I} \tag{13}$$

and compared to the threshold, $\text{SIR}_{\min}$, specified in Table 2. As mentioned above, we assume that the packet is successfully received if both

$$C \geq C_{\min} \tag{14}$$

and

$$\text{SIR} \geq \text{SIR}_{\min}. \tag{15}$$

The condition (14) in combination with the distance attenuation model used, results in a maximum distance $y_{\max}$ between the reference network units for possible packet transmissions. This distance can be calculated using (8) and (9) to be $y_{\max} = 31.5$ m. If

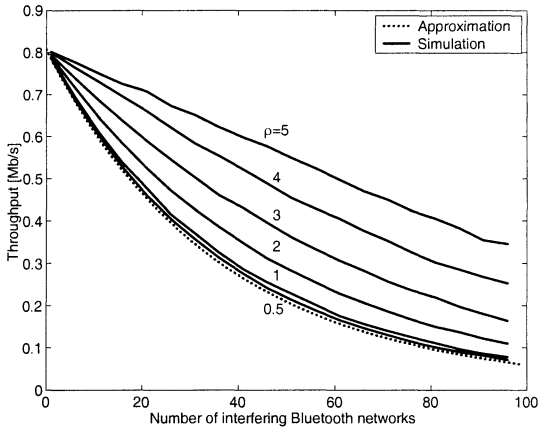


Figure 4: Throughput from both approximate expression and simulations as a function of number of interfering networks. The distance between the reference network units is $y = 1$ m. It can be seen that the agreement between simulations and analytical expression is high only for a small radius, ρ, of the area where the interferers are placed.

the reference network units are farther away from each other than $y_{\max}$, then throughput will be zero since the condition (14) will not be fulfilled for any packet transmissions within the reference network. In our simulations, the network units will always be closer than the maximum of 31.5 m, hence making the SIR the limiting parameter. The careful reader have noticed that 31.5 m is more than the stated 10 m reach of Bluetooth, which is a result of the simplified conditions under which we perform the simulations.

5 Results

The results of the simulations and the results from the approximative expression are shown in Figure 4. The distance y between the units of the reference network is 1 meter and it can be seen that the simulated throughput for the reference network deviates from the approximation quite substantially when the interfering networks are spread over the larger simulation areas, i.e., when the radius ρ is large. However, for smaller radii of the simulation area, the throughput is well approximated by the expression in (7). This is of course a consequence of putting more networks in the immediate neighborhood of the reference network, making each packet collision more severe in terms of SIR.

Common for the simulated curves is that the approximation seems to be a lower bound for the throughput, which agrees well with the assumptions used in the analytical derivation of throughput.

6 Conclusions

The results in Section 5 show that the approximate analytical expression for the throughput of packet-based slow frequency-hopping networks derived in [2] gives reasonable values when the impact of the interference from adjacent networks is severe. As we have seen, this is the case when the interfering networks are close to the reference network. For our setup, with 1 m between the reference units, the accuracy is very good if the interferers are within a radius smaller than about 1 m. In terms of Bluetooth user scenarios this implies that our simplified analytical model should be useful when analyzing, *e.g.*, body-area networks where we can expect the communicating units to be located within these dimensions. Further, the analytical expression is conservative in the sense that it does not over-estimate the throughput. This investigation is of course limited to a special case, but it still indicates the possibility of a more general use.

This initial study will be extended with more detailed comparisons and more realistic simulation setups.

References

[1] BLUETOOTH SPECIFICATION version 1.0. http://www.bluetooth.com.

[2] F. Florén, A. Stranne, O. Edfors, and B.-A. Molin. Throughput analysis of strongly interfering slow frequency-hopping wireless networks. In *Proc. of the 53rd IEEE Vehicular Technology Conference*, volume 1, pages 496-500, Rhodes, Greece, 2001.

Handover in existing Wireless Systems and a look forward

Henrik Persson[1] and Johan M Karlsson[2]

Department of Communication Systems, Lund Institute of Technology, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden E-mail: {henrikp[1], johan[2]}@telecom.lth.se.

Abstract

Future generation wireless systems will consist of several different kinds of systems. The mobile terminals are supposed to be able to communicate with local area networks as well as wide area networks, in a seamless manner. This is preferably done through the concept of handover, which is the technique to switch from one service area, cell, to another without loosing the connection. Handover is so far mostly performed within the same system, but handover between diverse network technologies are to be expected. The different handover techniques deployed to existing wireless systems are discussed. Further, some remarks regarding the handover procedure between different wireless systems are finally expressed.

1 Introduction

The future generations of wireless communication systems will most probably be a connection of several individual networks. The infrastructure will consist of wireless local area networks (WLANs) as well as systems covering large areas, such as universal mobile telecommunication systems (UMTS). Most of these systems use the concept of access points that divide the service area into different corresponding cells. One of the primary reasons to adopt cellular structure is to increase the effective total bandwidth by using different frequencies in different cells. This concept, known as frequency reuse, also makes it possible for mobile terminals to change between different frequencies as well as systems. Furthermore, frequency reuse not only saves the spectrum but also reduces transmission power by reducing cell size. A function that allows a mobile node to communicate with the access point in a cell and then switch to the access point in another cell is called handover. The purpose of the handover is to keep continuous or seamless service to mobile nodes through different cell coverage. Handover is consequently a special feature to deal with the mobility issue for wireless networks.

The increasing interest in the technologies beyond 3G has made the handover problem between wireless systems a hot spot in R&D. To take care of this new sort of handovers, it is important to evaluate the different types of handovers that already exist and how they work in the wireless environment of today. Todays wireless environment, that now sees a larger number of mobile terminals (MTs) and then consequently also a larger number of handovers.

The mobility in different wireless systems vary a lot, depending on if it covers an area outside or inside a building. The MTs in an office, covered by a WLAN, e.g., have no rapid differences in speed and direction, that might happen to MTs in a wireless system covering a highway.

The demand for mobility increase, e.g., cell phones and laptops are nowadays very common. A rising demand for mobile devices to go between different wireless systems have occured. This will increase the demands on the developers on equipment, e.g., will the MTs need double transmitters and support double technologies. The operators will also be forced to cooperate with other operators and also with owners of private WLANs. The cooperation is necessary, since a handover is made between different technologies or operators. Another problem that can occur is when a private WLAN is used inside a building and an external operator outside. The operators then want the MTs within the building to use their system, instead of the offices own free WLAN.

To neglect these demands could be devastating for developers. The mobility in cars and for pedestrians are well known, but now when the MTs will be handed over when they get transported through the front door it is a new challenge.

In Section 2 a brief discussion about handovers in general take place. Section 3 focuses on different types of handover, and Section 4 on handover in different wireless systems. In Section 5 handover between wireless systems are discussed and finally in Section 6 some conclusions are drawn.

2 Handover

Handovers can be necessary for many reasons, e.g., the MT moves out of reach of the current base station or the signal strength at another frequency is more favourable. These reasons are typical MT initiated, but a handover can also be initiated, e.g., by the network. The network can, e.g., initiate a handover if it detects too much congestion in one cell, then it order one or more MTs to be handed over to another cell, if possible.

A handover can be more or less difficult to make, depending on if it is just a handover within, e.g., the same access point (AP) or if it has to be rerouted to a higher level, e.g., mobile switching center (MSC). In the next generation of handover it will be more complicated when the handovers also occur between different systems, covering different areas. This type of handover can hand over MTs that want to have a faster, cheaper or more stable connection, to another type of wireless system, which will bring another dimension to handovers.

In [1] three phases of a handover are discussed, initiation, decision and execution. The initiation phase has the objective to recognize the need for handover and then initiate it. The decision phase compares information, user preferences and network polices, from a variety of sources. Considering the information, the best available handover is identified, then the execution phase is instructed to perform the handover. In the execution phase the new connection is made, from the information in the decision phase. Cell, code, technology or network is changed in a top-down approach, to make sure that the level with the best overview controls the handover.

An effect that might occur, during handovers is the ping-pong effect. It happens when several handovers occurs back and forth between two APs. This cause a heavy load on the network and does not really make the MT's radio signal stronger.

3 Types of handover

Different types of handover are used in the technologies evaluated. The article present different handover techniques and in this section three of the most common, hard, soft/softer and interfrequency handover, are presented.

A technique to provide "large" and "small" cells at a single location, is to use different heights on the antenna and different power levels. This is called the umbrella cell approach, that covers both the high speed vehicles and the slower users [2]. The technique ensures the users, with high speed, to be handed over as seldom as possible, using macro cells, and provides the users, with slow speed, with micro cell channels. The macro cell cover the smaller micro cells, as can be seen in Figure 1.

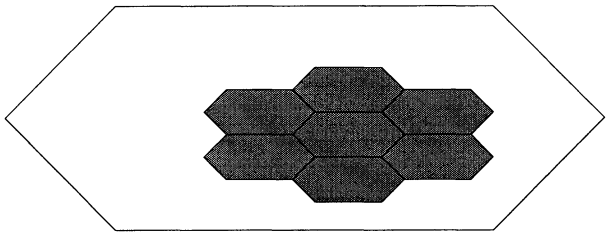


Figure 1: *Macro cell covering the micro cells*

3.1 Hard handover

In hard handover the connection to the current cell is broken before the connection to the new cell is made. This is called break-before-make. During hard handover the MT not only change cells, it also change transmission frequency and for a short moment, when the cell is changed, the MT is not connected to a traffic channel. At this moment it is most likely that the connection is being dropped.

3.2 Soft handover

In soft handover an MT is temporarily connected to both the original cell and the new cell. Instead of the hard handovers break-before-make, a make-before-break is used. Where the MT is connected to the old AP, during the time it takes to set up the connection with the new AP. Soft handover is, e.g., used in code division multiple access (CDMA) to reduce the interference into other cells [3]. A soft handover between two sectors, covered by the same base station, is called a softer handover.

The soft handover can last during the rest of the call if the MT remains in the area covered by two cells. The soft handover not only provides undetectable handover, it also cover the difficult areas between the base stations effectively [4].

Soft handover is most often related to CDMA systems, and can also be used in time division multiple access (TDMA) systems. Since the cost and power consumption will increase, though it would typically require two radio interfaces, the soft handover is not very common in TDMA systems [5].

3.3 Interfrequency handover

To better utilize hierarchical cell structures with macro, micro and indoor cells, interfrequency handover is needed. The micro cells will have different frequencies than the overlaying macro cell. Also a hot-spot cell could have more frequencies than the neighboring cells. It

will also be needed to take care of handovers from 3G-systems and systems beyond 3G to second generation systems and also handovers the other way around. The interfrequency handover has its origin in the fact that the MTs have to be able to measure the quality of another carrier frequency, while the current connection is maintained.

Two methods are considered for making measurements on other frequencies while still having the connection running on the current frequency, dual receiver and slotted mode [3]. In dual receiver, one receiver branch switch to another frequency to do measurements to see if the new frequency is more favourable than the current. The other receiver continues to receive at the present frequency. Dual receiver is considered attractive for mobiles with antenna diversity and the advantage of this approach is that the current frequency connection is not broken. The slotted mode approach uses idle periods to do interfrequency measurements and is suitable for MTs without antenna diversity.

4 Handover in Wireless Systems

To develop new handover strategies for the next generation of wireless systems, an investigation of the existing technologies is necessary, to avoid mistakes made in the past. A short survey of some proposed and existing handover techniques in a few well known wireless systems is presented in the following section.

4.1 GSM and GPRS

In the global system for mobile communications (GSM) standard about 40 reasons for a handover have been identified, but the two basic reasons are that the MT moves out of range for a base transceiver station (BTS) or a certain antenna respectively and when the handover occurs because of balancing the load in a cell [6].

In [6] four types of handover scenarios in GSM are identified, both internal and external, see Figure 2. When interference makes it impossible to transmit at a certain frequency, the base station controller (BSC) can change the carrier frequency. This is called an intra-cell handover (scenario A). Another internal handover, called inter-cell, intra BSC handover, can occur when the MT moves to another cell, without changing the BSC (scenario B). The inter-BSC, intra-MSC handover handle the switch from one cell to another, when the MT change cells controlled by different BSC (scenario C). The second external handover scenario is the intra MSC handover, where the MT change cell and the cells belong to different MSCs (scenario D).

General packet radio service (GPRS) uses the same scenarios that are used by GSM.

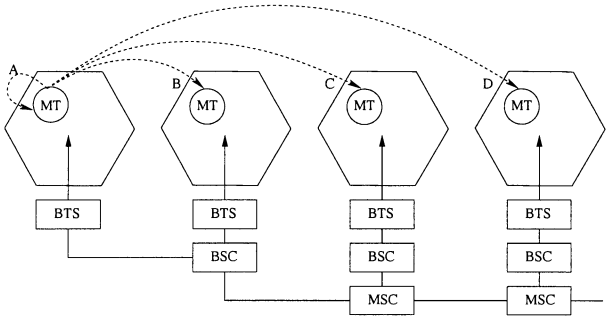


Figure 2: *Types of handover in GSM*

4.2 UMTS

In UMTS, both occuring as frequency division duplex (FDD) and time division duplex (TDD), soft, hard and interfrequency handover are supported (see Section 3). The handover decision is based on measurements made by the user equipment and the radio network controller and various cell specific parameters. This is the strategy used by the networks radio link control, but network directed handover can also occur for other reasons than the radio link control, e.g., to make sure that the congestion in one cell is not to high. Eight possible types of handover are possible [7].

- 3G-3G (handover within the UMTS network)
- FDD soft/softer handover (to support UMTS/FDD)
- FDD inter-frequency hard handover (to support UMTS/FDD)
- FDD/TDD handover (to support handover between UMTS/FDD and UMTS/TDD systems
- TDD/FDD handover (to support handover between UMTS/TDD and UMTS/FDD systems)
- TDD/TDD handover (handover to support UMTS/TDD)
- handover 3G-2G (e.g. handover to GSM)
- handover 2G-3G (e.g. handover from GSM)

4.3 Wireless ATM

In [6], three basic handover scenarios, for wireless asynchronous transfer mode (WATM), are mentioned. The scenarios remind about the ones for GSM and GPRS in Figure 2 and with some small adjustments,

in the figure, it can be used again, see Figure 3. Handover scenario A occurs when the MT moves from one radio transceiver (RT) to another RT, within the same AP and end-user mobility-supporting ATM switch-edge (EMAS-E). It is called the intra-EMAS-E/intra-AP scenario and reminds of the GSMs internal handovers. To switch between the two RTs is the APs main task. The handover scenario B is the intra-EMAS-E/inter-AP, where the MT moves between two RTs, not within the range of the same AP. In this scenario the RTs are connected to different APs, which are connected to the same EMAS-E. The EMAS-E hand over the connection from the old AP to the new. Scenario C, is the inter-EMAS-E/inter-AP, which handle handovers that involve different APs. Another switch called the cross over switch has to manage the handover between the switches.

Handover types that WATM support are, e.g., hard handover (see Section 3.1), network initiated and backward handover [6]. Network initiated handover occurs when the network decides that, e.g., the signal strength is to low or when the load on the current network is too high. In backward handover the MT continues to maintain the connection during the handover and switches over to a new AP when all entities involved are ready for the handover. WATM does not support soft handover.

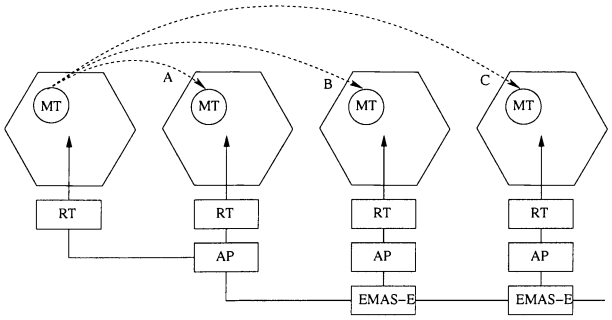


Figure 3: *Types of handover in WATM*

4.4 IEEE 802.11

Handover in an IEEE 802.11 system is, for example described in [8]. In an IEEE 802.11 system the MT is connected to an AP, which is connected to a wired backbone. Periodically (typical every 100 ms) the AP broadcast a beacon signal. After scanning the beacon signal, the MT associates with the AP with the strongest beacon. A lot of information, corresponding to the AP, is contained in the beacon signal and it is used by the MTs to distinguish the APs.

The MT observes the parameter received signal strength (RSS) from the beacon of its associated AP and when the RSS decreases the MT begins to scan, active or passive, the neighboring APs for a stronger beacon. Passive scanning is when the MT listens to available beacons. In active scanning the MT sends out a probe request to a selected number of APs, capable of receiving the probe. APs receiving the probe request respond with a probe response. Based on the strongest beacon or probe response the MT decides which the new AP will be. It sends a reassociation request, to the new AP, with information about the MT and the old AP. The new AP responds with a reassociation response. To indicate that a handover has taken place the inter-access-point protocol employs two protocol data units (PDUs). The old AP is informed by these PDUs, sent over the wired network, from the new AP.

4.5 HiperLAN/2

The main goal for HiperLAN/2 is that all applications, which today run over fixed infrastructure also can run over HiperLAN/2 network [9].

The MT uses the AP with the best signal to noise ratio. When the MT moves around it may order a handover to another AP, if it detects that the new AP has a stronger radio signal than the old AP. Packet loss can occur during handover.

HiperLAN/2 supports two forms of handover; reassociation and handover via the signalling across the fixed network [9]. Reassociation means that the MT starts all over again with listening to the broadcast channel from different APs and then chooses the AP with the best radio link quality. The second type of handover lets the new AP, to which the MT wants to be handed over to, get the association and connection information from the old AP. This is done by sending the information across the fixed network. Then the new AP is provided a fixed network address, e.g., an IP address, from the MT to make sure that there is a connection between the new and the old AP. This type of handover assures a fast handover and in the handover phase it minimize the losses of user plane traffic.

5 Handover between Wireless Systems

This is one of the most interesting areas in the new generation of communication. A couple of articles, e.g., [3, 8] present ways to integrate different wireless systems with each other and to make handover between the systems. To make this sort of handover seamless is one of the main issues for the research in this area. Seamless is really a key word, because if the connection is broken during the handover or has larger probability to be broken, than the handover between the same systems, nobody will make the attempt of changing systems.

One way to solve this sort of handover is to make umbrella cells. Where the macro cell has one technology, e.g., UMTS or GPRS that can be used by the high speed users in the macro cell and the low speed users use, e.g., HiperLAN/2 or IEEE 802.11a in the micro cell. An interfrequency handover between the technologies that form the umbrella cells can be used. With this model certain problems occur, e.g., who will charge the call, if the macro cell and the micro cells have different operators.

6 Conclusions

Since it already exist dual-band cell phones, that can be used in more than one frequency band, the step to manage two standards like UMTS and GSM or GPRS from the same cell phone, might not be so long. If the standardization groups can agree this might not be so far in the future. In the UMTS standard [7] there already exist certain handover strategies, that can be used for handovers between a GSM system and an UMTS system, in both directions. The next step will be to make hierarchical cell structures, like umbrella cells, with macro and micro cells. Where the macro cells typically can be UMTS systems and the micro cells can be covered by WLANs.

The handovers for future generations of mobiles faces a lot of problems, but they are not invincible, the possibilities for them are so great and the potential so high that the technologies will be brought to a new dimension. The handover will not be just to hand over a connection from one cell to another, it will be based on how fast the connection is, the price for the connection and the quality of the connection.

References

[1] P. Reynolds, "Mobility Management for the Support of Handover within a Heterogeneous Mobile Environment", 3G Mobile Communication Technologies, 2000. First International Conference on (Conf. Publ. No. 471), pp 341-346, 2000.

[2] T.S. Rappaport, "Wireless Communications Principles and Practice", Prentice Hall PTR 1996, ISBN 0-13-375536-3

[3] R. Prasad and T. Ojanperä, "An Overview of CDMA Evolution toward Wideband CDMA", IEEE Communications Surveys, Fourth Quarter 1998, Vol.1 No. 1.

[4] J.D. Gibson, Editor-in-Chief, "The Mobile Communications Handbook", CRC Press, 1996, ISBN 08493-8573-3.

[5] H. Mitts, H. Hansen, J. Immonen, S. Veikkolainen, "Lossless handover for wireless ATM", Mobicom'96, pp 85-96, Nov 1996.

[6] J. Schiller, "Mobile Communications", Addison-Wesley 2000, ISBN 0-201-39836-2

[7] ETSI TR 125 922 V3.6.0 (2001-09), "Universal Mobile Telecommunications Systems (UMTS); Interlayer procedures in Connected Mode (3GPP TS 25.303 version 4.3.0 Release 4)".

[8] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. Mäkelä, R. Pichna, J. Vallström, "Handoff in Hybrid Mobile Data Networks", IEEE Personal Communications Magazine, Vol.7 No.2, pp. 34-47, April 2000

[9] M. Johnsson, "HiperLAN/2 - The Broadband Radio Transmission Technology Operating in the 5 GHz Frequency Band", HiperLAN/2 Global Forum, 1999, Version 1.0

The Need for an Alternative to IPv4-Options

Pierre Fransson[1] and Andreas Jonsson[2]

[1]Luleå University of Technology 97187 Luleå, Sweden E-mail: pierre@sm.luth.se

[2]Luleå University of Technology 97187 Luleå, Sweden E-mail: aj@sm.luth.se

Abstract

IPv4 options are a generic and simple way of transmitting out-of-band information pertinent to network layer components, e.g., routers. Besides traditional options, such as loose source routing, record route and Internet timestamp, IPv4-options have also recently and in the past been proposed for several other uses. A small and non-exhaustive list of these uses include: conveying information between different layers in the networking stack, providing quality of service in ad hoc networks, propagating the priority of a real-time media stream to a receiver and as method of enabling Active Networks. However, unfortunately IPv4 packets containing options are reputed to be treated as second class citizens in the networking world, supposedly leading to the following problems: increased delay, increased jitter and dropped packets.

Despite the above alleged problems with IP options there does not seem to appear to exist any substantial amount of measurements on the impact of IP options. To verify the view that packets with options do indeed receive inferior treatment from routers, we have conducted a series of simple experiments, measuring delay and packet drops when using certain options. These measurements corroborate the general view that packets containing IP options are second class networking citizens.

Based on these measurements and the fact that there apparently exists a need for transmitting out-of-band data, we propose an alternative to IPv4 options that is not affected by delay or packet drops and which would therefore be more suitable to use.

1 Introduction

IPv4 options are a generic and simple way of transmitting out-of-band information pertinent to network layer components, e.g., routers and hosts. Besides the options specified in [1], such as loose source routing, record route and Internet timestamp, IPv4 options have also recently and in the past been proposed for several other non-standardized uses. A small and non-exhaustive list of these uses include: conveying information between different layers in the networking stack [2], providing quality of service in ad hoc networks[3], propagating the priority of a real-time media stream to a receiver [4], enabling truncation of packets [5] and as method of enabling Active Networks [6]. However, unfortunately IPv4 packets containing options are reputed to be treated as second class citizens in the networking world, supposedly leading to the following problems: increased delay, increased jitter and dropped packets. Despite these alleged short-comings of IPv4 options there does not appear to exist any recent measurements of their impact.

The generally referred to reason, as to why packets containing options are exposed to delay, jitter and drops, is that routers are unable to handle options using fast-path forwarding. A reason for this is that when router hardware was developed in the end of the 1980's, IP options were not commonly used. Therefore router designers optimized for the common case, i.e., packets without options, thereby reducing complexity. Furthermore since options were rarely used, there did not seem to be any reason to handle options at line speed [7].

Since we believe that there is a need of a generic mechanism to relay out-of-band data[1], which does not lead to adverse treatment of the packets, we have set out to examine the impact of using options on IPv4 packets. To this end a series of simple experiments have been conducted using a modified version of the ICMP-message *ping* [8] to measure the delay incurred by using options and to examine if routers actually drop packets carrying options to a greater extent than otherwise.

Even though our experiments are still at an initial stage, we belive that it is quite clear from our collected data that packets carrying options are in fact both delayed to a greater extent and dropped more often than packets without options. While a possible performance penalty, e.g., extra delay, may not always be of consequence, there are times when it does matter. Examples of such cases are the options used in *Hints and Notifications* [2] and *Selective Truncating Protocol* [5], which would typically be used for real-

[1]The above short list gives at least some support for this notion

time streaming media and would therefore suffer from extra incurred delay.

In this paper an alternative is suggested to the traditional IPv4 option, to be used mainly for options that are not yet standardized, which does not suffer the drawbacks of the current IPv4 options. Our proposal is based on the use of the reserved bit [1] in the IP-header, which is currently unused[2], to indicate the existence of options and a layout of the options themselves that is easier to examine and manipulate in hardware. There are two main reasons for using the reserved bit in the IPv4 header. First of all, in order for the hardware to be able to detect the existence of options rapidly, there should exist a way to indicate this in the first 20 bytes of the IP header. Otherwise it would be necessary to parse information outside of the IP header, contradicting the design principle *"optimize for the common case"*, which most likely is packets without options. Since the reserved bit is located in the first 20 bytes and is currently unused, it fits the bill nicely. Secondly, our initial experiments show that the reserved bit is not examined by routers. This is perhaps the most important aspect since this allows packets carrying options to be treated like packets without options.

The remainder of this paper is structured as follows. In section 2, the setup and results of the experiments are presented. Section 3 concludes by describing our proposal for an alternative to IPv4 options.

2 How Delay is Affected by the Presence of an Option

Some IP options, such as record route and source route, must be processed by every router. Our hypothesis is that setting an option in an IP packet will affect the delay of that packet negatively because routers must at least check the type of all options. In this section we will present an experiment which supports this hypothesis.

2.1 Setup

We use the well known "ping" application for sending ICMP echo request messages and receiving ICMP echo replies [8]. We have added the possibility to include the "no operation" IP option in the IP header in the echo requests [1]. This option does not require any processing by the routers nor by the host being "pinged." However, the routers and the pinged host needs to look at the option to realize that it does not need to be processed.

[2]It should be pointed out that there exist other suggested uses for the reserved bit and that suggesting to use it therefore is somewhat controversial

Successful runs	484
Router refusing options	97
Traceroute loops	140
Misc. failures	93

Table 1: Experiment runs.

An experiment run on a path to a destination is performed by using the "traceroute" application to track the routers on the path. Then we send each router in turn 100 echo requests without options followed by 100 echo requests with the "no operation" option included and note the average round trip time in each case. Consecutive echo requests are separated by 40 ms.

We have gathered a list of web-sites to a number of Universities around the world and performed the above procedure using these web-sites as destinations. The results are presented in the next subsection. All echo requests are sent from a host at Luleå University of Technology. Because of this, the fourteen first hops are the same in virtually all measurements.

2.2 Results

Table 1 sums up the experiment runs. We have gathered data from 484 different runs. Out of these, we observed 97 instances where a router forwarded packets without options but dropped packets with the no operation option. Unfortunately, we did not log whether any pair of these 97 instances involved the same router. Thus, we cannot draw any conclusion of the frequency of routers that drops packets with options, only that such routers exist.

The traceroute loops in table 1 is when the traceroute tool reports the same node more than once. We have not investigated why this occurs. In these cases, we have included results up to but not including the first duplicate node. Furthermore, some of the attempted runs did not produce any results at all, either due to a DNS lookup failure or because the network was temporarily down.

Figures 1, 2, and 3 shows the result from three different runs. All runs look similar. The average round trip time is consistently larger with the option in the IP header. Since each sample point represents the average of 100 echo requests, short term congestion can cause individual points to be significantly larger than the true average. We observe that the probability for this kind of variation in delay is significantly higher for packets with an option than for packets without options.

The increased round trip time may be caused by two things. Each router that forwards the packet may add a delay for looking at the option and the "pinged" node may add a delay for looking at the option before

replying.

Figure 4 shows the difference in average round trip time for each round between the 100 echo requests with option and the 100 echo requests without option. The fact that the slope of the least square fitted line is positive indicates that the performance degradation of using an option is additive with regard to the number of hops. This indicates that the delay each forwarding node adds for looking at the option is significant.

2.3 Discussion

Some nodes simply swap the source and destination addresses in the echo request packet, and modify the ICMP header to an echo reply and transmits this. As a result, the echo reply will also have the no operation option set. Other nodes do not set the option in the echo reply packet. Hence, the round trip time should be affected more when the echo reply also has the option set. However, we have not made any separation between these two behaviors in this analysis.

To determine how frequently routers that drops packets with options are encountered, we need to keep track of the path for each run. When a router is caught discarding packets with options, we should check if that router has been encountered in an earlier run.

The first hops are almost always the same in our data set. Therefore, we should repeat these measurements from a number of other points in the net to avoid bias that may come from characteristics which are unique to these fourteen routers.

From our collected data we can conclude that routers treat packets with an included option significantly worse than packets with no included option. It is very rare that a burst of 100 echo requests/replies without options has an average round trip time higher than the corresponding burst with options. In the cases where this happens, it is probably because of temporary congestion.

3 Alternative Solution

The IP options can be divided into two major types: hop-by-hop options and end-to-end options. The hop-by-hop options can be further divided into mandatory and optional options. The mandatory options cannot be ignored by any router, while the optional ones may be ignored. These could for example be used for implementation of hints and notifications.

In this section, we device an implementation which will enable the use of non-mandatory options without exposing the packets to unnecessary delay. In order for end-points, i.e., hosts, to able to use the new type of options, both end-points must be aware of them. Routers that are aware of the new type of options can

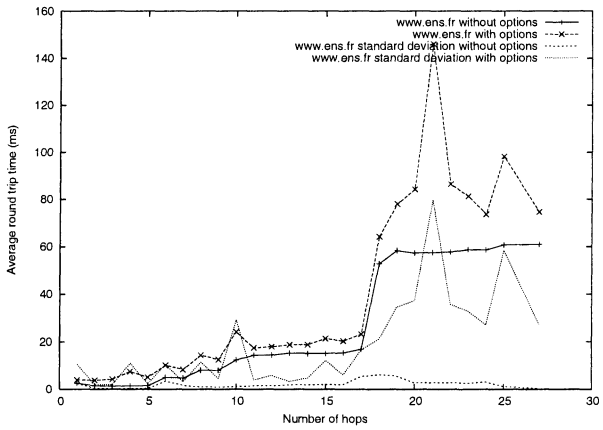


Figure 1: *Sample from the data set.*

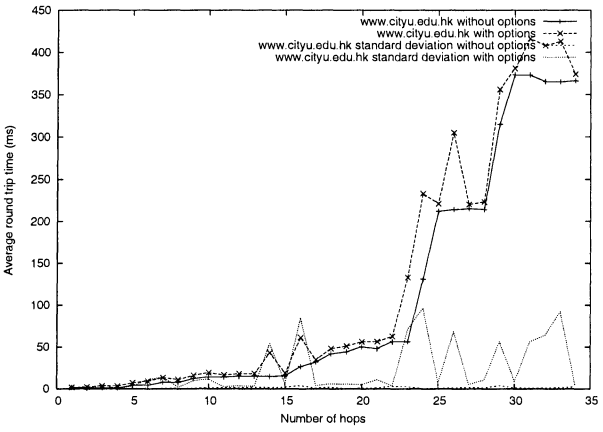


Figure 2: *Sample from the data set.*

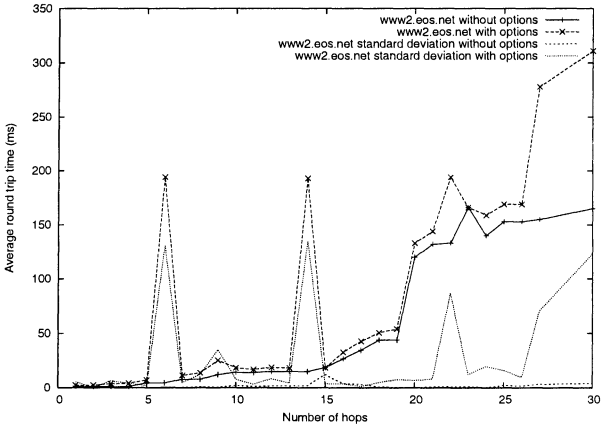


Figure 3: *Sample from the data set.*

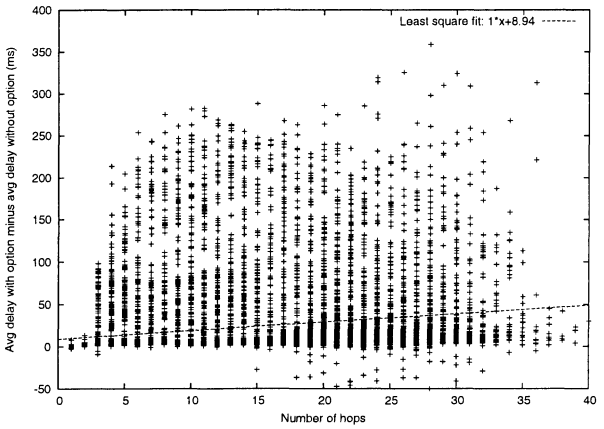


Figure 4: *Difference between the round trip times with and without the option.*

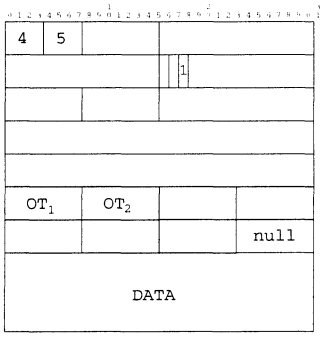


Figure 5: *IP header format for faster IP options in IPv4. ($OT_1, OT_2, \ldots$, *null*) comprise the table of option types.*

also use them. However, the new type of options are not suitable for implementation of mandatory options as this would require remaking and reconfiguring of more or less all existing routers; clearly a task bordering on the impossible. On the other hand, as long only non-mandatory hop-by-hop options are being implemented through the new type of options, routers can be unaware of them.

To enable faster implementation of IP options we propose arranging the option types in a table immediately following the IP header. The reserved bit in the IP header is used to indicate the existence of a table of IP-options directly after the IP header. The table of options is constructed accordingly. Each type of option is represented by one byte[3], which is placed back-to-back directly after the IP header itself. The end of the table is indicated a byte with a value of zero, i.e., the table is NULL-terminated. The options are divided into two classes: non-mandatory hop-by-hop options and end-to-end options. The difference between the two types of options should be discernible through the bit patterns, e.g., depending on whether the most significant bit is set or not. To ensure that routers can quickly detect whether or not there exist an hop-by-hop option of interest to the router, the hop-by-hop options have to be placed together in beginning of the table. This allows a router to stop examining the options included in the packet as soon as it encounters an end-to-end option.

Finally it is worth noting that it is also possible for future versions of IP to utilize this type of options. For instance IPv6 options could probably be processed faster by utilizing a table as mentioned above. In this case we could also use these options for mandatory hop-by-hop options.

[3]We assume that 8 bits are enough to represent the most important option types. To allow for more different kinds of options, we can reserve one option type to be used for extensions.

References

[1] Jon Postel. Internet Protocol. Standard RFC791, STD0005, Internet Engineering Task Force, September 1981.

[2] Lars Åke Larzon, Ulf Bodin, and Olov Schelén. Hints and notifications. In *Proceedings of Wireless Communications and Networking Conference, WCNC'02*, Orlando, Florida, March 2002.

[3] Chiang Lee, Chih-Horng Ke, and Chao-Chun Chen. Improving location management for mobile users with frequently visited locations. *Performance Evaluation*, 43(1):15–38, June 2001.

[4] T. Nakajima. Nps:user-level real-time network engine on real-time mach. Proceedings in Workshop, 1994.

[5] Mattias Engan, Pierre Fransson, and Olov Schelén. Selective truncating internetwork protocol: experiments with explicit framing. In *Proceedings of the Fourth International Conference on Computer Communications and Networks, ICCCN'00*, Las Vegas, Nevada, October 2000.

[6] David J. Wetherall and David L. Tennenhouse. The ACTIVE IP option. In *Proceedings of the Seventh ACM SIGOPS European Workshop*, Connemara, Ireland, September 1996.

[7] Steven M. Bellovin. Email correspondence, March 2002.

[8] Jon Postel. Internet control message protocol. Request for Comments 792, IETF, sep 1981.

Hints and Notifications

Lars-Åke Larzon, Ulf Bodin and Olov Schelén

Dept. of CS&EE, Luleå University of Technology, Sweden. E-mail:{lln,uffe,olov}@sm.luth.se

Abstract

With current Internet protocols, users may experience low and unpredictable forwarding quality at wireless links. This is due to varying link properties caused by changing radio conditions. Decreased forwarding quality can cause severe degradation in utilization. This is undesirable since forwarding capacity often is expensive at wireless links because of limited radio spectrum. Allowing the application and transport layers to communicate with wireless link layers can improve forwarding quality and utilization.

We propose to enable inter-layer communication by adding hints and notifications (HAN) to the Internet architecture. Hints can be introduced and used without notifications, while notifications needs hints or a similar mechanism to operate. By using IP options and ICMP messages to implement HAN, a backward-compatible partial deployment is possible. With HAN, the network layer becomes truly wireless friendly and radio spectrum can be used efficiently while supporting both real-time and traditional data applications.

1 Introduction

We propose introducing support for inter-layer communications to the Internet architecture in form of *hints and notifications (HAN)*. HAN enables spectrum-efficient data handling, which can improve service quality for individual users and utilization of wireless links. At wireless links, efficient usage of the radio spectrum is essential for reasons of cost-efficiency. HAN provides generic extensions to the network layer that improve transport of both real-time and traditional data over various wireless link technologies.

A strong trend is the increasing number of mobile devices connecting to the Internet using IP [6] over low- to medium-bandwidth wireless link technologies. Providing IP connectivity *all the way* to these devices is appealing since that purifies the infrastructure and enables simple access to the wide range of existing applications and link technologies in the Internet. As mobile devices become part of the Internet, they can communicate with other Internet hosts through a variety of wireless link layer technologies with different properties.

Constantly changing radio conditions can make the forwarding quality unpredictable at wireless links. Wireless link technologies may provide mechanisms for adaptive forward error correction (FEC) and automatic repeat request (ARQ). These mechanisms can introduce variations in delay and available bit-rate at higher layers. Moreover, the redundancy introduced by FEC and ARQ retransmissions may decrease the utilization of scarce forwarding resources at wireless links.

The radio spectrum available to any wireless link technology is a limited resource, and it is essential to utilize this scarce capacity as efficiently as possible to provide cost-efficient connectivity. Unnecessary transmissions and delays are desirable to avoid. We identify the following four problems associated with the efficiency of radio spectrum utilization:

- *Unknown payload properties.* Properties of packet payloads are unknown at the link layer. With payload information, link utilization can be improved for bit-error tolerant applications (i.e., re-transmitting or discarding packets due to acceptable payload errors can be avoided).
- *Unknown link properties.* Link layer properties such as available bandwidth and error-rates are unknown at higher layers. Using such information, applications can adapt dynamically to suitable codecs, frame rates, etc.
- *Misinterpreted loss.* Congestion-responsive applications interpret packet losses as signs of persistent congestion and cannot react properly to losses cause by bit-errors or temporary congestion (i.e., shorter than a typical RTT).
- *Temporary disconnection.* After a temporary disconnection, congestion-responsive applications are often late in detecting re-established connectivity, which can cause the link to be underutilized for a period.

To solve the problems listed above, messages need to be exchanged between non-adjacent protocol layers. The layered design of the Internet protocols considers the link layer to provide a simple frame delivery

service, which does not include such inter-layer communication. This makes IP suitable for internetworking over heterogeneous link technologies. It is important to preserve the layered design IP by allowing transparent operation over parts of the network where inter-layer communication are unsupported. Thereby, full backward compatibility with already installed IP infrastructure is ensured.

We propose adding *hints and notifications (HAN)* to the Internet architecture. By making HAN part of the network layer, it enables inter-layer communication without violating Internet layering principles. By getting *hints* from upper layers, link layers can use spectrum-conservative mechanisms to a greater extent. Hints can also improve flexibility for application designers by allowing data with different requirements to be placed in the same packet. By sending *notifications* from lower layers, applications and transport layers can react more accurately to variations in the physical medium.

There are several proposals for inter-layer communication aiming at enhancing performance in wireless networks. Most of them are however incompatible with each other and aim at specific improvements. Deploying all of these proposals would result in an unnecessarily complex system. The generic design of HAN supports any application, transport, or (wireless) link layer. HAN can be implemented using IP options and ICMP messages to simplify successive deployment without interfering with existing Internet nodes. Hints are independent of notifications, while notifications require usage of hints or some similar mechanism to transport notification requests with data packets.

2 Demands

Forwarding capacity of wireless links is expensive due to radio spectrum limitations. To achieve high utilization of the spectrum, error-rates are generally orders of magnitudes higher than in wired networks. This is because moderate increases in error-rate typically result in considerable higher bit-rates in higher layers. High bit-error rates can be hidden to higher protocol layers through local retransmissions at the link layer. This introduces however extra delay and degrades the bit-rate available in upper layers.

Radio conditions vary at different time-scales, ranging from milliseconds up to minutes. Upper layers will be exposed to these errors and must compensate for them. In this section, we discuss demands from protocols at higher layers on mechanisms that increase the utilization of the expensive forwarding capacity over wireless links.

2.1 Real-time applications

Today, there are applications that suffer more than others from packet losses in the Internet. Audio and video codecs can often cope with some degree of errors in the received data and still find it useful. For applications using such codecs, it is often preferable to receive packets with errors in the payload over losing packets or having them delayed. The packet will however be discarded if there are errors in the packet header and the error tolerance may differ among blocks of data in the payload. Consequently, information about data properties (e.g., error tolerance and delay requirements) is needed to support cost-effective usage of the radio spectrum allocated for wireless links.

Cost-effective usage of radio spectrum can be achieved through less protective FEC and ARQ strategies for error tolerant data blocks. Moreover, unequal error protection (UEP) [3] can be used when data blocks have different error requirements [2][7].

In the current Internet architecture, there is no straightforward way for applications using bit-error tolerant codecs to pass information about data properties to the link layer. Mechanisms for service differentiation allow for multiple levels of priority between packets (e.g., mechanisms specified in the Differentiated Services architecture [1]). These mechanisms do not however support multiple priority levels for data blocks within individual packets. We refer to this as the *unknown payload properties* problem.

Real-time applications could also benefit from getting information about current link layer conditions (e.g., available bandwidth, error rates, etc.). This could help a real-time transmission of video or audio to adapt dynamically to suitable codecs, frame rates, etc. We refer to this as the *unknown link properties* problem.

2.2 Congestion-responsive applications

An application is said to be congestion-responsive if it responds to packet loss by reducing its sending rate. Since congestion-responsive applications rely on feedback from receivers, they adjust sending rates at time-scales longer than their RTTs.

Common for congestion-responsive applications is that they interpret packet losses as signs of persistent congestion. However, at wireless links, a considerable number of packets are lost due to changing radio conditions. Moreover, packets may be interpreted as lost if they are delayed due to local retransmissions.[1] Consequently, congestion-responsive applications may underutilize wireless links with high bit-error rates, available bit-rates that vary at short time-scales (i.e., time-scales shorter than a typical RTT),

[1]For TCP, this might cause a timeout, which forces TCP into slow-start.

or both. We refer to this as the *misinterpreted loss* problem.

Another problem associated with wireless links is that hosts may be temporarily disconnected. Congestion-responsive applications that rely on feedback from receivers are late in detecting disconnections. Moreover, they might detect re-established connections late and need additional time to find the appropriate sending rate, causing the link to be under-utilized for a period. This becomes particularly problematic if the sending rate is reset after a re-establishment. We refer to this as the *temporary disconnection* problem.

To avoid the problems described above, congestion-responsive applications must be able to distinguish between losses related to persistent congestion, temporarily congestion, bit-errors, and disconnections. Applications should also be able to distinguish between packets being lost and packets being delayed. Moreover, as a wireless connection is re-established congestion-responsive applications should be able to immediately find the appropriate sending rate. The performance of these applications can be improved with a mechanism that notifies sources explicitly on the appropriate sending rate, why packets are dropped and whether they have been delayed due to local retransmissions.

3 Hints and notifications (HAN)

The Internet protocols were originally designed with the implicit assumption that bandwidth, delay, error-rates and jitter do not vary much at the link layer. When exposed to networks where these metrics vary, these protocols experience problems and perform poorly. HAN addresses these problems and aim at improving the performance of the Internet protocols when used in networks for which they were not originally designed.

The Internet network layer implemented with the Internet Protocol (IP) is the common denominator for several transport protocols and a large number of applications. This network layer provides global connectivity over a wide variety of link and network technologies. However, IP is not aware of local ARQ or FEC at the link layer and does not provide any input to optimize their operation. Moreover, IP is unable of providing upper layers with information of link actions and status.

The need for information exchange between the link and transport layers outlined in section 2 can be satisfied locally by violating layering principles. We believe however that the layering of the successful Internet architecture should be preserved to not increase the risk of malfunction due to many complex additions, where each addition only supports a few combinations of link layers and applications or transport layers. To preserve the layering, we propose to place information exchange mechanisms in the network layer itself. Thereby, link layers and transport layers or applications need not know each other's semantics in order to exchange information.

Hints and notifications are placed in the network layer headers where they are parsed by surrounding layers[2]. We propose to use an IP option [6][3] for hints and notification requests, and ICMP messages [5] for two different types of notifications. This simplifies a partial deployment of HAN in the existing Internet. Nodes capable of parsing hints do so, and provide their link layers with the information. For nodes unaware of the concept with hints, packets will be forwarded as usual with the IP option carrying the hints intact. Using ICMP enables deployment without inventing a new protocol to provide notifications.

3.1 Hints

Today, varying physical conditions on a link can prevent queued packets from being transmitted immediately. By providing hints from upper layers on how to treat packets at periods of bad radio conditions, unwanted and unnecessary packet losses and delays can be avoided. In its simplest form, a hint includes information about the boundaries between different parts of a packet and the acceptable error-rate and delay requirements for each part. The link layer can use this information to choose suitable modulation, error correction and verification mechanisms on a per-packet basis. This makes it possible to quickly react on short-time variations.

There are no strict rules of how the hints are used by the link layer, i.e., upper layers can not exactly instruct the link layer of how to treat frames. Neither is there any requirement of what mechanisms for error correction or recovery that the link layer must provide in order to use the hints. Instead, the hints are, as the name indicates, merely additional information of packet internals that the link layer may use as it finds suitable.

3.2 Notifications

As mentioned in Section 2.2, congestion-responsive applications might suffer from the misinterpreted loss and the temporary disconnection problems. To handle these problems, *link action notifications* can be used to inform sources of why packets are dropped and to inform sources of long packets delays caused by local retransmissions.

[2]In a practical implementation, the information processing could be done in the network layer and use out-of-band signaling between the network layer and other layers

[3]For IP version 6, this IP option is a per-link option.

Furthermore, as mentioned in Section 2.1, real-time applications might suffer from the unknown link-properties problem. To handle this, *link status notifications* can be used to inform sources of current and maximum available bit-rate at a wireless link. These notifications can also be used to detect re-established connectivity after a temporary disconnection.

In the following two sections, we present the two different types of notifications — link action notifications and link status notifications.

3.2.1 Link action notifications

Link action notifications are messages that origin from actions associated with individual packets, e.g., local retransmissions of link layer frames causing a packet to be delayed.

We have identified the following link actions that can be used to generate notification messages.

- *Packet delayed due to local retransmissions.* This information can, e.g., be used by a congestion-responsive source to avoid an unnecessary time-out, redundant retransmission at a higher layer, or both.
- *Packet dropped due to bit-errors* and *Packet dropped due to temporary congestion.* These can, e.g., be used by a congestion responsive source to avoid an unnecessary reduction of sending rate.
- *Packet dropped due to persistent congestion.* This can be used as a fast congestion notification (FCN).
- *Packet dropped due to that the wireless host is disconnected.* Upon reception of this information, an application can pause the data transmission until the the connection is re-established.

Link action notifications are put in separate packets to avoid having link layers monitoring transport and application layer headers, which would require per-flow states. To reduce the amount of notification traffic, notifications are sent only to notification-aware sources. Notification awareness is indicated in outgoing IP packets (e.g., as part of hints).

3.2.2 Link status notifications

Link status notifications are messages that contain information of static and dynamic link properties. For example, the maximum bit-rate of a wireless link is often a static parameter while the bit-error probability may vary.

We have identified the following link status notification messages.

- *Currently available bit-rate.* This information can for example be used by a TCP source to select an initial sending window after re-establishing a temporarily disconnected connection.
- *Maximum available bit-rate.* This information can be used as an alternative to probing to estimate link capacity. TCP, which normally reduce the sending speed significantly after exceeding the link capacity, could benefit from this.

3.2.3 Implementation

By defining notifications to be initiated upon request from sources aware of notifications, no states are needed in network nodes. Information about notification awareness and what types of notifications the source is interested in can be included in hints, or sent in separate packets. The stateless approach is attractive when users are highly mobile, since access points do not have to manage neither per-host or per-flow states, nor state migrations. If notifications are sent on request only, sources that are interested in notifications without any traffic to send must periodically probe the network, e.g., with ICMP packets.

4 Conclusions

In this paper, we define the concept of hints and notifications (HAN). HAN is justified by needs for spectrum-efficient data handling, improved service quality for individual users and high utilization of wireless links. Limited radio spectrum makes forwarding capacity expensive at wireless links, which emphasizes these needs.

HAN enables application and transport layers to communicate with various link layer technologies without violating layering principles. Thereby, HAN fits naturally into the layered Internet architecture. Inter-layer communication can improve forwarding quality and spectrum utilization. With hints, real-time applications can inform the link layer of delay requirements and error tolerance. In addition, notifications enables feedback of link layer actions and current status to application and transport layers. This information can help congestion-responsive applications increase link utilization by reacting properly to different types of packet loss.

By using IP options and ICMP messages to implement HAN, a backward-compatible and partial deployment in the current Internet is possible. Hints can be introduced and used without notifications, while notifications need hints or a similar mechanism to operate.

We plan to implement selected parts of HAN in the network simulator (NS) [4] to evaluate it further for different scenarios. Moreover, we intend to produce

an open implementation of HAN to perform application tests (i.e., to examine the interaction between HAN and its target applications).

References

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. Request For Comments RFC 2475, DARPA, December 1998.

[2] A. R. Calderbank and N. Seshadri. Multilevel codes for unequal error protection. *IEEE Trans. Inform. Theory*, 39:1234–1248, July 1993.

[3] B. Masnick and J. Wolf. On linear unequal error protection codes. *IEEE Trans. Inform. Theory*, IT-13:600–607, October 1967.

[4] Ns homepage. URL: http://www.isi.edu/nsnam/ns/.

[5] Jon Postel. Internet Control Message Protocol. Request For Comments RFC 792, DARPA, September 1981.

[6] Jon Postel. Internet Protocol. Request For Comments RFC 791, DARPA, September 1981.

[7] L.-F. Wei. Multilevel codes for unequal error protection. *IEEE Trans. Commun.*, 41:1439–1449, October 1993.

Information Flow in Radar

Lars Falk

Swedish Defence Research Agency (FOI), 172 90 Stockholm. E-mail: lars.falk@foi.se

Abstract

Information theory is a useful concept for describing radar measurements. The pioneering work was done by P. M. Woodward in his classical book "Probability and information theory, with application to radar" (1953). Radars are nowadays usually described in other terms, but an excellent understanding of their function can be obtained by calculating the total amount of information flowing through the sensor system. This approach allows one to locate bottlenecks and identify features in the system sensitive to electronic attack. Some examples will show that one can estimate the performance of a radar system accurately by considering the maximum amount of information available to the system without going into technical details. This useful pedagogical tool leads in advanced form to the Bayesian approach, which is frequently used to process high quality radar data.

1 Introduction

In 1948-49 Claude Shannon revolutionised the field of electronic communication by introducing his theory of communication [1]. Shannon's approach was based on a definition of information, which is directly applicable to engineering requirements. In addition Shannon proved a number of powerful theorems concerning the limits on communication imposed by noise. Shannon's predictions were far-reaching and his limits have only recently been approached by modern codes.

Shannon's theory of information became popular and even notorious, because many people neglected the technical restrictions emphasised by Shannon. His theory summarised the existing knowledge about predictions and signal processing and extended it in various directions. Several scientists tried to apply Shannon's results to their fields of interest. In 1957 E. T. Jaynes reinterpreted statistical mechanics as our degree of ignorance about a physical system. In 1953 P. M. Woodward wrote his classical book, "Probability and information theory, with application to radar" [2]. Woodward derived a number of results in a new way and introduced some novel concepts, like the ideal receiver and the ambiguity function.

2 Inverse probability

2.1 Bayesian methods

The most important result obtained from Shannon's theory is that noisy signals are in many respects similar to digitally quantised signals. This allows one to extend the ideas about how to store digital data to analogue signals. Shannon showed that for any message the information is related to the inverse probability of its appearance. The amount of information can be estimated by considering the probabilities of individual symbols and their nearest correlation.

Woodward showed that Shannon's argument is closely related to the method of inverse probability, which was promoted at the time by the mathematician Harold Jeffreys. Shannon may have recognised the connection, but probably avoided it to stay away from the quarrels, which bothered Jeffreys. Shannon expressed his theory in such a way that the results were acceptable to engineers and mathematicians. He successfully steered between the Scylla of loose reasoning and the Charybdis of intolerable rigor and entered a world full of unexpected results.

Many engineers doubted that Shannon's theory would be significant for radar. Communication is the art of detecting and separating symbols in a message. Shannon's crowning achievement is the coding theorem, which is not applicable to radar, where only one type of signal is used. The main problem is to detect pulses and measure their amplitude and delay. Pulses may be coded, but only to reduce power.

However, the detection process for radar is similar to that for communication symbols in noise. Shannon partly concealed the similarity by emphasising the statistics of infinitely long messages. His main objective was to prove the existence of efficient codes, while real codes have to be based on incomplete knowledge about the statistics of transmitted messages. Shannon's approach is as a result equivalent to Bayesian statistics, but he stayed away from the hotly disputed subject of prior knowledge. Ten years later this concept created a great rift in the statistical camp, which healed only after years of fierce battle.

2.2 The ideal receiver

Woodward was aware of the mathematical problems and extended Shannon's presentation by considering the probability of individual messages. He only performed the averaging at a later stage. This step is important from a conceptual point of view, since information obtained from an individual event can be negative, while Shannon's averaged information is always positive. An observer can be deceived by individual measurements, but Shannon's and Woodward's methods ensure that on the average uncertainty must decrease. This fact was first proved by Cox and widely publicised by Jaynes [3].

Woodward used the condition that in a rational decision no information may be neglected to define the concept of an ideal receiver. Previously it had been assumed that a radar receiver should maximise physical properties like signal-to-noise ratio or other conditions, but Woodward's idea is more fundamental: a receiver must not waste information. For a point target in Gaussian noise this condition leads to the correlation receiver, which is equivalent to the matched filter. This design was first introduced around 1940 and continued to be reinvented during the 1950-ies.

The advantage of Woodward's idea is that it is equally applicable to detection in complicated situations, like targets hidden in partly known clutter. The basic idea is to preserve knowledge about the complete probability distribution. Early radar engineers wanted to produce unique values, corresponding to the position of maximum probability, but in this way information is irretrievably lost. By insisting on the need for reversible processes Woodward explained the efficiency of many different detectors.

3 Practical applications

Woodward's approach is relatively little known, since once you possess an explicit model for target, clutter and noise you get excellent results by applying standard statistical methods. The model removes the need to consider the theoretical basis of the detection procedure. This means, paradoxically, that Woodward's theory is most useful, when there is insufficient knowledge about the radar or the received pulses, as commonly happens in planning or intelligence situations.

Woodward's approach is particularly valuable as a pedagogical tool, since it provides a language suitable for discussions with people without radar expertise. The idea that information should be measured in the number of bits is generally accepted, and communication between radar specialists and customers can be promoted by using such concepts. Information is the main constituent in a radar network and from the customer's point of view the desired product. This is particularly true, if different sensors are joined together in a network.

3.1 Conservation of information

Information theory allows one to estimate accurately the performance of a radar system without entering into technical details. The reason is easily understood by comparing with arguments used in other fields of physics. During the 19th century the scientific community gradually accepted the law of conservation of energy, and engineers started to develop mechanical and electric systems using this law as a guiding principle. It is often sufficient to know the limits of maximum performance to deduce the optimal design of a machine.

In particular, many machines are based on reversible operations to reduce the conversion of energy into heat. These considerations are a consequence of the second law of thermodynamics and information should similarly be conserved during a measurement. It seems to be a law of nature that information cannot be created within a closed system, just like entropy can not decrease. This condition determines the performance, which is achievable by heat engines and measurement systems.

An important difference must be observed between a general criterion and a calculation. One cannot obtain numbers from principles alone. This is illustrated by the problem of three doors, where you are presented with a choice between three doors and told that a car is hidden behind one of them. The question is how to act, if you are told to point at one door and then notified that the car is not hidden behind one of the other doors. Should one switch doors, if this is allowed?

Intuitively many people think that the choice does not matter, though a simple calculation shows that you double your chances by choosing the third door. An alternative way of arriving at this conclusion is to argue that the information offered can only be exploited if you act on it. This means that you must switch doors, though the argument does not explain how much you can gain in this way. You only know that on the average you cannot lose by acting rationally.



3.2 A SAR example

In radar work the missing numbers can fortunately be obtained by additional arguments. Modern electronic components are a few dB away from the best possible results and one can deduce the performance of a radar system by assuming the maximum result compatible with information theory. This is a simple and efficient method of estimating the performance of unknown radar systems.

As an illustration we will consider a SAR system. The synthetic aperture is obtained by flying a radar antenna along a straight track and collecting pulses until one has created an artificial antenna of suitable length. Wave theory shows that the maximum resolution on ground in the direction of flight is half the real antenna length, d/2 [4]. This is true for any range, which means you will have practical problems at large range, where the radar must provide enough power and emit the pulses sufficiently slowly to avoid range ambiguities.

Textbooks deduce these conditions by analysing the frequency modulation of target returns introduced by the motion of the aircraft relative to ground [4], but it is sufficient to consider the information obtained during a measurement to form an estimate. If a large antenna were used instead of the synthetic antenna the pulse repetition frequency (PRF) would be limited by the condition that the radar must complete each range measurement before moving on to the next resolution cell. The velocity V of the aircraft would then be limited by the condition,

$$V < PRF \cdot d/2$$

This agrees with formulas obtained from the sampling theorem. We have only exploited the condition that two equivalent radar systems must deliver information at the same rate. Little was learnt about the radar, since for this purpose you have to consider a real measurement.

3.3 The sampling theorem

The sampling theorem was rederived by Shannon, since the form due to Nyquist does not involve noise. Shannon showed that the sampling theorem has a simple form for long signals. The maximum number of bits required to characterise a signal of length T, bandwidth W and power P and white noise of power N is [1,2]:

$$I = WT \log_2 (1+P/N).$$

This formula shows, among other things, that it is not advantageous to produce strong radar returns. The noise acts as a quantising level, which defines the number of discernible amplitudes. If P=N the number of bits equals the time-bandwidth product, WT. This number provides a convenient definition of the rather elusive concept of radar gates. The maximum number of bits equals the number of gates, when the signal is barely detectable.

Consider a low PRF surveillance radar. There are WT bits of information and consequently WxPRF range gates. Each gate must have a length corresponding to the time interval 1/W. This interpretation is based only on basic signal analysis, though ultimately it depends on the fact that radar waves propagate in straight rays with constant velocity.

A similar argument applies to doppler gates. If a radar collects N successive pulses their relative phases measure the radial velocity of target in different range gates. The number of doppler gates equals N, since the information was increased by this factor. The frequency is uniquely determined only within PRF and thus the doppler gates have width PRF/N. This number is related to the total time of measurement, N/PRF, in well-known manner, but the point is that the method can be used to calculate the size of the gates in more complicated cases.

3.4 The number of cells

To demonstrate the meaning of the sampling theorem it is instructive to cite an investigation of the future of radar reproduced in a current textbook [4]. Radar energy is concentrated into definite directions by an antenna with main lobe angels $\Delta\theta$ and $\Delta\phi$. The lobe scans a sector Ω during time T, so the dwell time t in each position equals $t=T\ \Delta\theta\ \Delta\phi\ /\Omega$, while the number of beam positions is

$$n_1 = \Delta\theta\ \Delta\phi\ /\Omega.$$

The number of range gates is, as before,

$$n_2 = 1/PRF\tau.$$

Here τ is the pulse length. The number of doppler gates is

$$n_3 = PRF\ t.$$

The total number of cells is obtained by multiplying these numbers,

$$n = n_1\ n_2\ n_3 = T/\tau.$$

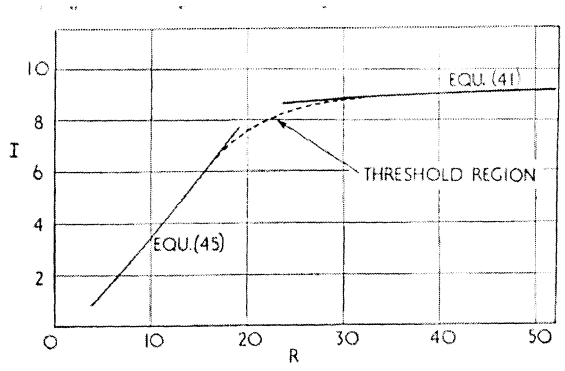
The authors do not mention that this result follows directly from the sampling theorem. Any simple radar pulse has $W=1/\tau$, and thus $n=WT$, but this is clearly a more general statement than the previous formula, since it does not depend on waveform. The available information is strictly limited by signal theory, while geometrical information about the scanning pattern must be treated separately.

The sampling theorem does not impose any particular restrictions on modern radar signals, which are wide both in time and frequency. This worried Woodward, who introduced the ambiguity function to analyse the various possibilities. He clarified the properties and restrictions on the function, but information does not affect this problem, since noise is not involved. Woodward ended his investigation with a classical understatement [2]: "The reader may feel some disappointment, not unshared by the writer, that the basic question of what to transmit remains substantially unanswered." A more positive

point of view would be that the designer has great freedom in his choice of waveform after the restrictions imposed by information theory have been taken into account.

3.5 Information flow

It may be asked whether the concept of information is really useful, since the benefits depend on which person is receiving the knowledge. An answer is suggested by the increase of information obtained when the energy from a target is integrated. The energy can be expressed in terms of a dimensionless number, R, by dividing with noise power per frequency interval. The figure shows Woodward's calculation of the flow of information for a stationary target. The curve is similar to the sampling formula, but contains some refinements due to the form of problem.



At low energies the integrated signal provides a steady increase in received information, but when the target is barely detectable the increase in information starts to slow down. This is natural, since the signal is now known and the radar is measuring fluctuations in the background noise of no interest to the observer.

The amount of useful information levels off, but this effect can be considered in a positive way. The signal works as a trigger, which tells the observer to adapt to a new situation. The information should be used to locate targets with increased accuracy, which can be done by activating a monopulse system or a system of range gates. This is the standard procedure adopted in air defence, but here the process is formalised in terms of information. Once a target has been detected you should stop the measurement or ask new questions by changing the receiver or emitted waveform. It is up to us to decide how to acquire additional information in an adaptive way.

4 Conclusions

The use of information as an analytical tool is particularly helpful in complicated situations, where simple estimates are required. A typical example would be the comparison between an established radar system and new candidates working on different principles, but designed to solve similar problems.

The investigation of information flow in radar allows one to identify bottlenecks in the existing communication equipment. This is important in order to estimate the sensitivity of radar systems to electronic attacks, since this can be difficult without a guiding principle which covers many different.

The examples presented here have deliberately been held simple to demonstrate the usefulness of information as a pedagogical tool. The main advantage is that optimal results can be estimated for many radar systems and the results transmitted to customers.

A detailed description of P. M. Woodward's career and pioneering work will be presented elsewhere [5].

References

[1] C. E. Shannon and W. Weaver. *The Mathematical Theory of Information.* The University of Illinois Press, Urbana 1953.

[2] P. M. Woodward. *Probability and Information Theory, with Application to Radar.* Pergamon Press, London 1953.

[3] E. T. Jaynes. *Probability Theory: The Logic of Science.* Textbook on http://bayes.wustl.edu/.

[4] S. Kingsley and S. Quegan. *Understanding Radar Systems.* McGraw-Hill 1992.

[5] L. Falk. *Information with Application to Radar – a Tribute to P. M. Woodward.* To be presented at Radar 2002, Edinburgh, October 2002.

Radio Navigation and Radio Positioning in Sweden

Rolf Zetterberg

Swedish Maritime Administration, 603 71 Norrköping, Sweden. E-mail: rolf.zetterberg@sjofartsverket.se.

Abstract

This paper gives an overview of the radio navigation and radio positioning systems that are available to users in Sweden. It also indicates the general trends in radio navigation and describes specific plans were those are known.

The paper is mainly based on information from the Swedish Radio Navigation Plan.

1 Available systems

1.1 Ground based systems

1.1.1 Common systems

Loran C and the Russian equivalence Chayka are the only ground based general systems which remains in operation. Loran C is a pulsed, hyperbolic system operating in the 90 to 110 kHz frequency band. The system is based upon measurement of the difference in time of arrival of pulses radiated by a chain of synchronized transmitters that are separated by several hundreds of km.

There is no transmitter in Sweden and Sweden is only partially covered by signals from this system. Northwest Europe is covered by a chain operated by NELS, North European Loran Systems, financed by 8 countries. Four of these have now declared their intention to withdraw from this co-operation so the future for the system in Europe is uncertain.

The Loran C chains in US are now being updated with solid state transmitters and remote control systems. It is likely that they are kept in operation for the foreseeable future mainly due to the fact that the Loran C signal is less sensitive to jamming and interference.

1.1.2 System for aviation

Navigation systems for civil aviation are to a great extent subject to regulation by ICAO and introduction of new systems is a slow process with regulation, certification and implementation.

NDB, Non Directional Beacons, the oldest radio navigation aid, is still used but other and better systems are also available and more important today.

VOR, Very High Frequency Omnidirectional Range, is expected to cease operation within 10 years while DME, Distance Measuring Equipment, will be kept in operation for the foreseeable future.

For approach and landing is in addition ILS, Instrument Landing System, used. ILS comes in three different accuracy categories, which allows landing in different conditions.

The methods used so far to define required navigation capability has been to require a specific type of equipment. ICAO has now developed a new concept, Required Navigation Performance, RNP, to define requirements for navigation capability. RNP requires that the user is able to navigate with a defined level of precision within a certain airspace.

1.1.3 Marine systems

The old DECCA and radio beacon systems have been taken out of operation recently. Still remaining are a few Fixed Direction Finders, mainly used in search and Rescue operations. There is also ground based support system for radar navigation, RACONS, which gives an coded answer on the same frequency when it detects a radar pulse. This gives the navigator a positive identification of the radar echo.

1.1.4 System for land navigation

Position fixing with GSM based on direction and distance to a base station is expected to increase. Accuracy will be increased by use of several base stations in the calculation.

1.2 Onboard systems

Navigation radar is widely used in marine navigation and is a requirement from IMO for commercial shipping . It is to some extent even used in aviation.

For military use has different map matching systems been developed. A system called TERNAV, developed by SAAB, utilize a radar altimeter and determines the position by correlating the measured height profile with a terrain database.

Simpler map-matching systems are also used in some car navigation systems.

1.3 Satellite based systems

The satellite based systems, mainly GPS, are dominating the radio navigation area and gives unsurpassed characteristics for many users.

1.3.1 Global Positioning System

Since the intentional degradation of the signal for civil users (Selective Availability) ceased in May 2000 has a new specification of the characteristics been presented. The "normal" accuracy is better than 10 m but it can be up to 36 m + errors from the receiver.

The plans for the future includes several steps. The first, 2003 - 2006, with the so called GPS II R will give a civil C/A-code on the frequency L2.. The satellites in block GPS II F add a third civil signal on L5 (1176,45 MHz) and is planned for 2005 - 2010. This will result in an increased accuracy.

1.3.2 Glonass

The Russian satellite navigation system has theoretically very good characteristics but has never reached a full satellite constellation. At present are only six satellites available and the future of the system seems uncertain.

1.3.3 Galileo

After years of discussions within the European Union, the Transport Council decided in March 2002 to enter the development phase of the Galileo project which aims at a satellite navigation system under European, civil control to be operative in 2008. The cost for this phase of the project is estimated to 10 billion Euros and is shared between EU and ESA.
The plans are not finalised but preliminary include several different services, such as a free "Open Access Service", a "Safety of Life Service" including integrity information and a "Commercially Available Service" with high accuracy.

In general terms the idea seems be to offer a better service than GPS and also to introduce some kind of cost recovery mechanism for the more advanced services.

1.3.4 COSPAS/SARSAT

This system is based on an international co-operation and is intended for localization of people and crafts in distress. A emergency transmitter (called EPIRP on sea respectively ELT in air) transmit on 406,025 MHz and signal is detected by the satellites which use the doppler effect to calculate the position of the transmitter. The coverage is global and the accuracy normally better than 2 Nautical Miles.

2 Support systems

Several support systems intended to augment the satellite systems, mainly GPS, are available or planned.
For maritime users has a system, defined by IALA (International Association of Aids to Navigation and Lighthouse Authorities), been used for nearly 10 years. Differential correction and integrity information for GPS are transmitted on the traditional radio beacon band (285 –315 KHz). The Swedish Maritime Administration operates 10 such stations, which cowers all Swedish waters. The resulting accuracy for the user is 1 – 2 m.
Based on information from the reference stations operated by the National Land Survey are differential corrections distributed on RDS and DARC channels on the FM broadcasting network.

The distribution of corrections for carrier phase measurements (Real Time Kinematics) has also been introduced. First as local corrections but now also multi-station information , which allows the user to calculate his position with centimetre accuracy in real time.

Under test is also the joint project between EU, ESA and Eurocontrol called EGNOS (European Geostationary Navigation Overlay system). This system includes a number of reference and control station on ground and a couple of geostationary satellites for distribution of wide-area corrections and integrity information for GPS. The satellites shall also transmit a ranging signal of GPS-type and can thus be used as additional navigation satellites by GPS-users. Similar systems are developed in USA and Japan and these systems will give a global coverage except for areas at high latitudes.

3. Summary

Radio navigation, based on satellite navigation systems is expected to become mass market service during the next 10 years. It will be utilized in different kind of position related services as well as for traditional "navigation". The future systems will be truly multi-modal with both professional users in transport, timing, survey, science etc and large user groups for personal services, car navigation etc.

Networks on Chip

Axel Jantsch

Laboratory for Electronics and Computer Systems, Royal Institute of Technology, E-mail: axel@imit.kth.se

Abstract

Future single chip systems will resemble more traditional computer networks than traditional central processors. The main reasons for this trend are (A) the infeasibility of global synchrony on a single chip, (B) the necessity of reuse of existing hardware and software components as much as possible, and (C) the heterogeneity of system functions and features. The consequences of this trend are far reaching and imply the shift in concern from computation and sequential algorithms to concurrency, communication and interaction in every aspect of design and development of hardware and software. A concrete example of this shift is the expected replacement of purely sequential computer languages by languages that contain concurrency as a first order object.

1 Motivation

IC manufacturing technology will provide us with a a few billion transistors on a single chip within a few years. Assuming that these predictions hold and that the market will continue to absorb ever higher volumes of ICs, the key questions are: how will the future chips be organized and how will future systems, which include these chips, be designed? One possible answer is that single CPUs will still occupy entire chips and will exhibit correspondingly higher performance. The instruction set will essentially remain unchanged to provide backwards compatibility and the will still be implemented in C or C++. However, there are a few factors which make this scenario unlikely:

1. Physical effects of deep sub-micron technology make it increasingly difficult to maintain global synchrony among all parts of the chip. The clock signal will soon need several clock cycles to travel across the chip and the clock distribution tree is already today a major source of power consumption and cost. The trends of scaling to smaller geometric dimension and higher clock frequency make these problems more significant every year.

2. Synthesis and compiler technology development do not keep pace with IC manufacturing technology development. As a consequence, which is called the *design productivity gap*, we need either exponentially growing design teams or design time to design and implement systems which fit onto a single IC. Since both alternatives are unrealistic we have in the past escaped from the problem by using ever more complex components as primitive design units. These primitive design units have evolved from individual transistors to logic gates to entire ALUs, multipliers and finite state machines. This trend will likely continue with CPU and DSP cores and blocks for compression, encryption and similar functions being the primitive design units. These design units have however asynchronous interfaces to the outside and vastly different internal clocking regimes. As a result a *globally asynchronous and locally synchronous (GALS)* design style emerges already today.

3. Obviously, systems that can be implemented on a single chip become increasingly more complex. As a result different functions and features with vastly different characteristics and history reside on the same chip. signal processing algorithms which recover and generate radio signals will coexist with global control, maintenance and accounting functions as well as with natural language comprehension and generation functions. These functions are developed in different contexts, by different teams, with different design languages and tools. However, they need to be integrated into a single chip.

Taking these current trends and facts together it is natural to contemplate a design paradigm where a set of interacting functions and features are implemented on a set of asynchronously communicating resources, such as CPU cores and specialized hardware blocks. In this scenario the mapping and implementation of system functions onto resources is covered by traditional design and synthesis methods. It may even an integral part of the reuse of a given system function and a resource. For instance, the purchasing of a bluetooth protocol stack may include its implementation on a combination of a custom hardware block and an ARM processor core. However, providing a chip level communication infra-structure and

mapping of system level interactions onto the communication infra-structure is not covered by any traditional design methodology and will become the focus of research and tool development in the near future. In fact, in the last two years we have seen several concrete proposals for on-chip network architectures.

in 2000 Hemani and al. [1] have proposed a packet switched architecture with switches surrounded by six resources and connected to 6 neighboring switches. The architecture is called a Honeycomb due to the hexagon based pattern of switches and resources. The concept of packet switching re-appears in other consecutive approaches but the topology simplifies in most proposals to a mesh of resources and switches. In 2001 Dally and Towles [2] propose a mesh based packet switched network with very simple switches which require less than 10% area overhead. MicroNetworks proposed by Drew Wingard [3] is another packet switched on-chip interconnection mechanism proposed recently. Keutzer et al. [4] and Sgroi et al. [5] provide general discussions and motivations for communication centric on-chip architectures and platforms and for the strict conceptual separation of computation from communication. More recently, Kumar et al. [6] put forward a detailed packet switched, mesh based on-chip communication infra-structure together with a design methodology. The proposed concept of a region breaks the strict mesh-based geometry. A region can cover an arbitrary number of switches and resources and allows to accommodate larger resources such as FPGA areas and memory banks in a flexible way. This architecture we will also outline in the consecutive sections of this paper. Valtonen et al. [7, 8] propose an on-chip interconnected network of resources or cells but put the main emphasis on fault tolerance and unlimited scalability.

2 Network on Chip Architecture

The Network-on-Chip (NOC) architecture, as outlined in figure ??, provides the communication infrastructure for the resources. In this way it is possible to develop the hardware of resources independently as stand-alone blocks and create the NOC by connecting the blocks as elements in the network. Moreover, the scalable and configurable network is a flexible platform that can be adapted to the needs of different workloads, while maintaining the generality of application development methods and practices.

A simple mesh interconnection topology is simplest from a layout perspective and the local interconnections between resources and switches are independent of the size of the network. Moreover,routing in a two-dimensional mesh is easy resulting in potentially small switches, high bandwidth, short clock cycle, and overall scalability. A NoC consists of resources and switches that are connected using channels as a mesh (Manhattan- like structure) so that they are able to communicate with each other by sending messages. A resource is a computation or storage unit. Switches route and buffer messages between resources. Each switch is connected to four other neighboring switches through input and output channels. A channel consists of two one-directional point-to-point buses between two switches or a resource and a switch. Switches may have internal queues to handle congestion. The precise layout and geometry depends on the technology generation. We expect that the area of a resource is the maximal synchronous region in a given technology. It is expected to shrink with every new technology generation. Consequently the number of resources will grow, the switch-to-switch and the switch-to-resource bandwidth will grow, but the network wide communication protocols will be unaffected. Figure ?? illustrates the principles of the physical floor plan within the NOC. Consider a 60nm CMOS technology expected in 2008, a 22mm 22mm chip size, a resource size of 2mm 2mm and a minimum wire pitch of 300nm. A NoC would accommodate 10 10 resources, each switch would occupy $30\mu m \times 30\mu m$ and the channels would be can use 3 metal layers for we have space for 300 wires. Since we need control, handshaking and signaling, this scenario would yield an effective data bus width of 256 wires.

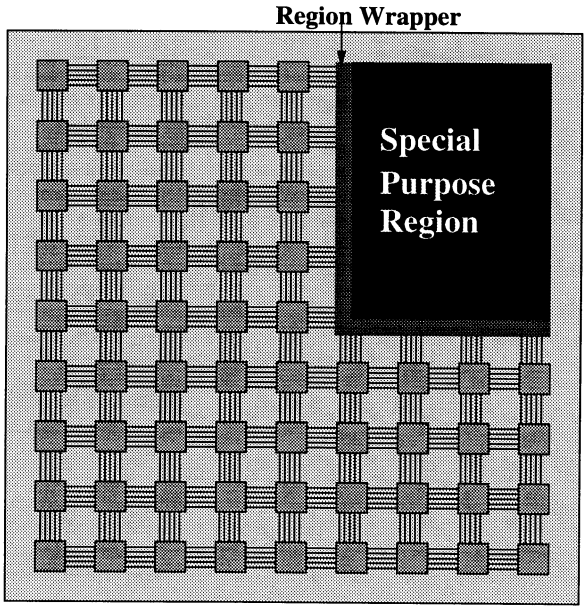


Figure 1: Each node in the mesh contains a switch and a resource.

A *Region*, as illustrated in figure ??, is an area inside the NoC which is insulated from the network by a wrapper. The internal architecture and organization of a region is invisible to the network and can be arbitrary. The network is not even aware of the presence

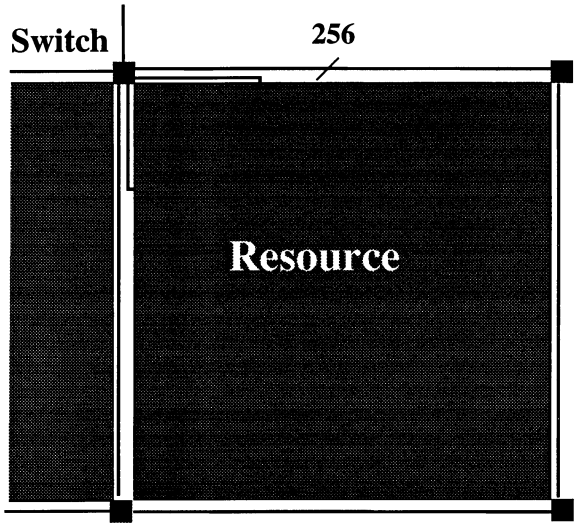


Figure 2: The expected footprints of resource, switch and channels in 60nm CMOS technology.

of the region because the region wrapper maintains the illusion of an undisturbed network to the outside by, for instance routing packets around rather than through the region if the packet's destination is on the other side. This concept allows to accommodate arbitrary sized resources in a flexible way.

3 NoC Implications

There are many compelling reasons for the emergence of NoCs although the architecture and organization may be different from everything proposed today. However, the precise application area and deployment mechanism is still very uncertain. Two scenarios can be envisioned. Scenario (A) holds that there will be a rather big number of NoC based platforms specialized for a particular product range. For instance, the NoC platform for cellular phones will be optimized for power efficiency and equipped with a limited but highly targeted set of functions and features. The NoC platform for base station will be optimized for optimized for performance and scalability to allow painless growth when higher performance or extended functionality is requested. NoC platforms for automotive and medical applications will be highly fault tolerance and safety critical while multi-media consumer devices will provide a variety of compression/decompression and encryption/decryption built-in functions. Development and verification cost for the platform, the communication infra structure, operating system services and design methodology and tools will be shared among all the products in an application domain. Conceptually, a NoC platform will be similar to dedicated embedded systems of today.

In contrast, scenario (B) holds that the volumes in individual application areas such as base stations, mobile phones, automotive devises, etc. are not sufficient to amortize the significant investment in platform development and manufacturing sites. Only the combined volumes of several of these application areas will justify the costs of 60nm-10nm CMOS fabs and the development and maintenance of a NoC platform architecture and design tools. The result would be a general purpose NoC assuming the role of today's CPUs as general computing engines for a wide range if not for all application areas including the PC. The consequences would be far reaching. Many of today's computer science and computer engineering disciplines have a sequentially operating single CPU as implicit or explicit fundamental assumption. These disciplines had to broaden their scope considerably and entire curricula had to be re-written. The bulk of theory and practice in disciplines such as computer languages, compiler construction, operating systems, programming, algorithm design, simulation, etc. are geared towards single CPU implementations. This is not to say that there is not a significant amount of knowledge and expertise on parallelism and concurrency in all these disciplines. But replacement of current CPUs by NoC based computing engines would bring concurrency to the mainstream in all disciplines and downgrade sequentiality to a limited and special case.

Several key developments are a prerequisite for scenario (B) to become a viable alternative.

NoC Architecture Since the resources in a NoC can be of different types from CPU cores, DSP cores, FPGA blocks to dedicated hardware blocks, a right mixture has to be found which suits all the different application areas sufficiently well. Furthermore, the communication network must have a sufficiently high bandwidth in addition to be able to accommodate various traffic types such as real-time traffic, regular data streams and irregular control messages.

NoC Assembler Language A standardized way to configure and program a NoC has to be developed which is independent from the particular NoC instance with a fixed number of resources. It has also to be independent from the various ways to specify and model the functionality of systems at a high level. In addition to capture the computation of resources it has to capture the acts of communication between resources.

NoC Operating System A standardized set of operating system services and interfaces has to be developed. A NoC operating system is a generalization of traditional operating systems and will include, among other features, the internal

and external communication, run-time error diagnostics and recovery, load balancing and dynamic task migration.

NoC Design Methodology A sufficiently robust and predictable design methodology, including the detailed methods and tools, has to include the mapping and verification of the functionality as well as of all the relevant non-functional requirements and constraints.

The great potential of NoC based platforms notwithstanding, it is far from obvious that this potential can be tapped. For this list illustrates clearly that if even one of the necessary components of a NoC platform is not developed well enough, the application of NoC concepts will be limited to rather small niches and not become mainstream.

4 Conclusion

We have briefly discussed the emerging concept of Networks-on-Chip (NoC) and we have described one concrete NoC architecture. This led us to speculate on the future and potential impact of NoC platforms. We have concluded this discussion with the observation that, even though the potential of NoC concepts may be tremendous, it is not likely that NoC can fulfill its promises due to the significant number of key components that have to be developed before NoC products can be deployed on a larger scale.

References

[1] Ahmed Hemani, Axel Jantsch, Shashi Kumar, Adam Postula, Johnny Öberg, Mikael Millberg, and Dan Lindqvist. Network on chip: An architecture for billion transistor era. In *Proceeding of the IEEE NorChip Conference*, November 2000.

[2] William J. Dally and Brian Towles. Route packets, not wires: On-chip interconnection networks. In *Proceedings of the 38th Design Automation Conference*, June 2001.

[3] Drew Wingard. MicroNetwork-based integration of SOCs. In *Proceedings of the 38th Design Automation Conference*, June 2001.

[4] Kurt Keutzer, Sharad Malik, Richard Newton, Jan Rabaey, and Alberto Sangiovanni-Vincentelli. System-level design: Orthogonalization of concerns and platform-based design. *IEEE Trasnactions on Computer-Aided Design of Integrated Circuits and Systems*, 19(12):1523–1543, Decmber 2000.

[5] Marco Sgroi, M. Sheets, A. Mihal, K. Keutzer, S. Malik, J. Rabaey, and A. Sangiovanni-Vincentelli. Addressing the system-on-a-chip interconnect woes through communication-based design. In *Proceedings of the 38th Design Automation Conference*, June 2001.

[6] Shashi Kumar, Axel Jantsch, Juha-Pekka Soininen, Martti Forsell, Mikael Millberg, Johnny Öberg, Kari Tiensyrjä, and Ahmed Hemani. A network on chip architecture and design methodology. In *Proceedings of IEEE Computer Society Annual Symposium on VLSI*, April 2002.

[7] T. Valtonen et al. Interconnection of autonomous error-tolerant cells. In *Proceedings of the International Symposium on Circuits and Systems*, Scottsdale, AZ, USA, 2002.

[8] T. Valtonen et al. An atonomous error-tolerant cell for scalable network-on-chip architectures. In *Proceedings of the 19th IEEE NorChip Conference*, Kista, Sweden, November 2001.

Layout Generation of Matched Capacitors

Emil Hjalmarson, Robert Hägglund, Per Löwenborg, and Lars Wanhammar

Department of Electrical Engineering
SE-581 83 Linköping, Sweden
E-mail: {emilh, roberth, perl, larsw}@isy.liu.se

ABSTRACT

In this paper we discuss layout generation of on-chip matched capacitors with accurate capacitance ratios. Capacitor sizing and placement suitable for design of error insensitive capacitance ratios are discussed including a proposed algoritm for common-centriod placement. Also, a regular layout structure is proposed that is suitable for implementation of capacitors with well matched capacitance ratios.

1. INTRODUCTION

Capacitors are common building blocks in analog circuits. Often, not only single capacitors with accurate capacitance values are used but also groups of two or more capacitors with specified capacitance ratios. It is therefore important that the capacitors are matched. Two capacitors are said to be *matched* if their capacitance ratio is determined to a precise value that is robust against changes of the conditions for which the value is achieved.

Accurate capacitance ratios are essential in, for example, switched-capacitor (SC) filters, reference level circuits in data converters, and in fully differential analog circuits in general. In SC filters, the time constants are functions of capacitance ratios rather than absolute values. Accurate capacitance ratios are therefore necessary in order to realize high accuracy SC filters. In data converters the linearity depends on the accuracy of the reference levels and accurate capacitance ratios are needed in order to preserve the symmetric property of a differential analog circuit. Poor symmetry leads to increased distortion. In general, well matched capacitors lead to analog circuits with more predictable characteristics.

For integrated analog circuits, the properties of the processing technology allow, if properly done, capacitors with precise capacitance ratios to be implemented. This is one of the major reasons for why SC filters have become so successful.

This paper addresses the problem of generation of on-chip polysilicon-to-polysilicon capacitor matrices of matched capacitors in a standard CMOS process. Even though the accuracy of the resulting capacitance ratios is emphasized, the capacitors are also designed to have certain absolute capacitance values. However, the achievable accuracy of the latter is substantially lower than that of capacitance ratios. Further, a regular layout structure is proposed that is suitable for implementation of matched and interference noise protected capacitors.

Section 2 reviews the major fabrication induced error sources for on-chip capacitors and their impact on matched capacitor design. Section 3 discusses suitable design and placement of the elementary building blocks, i.e., the unit-capacitors. Section 4 presents a proposed layout suitable for generation of matched on-chip capacitors. The proposed scheme follows the design and placement rules discussed in Section 3, yielding a design of matched capacitors which are highly robust against fabrication induced errors.

2. FABRICATION INDUCED ERRORS

In the following, the capacitors consist of parallel connections of several smaller so called *unit-capacitors*. The reason for considering such configurations is that the use of unit-capacitor structures is a successful way of overcoming or reducing the effect of various errors introduced during fabrication.

In the fabrication process of on-chip capacitors the capacitance value of a typical single capacitor can vary with up to 10 to 30 percent from the desired value [1]. Because of this it is difficult to produce high accuracy capacitors in a standard CMOS process as well as integrated circuits which rely on the accuracy of a single capacitance value. If, instead, capacitance ratios are the critical parameters, several techniques can be used in order to make the capacitors matched. These techniques require the understanding of the error sources of the fabrication process which can be categorized into random errors and systematic errors. Some of the most important error sources, regarding capacitor design, are discussed below.

2.1 Random errors

The random errors can be grouped into peripheral fluctuations and areal fluctuations. The peripheral fluctuations effect the perimeter of a surface, for example a polysilicon area, and may cause irregularities at the edge. The areal fluctuations are random variations in the thickness of layers in the design, for example the silicon oxide thickness between two polysilicon plates and the plates

themselves.

It has been shown that the relative capacitance error of a square shaped capacitor, C_k, due to random peripheral fluctuations is proportional to $C_k^{-3/4}$ while the relative error due to areal fluctuations is proportional to $C_k^{-1/2}$ [2].

It can also be shown that a realization of one large capacitor with capacitance C_k and a realization of N_k parallel connected smaller capacitors with capacitance C_k/N_k will have the same random relative error due to areal fluctuations. Parallel connected capacitors will, however, have a factor of $N_k^{1/4}$ larger relative errors due to peripheral fluctuations [2].

2.2 Systematic errors

The most important systematic errors that contribute to mismatched capacitors can be divided into the following groups [1], [3].

2.2.1 Oxide thickness variation

The capacitance value from a unit-capacitor depends on the thickness of the oxide layer between the capacitor plates. Usually, on a die surface, the gradient of the oxide thickness is slowly varying and can locally be approximated to be constant [1]. With this assumption, the effect of oxide thickness variation can be eliminated to a first order approximation by using a regular common-centroid layout scheme. For more complex thickness variations, common-centroid layout techniques will in general reduce the systematic error [4].

2.2.2 Thermal gradients

On-chip power devices cause thermal gradients. These thermal gradients will cause variations of the unit-capacitor values. To achieve high precision matching between capacitors they need to be placed as far away from the high power devices as possible. Further, effects of the thermal gradients can be decreased by using a regular layout style with well distributed unit-capacitors [1].

2.2.3 Process bias

The capacitance values of the fabricated capacitors will differ from the nominal values since the feature sizes of the photolithographic masks are scaled with some errors. A so called process bias occurs. To decrease the effect of the process bias it is necessary to use capacitors with identical area-to-perimeter ratio. Hence, the ratio between capacitances when unit-capacitors with equal capacitance values and geometries are used will be insensitive to uniform process bias.

2.2.4 Variable etch rate

Due to variable exposure of photoresist, the shape of the devices will vary depending on the surrounding geometries. It has been shown that these effects are much more significant to devices placed at the edge of the capacitor matrix [1], [3]. Etch rate variations of a surface are minimized if the surrounding geometries at all directions appears identical. This can be achieved by inserting so called dummy elements whose purpose is to maintain the symmetry at a non-symmetric direction. The dummy elements should be connected either to ground or a high impedance node.

2.2.5 Fringing capacitances

Fringing capacitances occur between the side-walls of the capacitor plates and between a side-wall and the layers in the near vicinity of the capacitor. Fringing capacitances will increase the effective capacitance and introduce cross-talk to elements surrounding the capacitor. The cross-talk can be decreased by using electrostatic shields. Such shields also decrease the fringing fields from the top capacitor plate to adjacent structures.

2.2.6 Interconnection capacitances

When a wire is connected to the top plate of a capacitor the wire may have to pass over the lower polysilicon plate. An interconnection capacitance is then introduced between the wire and the lower polysilicon layer, that is included to the total capacitance. Other interconnection capacitances occur between the wires to the top and bottom plate of the capacitor.

The effect of interconnection capacitances on matching of the capacitors can be reduced by designing the connecting wires to share a common-centroid point. In this way the interconnection capacitance contribution will be approximately equal for all capacitances when the oxide thickness is varied. If, further, the overlapping parts of the interconnections have identical area-to-perimeter ratios, the interconnection wires will be insensitive to process bias.

3. CAPACITOR DESIGN USING UNIT-CAPACITORS

In this section we discuss the design of capacitors using parallel unit-capacitors in such a way that the capacitors are matched and meet certain capacitance ratio requirements. For each unit-capacitor, a square chip area segment with sides of length L is utilized. The geometry of the unit-capacitors considered here are shown in Fig. 1.

3.1 Unit-capacitor geometry

Since the effect of peripheral fluctuations increases with the perimeter of the unit-capacitors, the perimeter should be made small. The unit-capacitors should also at the same time be designed to maximize the chip area utilization. If the unit-capacitor is designed as a square with side L the area-to-perimeter ratio will be $0.25L$ (This is the same area-to-perimeter ratio as that of a circle and an octagonal polygon with equal side lengths). If Available, the usage of 135° corners can further improve the area-to-perimeter ratio. For a typical unit-capacitor inscribed within a square with side L as shown in Fig. 1, there exist an optimum choice of the length x, where $0 \le x \le L$, that maximizes the unit-capacitor area-to-perimeter ratio.

One can show that the area-to-perimeter ratio is maximized for the unit-capacitor geometry shown in Fig. 1, if the length x is chosen according to

$$x = (\sqrt{1+\sqrt{2}})(2-\sqrt{1+\sqrt{2}})L \quad (1)$$

or $x \approx 0.693L$. With the choice of x as in (1) the resulting unit-capacitor area is given by

$$A_{\text{unit}} = 2\sqrt{2}(1+\sqrt{2})(\sqrt{1+\sqrt{2}}-\sqrt{2})L^2 \quad (2)$$

or $A_{\text{unit}} \approx 0.953L^2$. The resulting area-to-perimeter ratio is improved by about 5 percent compared with a square sized unit-capacitor.

3.2 Unit-capacitor placement

When designing well matched capacitors consisting of parallel unit-capacitors, the unit-capacitors are required to be properly placed in order to reduce the influence of the systematic error sources that can otherwise degrade the matching. As discussed above the use of unit-capacitors with the same capacitance values and identical geometries will cause the capacitance ratios to be insensitive to process bias errors. The unit-capacitors that are laid out are here, then, restricted to be identical.

In general, with M different capacitors C_k, $k = 0, 1, ..., M-1$, to be matched, the placement task can be formulated as placing N_k unit-capacitors, each of identical shape and size C_{unit}, in such a way that the influence of systematic errors is minimized. This is achieved here by utilizing the following four layout rules.

3.2.1 Common-centroid placement

When the average thickness of the oxide layer is different for each unit-capacitor, the total capacitance of all N_k unit-capacitors implementing C_k will change and thereby the different capacitance ratios will be altered. For a certain placement of the N_k unit-capacitors of a particular capacitor C_k, the effective capacitance will approximately be inversely proportional to the average oxide layer thickness. If the oxide layer thickness gradient is constant and if the unit-capacitors of a particular capacitor C_k are placed such that the placement is symmetric around a certain point (called the common-centroid point), the total capacitance C_k is determined by the oxide layer thickness at that particular point. The oxide layer thickness at this point approximates the average oxide layer thickness of C_k. Further, if the common-centroid points of all C_k's coincide all capacitance ratios are insensitive to oxide layer thickness variations (to a first order approximation), since the C_k's all share the same average oxide layer thickness.

3.2.2 Interdigitated placement

If unit-capacitors belonging to the same capacitor C_k, are placed in close vicinity to each other the capacitance value of C_k will become more sensitive to systematic errors occurring locally at that particular part of the capacitor matrix. Such error sources could, for example,

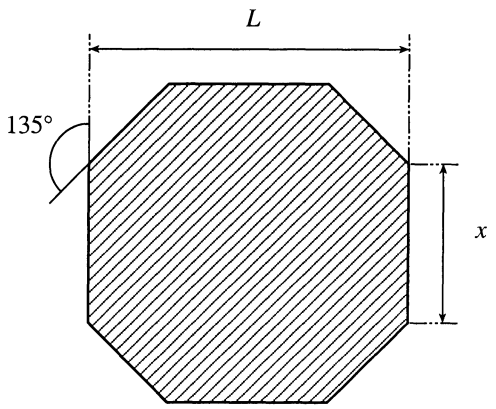


Figure 1. Geometry of the top plate of a unit-capacitor.

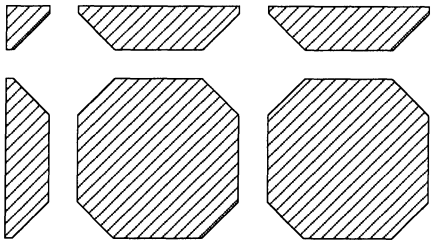


Figure 2. Upper left corner of a unit-capacitor matrix with inserted dummy elements.

be a large thermal gradient or higher order variation in the oxide layer thickness. If the unit-capacitors belonging to different C_k's are distributed as uniformly as possible throughout the available unit-capacitor area and placed in an interdigitated manner, the influence of errors tends to average out on all C_k's and thereby the capacitor mismatch is reduced.

3.2.3 Dummy element insertion

As mentioned in Section 2 the etch rate variations of the matrix of unit-capacitors are minimized if the surrounding geometries of each unit-capacitor appears identical at all directions. A unit-capacitor at the edge of the capacitor matrix will not experience the same surrounding geometries as a unit-capacitor at the interior. To circumvent such irregularities, so called dummy elements can be inserted as illustrated in Fig. 2. It should be noted that the dummy elements need not be of the same size as the cells used in the capacitor matrix as long as the symmetry is preserved.

3.2.4 Compact placement

The unit-capacitor matrix should be made as compact as possible. One compact structure, that is used in this paper, is a square shaped unit-capacitor matrix which has reasonably small maximum distance between two elements. Having as small maximum distance between unit-capacitors as possible reduces the effect of thermal gradients and makes the assumption of a constant oxide layer thickness gradient more accurate. However, restricting the total

number of unit-capacitors N_{tot} given by

$$N_{tot} = \sum_{k=0}^{M-1} N_k \tag{3}$$

to equal r^2, where $r > 0$ is an integer, can result in reduced accuracy of the designed capacitance ratios due to finite granularity of the fabrication process, as is discussed in the next section.

3.3 Unit-capacitor sizing

Since the effects of systematic errors such as variations in the oxide thickness will increase if the unit-capacitors are made too large, the side length of the square sized chip area segment, L, should not be made too large. At the same time, since the random peripheral fluctuations increase as $N_k^{-1/4}$ with the number of unit-capacitors, N_k, the unit-capacitor size should not be too small, since this results in a large N_k in order to realize a given total capacitance value. Hence, there exists a trade-off when choosing unit-capacitor size. Typically, for a quadratic shaped capacitor the side L is chosen to be between 20 and 50 μm [2].

Due to the finite granularity of the fabrication process the unit-capacitors can not be designed with arbitrary side lengths L, which means that it will belong to a finite set of values and the realizable unit-capacitor values belong to a discrete set of possible capacitance values. This, further, means that when realizing the desired capacitance values C_k with unit-capacitors there will be some errors. The absolute error ε_k of capacitor k is given by

$$\varepsilon_k = |N_k C_{unit} - C_k| \tag{4}$$

where C_{unit} is the unit-capacitance value, N_k is the number of unit-capacitors used, and C_k is the nominal desired capacitance value. The error of the capacitance ratio between two capacitors $\varepsilon_{k,l}$, is given by

$$\varepsilon_{k,l} = \left| \frac{N_k}{N_l} - \frac{C_k}{C_l} \right| \tag{5}$$

The design of the capacitors, i.e., finding the N_k's and a suitable unit-capacitor size C_{unit} given a set of nominal capacitance values C_k, can be made in several ways. For example, the absolute error in (4) could be minimized for all capacitors k, while constraining the relative error in (5) to be smaller than some specified value. Another possibility could be to minimize the relative error while constraining the absolute error. Also, other norms could be used to measure the errors that should be minimized.

Further, since the capacitor matrix should be square shaped the total number of unit-capacitors N_{tot} given by (3) could be specified to be a equal to r^2, where the integer $r > 0$. With such a constraint the resulting worst case accuracy of the capacitance ratios or absolute values could be unacceptably large. In such cases it might be favourable to let a percentage p, of the unit-capacitors in the capacitor matrix be unused, thereby specifying the total number of unit-capacitors to be within a specified range $(1-p)r^2 \le N_{tot} \le r^2$. These unit-capacitors should preferably be placed in the corners of the capacitor matrix in order to make the unit-capacitor placement as compact as possible. However, by doing so the capacitance-to-chip area ratio decreases. Therefore there exist a trade-off between utilization of chip area and accuracy.

3.4 Common-centroid algorithm

When the number of unit-capacitors N_k of size C_{unit} is determined for each capacitance value C_k, the unit-capacitors should be placed in such a way that they all share a common-centroid point in order to make the capacitance ratios less sensitive to first-order variations in the oxide layer thickness. This common-centroid point occurs at the centre of the unit-capacitor matrix.

An algorithm for placing the unit-capacitors with a common-centroid point can be established if considering the following rules:

1) If a capacitor k, consists of only one unit capacitor, that is one $N_k = 1$, the unit-capacitor matrix must consist of an odd number of unit capacitors thus r has to be an odd number. That particular unit-capacitor is placed in the centre of the matrix.
2) At most one capacitor must have $N_k = 1$. It is not possible to achieve a common-centroid placement for all capacitors if more than one capacitor consists of only one unit capacitor.
3) If r is an even number, N_k may not be odd for any capacitor.

In order to follow the rules mentioned above the number of unit-capacitors, as well as the unit-capacitor size, is needed to be restricted in order to be able to create a common-centroid placement. This can be done by including additional constraints to the sizing problem addressed in Section 3.3.

If the three rules above are followed, all capacitors can be realized as unit-capacitors with a common-centroid point by using the following procedure:

1) If there is one capacitor consisting of only a single unit-capacitor, place this capacitor in the centre of the structure and proceed to 3). Otherwise go to 2).
2) If there are capacitors consisting of an odd number of unit capacitors (any N_k is odd) place one of these unit-capacitors at the centre of the structure. Decrease N_k by one. The number of unit-capacitors remaining for this specific capacitor will now be even. Go to 3).
3) If there are capacitors with an odd number of unit-capacitors left, then place three unit-capacitor from each one of these so that these three unit-capacitors share the same common-centroid point. Decrease the corresponding N_k's by three. Go to 4).
4) Since all remaining N_k's are even, each pair of unit-capacitors belonging to the C_k's can be placed symmetrically with respect to the common-centroid point.

As mentioned in Section 3.2, the unit-capacitors

belonging to different C_k's should be distributed as uniformly as possible throughout the available unit-capacitor area. This can be achieved by placing the unit-capacitors according to the algorithm above and simultaneously restrict the placement of unit-capacitors belonging to the same C_k to be as distributed as possible over the entire capacitor matrix.

4. PROPOSED LAYOUT SCHEME

The proposed layout scheme for a grounded matched capacitor array is shown in Fig 3. In this case a common bottom plate is used. The black boxes indicate the routing area where each capacitor is connected to the output node with its own wire. The proposed layout scheme requires a process with at least three metal layers where the lowest metal layer is used as a shield. On top of the shield the remaining two metal layers are used for routing. Both the shield and the bottom plate are connected to ground along the edges of the structure.

To connect to the capacitors a hole is made in the shield layer on top of each capacitor. If the process do not allow stacked vias on top of the capacitor plate, the connection can with some loss of symmetry, be made at the edge of the plate instead. A guard-ring, not shown in the figure, surrounds the capacitor array. To further reduce the substrate noise a shielding well connected to a "quiet" voltage can be placed beneath the capacitor structure [4].

In the structure shown in Fig. 3, 16 unit-capacitors are used. The interconnection of the unit-capacitors is however not restricted to be applicable to an even number of cells. This can be realized by reducing the number of unit-capacitors to 9 by removing, for example, the left column and the bottom row of the unit-capacitor matrix while keeping the structure of the interconnection network unchanged for the remaining capacitors.

Due to the H-tree structure of the interconnection wires, all capacitors will experience the same RC-delay to the output node. This property can be of importance, for example when the capacitors are used as sampling capacitors in parallel sampling interfaces, in order to minimize skew of the sampling instants.

Further, to ensure good matching the interconnection capacitances must also be matched. For the proposed structure almost all wires will share a common-centroid point when r is even. Thereby the paracitics introduced in the interconnect wires will be the same for all capacitors in the array, and thus the capacitive ratio will to a first order not be effected by varying interconnection capacitances.

The top shield provides several important properties such as reducing the fringing fields between unit capacitors themselves and the surrounding structures [1]. Since all routing is done above the shield the wires will not have a direct capacitive coupling to the substrate, an thus the sensitivity to substrate noise will be reduced. The total area required for the capacitor array can also be reduced since the routing is done on top of the capacitors instead of between them. This decreases the systematic errors. The shielding layer will also act as an extra capacitor plate, forming a sandwich structure. This will increase the capacitance per unit area.

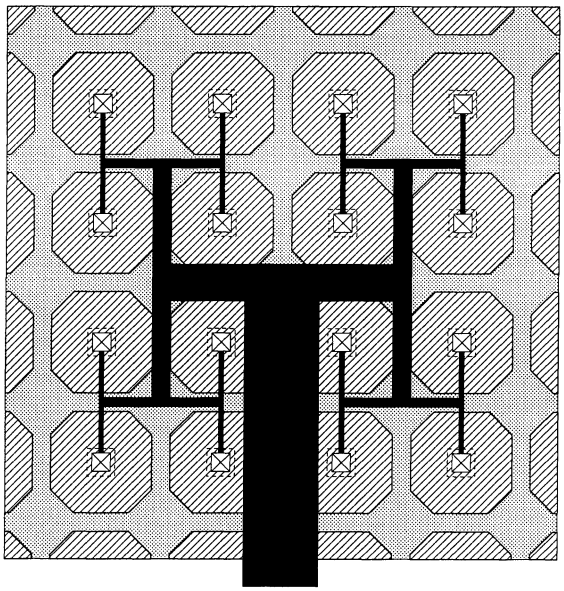


Figure 3. Proposed layout scheme.

5. CONCLUSION

We have discussed layout generation of on-chip matched capacitors with accurate capacitance ratios. Random and systematic error sources that are introduced during fabrication have been reviewed and proper sizing and placement techniques to reduce these errors have been discussed, including suitable capacitor sizing and an algorithm for capacitor placement. Also, a regular layout scheme and an interconnection strategy have been proposed that are suitable for implementation of capacitors with well matched capacitance ratios.

REFERENCES

[1] A. Hastings, "*The Art of Analog Layout*," Prentice Hall, 2001.

[2] J.-B. Shyu, G. C. Temes, and K. Yao, "Random Errors in MOS Capacitors," *IEEE J. Solid-State Circuits*, vol. SC-17, no. 6, Dec. 1982.

[3] M. J. McNutt, S.LeMarquis, J. L. Dunkley, "Systematic Capacitance Matching Errors and Corrective Layout Procedures," *IEEE J. Solid-State Circuits*, vol. SC-29, no. 5, May 1994.

[4] M. Ismail, T. Fiez, "*Analog VLSI Signal and Information Processing*," McGraw-Hill 1994.

[5] R. Singh, A. B. Bhattacharyya, "Roles of Corners in Matching of Linear MOS Capacitors," *IEEE Trans. Circuits Syst.*, vol. 36, no. 3, Mar. 1989.

Automatic Device Sizing in Analog Circuit Design

Robert Hägglund, Emil Hjalmarson, and Lars Wanhammar

Department of Electrical Engineering
SE-581 83 Linköping, Sweden
E-mail: roberth@isy.liu.se, emilh@isy.liu.se, larsw@isy.liu.se

ABSTRACT

In this paper we describe an optimization-based method for sizing the devices in an analog circuit to meet a set of given constraints. The current optimization method does not guarantee to find an optimal solution but finds a set of good solutions within a short time. To decrease the computation time no external simulators are used, instead the transistor model, BSIM3v3, has been integrated into the program. The motivation to use an optimization-based device sizing program for analog circuit design are to increase the design efficiency and use more complex cost functions.

1. INTRODUCTION

It is well known that a considerably amount of time is spent on the iterative process of device sizing using simulations of analog circuits. During this process it is difficult to obtain optimal performance since many of the performance metrics are nonlinear functions of the design parameters, i.e., transistor sizes and bias currents. Optimization-based methods can in this case be used to achieve good trade-offs between several performance metrics. Furthermore optimization-based device sizing can be used to improve the design efficiency for analog circuit design since the designer can work at a higher level of abstraction.

There are a multitude of optimizing techniques that can be used for designing analog circuits [1], for example knowledge-based methods, global optimization methods, geometric programming, and gradient-based optimization methods.

A knowledge-based system incorporate the know-how of an experienced analog designer. The advantage of this approach is the short computation time. Its drawback is the long set-up time for optimizing new structures since all circuits requires their own design rules. Further there is no guarantee that a global optimum is obtained. An example of a knowledge-based system is the IDAC program [2].

Global optimization methods on the other hand use optimization algorithms such as simulating annealing, genetic evolution, branch-and-bound, or exhaustive search. However, in practice non of these methods are guaranteed to find the global optimum since the optimization time is limited [3]. In most cases, however, a good solution is found within a short time. In our case we do not consider the computation time as a major problem.

An optimization method that has become popular during the last few years is geometric programming [3], [4]. All design equations are formulated as posynomials, transforming the problem into a convex problem for which a global optimum is easily found. The drawback is that accurate design equations cannot be formulated as posynomials, thus, the use of posynomials restricts the transistor model to be of low accuracy. Furthermore, it is difficult to incorporate certain types of constraints and the use of simplified models does not ensure that all transistors operate in the proper region.

Most classical optimization methods rely on gradient methods to find a local optimum. These can also be used for optimizing the performance metrics of an analog circuit. In [6], [7] simplified models together with a minimax formulation of the problem were used to find the DC operating points. Complete SPICE simulations can also be used in the iteration loop at a cost of very long simulation times [8].

Here we present on-going work to incorporate high accuracy transistor models in order to ensure accurate results. A minimax formulation of the optimization problem is currently used. The optimization scheme does not guarantee to find the optimal solution, but a set of good solutions are found. Future work involves the implementation of more sophisticated optimization methods.

2. THE OPTIMIZATION PROCEDURE

A flow chart of the optimization algorithm is shown in Figure 1. The optimization procedure is divided into two steps: first DC operating points meeting the DC constraint are found. Second, using the sizes generated by the first step, the AC performance metrics are optimized. Both the optimization steps are currently formulated as minimax problems. These problems can be solved using gradient based methods to minimize the objective function. Since this formulation does not result in a convex problem we cannot guarantee to find the global optimum. However, by initializing the optimization algorithm with several different starting points a set of good solution can be found.

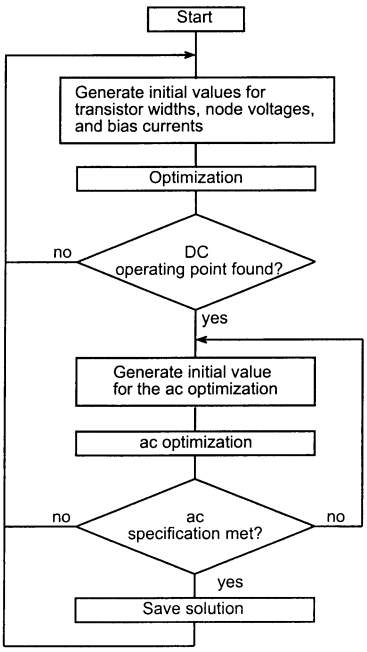


Figure 1. A flow chart of the optimization procedure.

2.1 Finding the DC operating point

To find suitable DC operation points an optimization technique are used to solve a non-linear system of equations. Each transistor are modelled as a black box. Given the node voltages V_G, V_S, V_D, and V_B along with the width and length of the transistor we can compute the drain source current, I_D, the threshold voltage, V_{TH}, and the saturation voltage, V_{DSAT}. These parameters are in this case computed using the BSIM3v3 transistor model [9]. The transistor models are implemented directly into the program to avoid the time consuming task of calling an external simulator program like for example Spectre™. Further, using the complete BSIM model for DC signals also provide high accuracy during the optimization process, ensuring that all transistors are biased in the proper operation region. Using a too simplified model results in large deviations in performance compared to simulations with high accuracy models. This phenomena can be encountered in programs using simple, but computational efficient, models used in for example [3].

The circuit topology itself introduces constraints on the node voltages and branch currents. For example, in Figure 2 the voltage at the drain of transistor M_1 and M_2 must be equal. Furthermore, in every node Kirchhoff's current law must be met.

Starting from an initial guess of node voltages, transistor widths, and bias currents a minimax optimization routine is used to determine the currents in all branches. For example, in the circuit shown in Figure 2 the current going out of M_1 must be the same as the current going into M_2. The objective function is therefore to minimize the difference between those currents by adjusting the width of the transistors, the node voltages and the available bias currents. Applying these equations together with constraints on the operation region of the transistors will result in a DC solution where all transistors are operating in their desired operation region.

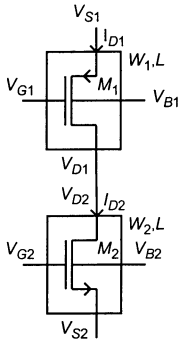


Figure 2. An example circuit

2.2 The AC optimization procedure

When the DC operation points are found a second optimization step is used to assure that the AC constraints is met. The optimization parameters are the unity-gain frequency, the values of the components in compensation network. A simplified model for the small signal characteristics is currently used to limit the number of parameters and thereby the computation time in the AC optimization step. The drawback of this is naturally some loss in accuracy. The errors introduced by using a simplified model can be decreased by applying a model where more capacitors are included and that their capacitance are more accurately modelled [10]. However, as will be shown later, the errors introduced by using a simple model will not have a major impact on the ability to determine whether a solution meet the specification or not.

When choosing what parasitics to include in the small signal model the size of the errors introduced in the cost function have been taken into consideration. For example a capacitor like C_{ds} can be ignored since it is very small compared to the other capacitors and, thus, have small impact on the performance metrics.

The transistors small signal parameters g_m, g_{mbs}, and g_{ds} have been computed by taking the numerical derivative of the drain current modelled by the DC equations in the BSIM3 model. This model is used for an accurate computation of the performance metrics. The small signal parasitic capacitances used here are the C_{gs} and the C_{gd}. They have been modelled using a low-order approximation [11] adjusted to the characteristics of the capacitances from simulations. The capacitor models are expressed as

$$C_{gs} = \frac{2}{3} C_{ox} W L k_{gs} + a_{gs}$$
$$C_{gd} = k_{gd} \Big/ \left(\sqrt{1 + \frac{V_{db}}{\Phi_s}} \right) \quad (1)$$

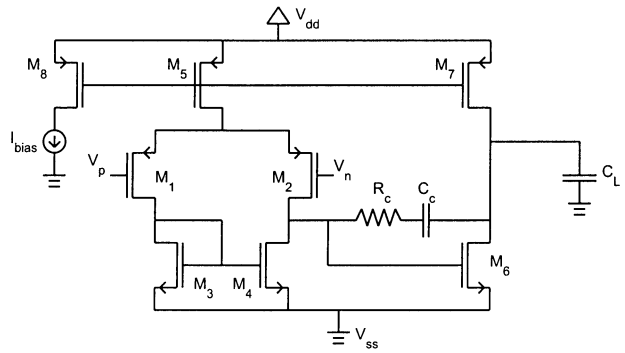


Figure 3. Single-ended two-stage operational amplifier.

where the factors k_{gs}, a_{gs} and k_{gd} are used as fitting parameters. Based on these AC parameters the frequency dependent properties of the circuit is computed.

A large number of starting points can be evaluated in the optimization process. At the end the solution that best fits the design objectives is chosen.

3. DESIGN EXAMPLE

The differential input single-ended output two-stage operational transconductance amplifier, OTA, used in this example is shown in Figure 3. This OTA is chosen in this design example since it is a commonly used all-purpose OTA. It has also a large output swing, high open loop voltage gain, and large common-mode input swing. The non-dominating pole, originating from the parasitic capacitance at the gate of transistor M_6, is its main drawback. The load capacitor C_L is set to be 5 pF and the power supply voltage is 3.3 V. Symmetry is obtain by setting the widths to be equal for transistors M_1, M_2 and M_3, M_4 respectively. The channel length, L, is set to 0.7 μm for all transistors. Currently all constraints used in this design example are derived using simplified transistor models.

3.1 DC Constraints

To meet different types of performance measures a set of constraints must be specified. A short description of the constraints that where used in this example is given below.

A constraint is that all transistors must operate in the saturation region. To guarantee this the following well-known relationship must be met for each transistor

$$\begin{aligned} V_{GS} &> V_T \\ V_{DS} &\geq V_{DSAT} \end{aligned} \tag{2}$$

where V_{DSAT}, is the minimum drain source voltage for each transistor. The quiescent power dissipation of the circuit is given by

$$P_{max} \geq (V_{dd} - V_{ss})(I_{D5} + I_{D7} + I_{D8}) \tag{3}$$

In, for example, analog filters where a single bias network are used for several OTAs the power dissipation in the bias network can be neglected. The output range is computed by the following expression.

$$\begin{aligned} OR_{max} &\leq V_{dd} - V_{DSAT7} \\ OR_{min} &\geq V_{DSAT6} \end{aligned} \tag{4}$$

The minimum and maximum voltage at the input of the amplifier is given by

$$\begin{aligned} CMR_{max} &\leq V_{dd} - V_{DSAT5} - V_{TH1} \\ CMR_{min} &\geq V_{TH3} + V_{DSAT1} - V_{TH1} \end{aligned} \tag{5}$$

In order to minimize the input offset voltage of the circuit the following relationship must be met.

$$\frac{2W_4W_7}{W_5W_6} = 1 \tag{6}$$

The common-mode rejection ratio at low frequencies is approximated by

$$CMRR_{min} \leq 2\frac{g_{m1}g_{m3}}{(g_{ds1} + g_{ds3})g_{ds5}} \tag{7}$$

The power supply rejection ratio from the negative supply can approximately be expressed as

$$negPSRR_{min} \leq \frac{g_{m1}g_{m6}}{(g_{ds1} + g_{ds4})g_{ds6}} \tag{8}$$

which is the PSRR at low frequencies. Constraints on the CMRR and PSRR for higher frequencies can easily be introduced if required, as for example in mixed-signal circuits where high frequency clock noise, etc., is injected into the power supplies. These high frequency constraints must be implemented in the AC optimization since they depend on the poles and zeros of the OTA.

3.2 AC Constraints

In the design of an amplifier it is common to have a specification on for example DC gain, unity-gain frequency, phase margin and so on. To meet the specification, constraint must be set during the AC optimization phase.

A simplified model for the transfer function of a two-stage operational amplifier is given by

$$A(s) = \frac{A_0(1 + s/z_1)}{\left(1 + \frac{s}{p_1}\right)\left(1 + \frac{s}{p_2}\right)\left(1 + \frac{s}{p_3}\right)\left(1 + \frac{s}{p_4}\right)} = \frac{A_0N(s)}{D(s)} \tag{9}$$

where A_0 is the DC gain, z_1 is the zero, p_i are the poles, $N(s)$ is the numerator and $D(s)$ is the denominator. The DC gain can be expressed as

$$A_0 = \frac{g_{m2}}{g_{ds2} + g_{ds4}} \cdot \frac{g_{m6}}{g_{ds6} + g_{ds7}} \tag{10}$$

if transistor M_5 is assumed to have zero output conductance. Furthermore, the dominant pole is given by

$$p_1 = \frac{g_{m2}}{A_0C_c} \tag{11}$$

The pole originating from the output stage is

$$p_2 = \frac{g_{m6}}{C_c(C_1 + C_{TL}) + C_1 C_{TL}} \quad (12)$$

where C_1, the parasitic capacitance at the gate of transistor M_6, can be expressed as

$$C_1 = C_{gs6} + C_{db2} + C_{db4} + C_{gd2} + C_{gd4} \quad (13)$$

and C_{TL}, the total capacitance at the output node, is given by

$$C_{TL} = C_L + C_{db6} + C_{db7} + C_{dg6} + C_{dg7} \quad (14)$$

Only C_{gs} and C_{db} are implemented in the optimization program and therefore the AC properties will not have the same accuracy as for example in the Spectre™ simulator.

The third pole is introduced in the current mirror and it is given by

$$p_3 = \frac{g_{m3}}{C_2} \quad (15)$$

where the capacitance C_2 at the gate of transistor M_3 is given by

$$C_2 = C_{gs3} + C_{gs4} + C_{db1} + C_{db3} + C_{db4} \quad (16)$$

The compensation pole is given by

$$p_4 = \frac{1}{C_1 R_c} \quad (17)$$

and the compensation zero is

$$z_1 = -\frac{1}{\left(\frac{1}{g_{m6}} - R_c\right) C_c} \quad (18)$$

The zero can be placed at different locations. For example far away in the right hand plane, at infinity, or in the left hand plane at $z_1 = -1.2\omega_u$, depending of the value of R_c.

In this design example the zero is chosen to be in the left hand plane about 1.2 times the unity-gain frequency of the amplifier. This position of the zero will increase the phase margin and the unity-gain frequency [11].

At the unity-gain frequency the magnitude of the gain equals unity. This can be reformulated as

$$A_0|N(jw_u)| - |D(j\omega_u)| = 0 \quad (19)$$

The phase margin can be found according to

$$\phi_M = \pi + \text{atan}\frac{\omega_u}{z_1} - \text{atan}\frac{\omega_u}{p_1} - \text{atan}\frac{\omega_u}{p_2} - \text{atan}\frac{\omega_u}{p_3} - \text{atan}\frac{\omega_u}{p_4} \quad (20)$$

Another performance measures of interest is the slew rate which can be approximated as

$$SR = min\left(\frac{I_{D5}}{C_c}, \frac{I_{D7}}{C_c + C_{TL}}\right) \quad (21)$$

One problem with the approach given above is to find the poles and zeros of the circuit. Usually it is easier to derive the transfer function than factorizing the numerator and denominator into accurate expressions for the poles and the zeros. In our case it is possible to compute the unity-gain frequency and the phase margin directly from the transfer function [10].

4. PERFORMANCE OF THE OPTIMIZATION

All simulations and optimizations where carried out with parameters from a 0.35 µm CMOS process. The simulations were performed in the Cadence Spectre™ simulator. The two-stage amplifier have been optimized to meet one specification. Several solutions were found and simulated in the Spectre™ simulator to evaluate the performance of the optimization program. At the end the best solution, i.e., the solution with the best value of the objective function is chosen.

4.1 Specification

The specification together with the simulation results is shown in Table 1. In this case a low-power specification was chosen. The objective function was to maximize the unity-gain frequency. The sizes of the transistors in the OTA is shown in Table 2.

Table 1. Specification of the two-stage OTA.

Performance measure	Specification	Optimized circuit
DC gain	> 80 dB	83.6 dB
Unity-gain frequency	>100 MHz	105.5 MHz
Phase margin	> 60°	61.3°
Slew rate	>100 $V/\mu s$	126 $V/\mu s$
Common-mode rejection ratio	> 60 dB	84.6 dB
negative PSRR	> 80 dB	92.2 dB
Power dissipation	< 2 mW	1.9mW
Output range	0.4 - 2.8 V	0.15 - 3.15 V
Common-mode range	0.0 - 2.0 V	0.0 - 2.4 V

4.2 Device sizes

The sizes of the transistors, compensation capacitor and resistor, and the bias current are shown in the Table 2 for the best solution.

Table 2. The device sizes of the OTA.

Variable	Circuit	Variable	Circuit
W_1=W_2	252.9 µm	W_8	178.5 µm
W_3=W_4	79.6 µm	C_c	1.98 pF
W_5	373.1 µm	R_c	918.2 Ω
W_6	187.4 µm	I_{bias}	104.7 µA
W_7	409.5 µm	L	0.7 µm

5. COMPARISON WITH SPECTRE™ SIMULATIONS

Due to the use of simplified the small signal model some differences between the resulting performance of the optimization program and the Spectre™ simulator is

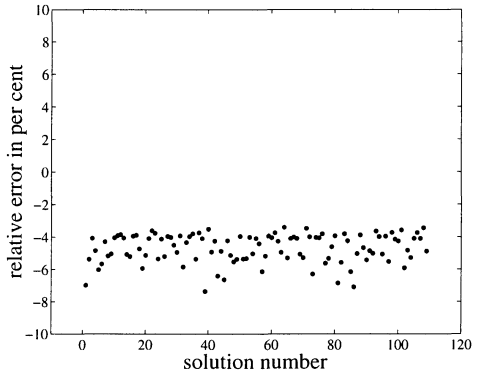


Figure 4. Deviation of the unity-gain frequency between the optimization program and simulator.

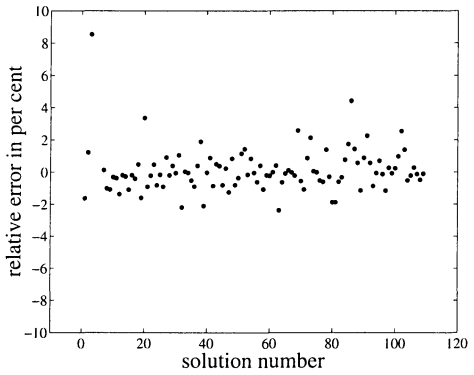


Figure 5. DC gain deviations between optimized and simulated performance.

obtained. The discrepancy for the unity-gain frequency and the DC gain have been evaluated for about 100 different solutions. The results are shown in Figure 4 and Figure 5, respectively.

The deviation in the unity-gain frequency can be explained by three causes. The first one is the simplified models used for the poles and zeros of the circuit. Second, all parasitic capacitances in the small signal model for the transistor are not incorporated in the optimization program. The last cause is the simplified modelling of the capacitors used in the current version of the optimization program.

The deviation in the DC gain can be explained by the use of the simplified model for the DC gain. Furthermore, the computation of the g_m, g_{mbs}, and g_{ds} values differs a little with the one used in the simulator. These parameters are derived by taking the numerical derivative of the current with respect to V_{GS}, V_{BS}, and V_{DS} respectively in the operation point.

6. FUTURE WORK

The optimization procedure used in this paper is divided into two different steps. On-going work is to integrated both optimization steps into a single step. A single optimization step makes it possible to minimize the power dissipation for the specification. Future work will include optimization of more complex performance metrics. The implementation of a corresponding layout generator will allow back annotation of parasitic properties introduced in the layout phase and improve the design efficiency significantly.

7. CONCLUSION

A first version of an optimization-based device sizing program has been implemented. The program is capable to size the transistors of analog circuits, such as operational amplifiers, to meet certain performance metrics. We can not guarantee that the optimizer finds the global optimum since the problem is not convex, but it can generate a set of good solutions. Reconfigurating the optimization program for a new circuit can be done in a reasonable short time. The program uses the complete BSIM3v3 transistor model for the DC operation points to ensure that each transistor are biased to operate in the desired operation region. Simplified small signal model for the transistors are used to decrease the AC optimization time. The designed circuits have been verified in a Spectre^{TM} simulation to evaluate its performance.

REFERENCES

[1] R.K. Brayton, G.D. Hachtel, and A.L Sangiovanni-Vincentelli, "A survey of optimization techniques for integrated-circuit design," *Proc. IEEE*, vol. 69, pp. 1334-1362, Oct. 1981.

[2] M. G. R Degrauwe, et. al., "IDAC: An interactive design tool for analog circuit design," *IEEE Trans. on Computer-Aided Design*, vol. 8, pp. 1247-1265, June 1989.

[3] M. M. Hershenson, S.P. Boyd, and T.H. Lee, "Optimal Design of a CMOS Op-Amp via Geometric Programming," *IEEE Trans. Computer-Aided Design*, vol. 20, pp. 1-21, Jan. 2001.

[4] P. Mandal and V. Visvanathan, "CMOS op-amp sizing using a geometric programming formulation,", *IEEE Trans. Computer-Aided Design*, vol. 20, pp. 22-28, Jan. 2001.

[5] A.R. Conn, et. al., "Optimization of custom MOS circuits by transistor sizing," *Proc. Int. Conf. Computer-Aided Design*, pp. 174-180, 1996, San Jose, USA.

[6] P. Heikkilä, M. Valtonen, and K. Mannersalo, "CMOS op amp dimensioning using multiphase optimization," *Proc. IEEE Int. Symp. Circuit Systems*, pp. 167-170, 1988, Helsinki, Finland.

[7] F. Leyn, W. Daems, G. Gielen, and W. Sansen, "Analog circuit sizing with constraint programming modelling and minimax optimization,", *Proc. Int. Symp. Circuits Systems*, pp. 1500-1503, 1997, Hong Kong.

[8] W. Nye, et. al., "DELIGHT.SPICE: An optimization-based system for the design of integrated circuit," *IEEE Trans. Computer-Aided Design*, vol. 7, pp. 501-519, 1988.

[9] Y. Cheng, et. al., "BSIM3v3 manual (final version)," 1996.

[10] R. Hägglund, E. Hjalmarsson, and L. Wanhammar, "Optimization-based device sizing in analog circuit design," *Proc. Swedish System-on-Chip Conf.*, Falkenberg, Mar. 2002.

[11] D.A. Johns and K. Martin, *Analog integrated circuit design*, Crawfordsville, John Wiley & Sons, 1997.

Optimal logic level pipelining for digit-serial implementation of maximally fast recursive digital filters

Oscar Gustafsson and Lars Wanhammar

Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, SWEDEN
E-mail: {oscarg, larsw}@isy.liu.se

Abstract

Although recursive filters are not possible to pipeline at the algorithmic level, it is possible to introduce logic level pipelining and thereby increase the maximal sample rate. This is because the critical path gets shorter so that the circuitry can be clocked faster. However, the latency of the operations increases with increased pipelining. It is of interest to find the optimal degree of pipelining to be able to maximize the sample rate of the implementation. In this paper the optimal degree of logic level pipelining is derived for maximally fast implementations of recursive digital filters.

1 Introduction

Frequency selective digital filters are used in many contemporary digital signal processing systems. As more and more equipment is portable, the power consumption is becoming an important design criterion. Maximally fast implementations and bit-serial arithmetic has been shown to be an efficient way to implement recursive digital filters with low power consumption [1] [2]. With maximally fast (or rate optimal) we denote implementations of recursive filters which obtain the maximal sample frequency. The maximal sample frequency of a recursive filter is

$$f_{max} = \min_i \left\{ \frac{N_i}{T_{op,i}} \right\} \tag{1}$$

where N_i is the number of delay elements and $T_{op,i}$ is the total latency of the operations in the recursive loop i [3].

For bit- and digit-serial arithmetic the latency of an operation is proportional to the number of fractional bits of the coefficient as

$$T_{op,mult} = \left\lceil \frac{W_f}{d} \right\rceil T_{clk} \tag{2}$$

where W_f is the number of fractional bits for the multiplier coefficient, d is the digit-size ($d = 1$ for bit-serial), and T_{clk} denotes the clock frequency. Hence, the latency of a multiplication depends on both the coefficient of the multiplier and the clock period of the system. The clock period of the system depends on the longest logical path between two registers, and, thus, depends on the interconnection between different operations, while the number of clock periods only depends on the number of cascaded multipliers and, thus, the filter algorithm. It is therefore convenient to divide the latency in algorithmic latency, L_{op}, and clock period, T_{clk}, as $T_{op} = L_{op} T_{clk}$. In the same way is it possible to divide T_{min} in L_{min} and T_{clk}.

In earlier work three different logic models for bit-serial arithmetic has been defined depending on level of pipelining; one for static logic, one for dynamic logic and one for pipelined dynamic logic [1]. In this work the pipelining is generalized to be performed after a number of operations. Further, digit-serial arithmetic is also considered.

2 Cyclic Scheduling

To obtain a maximally fast implementation using bit- or digit-serial arithmetic the operations must be scheduled over more than one sample period [4]. The minimum number of periods to schedule for should be selected as

$$m \geq \left\lceil \frac{\max_j (T_{exe,j})}{T_{min}} \right\rceil \tag{3}$$

where $T_{exe,j}$ is the time between two adjacent inputs for operation j. The total scheduling period is then mT_{min}. This gives that m must be selected so that mT_{min} yields a schedule over an integer number of clock periods.

To schedule over m sample periods the algorithm must be unfolded m times. An illustration of unfolding is shown in Fig. 1.

The cyclic scheduling formulation exploits the fact that the algorithm is recursive. The scheduling formulation can be viewed on the circumference of a cylinder as shown in Fig. 2. One property of a maximally fast cyclic schedule is that all operations in the critical loop are scheduled adjacent, without any shimming delays, to each other. As it is possible to introduce pipelining in all non-critical operations, the maximal clock frequency will be determined by a critical path going through the operations in the critical loop only.

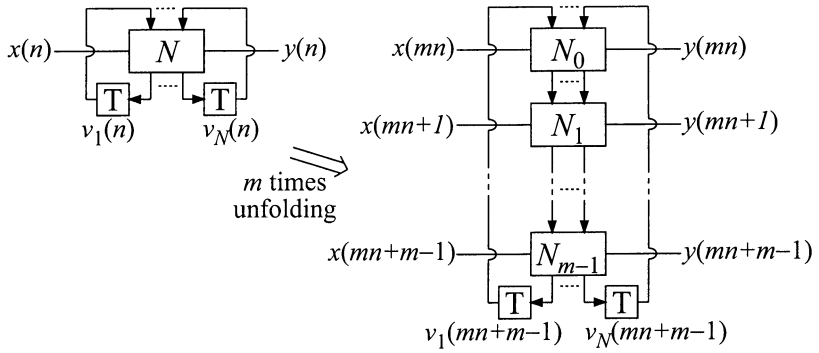


Figure 1: *Illustration of unfolding an algorithm m times.*

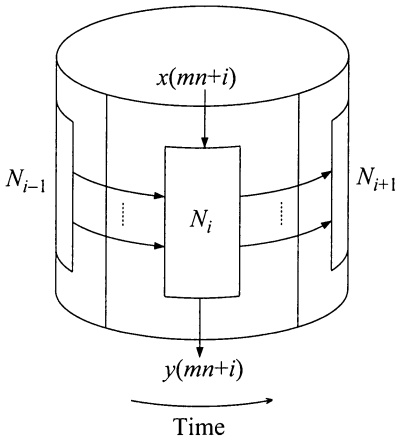


Figure 2: *Illustration of cyclic scheduling.*

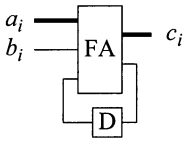


Figure 3: *Bit-serial adder with critical path marked.*

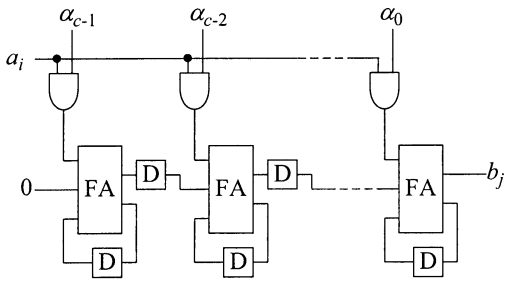


Figure 4: *General coefficient bit-serial serial/parallel multiplier.*

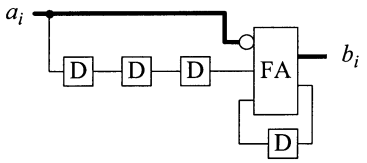


Figure 5: *Simplified bit-serial multiplier with coefficient 0.4375. The critical path is marked with a thick line.*

3 Bit- and Digit-Serial Operations

In bit-serial processing one bit of the input word is processed each clock cycle. A bit-serial adder is shown in Fig. 3. The algorithmic latency of a bit-serial adder is zero clock cycles. However, the critical path of the adder will affect the maximal clock frequency. Bit-serial multiplication can be performed using a serial/parallel multiplier as shown in Fig. 4. For a fixed coefficient, these multipliers can be simplified to reduce the hardware complexity [5]. A simplified serial/parallel multiplier for coefficient $0.4375 = 0.0111_{2C} = 0.100\text{–}1_{CSD}$ is shown in Fig. 5. As stated previously the algorithmic latency is proportional to the number of fractional bits of the coefficient. The critical path of the multiplier is at most one full adder as can be seen from Fig. 5. If the coefficient only has one non-zero bit, no full adder is required, and, thus, that operation will not alter the critical path. However, it must still be included for the algorithmic latency, off course.

A digit-serial adder can be designed by unfolding a bit-serial adder [6]. In Fig. 6 a digit-serial adder with digit-size 3 is shown. The critical path from input to output is one full adder, but for some operation there will be a critical path from the input to the carry register of the adder. Digit-serial serial/parallel multipliers can be designed and simplified in a similar way as bit-serial serial/parallel multipliers. The latency is as stated in (2), while the critical path is the same as for a digit-serial adder.

4 Pipelining

Although it is not possible to pipeline recursive loops at the algorithm level, we can introduce pipelining at the logic level. The critical path goes through the operations in the critical loop unfolded m times. This limits the sample period to $T_s = T_{min} = L_{min} T_{clk}$. Introducing a pipelining stage corresponds to increasing the latency for an operation with one clock cycle. However, it may also shorten the critical path and thereby allow a decrease of the clock period. Hence, the minimal sample period increase and the clock period decrease. It is therefore possible to obtain an optimal level of pipelining that maximizes the sample rate.

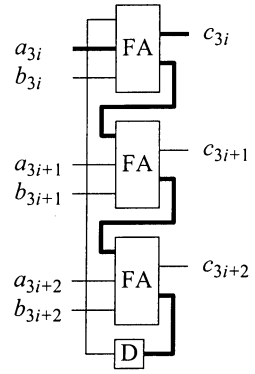


Figure 6: *Digit-serial adder with digit-size 3. The critical paths are marked with a thick line.*

In this section we derive the optimal level of pipelining for maximally fast bit- and digit-serial implementations of recursive algorithms.

4.1 Bit-serial implementation

Assume that the critical loop has N_{op} operations and the algorithm is unfolded m times according to (3). Only operations that adds to the critical path should be included in N_{op}. Multiplications with only one non-zero bit in the coefficient should therefore not be included. For bit-serial arithmetic each operation will add a time of T_{FA} to the critical path, where T_{FA} is the gate delay of a full adder. Thus, the minimal clock period is $T_{clk} = mN_{op}T_{FA}$. If the algorithmic latency in the critical loop is $L_{tot} = \Sigma W_{f,i}$ and there is one delay element in the critical loop, the minimal sample period is $T_s = L_{tot}mN_{op}T_{FA}$.

If pipelining is introduced after every M operations the critical path is reduced to

$$T_{clk} = MT_{FA} + T_D \tag{4}$$

where T_D is the setup time of the pipelining registers plus the propagation delay from the clock edge to the output of the registers. The algorithmic latency of the critical loop is now

$$L_{min} = L_{tot} + \frac{N_{op}}{M} \tag{5}$$

The resulting minimal sample period is

$$\begin{aligned} T_s &= L_{min}T_{clk} \\ &= \left(L_{tot} + \frac{N_{op}}{M}\right)(MT_{FA} + T_D) \\ &= (ML_{tot} + N_{op})T_{FA} + \left(L_{tot} + \frac{N_{op}}{M}\right)T_D \end{aligned} \tag{6}$$

Taking the derivative of the minimal sample period with respect to M we find

$$\frac{\partial T_s}{\partial M} = L_{tot}T_{FA} - \frac{N_{op}}{M^2}T_D \tag{7}$$

and the optimal level of pipelining is found to be

$$M = \sqrt{\frac{N_{op}T_D}{L_{tot}T_{FA}}} \tag{8}$$

For bit-serial arithmetic we can utilize the logic-level implementation to transfer from a static CMOS implementation to a dynamic implementation. Is is possible to merge the full adder and the flip-flop to obtain a more efficient implementation [1]. This corresponds in the simplest case to $M = 1$, i. e., one flip-flop after each full adder. However, it is possible to introduce two flip-flops and pipeline the full adder internally as in Fig. 7. This corresponds to $M = 1/2$. By allowing pipelining inside the adder it should also be possible to obtain a resolution of 1/2 for M.

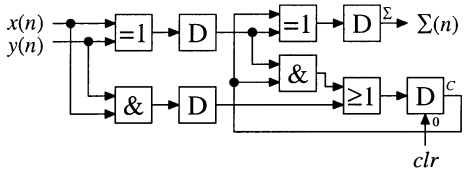


Figure 7: *Bit-serial adder with two pipeline stages.*

4.2 Digit-serial implementation

The minimal clock period for an implementation with digit-size d is $T_{clk} = (mN_{op} + (d-1))T_{FA}$, i.e., mN operations plus $d-1$ extra full adders at the last operation as shown in Fig. 6. The algorithmic latency of the critical loop is now $L_{tot} = \Sigma\lceil W_{f,i}/d \rceil$. Introducing pipelining after every M operations reduces the critical path to

$$T_{clk} = (M + d - 1)T_{FA} + T_D \tag{9}$$

while the algorithmic latency increases to

$$L_{min} = L_{tot} + \frac{M}{N_{op}}. \tag{10}$$

The minimal sample period is now

$$\begin{aligned} T_s &= L_{min}T_{clk} \\ &= \left(L_{tot} + \frac{N_{op}}{M}\right)((M + d - 1)T_{FA} + T_D) \\ &= \left((M + d - 1)L_{tot} + N_{op} + \frac{N_{op}(d-1)}{M}\right)T_{FA} + \\ &\quad \left(L_{tot} + \frac{N_{op}}{M}\right)T_D \end{aligned} \tag{11}$$

Again, taking the derivative of the minimal sample period with respect to M yields

$$\frac{\partial T_s}{\partial M} = \left(L_{tot} - \frac{N_{op}(d-1)}{M^2}\right)T_{FA} - \frac{N_{op}}{M^2}T_D \tag{12}$$

and the optimal level of pipelining is

$$M_{opt} = \sqrt{\frac{N_{op}T_D + N_{op}(d-1)T_{FA}}{L_{tot}T_{FA}}} \tag{13}$$

To see how the optimal pipelining level M_{opt} is affected by the digit-size d we first assume

$$T_D = T_{FA} \tag{14}$$

which yields

$$M_{opt} = \sqrt{\frac{N_{op} d}{L_{tot}}} \tag{15}$$

Hence, the optimal level of pipelining is proportional to the square root of the digit-size. However, as the algorithmic latency is proportional to the digit-size we notice that

$$L_{tot} = \sum_i \left\lceil \frac{W_{f,i}}{d} \right\rceil \approx \frac{\sum_i W_{f,i}}{d} \tag{16}$$

(equality if all $W_{f,i}$ is an integer multiple of d). We now obtain

$$M \approx d \sqrt{\frac{N_{op}}{\sum_i W_{f,i}}} \tag{17}$$

Thus, the optimal level of pipelining is directly proportional to the digit-size.

5 Examples

In Fig. 8 a second-order lattice wave digital filter allpass section is shown with its critical loop marked. The reason for using this structure is that it is suitable for VLSI implementation as it has a regular, low complexity structure, low coefficient sensitivity, and yields a regular implementation. Is is clear that the number of operations in the critical loop, N_{op}, is six. The multipliers α_1 and α_2 will both contribute to the algorithmic latency of the loop, L_{tot}. A cyclic scheduling formulation for this filter is found in [1]. Cyclic scheduling formulations for other types of wave digital filters are found in [7]–[9].

If $T_{FA} = T_d = 1$ ns is assumed then the estimated sample period is plotted in Fig. 9 for different number of fractional bits and bit-serial arithmetic. Figure 9 shows that the longer coefficients present in the loop, the more pipelining stages can be introduced with a decrease of the sample period. Further, it can also be seen that more pipeline stages can be introduced if the number of fractional bits are high compared with a lower number of fractional bits.

Considering digit-serial arithmetic with the same assumptions and four fractional bits for both coefficients, the sample periods shown in Fig. 10 are obtained. From Fig. 10 it can be seen that it is important that the digit-size is an integer multiple of the number of fractional bits. The minimal sample period is obtained for a digit-size of one, two and four bits. This is due to the fact that the latency of a digit-serial multiplication is $\lceil W_f/d \rceil$. Thus, if the digit-size is not an integer multiple of the number of fractional bits, algorithmic latency will be rounded towards the next larger integer. It can also be seen that the optimal level of pipelining is dependent on the digit-size. This is an expected result as can be seen from (13). This is shown more clearly in Fig. 11 where the optimal level of pipelining is plotted versus digit-size with three and five fractional bit coefficients for the multipliers, respectively.

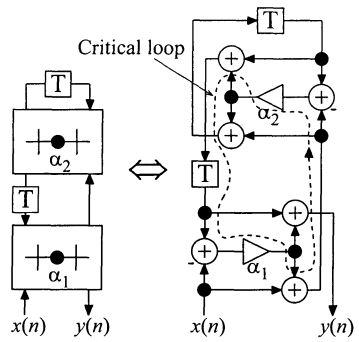


Figure 8: *Second-order lattice wave digital filter allpass section.*

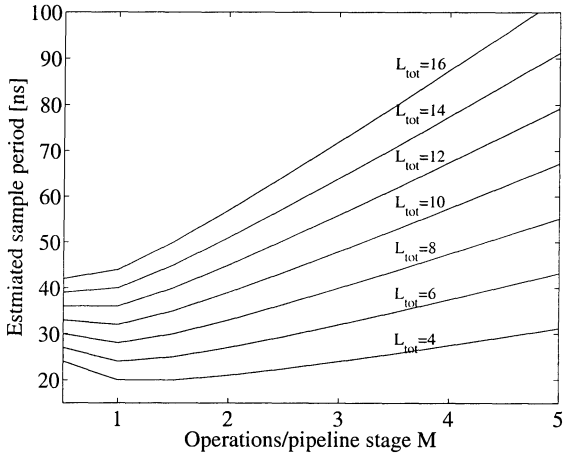


Figure 9: *Sample period for second-order allpass sections using bit-serial arithmetic as a function of the degree of pipelining.*

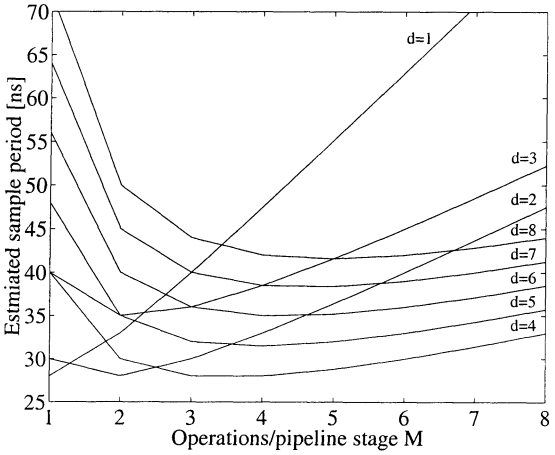


Figure 10: *Sample period for second-order allpass sections using digit-serial arithmetic as a function of the degree of pipelining.* $W_{f,\alpha 1} = W_{f,\alpha 2}$

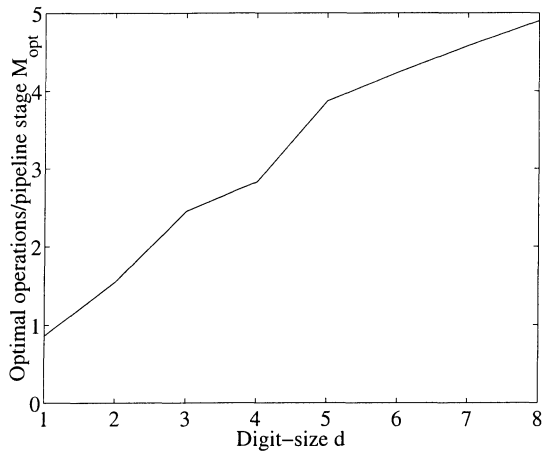


Figure 11: *Optimal level of pipelining as a function of digit-size for second-order allpass section with $W_{f,\alpha 1}$= 3 bits and $W_{f,\alpha 2}$ = 5 bits.*

6 Conclusions

In this work the optimal degree of logic level pipelining was derived for maximally fast recursive digital filters using bit- and digit-serial arithmetic. Even though it is not possible to introduce pipelining on the algorithmic level, it is possible to pipeline the operations at the logic level. Pipelining will increase the algorithmic latency, but the maximal clock frequency will also increase.

References

[1] M. Vesterbacka, *On Implementation of Maximally Fast Wave Digital Filters*, Linköping Studies in Science and Technology, Dissertation No. 487, Linköping University, Sweden, 1997.

[2] L. Wanhammar, *DSP Integrated Circuits*, Academic Press, 1999.

[3] M. Renfors and Y. Neuvo, "The maximum sampling rate of digital filters under hardware speed constraints," *IEEE Trans. Circuits Syst.*, vol. 28, no. 3, pp. 196–202, March 1981.

[4] K. Palmkvist, M. Vesterbacka, P. Sandberg, and L. Wanhammar, "Scheduling of data-independent recursive algorithms," *Proc. European Conf. Circuit Theory Design*, Istanbul, Turkey, Aug. 27-31, 1995, vol. 2, pp. 855–858.

[5] M. Vesterbacka, K. Palmkvist, and L. Wanhammar, "Realization of serial/parallel multipliers with fixed coefficients," *Proc. National Conf. Radio Science (RVK)*, Lund, Sweden, April 5–7, 1993, pp. 209–212.

[6] K. K. Parhi, "A systematic approach for design of digit-serial signal processing architectures," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 358–375, Apr. 1991.

[7] O. Gustafsson and L. Wanhammar, "Implementation of maximally fast ladder wave digital filters using a numerically equivalent state-space representation," *Proc. IEEE Int. Symp. Circuits Syst.*, Orlando, Florida, May 30–June 2, 1999, vol. 3, pp. 419–422.

[8] O. Gustafsson and L. Wanhammar, "Maximally fast scheduling of bit-serial lattice wave digital filters using constrained third-order sections," *Proc. IEEE Int. Conf. Electronics Circuits Syst.*, Paphos, Cyprus, Sept. 1999, pp. 729–732.

[9] O. Gustafsson and L. Wanhammar, "Maximally fast scheduling of bit-serial lattice wave digital filters using three-port adaptor allpass sections," *Proc. Nordic Signal Processing Symp.*, Norrköping, Sweden, June 13–15, 2000, pp. 441–444.

Reduction of clock noise in mixed-signal circuits

Erik Backenius, Mark Vesterbacka, and Robert Hägglund

Department of Electrical Engineering
SE-581 83 Linköping, Sweden
E-mail: {erikb, markv, roberth}@isy.liu.se

ABSTRACT

A major concern in mixed-signal circuits is the noise injected by the digital circuits into sensitive analog circuits. Of particular interest in this work is the problem with large capacitive coupling between the digital clock network and the substrate shared with the analog circuits. It is in general more easy to reduce low frequency noise compared with high frequency noise. Therefore, we have developed a strategy where we reduce the high frequency content of the clock by using smooth clock edges, and a special digital flip-flop circuit. This strategy will be evaluated in a test chip where we can control the rise and fall time of the clock edges of a high-performance digital FIR filter, and measure the performance of a fifth-order analog active-RC filter.

1. INTRODUCTION

Microelectronics is heading towards integration of full systems on a single chip (SoC), including both analog and digital circuits (mixed-signal). A major problem related with mixed-signal circuits is the noise injected into the analog circuits by the switching digital circuits, which is illustrated in Fig. 1. This noise is degrading the performance of the analog circuits [1]. Current research on noise in mixed-signal circuits is focusing on modeling techniques and strategies for reducing noise. This work is focusing on the latter.

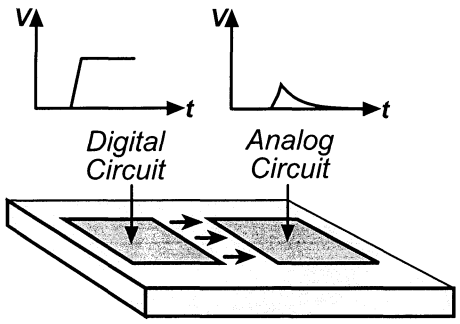


Figure 1. Noise injected into the analog circuit due to the switching of the digital circuit.

In this introduction we will briefly review noise injection mechanisms, noise reception mechanisms and a few common techniques for reducing noise in mixed-signal circuits. In section two a simple model of a mixed-signal circuit is presented. A noise reduction strategy is developed based on simulations of the model, which is described in section three. The strategy aims at reducing the high frequency content of the clock signal. In section four a test chip for evaluation of the proposed noise reduction strategy is described. Finally, we conclude the work in section five.

1.1 Noise injection mechanisms

A pn-junction between the transistor and the substrate is mainly capacitive. Hence, a signal transition causes noise to be injected from the transistors into the substrate via the capacitive coupling. The noise is then spread in the substrate due to the conductivity of the substrate. Another source of noise is the interconnect, that also is capacitively coupled to the substrate. Hence, a transition in a digital node will be capacitively coupled to the substrate by both the pn-junctions and the interconnect in that node.

The switching of digital circuits produces current peaks, which results in fluctuations on the power supply lines due to the inductive behavior of the wiring from chip to off-chip. This kind of noise is known as simultaneous switching noise (SSN). The SSN in the power supply lines in digital circuits is directly injected into the substrate through the biasing substrate contacts.

The strong electrical field between drain and source in submicron transistors makes the carriers (electrons in nMOS, holes in pMOS) to acquire enough energy to become what is called hot. In nMOS devices these hot electrons impact the drain of the transistor and produce ionization and dislodging holes that are spread in the substrate [1]. The result is a current flowing from the bulk of the transistor into the substrate, leading to voltage fluctuations in the substrate.

1.2 Noise reception mechanisms

Since an analog circuit similarly to a digital circuit contains transistors, interconnect, and biased substrate contacts, the noise in the substrate will be received by the analog circuits via the same type of capacitive coupling as

was described in section 1.1. Another noise reception mechanism is the threshold voltage dependence of a transistor on the bulk-source voltage V_{BS}. When substrate noise reaches the transistor, the voltage V_{BS} is affected, and hence the drain current and the drain-source voltage V_{DS}. This is observed as noise in the analog signal.

1.3 Noise reduction techniques

In lightly doped substrates it is possible to reduce the substrate noise in the analog circuit by increasing the distance to the digital circuit. This is due to the increased impedance. In heavily doped substrates, increasing the distance between the digital and the analog circuit is useless as soon as the distance is more than four times larger than the thickness of the epitaxial layer [2]. This is due to that the noise is spread via the highly conductive p^+ layer. A drawback of increasing the distance between the circuits is the increased area cost.

By inserting a guard ring between the circuits, the substrate noise can be suppressed if a lightly doped substrate is used. This is due to that a guard ring is a low impedance path to ground or V_{dd}. In the case of a heavily doped substrate the effect of guard rings is limited due to the high conductive p^+ layer.

SSN may be reduced by the use of an on-chip decoupling capacitor. The design of the decoupling capacitor is critical due to the fact that a poorly sized capacitance can result in an increased SSN.

To prevent simultaneous switching noise at the analog power supply lines, it is common to separate the digital power supply lines from the analog. Even if a common supply voltage source is used, it is preferable to use separate pins when taking the parasitic impedance of the interconnect from on-chip to off-chip into account.

In a mixed-signal circuit it is recommended to use lightly doped substrates instead of heavily doped, due to the fact that noise is easily spread over the entire chip via the highly conductive layer of p^+ [1].

2. A SIMPLE MODEL OF A MIXED-SIGNAL CIRCUIT

From a simple model of a mixed-signal circuit we will see that the magnitude of the received noise in the analog circuit is highly dependent on the frequency content of the substrate noise.

In Fig. 2, a mixed-signal circuit consisting of an analog circuit and a digital circuit occupying equally large area is shown. The biasing of the analog and the digital substrate is assumed to be uniform within their respective areas of the circuit. The substrate is also biased by a highly conductive backside contact. In our simple model, we only consider three nodes, namely the backside contact, the substrate in the analog circuit, and the substrate in the digital circuit. A standard 0.35 μm CMOS process with a lightly doped substrate is assumed. The thickness of the substrate is 725 μm and it is considered to be uniformly doped. The substrate can accurately be modeled as resistive for frequencies from DC up to a cut-off frequency, f_c, which is defined as

$$f_c = \frac{1}{2\pi r_{sub} C_{sub}} = \frac{1}{2\pi \rho_{sub} \varepsilon_{Si}} \quad (1)$$

where r_{sub} and C_{sub} are the resistance and the capacitance of the substrate, ρ_{sub} and ε_{Si} are the resistivity of the substrate and the permittivity of silicon, respectively [3]. The resistivity of the substrate is 0.1 Ωm which leads to that the behaviour of the substrate is mainly resistive for a frequency below 15 GHz.

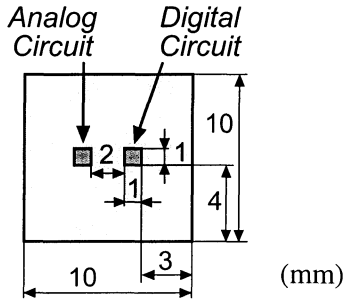


Figure 2. A mixed-signal circuit.

The finite element method is used to simulate the circuit. With the tool FEMLAB, a mesh of the substrate shown in Fig. 2 can be created. In our simulations, the generated mesh consists of $8 \cdot 10^4$ nodes that are reduced to the three nodes of interest. In Fig. 3, the model of the substrate and the parasitic inductances of the interconnect from chip to off-chip is shown. The three resistors forms a full model of the resistive coupling between these nodes. Using the finite element method, the values of the resistors can be estimated to R_1 = 13.8 kΩ and $R_2 = R_3$ = 27.7 Ω. The values of the inductances L_1, L_2 and L_3 are all assumed to be 5 nH.

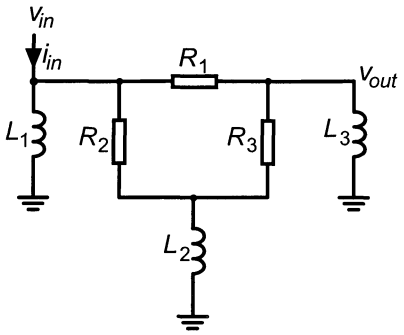


Figure 3. A simple model of a mixed-signal circuit.

With the assumption that a disturbance has been coupled into the substrate in the digital part as a voltage fluctua-

tion, it is of interest to determine the transfer function from the digital substrate to the analog substrate. The transfer function can be written as

$$H(s)=\frac{sL_3(R_2R_3+sL_2(R_1+R_2+R_3))}{R_1R_2R_3+s(L_3R_2(R_1+R_3)+L_2R_1(R_2+R_3))+s^2L_2L_3(R_1+R_2+R_3)} \quad (2)$$

The magnitude response of $H(s)$ is shown in Fig. 4. From the figure it is clear that a high frequency is less attenuated in the substrate than a low frequency. It can also be seen that the noise increases with 40 dB per decade in the frequency range of 2 MHz to 300 MHz. Hence, it is desirable to reduce the high frequency content of the substrate noise.

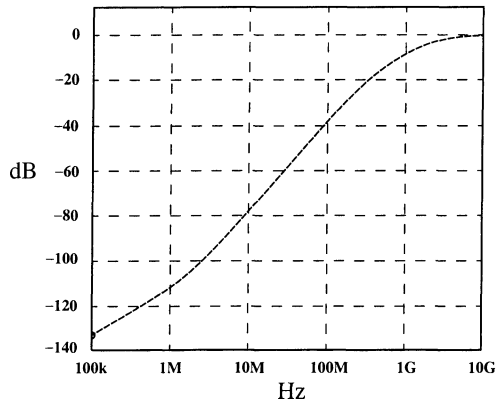


Figure 4. Magnitude response of the substrate.

3. A NEW STRATEGY FOR REDUCING CLOCK NOISE

Digital circuits commonly require a clock for global synchronization. When a large clock distribution network is used, the clock is strongly coupled to the substrate via the parasitic capacitance, and it will also require large drivers that are capacitively coupled to the substrate. When the clock buffer charge or discharge its large capacitive load, a large current floats through the parasitic inductance of the interconnect from chip to off-chip and SNN occurs. Hence, the distribution of the clock generates a considerable amount of the substrate noise that reduces the performance of the analog circuits.

A reduction of the high frequency content of the clock would reduce the noise in the substrate of the analog circuit. Therefore we propose a strategy for reducing the noise by reduction of the high frequency content in the clock. Reducing the rise and fall time of the clock, or ultimately using a sinusoidal waveform would achieve this. The proposed strategy requires a flip-flop that can operate with long rise and fall times of the clock. Such a flip-flop is presented in [6], which is used in this work. The efficiency of this strategy will be evaluated in a test chip that is currently being designed.

4. TEST CHIP

We have designed an experiment to evaluate the proposed strategy. A test chip is currently being designed that contains a digital circuit, a clock driver with programmable rise and fall time, and an analog circuit. The test chip will be manufactured in a 0.35 µm CMOS process with two poly layers and three metal layers. The substrate is a lightly doped p-substrate.

4.1 The digital circuit

An FIR filter was chosen for the digital part of the test chip, since this is a common component in signal processing applications. The filter is a low pass filter of order 13 with an external word length of 16 bits. The filter is realized in direct form, which can be described as

$$y(n) = \sum_{k=0}^{13} h(k)x(n-k) \quad (3)$$

where $h(k)$ is the impulse response of the filter and also the coefficients of the multipliers [4]. Each delay element in the structure, shown in Fig. 1, is implemented with 16 parallel flip-flops. The addition and multiplication are implemented by hard wired shifts and a carry-save adder-tree which is an efficient implementation approach [5]. A pipelined ripple-carry adder implements the vector-merging adder (VMA), which is the final adder in the carry-save adder-tree.

The carry-save tree is not pipelined. With pipelining the throughput could be increased further. In this case we are interested of keeping the amount of flip-flops at a reasonable level to have realistic results.

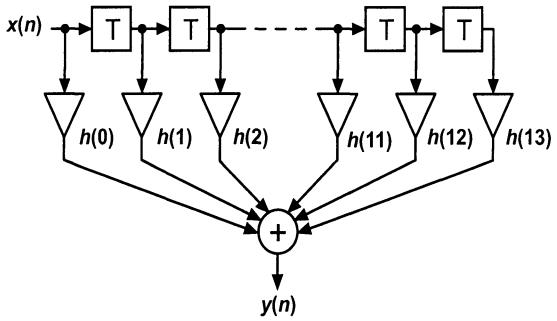


Figure 5. The FIR-filter in direct form.

The full adders in the carry-save adder tree are realized using static differential cascode voltage switch logic (DCVS). Note that the design effort have been put on designing a high-performance digital filter, with a balanced level of pipelining, and hence the number of flip-flops compared to the amount of logic should be characteristic for a high-performance filter. Thereby it should be possible to generalize the results of this experiment to other high performance DSP circuits.

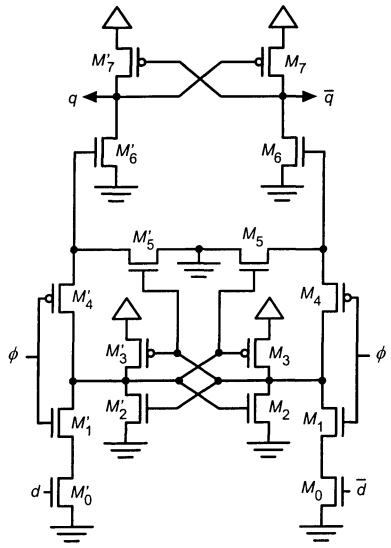


Figure 6. The robust differential flip-flop.

The differential flip-flop presented in [6] is chosen due to its robustness against long rise and fall times on the clock. The only difference between the flip-flop in [6] and the flip-flop in the test chip is that the multiplexer at the input is substituted by the two pull-down transistors M_0 and M'_0 in Fig. 2.

4.2 The adjustable clock buffer

The last stage in the adjustable clock buffer consists of 256 unit buffers. The unit buffer (Fig. 7) is realized using a tri-state like buffer where the pMOS and nMOS devices can be activated independently. Two binary vectors with a word length of 8 bits are used to control the number of active nMOS and pMOS devices. This makes it possible to individually adjust the rise and fall time of the clock signal.

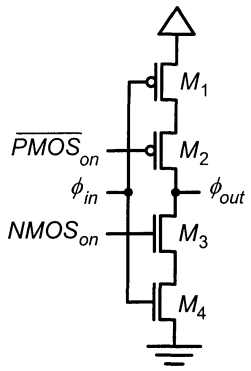


Figure 7. The unit buffer.

4.3 The analog circuit

The analog filter is a fifth-order low-pass leapfrog filter without finite zeros implemented in the active-RC technique. The filter is derived from a doubly resistively terminated LC ladder network to achieve low sensitivity for variation in the component values [7]. Further, the variations in the pass band of the LC network is proportional to the product of the frequency, ω, and the group delay, τ_g, [7]. This sensitivity of the LC network is minimized by minimizing the product $\omega\tau_g$ given that the filter specification is met. The minimization of the sensitivity of the LC ladder network will also decrease the sensitivity of the pass band variations in the leapfrog implementation due to finite Q-values of the integrators [8].

4.3.1 Implementation

The integrators in the active RC filter consist of two operational amplifiers (op-amps). Using two op-amps in an integrator will increase its Q-value and decrease the phase and gain errors of the integrator [9]. In the filter the inverting and non-inverting integrators are implemented with the proposed structures in [9].

The op-amps are implemented as ordinary two-stage operational amplifiers with a common drain amplifier as an output buffer. This will decrease the output resistance at the cost of a decreased output swing. Using pMOS transistors as input devices will increase the unity-gain frequency, slew rate, and decrease the flicker noise of the amplifier [10]. Further, using pMOS input transistors will reduce the effect of substrate noise injected to the input transistors since all pMOS transistors are placed in an n-well.

The cut-off frequency of the filter is programmable and can be adjusted off-chip. The possible range for the cut-off frequency is 1.5 to 3 MHz. This frequency is adjusted by placing additional capacitors in parallel with the integrator capacitor.

4.3.2 Power supply rejection ratio

The power-supply rejection ratio (PSRR) expresses how much noise that are suppressed through the power supply lines to the circuit's output. We must guarantee that the PSRR is high enough at crucial frequency ranges. Typically, we see strong noise components, e.g., glitches and voltage drops at multiples of the clock frequency spreading through power lines and substrate. Hence, we must design the circuit to attenuate these components to meet the specification. The transfer function from the input and from the negative power supply line to the output of the analog filter are shown in Fig. 8, indicated with solid and dashed lines, respectively. These transfer functions are achieved when the gate and bulk are connected to the same potential and thereby the bulk effect of the nMOS transistors are ignored. The clock frequency of the digital part will be around 40 MHz. For this frequency and multi-

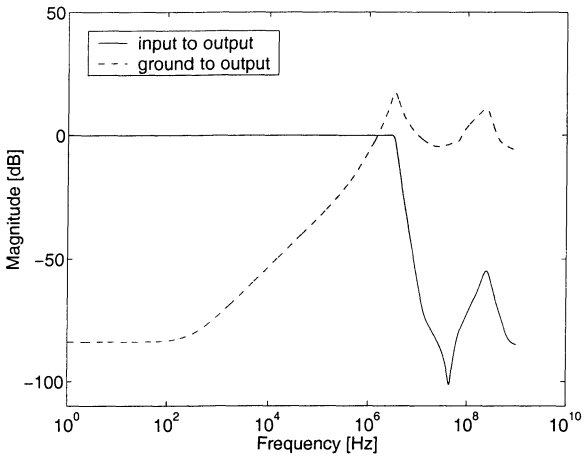


Figure 8. The transfer function from the input to the output and from negative power supply to the output, indicated with solid and dashed lines, respectively.

ples of it, the noise injected into the negative power supply will be amplified to the output.

4.3.3 Noise reduction techniques

To reduce the effect of the substrate noise, the analog filter is shielded from the digital part utilizing p^+ guard rings connected to the analog negative power supply. This choice of guard ring will decrease the noise introduced in the analog part of the circuit [11], [12]. Further, substrate contacts are connected to each transistor to decrease the variations in the threshold voltage.

Separate power supply lines are used for the analog and digital parts to avoid direct injection of SSN from the digital circuit into the analog circuit.

4.4 Test setup

The test chip will be evaluated in a test setup where a DC voltage will be used as the input signal to the analog filter. A series of measurements is planned to be performed where the rise and fall times of the clock connected to the digital FIR filter are varied. The noise and the tones at multiples of the clock frequency are going to be measured at the input and output of the analog filter. Also different clock frequencies for the filter will be tested.

5. CONCLUSION

In this paper we have briefly described the nature of noise in mixed-signal circuits. Further, a simple model of a mixed-signal circuit has been used to estimate a resistive model of the substrate. The simulation of the model gives that a high frequency signal is less attenuated in the substrate compared to a low frequency signal. With the knowledge that the clock distribution network generates a considerably amount of substrate noise that reduces the performance of analog circuits, and the results obtained from the mixed-signal model, a strategy for reducing the noise caused by a digital clock has been proposed. The strategy uses a clock with long rise and fall times, and a special flip-flop that can operate with the special clock. Using longer rise and fall times decreases the magnitude of high frequency components in the clock signal. A test chip, containing an FIR filter and an analog filter, is currently being designed in a 0.35 µm CMOS process in order to evaluate the performance of the long rise and fall times. Results from measurements on the fabricated circuits will hopefully be available in the near future.

6. REFERENCES

[1] X. Aragonès, J. L. González & A. Rubio, *"Analysis & solutions for switching noise coupling in mixed-signal ICs,"* Kluwer Academic Publishers, Netherlands, 1999.

[2] D. K. Su, M. J. Loinaz, S. Masui, B. A. Wooley. "Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrated Circuits," *IEEE J. Solid-State Circuits,* vol. 28, pp. 420-430, 1993.

[3] R. Singh, "A Review of Substrate Coupling Issues and Modelling Strategies," *Proc. IEEE 1999 Custom Integrated Circuits,* pp. 491-498, 1999.

[4] L. Wanhammar, H. Johansson, *"Digital filters,"* Dept. of Electrical Eng., Linköping University, 1999.

[5] T. G. Noll, "Carry-save architectures for high speed digital signal processing," *J. VLSI Signal Processing,* vol. 3, pp. 121-140, 1991.

[6] M. Vesterbacka, "A robust differential scan flip-flop," *Proc. 1999 IEEE Int. Symp. on Circuit and Systems,* vol. 1, pp. 334-337, 1999.

[7] G. C. Temes and H. J. Orchard, "First-order sensitivity and worst case analysis of doubly terminated reactance two-ports," *IEEE Trans. Circuit Theory,* vol. CT-20, no. 5, Sept. 1973.

[8] P. O. Brackett and A. S. Sedra, "Direct SFG Simulation of LC ladder networks with applications to active filter design," *IEEE Trans. Circuit Syst.*, vol. CAS-23, no. 2, Feb. 1976.

[9] R. L. Geiger and G. R. Bailey, "Integrator design for high-frequency active filter applications," *IEEE Trans. on Circ. Syst.*, vol. CAS-29, no. 9, Sept. 1982.

[10] D. A. Johns and K. Martin, *"Analog integrated circuit design,"* John Wiley & Sons, USA, 1997.

[11] N. K. Verghese and D. J. Allstot, "Rapid simulation of substrate coupling effects in mixed-mode ICs," *IEEE Custom Integrated Circuit Conf.*, pp. 18.3.1-18.3.4, 1993, San Diego, USA.

[12] H. Chen, et. al., "The study of substrate noise and noise-rejection-efficiency of guard-ring in monolitic integrated circuits," *IEEE Int. Symp. Electromagnetic Compatibility,* vol. 1, pp. 123-128, 2000, Washington, USA.

The Wireless IP Project

Mikael Sternad[1]

[1]Signals and Systems, Uppsala University, PO Box 528,SE-751 20 Uppsala, Sweden.

Abstract

The optimization of resources in wireless packet data systems is challenging when users are mobile with time-varying link quality. Within the SSF PCC program, we have since year 2000 formed the Wireless IP project, which studies such issues. We perform research towards a flexible packet data system with wide area coverage for mobile users. This system should have performance equivalent to a 100Mbit/s Ethernet connection, with support for Quality of Service and fairness between wireless users.

This will require an optimization that involves many levels in the OSI stack. In the project, we develop and integrate a set of key technologies, consisting of prediction of radio channels, fast link adaptation, scheduling of several packet data streams over time-varying channels, fast resource allocation using antenna arrays and the use of information from lower layers in higher layers, in particular the transport layer. With these methods, the transmission of packet data to/from mobile users can be radically improved, without expensive use of additional spectral bandwidth.

1 Introduction

Wireless systems are evolving rapidly and most traffic to mobile terminals will be IP-based end-to-end in a not too distant future. This gives rise to many performance tradeoffs and design issues, and we need to reexamine our approach to the design of wireless systems.

Wideband wireless links are used today to transport packet data to stationary users. A main challenge is to improve the economy, spectral efficiency and bitrates provided in wide-area coverage for mobile (vehicular) users. Potential markets exists, as indicated e.g. by the increased interest in "telematics" and "infotainment" in the automotive industry, but severe problems remain, as illustrated by the economic and technical difficulties accompanying the birth of 3G systems all over the world. The spectral efficiency has to improve to decrease the cost per bit in any proposed novel system beyond 3G.

Within PCC, the Wireless IP project [1][1] performs research with the aim of proposing a flexible and affordable general packet data system, with wide area coverage for mobile users. It should have *performance equivalent to a 100Mbit/s Ethernet connection, with support for Quality of Service (QoS) and fairness between wireless users.*

This goes far beyond what is attainable today with e.g wireless LAN systems such as Hiperlan II, which are appropriate only for short range and slow users, or GPRS/WCDMA 3G systems, that offer lower capacity at rather high cost. The time-varying radio propagation conditions encountered by vehicular users represent a major challenge. However, we believe our goal to be attainable at reasonable cost. Long-standing basic design principles for data communication systems and wireless systems have conspired to create inefficiencies in mixed fixed and wireless mobile environments. By reducing these inefficiencies, packet transmission to mobile users may be radically improved without the use of additional and expensive spectral bandwidth.

Systems for data communication are conceptually designed in multiple layers, from applications down to the physical layer. Design and standardization is simplified when layers are handled separately, but this may lead to inefficient resource utilization.

In lower layers, the design of wireless systems traditionally has the goal of hiding the effect of time-varying channel conditions by averaging: variations in channel quality are counteracted by coding and interleaving. Spread spectrum techniques are used partly to average out and reduce variations in the interference environment. The resource allocation is designed to guarantee a modest level of performance for all users, rather than maximizing the spectral efficiency. This leads to conservative designs: the received power and channel quality in general becomes higher than required by the services, which wastes resources.

The Wireless IP project develops methods that depart radically from the principles outlined above. We investigate intelligent resource utilization for wireless packet data systems, with the following characteristics:

1. Wherever possible, resources should be adjusted to the instantaneous channel quality, to prevent waste. The use of channel coding should be minimized. We systematically explore methods that use channel

[1]The Wireless IP project is a collaboration between researchers from Uppsala University, Chalmers University of Technology, Karlstad University and Politechnico di Bari, Italy, with Ericsson Research, Ericsson Mobile Data Design and Nera, Norway as industrial partners.

Participants funded by SSF PCC are Arne Svensson, Tony Ottosson, Sorour Falahati and Bartosz Mielczarek from Chalmers and Mikael Sternad (project leader), Anders Ahlén, Nilo Casimiro Ericsson and Anna Ewerlid from Uppsala. External collaborators are Anna Brunström, KaU, Saverio Mascolo, Politechnico di Bari and Torbjörn Ekman and Mathias Johansson, both at Uppsala University.

quality feedback and optimize resources, instead of using averaging.

2. We develop and use methods that *take advantage of the varying channel properties*, to obtain "trunking gains", or multiuser selection diversity, for multiple users who share the radio bandwidth.

3. Multiple antennas, at receivers as well as transmitters, reduce interference, increase the range and limit the required transmitted power. Their use and design is integrated in the proposed schemes.

4. Finally, we strive to make congestion control more efficient by using information from lower layers.

A careful application and combination of these methods lead to accumulated gains in efficiency: improved spectral efficiency as well as increased efficiency at higher layers.

Sections 2-4 below provide an overview of the research conducted within the project and of our present results. The most recent results are presented in more detail in the papers [2, 3, 4, 5, 6] at this conference. See also our webpage [1] and the Licentiate Theses [7], [8] and [9].

2 Prediction of Fading Channels

A main difficulty in data transmission over wireless links to mobile users is the fading, which introduces variations in the channel quality on all time-scales. At the lowest layer, we therefore develop predictors of the time-varying impulse responses of fading channels to mobile users, and of the signal-to-interference and noise ratio (SINR). We have so far focused on prediction of short-term fading, which is crucial for link-level adaptive modulation and scheduling. Predictions of 2-5 ms ahead are required, and will at vehicular speeds correspond to a considerable fraction of the radio wavelength.

Channel prediction over such horizons is difficult, and has only recently begun to be investigated. We have developed predictors that use channel tap estimates as regressors to predict the power of all significant channel taps. The predictors have been validated on an extensive set of broadband channel measurements [10, 11], see Figure 1 for an illustration. A recently developed unbiased quadratic predictor [3] provides the so far best performance, see Figure 2. We have also obtained a good theoretical understanding of the limits of attainable performance in different propagation environments.

3 The Link- and MAC Levels

Channel prediction can be used to adapt the data rate to the channel properties, via *adaptive modulation and coding* (link adaptation). With our recent results [4], the channel prediction error can be taken into account when optimizing the adaptive modulation.

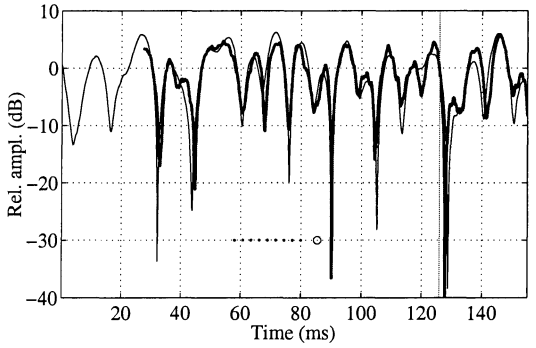


Figure 1: Typical prediction result for the fast fading of one channel tap in a channel of bandwidth 5MHz at 1880MHz. The prediction horizon is 5.5ms. The prediction (solid) is overlaid on a least squares estimate of a true tap (thin line). Measured channel sounding data is used. Note that it is possible to predict beyond a fading dip, so that the transmission can be planned accordingly. For details, see [3], [10],[11].

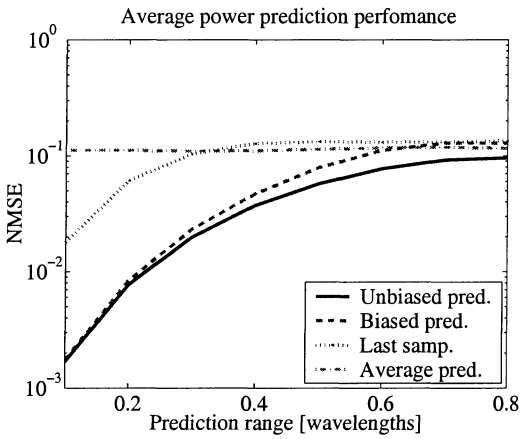


Figure 2: Power prediction on 5 MHz channels at 1880MHz. The normalized MSE is evaluated as a function of the prediction range, averaged over 37 measurement locations. The average predictor extrapolates an estimated average power level, while the last sample predictor extrapolates the last measurement [3].

Prediction is also of crucial importance for improving the MAC (Media Access Control) layer. It can there enable the *scheduling* of several data streams to/from mobile users whose channels fade independently: instead of counteracting the fast fading with coding and spreading, one should instead, if possible, delay transmission of a particular user's data until the channel is good. This principle, which has been denoted multiuser diversity, represents a powerful tool for improving the system capacity.

We have studied and developed schemes for using multiuser diversity by combining adaptive modulation and predictive scheduling [7, 16, 17, 20]. An important aspect of our work is that packets with different payloads should reliably be provided with different Quality of Service (delay and error probability), despite channel variations.

3.1 Link Adaptation and Scheduling

We have so far primarily considered downlink (base to mobile) transmission from a single base station over a shared bandwidth to several mobile users. Per-flow buffering of data streams is assumed at the interface between the wired system and the wireless links, see Figure 2. Our proposed solution involves the use of channel power prediction at the mobiles. All active mobiles transmit quantized channel quality prediction reports to the base station.

Based on this information, a scheduler, located at the interface between the fixed network and the radio links, should distribute a set of shared radio resources (time-slots and perhaps also frequencies and base stations, see Section 3.2 below). The task of the scheduler is then to empty the buffers, taking into account the predicted channel capacities for each user several time-slots ahead. The scheduler should also take the buffer contents and the quality of service requirements (packet error probabilities and delays) for each data stream into account.

When the scheduling for a time interval is complete, bits from appropriate queues are formed into link level packets, modulated with a modulation alphabet appropriate for the predicted channel quality and are then transmitted in their appropriate time-slots.

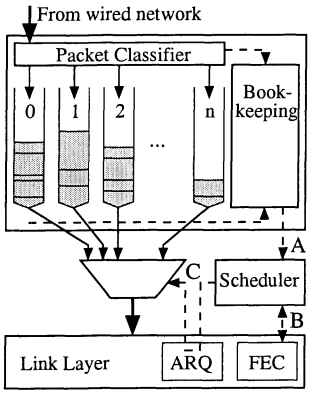


Figure 3: The downlink buffer and the scheduler. Packets arrive at the top and are inserted into their respective queues, restoring order among occasionally arriving out-of-order packets. The buffer regularly submits a status report (A) to the scheduler, containing info about the priorities, the size of the queues, and the required link service, some of which is also passed to the link layer (B). The scheduling decision (C) is updated by the link layer ARQ, and is then used to drain the queues, with a bit-level resolution of the outflows. For details, see [7, 17, 19].

3.2 Scheduling in Multiple Dimensions: An Adaptive OFDM System Proposal

We have until now mainly studied adaptive modulation and scheduling in a TDMA time slot reservation framework. The whole slot is then assigned to one user, as will also mostly be the case in HSDPA for release 5 of WCDMA [12, 13].

Depending on the scattering environment, the fading at different frequencies within a 5 MHz UMTS bandwidth may be almost uncorrelated, as exemplified by Figure 4. This provides another dimension over which to perform optimization and scheduling, and thus another tool for improving the total system capacity. To explore these possibilities we consider a slotted OFDM system in which the scheduler reserves time-frequency slots for multiple users.

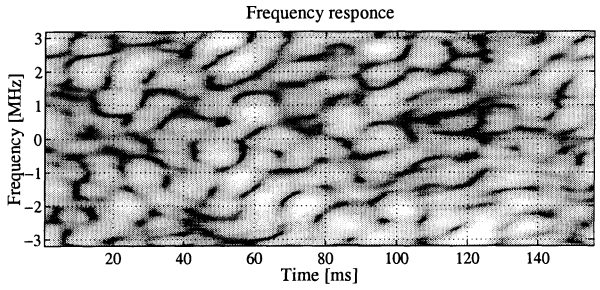


Figure 4: Time-frequency representation of an estimated channel obtained from real measurement data on a 6.4MHz channel. White color denotes high power whereas dark color denotes low power. The dynamic range and the speed of the mobile is approximately 40 dB and 50 km/h, respectively. The coherence bandwidth is 0.6 MHz.

The efficiency is in this system increased further by including the slot allocations within neighboring transmission beams in the scheduling. This limits interference and we may attain a frequency reuse close to a factor of 1.

We are at present performing research into the possible capacity attainable by such a hypothetical 4G system [2, 14, 15]. A preliminary proposal for the downlink, and an estimate of its spectral efficiency is presented at this conference in [2]. Most time-frequency slots can be utilized by mobile terminals with good channels even with rather few active users within an antenna beam/sector. High order modulation is then used in most slots.

For example, Figure 5 considers downlink transmission with ARQ of best effort traffic to K mobile users per sector, to mobiles that each have L antennas with 16dB SINR per antenna and that use maximum ratio combining (MRC). An efficient fast link adaptation scheme is used, which provides a spectral efficiency of 1.8 bits/s/Hz when one user with one antenna uses the channel exclusively. The estimated spectral efficiency improves to 3.6 bits/s per Hz per sector when five active users, each with two receiver antennas, share the channel.

3.3 Scheduling Methodologies

We have proposed to base the scheduling on a criterion that represents the total user satisfaction [18]. We are developing algorithms that optimize this criterion and make efficient use of the available bandwidth, at reasonable computational complexity [7].

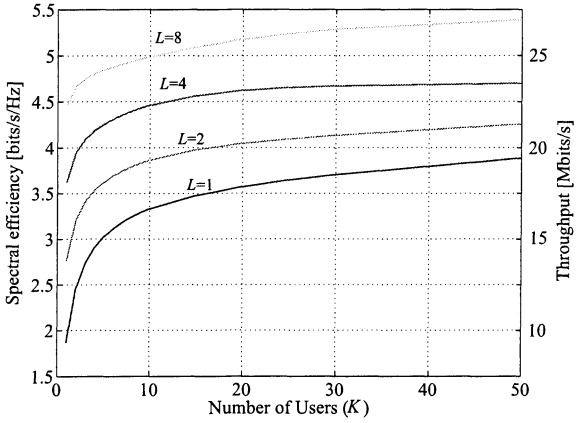


Figure 5: Spectral efficiency using AMS and simple ARQ with Lth order MRC diversity in the mobile and Kth order of selection diversity between the users. The symbol energy to noise ratio per receiver antenna is 16 dB. Right-hand scale: The capacity within one sector or beam in a 5MHz downlink [2].

Furthermore, we combine scheduling with Hybrid type-II ARQ [8], a link level retransmission scheme that uses incremental redundancy. Link-level packets are first transmitted uncoded, with additional coding information transmitted later, if required due to an unsuccessful transmission. This combination attains high throughput, low delay and it makes the performance relatively insensitive to the unavoidable errors in channel predictions [16, 17, 19]. We use rate compatible convolutional codes, that enable soft decoding with acceptable computational complexity.

We also study Turbo codes, of interest in particular in environments with high interference levels [9, 23, 24, 25]. The critical issues of rate compatibility, channel estimation and header encapsulation are treated in [5].

A novel scheduling methodology, that offers high performance and flexibility at the price of higher computational complexity as compared to the methods mentioned above, has recently been derived [21, 22]. In this method, Bayesian probability theory and the maximum entropy principle are used to optimize a throughput criterion in a way that handles fairness constraints as well as different amounts of knowledge about buffer inflow statistics, channel capacities and prediction errors.

4 Transport Protocols over Fading Wireless Links

On a higher level, we study the behavior of transport protocols, in particular TCP. Fading wireless links are often characterized by a sporadic high bit error rate and a rather high and time-varying bandwidth-delay product. TCP performance is then often reduced.

An ingenious end-to-end scheme that requires modification of the bandwidth estimation algorithms only at the sender side of a TCP connection is TCP Westwood [27], recently proposed by Saverio Mascolo, who is engaged as a researcher in our project. Preliminary investigations have shown TCP Westwood to provide a large performance improvement over wireless links. We are evaluating TCP Westwood in a simulation environment with link level retransmissions, realistic fading models and erroneous feedback of ACK's due to wireless return channels. Results are presented in the poster [6].

4.1 Interaction of Link Level Retransmissions with Transport Level Protocols

Link level retransmissions are used in GPRS, and will probably have to be used in all future systems aimed at mobile users with fading channels. They improve the channel quality but may adversely affect higher level transport protocols if many retransmission attempts are allowed.

With mobile users traveling through complicated propagation environments, fading will generate channel quality variations on all time-scales. An important aspect here is how transport level congestion control reacts in such environments when link level retransmissions such as our Hybrid type II ARQ are active.

Split connections, with a separate transport-level protocol over the wireless link, offer potential advantages. They enable the introduction of performance-enhancing algorithms, that use information from lower layers, at a gateway to the wireless part of the system. There are also potential drawbacks with split connections. They break the end-to-end semantics. They may offer only limited improvements when link-level retransmission is already in place. They will also require fast migration of information between nodes at hand-over. A further aspect that may limit the use of split connection schemes would be an increased use of IP-level security and encryption [28].

4.2 A Simulation Environment for Studying Multilevel Optimization and Interlayer Interaction

It is not at all simple to obtain reliable and relevant results when investigating multilayered communication systems working on different time-scales. We are therefore making a major effort to create an appropriate simulation environment [6, 29, 30, 31]. Using real TCP and UDP stacks, it generates packets that propagate in real time over simulated wireless links, so that important aspects and time-scales of lower layers, and inter-layer interactions can be studied. Radio channel properties are obtained from ray-tracing simulations, which are validated against channel sounding measurements provided by industry. This simulator will be a valuable tool in our continued research.

References

[1] The Wireless IP Project within the Swedish SSF PCC (Personal Computing and Communication) program: www.signal.uu.se/Research/PCCwirelessIP.html

[2] T. Ottosson, M. Sternad, A. Ahlén, A. Svensson and A. Brunström, "Towards a 4G IP-based Wireless Systems Proposal," *Radiovetenskap och Kommunikation, RVK 02*, Stockholm, June 10-13, 2002.

[3] T. Ekman, M. Sternad and A. Ahlén, "Unbiased power prediction on broadband channels," *RVK 02*, Stockholm.

[4] S. Falahati, A. Svensson, T. Ekman and M. Sternad, "Effect of prediction errors on adaptive modulation systems for wireless channels," *RVK 02*, Stockholm.

[5] B. Mielczarek and A. Svensson, "Packet transmission with turbo codes on fading channels," *RVK 02*, Stockholm.

[6] A. Ewerlid and N.C. Ericsson "Real time simulation of TCP/IP over wireless fading channels," *RVK 02*, Stockholm.

[7] N. C. Ericsson, *On Scheduling and Adaptive Modulation in Wireless Communications*. Licentiate Thesis, Signals and Systems, Uppsala University, Sweden, June 2001.

[8] S. Falahati, *Convolutional Codes for Wireless Packet Data Systems*. Licentiate Thesis, Chalmers University of Technology, Feb. 2000.

[9] B. Mielczarek, *Synchronization in Turbo Coded Systems*. Licentiate Thesis, Chalmers University of Technology, April 2000.

[10] T. Ekman, *Prediction of Mobile Radio Channels*. Licentiate Thesis, Signals and Systems, Uppsala University, Sweden, Dec. 2000.

[11] M. Sternad, T. Ekman and A. Ahlén, "Power prediction on broadband channels," *IEEE Vehicular Technology Conference-Spring*, Rhodes, Greece, May 6-9 2001.

[12] 3rd Generation Partnership Project (3GPP), Technical Specification Group Radio Access Network: "UTRA High Speed Downlink Packet Access (Release 4)," *3GPP TR 25.950 V2.0.0.*

[13] S. Parkvall, E. Dahlman, P. Frenger, P. Beming, M. Persson, "The evolution of WCDMA towards higher speed downlink packet data access," *Nordic Radio Symposium 2001 (NRS 01)*, Nynäshamn, Sweden, April 3-5, 2001.

[14] T. Ottosson, A. Ahlén, A. Brunström, M. Sternad and A. Svensson "Toward 4G IP-based Wireless Systems," *Future Telecommunication Conference* Beijing, China, November 2001. Also presented at the *4th Wireless World Research Forum Meeting*, Paris, Dec. 2001.

[15] T. Ottosson, A. Ahlén, A. Brunström, M. Sternad and A. Svensson, "Toward 4G IP-based wireless systems: A proposal for the uplink." *5th Wireless World Research Forum Meeting*, Temple, AR, March 2002.

[16] N.C. Ericsson, S. Falahati, A. Ahlén and A. Svensson, "Scheduling and adaptive transmission for the downlink in 4G systems," *Future Telecommunication Conference* Beijing, China, November 2001. Also presented at the *4th Wireless World Research Forum Meeting*, Paris, Dec. 2001.

[17] N. C. Ericsson, S. Falahati, A. Ahlén and A. Svensson, "Hybrid type-II ARQ/AMS supported by channel predictive scheduling in a multi-user scenario," *IEEE VTC2000-Fall*, Boston, MA, USA, Sept. 24-28, 2000, pp. 1804-1811.

[18] N.C. Ericsson, "Scheduling as an optimization problem," *Nordic Radio Symposium 2001 (NRS 01)*, Nynäshamn, Sweden, April 3-5 2001.

[19] S. Falahati, N.C. Ericsson, A. Ahlén and A. Svensson, "Hybrid type-II ARQ/AMS and scheduling using channel prediction on fading channels," *Nordic Radio Symposium 2001 (NRS 01)*, Nynäshamn, Sweden, April 3-5 2001.

[20] N. C. Ericsson, "Adaptive modulation and scheduling over fading channels," *IEEE GlobeCom99*, Rio de Janeiro, December 5-9, 1999. pp. 2668-2672.

[21] M. Johansson, "Probabilistic resource allocation and scheduling for multiuser communication systems". *IEEE GlobeCom01*, San Antonio, November 2001.

[22] M. Johansson, "Multiuser Diversity and Maximum Entropy Scheduling," *Radiovetenskap och Kommunikation RVK 02*, Stockholm, June 10-13, 2002.

[23] B. Mielczarek and A. Svensson, "Fast header decoding using turbo codes on Rayleigh fading channels," *International Conference on Information, Communication, and Signal Processing*, Singapore, Oct 2001.

[24] B. Mielczarek and A. Svensson, "Joint adaptive rate Turbo decoding and synchronization on Rayleigh fading channels," *IEEE Vehicular Technology Conference 2001 (VTC2001-Spring)*, Rhodes, Greece, May 2001.

[25] B. Mielczarek and A. Svensson "Improved Turbo decoder structures for burst error channels", *IEEE Vehicular Technology Conference 2001 (VTC2001-Spring)*, Rhodes, Greece, May 2001.

[26] A. Brunström and T. Ottosson, " The introduction of soft information in IP-based wireless networks," *Nordic Radio Symposium 2001 (NRS 01)*, Nynäshamn, April 3-5 2001.

[27] S. Mascolo, C. Casetti, M. Gerla, M.Y. Sanadini and R. Wang, "TCP Westwood: Bandwidth estimation for enhanced transport over wireless links. *ACM MobiCom01*, Rome, July 16-21 2001.

[28] A. Ewerlid, "Reliable communication over wireless links," *Nordic Radio Symposium 2001 (NRS 01)*, Nynäshamn, Sweden, April 3-5 2001.

[29] A. Ewerlid and N.C. Ericsson "A tool for real time simulations of TCP/IP over wireless fading channels," *SMSI Advanced Simulation Technology Conference*, San Diego, April 14-18. 2002.

[30] M. Berlin, *Channel Modeling and Simulation for the Mobile Internet*. Master Thesis, Signals and Systems, Uppsala University, March 2002.

[31] M. Wännström, *Simulation Models for Hybrid ARQ Adaptive Coding Schemes*. Master Thesis, Signals and Systems, Uppsala University, April 2002.

Achieving Non-intrusive Notifications in a Ubiquitous Service Environment

Patrik Werle[1] and Carl Gustaf Jansson[2]

[1]The fuse Research Group, DSV/Wireless@KTH, KTH, Forum 100, SE-164 40 Kista. E-mail: werle@dsv.su.se
[2]The fuse Research Group, DSV/Wireless@KTH, KTH, Forum 100, SE-164 40 Kista. E-mail: calle@dsv.su.se

Abstract

The use of computing artefacts for collaboration is problematic since the artefacts often are designed mainly for individual purposes. We identify some sources of intrusion on the collaborative task imposed by the trends towards more and more computer dense environments and we put forward some general design principles to diminish them. Based on these principles we have designed a non-intrusive system for the delivery of notifications. A prototype of the system is presented where we explore different design aspects of the output terminals, such as personal vs. public and in focus vs. peripheral.

1 Introduction

In the **fuse** research group at DSV, we design architectures and infrastructures for ubiquitous computing environments [1] and develop prototypes of interactive services that support collaboration. We also study the use of wearable computers and context-sensitive equipment and we explore ad-hoc environments and wireless networks.

A ubiquitous service environment, or USE, [2] is a physical and logical space where users can collaborate supported by multiple services providing adequate computing and communication facilities. The physical space is dominated by the material manifestation of computing resources, such as computers, screens, keyboards, pointing devices, PDAs, sensors and so on.

The logical space is to a large extent shaped by the amount of communication available between the physical devices. None or very little communication tends to partition the logical space into cells which more or less equate the confines of a particular device. In this situation, each attempt to transfer data of effect an operation becomes hampered by the technicalities involved [2].

One example of USE services that we have developed in the **fuse** group is Active Documents. The idea of Active Documents is to take off from the agent-programming paradigm [3], and turn documents into autonomous mobile agents and by that give them some useful qualities. A document should, for example, be aware of its content and the intention with it, and be aware of the context it is operating in, e.g. its receivers (who, why, preferences about formats, physical surroundings, etc.). The documents are active in the sense that they are autonomous (act independently), reactive (react on changes in the environment), and proactive (have their own goals and plans).

In [2] we presented a ubiquitous service environment for teamwork, supported by Active Documents. The USE consists of a physically dressed conference room, a set of services, and a software architecture based on Java and Jini [4]. We also presented a prototype and preliminary results from observations in a meeting scenario. The goal was to provide the meeting participants with support for their collaborative task by, for example, offer them services that give them access to each other's computers and to the resources that reside within the room. One purpose of the Active Documents in this prototype was to utilize context information for detecting a meeting and then "enter" the meeting room and present itself. The main idea was that the documents actively should participate in the work and thereby support the work process.

In this paper, we present some design principles for non-intrusiveness environments and a prototype for handling the presentation of incoming external notification messages, also in a meeting scenario.

2 Designing Non-intrusive Environments for Local Collaboration

The way people use computing artefacts in local collaborative settings is problematic. The main reason for this is that the tools that are being used to a large extent are designed as personal tools. Constant intrusions on the collaborative task are imposed from lacking support for collaboration, external non-related events, parallel non-

related work and the breaking of different kinds of social norms. We define *intrusion* as an occurrence of a process or task that is not intimately related to the current task of a group and that interferes with the realization of that task. We have identified four major sources of intrusion for this kind of settings:

- Insufficient collaboration support
- External events and notifications
- Parallel individual activities
- Breaking social rules

2.1 Design Principles

We have identified some general design principles that can prevent or diminish some of the intrusions in ubiquitous computing environments.

The first principle is to create good models and explicit representations of the tasks and the work situations. The core of mechanisms to diminish intrusiveness phenomena is good co-ordination strategies and these in turn must rely on adequate models of all the tasks being performed in each work environment. One particular problem with the handling of parallel tasks is the cognitive load when switching between tasks. If the services supporting different tasks can minimise the effort in switching between tasks, the effect of intrusions will decrease.

Another principle is to try to model what should be characterized as intrusive and non-intrusive in the current circumstances. Such model include the social rules valid for a certain situation, the normal behavioural and perception patterns within a group, and rough operational initial mechanisms such as the notion of *message impact* used in the prototype described below.

Yet another general principle is to offer the users a better support for the local collaboration, for example by facilitate shared focus, simultaneous and transparent interaction, ad-hoc device integration, and personalisation of the workspace.

In the prototype presented in this paper, we also explore some specific design aspects that we have identified. These aspects (or dimensions) are:

- Personal vs. public: Viewing all public and private devices available in a work environment as potentially shared and useful for all parallel tasks by all participants
- In focus vs. peripheral: Utilising the periphery of the work environment as much as the current focus point, making devices less focus demanding and allowing for easy transitions between periphery and focus
- Degree of explicitness: Utilising mixed implicit and explicit representations of information
- Choice of modality: Providing a rich set of modalities both for input and output and utilising mixed modalities for efficient interactions
- Information filtering: Providing a set of abstraction, sorting and filtering techniques
- Co-ordination and frequency modulation: Providing rich sensory observations as a basis for context dependency mechanisms
- Degree of context-awareness: Unifying the access to personal and collaborative support services, allowing for easy ad-hoc configuration of all available hardware and software components and implementing efficient co-ordination strategies to allow for temporal, physical personalised synchronisation. This also includes negotiation schemes to create co-ordinations optimal for the whole group.

In an attempt to deal with one of the issues causing intrusions: the external event, a prototype environment for non-intrusive notifications was implemented. This environment consists of both hardware and software components and makes it possible to deliver short messages in different modalities, on public or private displays, and with different renderings.

3 Non-intrusive Notification Prototype

We have chosen to explore the design issues from above by targeting a certain type of intrusions. These intrusions are disturbances caused by the process of delivering notifications or short messages to members of a group currently involved in a meeting. With notifications we mean short messages of several different types, such as calendar event notifications, short or truncated email, SMS messages, etc.

A typical such intrusion might be the delivery of an email to one of the persons in the meeting. This event interferes with the meeting in several ways. One such interference is the loss of attention from the receiver of the message. His or hers attention is required firstly to manipulate some artefact to open an email program, and then to read the actual message. Another interference is the effect on the other participants in the meeting. They might first be disturbed by some notification sound or visual cue, and later by the fact that the receiver's attention is not focussed towards the meeting. The complete intrusion comprises the effects on both the receiver of the message as on the other people in the room.

A prototype environment was developed where hardware and software was carefully designed in order to try and minimize the intrusions from incoming notifications. In the scenario for the prototype a small number of persons are having a work meeting; working

on a document on a shared screen. While they work they are constantly given notifications of different kinds. These notifications appear both on personal devices and on public devices in the room. The participants can collectively affect with what impact the messages are being presented in the room by turning a knob residing in the middle of the table.

The physical setting consists of a meeting room with a table, some chairs and a wall display. Embedded in the table are small screens, one for each place around the table. On the middle of the table there is a "knob". To personalize the embedded screens, a user pushes a personal identifier device into a receptor connected to the screens. A projector is also used to project information against one of the walls. A device that resembles an analogue clock (the iClock) is hanging on another wall. A pair of speakers is also present in the room. A number of laptops are used to run different software. All computers are interconnected via wireless lan.

The different services in the prototype are:

- **Notification message originators** – where messages for human users are created
- **Impact controller service** – a "knob" used by the participants to influence the impact with which the notification messages should be displayed
- **Message terminal services** – where notification messages are displayed
- **Active Document service** – a context sensitive message routing service.

The services will be described more in detail below.

3.1 Notification message originators

The message originators provide the input of messages to the system. The originators can be self-contained programs, such as a calendar application or parts of a larger system, such as an email client. Typically they would build upon some existing messaging application.

The message originators try by using Jini to find a suitable Active Document that could route the messages to the proper destination. If there are several routers available, there is a problem of selection. A solution to this problem is presented in [2], where location information is used to choose Jini services in the same room as the user.

3.2 Impact controller service

The "knob", situated in the middle of the table is really a potentiometer connected to an Ethernet enabled embedded Java computer (a TINI-board from Dallas Semiconductors [5]). The knob's position is intended to represent the participant's common agreement on how willing they are to be disturbed. The idea is that the participants should be offered a possibility to control with which *impact* the messages should be displayed in the room. This could depend on e.g. the type of meeting, the state of the meeting, who the participants are, etc. The level of impact that the knob generates is called the *I-value.*

3.3 Message terminal services

In this prototype we have developed four different message terminal services:

The **iClock** service is running on a computer that hangs on a wall. The iClock shows a clock most of the time but can also fade in text messages every now and then. With the iClock service we want to examine a service that is *public, peripheral, visual,* and that displays the message *explicitly*. Our hypothesis is that a glance at the iClock is just as little intrusive as a glance at an ordinary wall clock (often present in meeting rooms). And therefore, the iClock has been assigned very low I-values, resulting in that the iClock is one of the few terminals that display messages when the I-value is set to very low. If the I-value rises above a certain level, the clock starts to chime discretely each time a new message is being displayed. The service thereby has several "modes" with different corresponding I-values (all the different terminal services have different modes in a similar way) (Fig. 1).

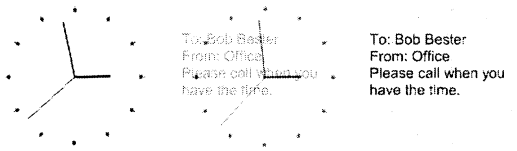


Figure 1: The iClock, first showing the time and then fading in a message for one of the participants in the room.

The **sBanner** is a service that uses the upper part of the projector display area for displaying messages. The messages body scrolls over the screen while the receiver info blinks slowly. With the sBanner service we want to examine a service that is *public, in focus, visual,* and that displays the message *explicitly*. Our hypothesis is that a message being displayed on the sBanner is very intrusive, since the projector display area often is where the participants have their shared focus. The sBanner service therefore has been assigned high I-values. The sBanner service can change its behaviour with rising I-value, by increasing scroll speed and the frequency of the blinking of the text (Fig. 2).



Figure 2: The banner service displaying a message.

The **WISP** service, also described in [6], is the source of the sounds you hear from the speakers. The speakers are constantly playing a "soundscape" where small shifts in the sound will be used to convey messages. Each recipient is associated with a specific set of sounds and the idea is to point out to the recipient that there is a message for her. With the WISP service, we want to examine a service that is *public, peripheral, audio based,* and that has a more *ambiguous* representation of the message than if it was spelled out in plain text. In this prototype, the Wisp service works as the last way out: if there are no other services suitable for presenting the message, the message is being routed to the Wisp service. The Wisp covers the whole range of possible I-values. The speaker volume is adjusted accordingly to the current value and some sounds are also exchanged to better fit the current I-value.

On the screens embedded in the table there is a service called **iBanner** running. The screens are actually Compac iPAQ computers equipped with Wireless LAN cards. The iPAQs (hence the iBanners) are 'personaliseable' by pressing a personal identifier into a receptor connected to the iPAQ. The person identifier devices are iButton memory capsules from Dallas Semiconductors [5]. The iBanner service is similar to sBanner; a service that on the upper part of the screen shows scrolling text, with the difference that the iBanner service will ring at very high I-values. With the personal banner service we want to examine a service that is *private, visual,* and that displays the message *explicitly.* Our hypothesis regarding the iBanner service is that it is quite intrusive when a participant read messages on her personal device, thus it has been assigned high I-values.

3.4 The Active Document Service

The Active Document service is responsible for the routing of the messages. The service receives all incoming messages and for each message makes a decision regarding which terminal service the message should be displayed on.

During its operation, the message router uses Jini to subscribe to notifications of terminals as they come on-line or exit. The message terminal services appear to the message router as a set of services all implementing the same service interface. All information about the terminal, and hence any routing decision, is obtained through this interface. Since the terminals appear as Jini services, terminals can be started or stopped at will. The router is automatically informed of the changes and can adapt appropriately.

The Active Document and the terminal services have also been written with an interoperability policy that can be paraphrased as 'mutual suspicion'. This means that each part supports the protocols of interaction but does not depend on the other part to honour it correctly. We believe that this approach, although slightly demanding, will be rewarding in the long perspective as it increases the robustness of the system.

When a message is delivered to a terminal it may take several seconds to actually render. It is therefore important for the router to learn when a terminal is ready to receive the next message. The decision regarding which terminal that should display a specific message depends on several conditions. For example, the routing of the messages depends on the position of the knob, e.g. the message will only be sent to a terminal that supports an impact that does not exceed the wanted value.

There could also be meta-data associated with the message that could influence the routing decision, such as a private flag (the message will only be routed to a personalised device and not to a public). A terminal must also support the message's content type etc.

The Active Document persistently looking for new possible terminal services suitable for displaying the queue of incoming notification messages. It also get informed about changes in the context, such as that a person has entered the room or that the knob's position has been changed, and react accordingly to this information.

4 Summary

In this paper we have put forward some design issues for systems that support local collaboration in ubiquitous computing environments. Based on the exploration of theses issues, we present a prototype implementation for non-intrusive notifications. This prototype system has been taken to an initial user testing and extensive demonstration. The research team's own initial use of the system and the possibilities of demonstrating it to a wide array of users has given us some interesting input on how to further design a non-intrusive, collaborative environment, both on the physical and the logical level.

The prototype also illustrates some aspects of the concept of Active Document used in a USE. The prototype is designed to work in a dynamic ad-hoc environment where user terminals, message generators and terminal services etc. come and go. The use of context information for adapting the service has been explored. The knob represents one context sensor to control the routing. We mean that the use of Active Documents to support collaboration, work processes, and individual users in different situations has once more turned out to be a promising approach and will be further developed.

5 References

[1] Mark Weiser. The Computer for the 21 st Century. *Scientific America,* 265(3) 94-104, 1991.

[2] Patrik Werle, Fredrik Kilander, Martin Jonsson, Peter Lönnqvist and Carl Gustaf Jansson. A Ubiquitous Service Environment with Active Documents for Teamwork Support, in *The Proceedings of the Ubicomp 2001 Conference,* Atlanta, Georgia, September 2001.

[3] N. R. Jennings and M. Wooldridge. Applications of Intelligent Agents in *Agent Technology: Foundations, Applications, and Markets,* Springer-Verlag, 1998.

[4] Java and Jini is products from Sun Microsystems <http://www.sun.com/>.

[5] TINI and iButton is products of Dallas Semiconductor Corp http://www.ibutton.com/.

[6] Fredrik Kilander and Peter Lönnqvist. A Weakly Intrusive Soundscape for Intuitive State Perception. Position paper at *the continuity in Future Computing Systems workshop* at the I3 2001 Spring Days, Portugal, 2001.

Authorization and Access Control System in Loosely Connected Distributed Environments

Matei Ciobanu Morogan[1]

[1] DSV KTH/SU, Forum 100, 164 40, Kista, Sweden. E-mail: matei@dsv.su.se.

Abstract

Insuring certain required level of security and confidentiality in computer network applications usually leads to a compromise between functionality (like accessibility, ease of use, etc) and security restrictions (e.g. only allowing access from certain computers). With computer networks becoming ubiquous and with a large variety of terminals and access points available, the restrictions imposed by traditional approaches become very limiting.

For an authorization system to be useful and to provide high level of functionality in such environments without compromising security, it has to be specifically designed, taking into account the differences that exist and specific situations that can arise.

In this paper we are investigating how such a system should be designed, by first setting up a list of requirements it has to satisfy, then creating a model for a solution.

1 Introduction

Current IT trends like mobile computing and wireless networks lead to new form of computer networks which is dynamically created by the devices themselves when in proximity to each other. This is usually referred to as ad–hoc proximity networks [1], [2].

Ad-hoc proximity networks differ from wired local area networks in several important aspects [3]. One aspect is that access to the global network can not always be established, either because of lack of coverage or because of high cost of communications. Therefore, transactions and control mechanisms in such networks must be as independent from outside resources as possible. Another specific aspect is unpredictability of the network, that is, devices which form it are not preregistered or known in advance. Networks are formed dynamically by users and devices in a proximity. For instance, when several people meet in a public place, a network would be formed by their PDA:s, mobile phones, laptops, and maybe some public terminal in their vicinity. Finally, the control of devices in the network can not be as strict as in wired LANs. In wireless networks it is quite easy to connect from the outside without knowledge of other users, since proximity is quite difficult to specify and use as an access control criteria.

The following aspects are key differences between ad-hoc proximity networks and classic, wired LANs:

- Access to servers or resources in the global network may be expensive or impossible.
- Network configuration is unpredictable, due to unknown and unregistered users.
- Access to a network is open to all devices and users in physical proximity.
- Devices that form an ad-hoc network, ranging from mobile phones to laptop computers and networked speaker systems, are usually heterogeneous with limited computing and memory resources.
- Devices that form ad-hoc networks are more exposed to failure and theft than wired stationary computers.

Because of these specific characteristics of formation and usage of ad hoc proximity networks, security for them is more important and also not completely equivalent to security in standard wired LANs. Classical security solutions for computer networks may not work in these environments or may pose constraints which highly reduce their functionality. Therefore, security solutions specific for ad hoc proximity networks need to be created providing the functionality found in classical computer networks.

For users to be able to share resources in a network, some form of *authorization system* needs to be in place. Authorization systems are used to regulate actions that users can perform with objects in the system. For that, a *policy* is created, specifying the conditions that have to be fulfilled to allow a user to perform an action on a particular object. Based on this policy, the authorization system makes and enforces *authorization decisions*.

Previous work has been done on authorization systems for distributed networks. One interesting example is the Akenti system [4]. This system is based on use-condition certificates and attribute certificates which may be fetched by a target server from their owners at the time when the access decision has to be made. While offering great flexibility, it is designed with

the idea of network access being always available, something that is not guaranteed in ad hoc proximity networks. Another example is [5], a system that uses certificates to carry access permissions with Java code. These certificates specify the access rights that a certain Java class has on systems. If the target host accepts the signer of the certificate, the class is allowed to perform the actions. In this way, it makes it possible to set up authorizations and capabilities of code, but there is no way to specify access conditions for a particular object.

The goal of this paper is to elaborate the model of a system for secure authorization in ad-hoc proximity networks.

2 Design Requirements

As a consequence of the characteristics of ad-hoc networks described in the introduction, the following features are required in a security system for such environments. The system has to be:

- de-centralized – to support disconnected operations;
- flexible – to support varying characteristics of hardware platforms;
- dynamic – to accommodate unpredictable network structure;
- autonomous – to work without network access;
- ad-hoc – works with no prearrangements whatsoever.

It should be possible to specify access conditions for each object in a system, where object means either data object or resource objects (i.e. files, documents, printers, screens). Users should have different capabilities, and the authorization system should decide whether to allow or not an action to be performed, based on comparison of access conditions for an object and the capabilities (authorizations) of a user. This design gives flexibility, by allowing access conditions to be specified either generally, based on user properties (all users with certain characteristics are allowed to do a certain action) or specifically based on the user identity.

A functionality that is not found in classical authorization systems but would be useful in ad hoc networking environments is to make it possible to specify access conditions based on proximity, or on physical presence in a particular area (for example, access to the printer permitted to all users in a room). This implies the existence of a service that provides secure location information.

3. System Model

Based on the requirements listed in section 2 the following model for the authorization system has been created.

In this model, both data and resources are represented as objects. Objects provide the possibility to perform a number of actions. Each object contains a list of conditions that a user must fulfill to be allowed to perform an action. Furthermore, individual objects can be part of one or several authorization domains. In each authorization domain there is a central authority that all objects in the domain trust. Access conditions for an object may differ between domains (see Figure 1).

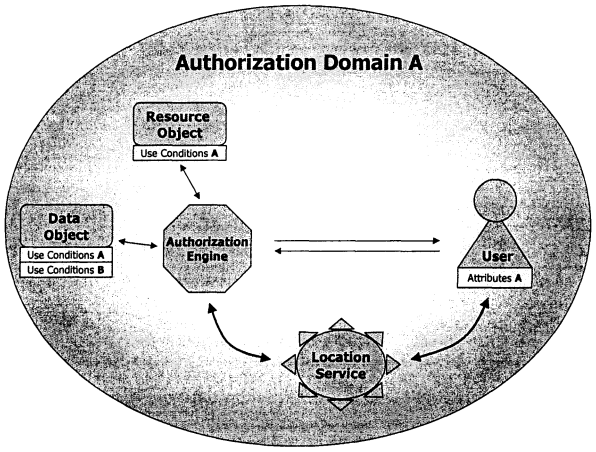


Figure 1 Authorization Domains

Users have one or several attribute certificates [6] signed by the central authority of the domain. These certificates state the capabilities or access rights a user has, or may state that the user belongs to a certain group. The attribute certificates are bound to the identity of the user, so in order to use them, the user needs to be authenticated. For this, X.509 certificates are used [7].

Access decisions are taken by an Authorization Engine (AE). This is a service that compares access conditions for objects and attributes of users and decides if an action is allowed or not. When a user wants to perform an action with an object, it needs to first contact an AE to get an access decision. The AE authenticates the user, receives user's capabilities, then compares them with the object's access conditions. Depending on the those conditions, the AE may contact the location service. The AE also caches certification information, so that authentication and X.509 certificate exchange can be performed. If the decision is to approve the action, the AE returns a ticket that the user can present to the object.

4. System Components

In this section we describe components of the system.

4.1 Data and Resource Objects

All data and resources in the system are represented by objects. They support a number of actions that can be performed with them. For example, an object representing a document may provide "read" and

"update" actions, and an object representing a printer may provide "print" and "get status" actions. Users are allowed to perform these actions only if they can present a time-stamped ticket from the AE that states that the user has fulfilled the necessary conditions. When a user contacts an object without a ticket, the object replies by indicating the AE to be used.

4.2 Use Conditions

Objects contain a list of access conditions specific for each domain and each individual object. Access conditions describe the access policy for each of the possible actions that can be performed on the object. The use conditions for a certain domain require user attributes certified by the central authority of that domain. Examples of such conditions are that the user needs to be a member of a certain group to be permitted to read the object, or that the user has to be in a certain location.

4.3 User Attribute Certificates

Each user has a number of attribute certificates. These state what capabilities a user has, or if the user is a member of a particular group. The attribute certificates for a domain are signed by the domain authority.

4.4 Location Service

The Location Service provides secure location information. This information can have several levels of trust, depending on the reliability of the method used to obtain it. For example, information based on the device being in range for a wireless network may be used to regulate access to a printer, but not in more sensitive applications. The Location Service returns location information upon request of Authorization Engines certified in the domain.

4.5 Authorization Engine

When a user wants to perform an action on an object, it needs to obtain a ticket from the Authorization Engine (AE) accepted by the object. This may be, depending on the particular security policy, any AE in the domain, or a specific AE, for example the one on the device the object is stored on.

The AE's are certified by the domain authority. When a user contacts an AE to obtain an usage ticket, a challenge-response authentication is performed to verify both user's and AE's identities. Then, the required attribute certificates are presented to the AE. The AE contacts the object that is being accessed and retrieves its access conditions. Depending on the access conditions, it may also contact the Location Service and inquire about the location of the user and of the object. Based on the access conditions and also on the general security policy of the domain, the AE makes an access decision. If the decision is to grant access, a ticket is returned to the user.

Since AE's may not be able to reach other servers in the domain, they need to store all relevant information locally. Therefore, root certificate for the local X.509 certification structure is stored, and also the certificate of the domain authority. Other information that has to be stored is the security policy file of the system. All the stored information should be updated at regular intervals, when the authorization engine is in contact with other authorization engines in the system.

4.6 Domain Authority

An authorization domain is a logical domain defined by a common domain authority. All attribute certificates that are valid in this domain are signed by the local domain authority. The domain authority also signs the security policy file for the domain. The objects in the domain accept tickets from Authorization Engines certified by the same domain authority. An example of a domain would be a company, with the security administration being the domain authority. The administration issues certificates to users stating that they are part of a particular group or that they have certain capabilities. All objects in that domain accept the authority of the administration.

5 Other Issues

To make the system usable on a large scale, issues relating to authorization across domains have to be addressed. Also, if an authorization domain is large, there may be a need to delegate the authority to issue attribute certificates to several authorities then the central domain authority.

5.1 Cross-domain Authorization

The system model presented here works as long as the user has been previously registered in a domain and possesses specific attribute certificates. If a user changes domain, the capabilities and group memberships from the original domain are no longer valid. While this is the correct behavior in most cases, in certain situations it may be useful to make some attributes from one domain valid in another domain as well.

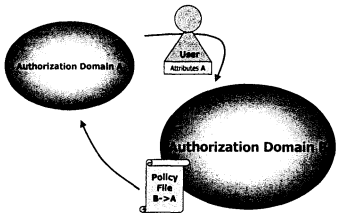


Figure 2 Cross-domain Authorization

Cross-domain authorization has to be set up by the local domain authorities. If a domain authority decides to accept attribute certificates from another domain, it specifies that in the domain security policy file (Figure 2). Each Authorization Engine should update this file at regular intervals. The Authorization Engines should also cache the certificates for all the other domain authorities that are specified in the security policy file.

5.2 Delegation of Authority

If a domain grows over a certain size, it may become unfeasible that all attribute certificates are issued by one single domain authority. In that case, the domain authority needs to be able to invest other authorities with the power to issue attribute certificates to the users. Therefore, instead of having just one unique domain authority, a hierarchy of authorities would be created, with the original domain authority at the top. Certificate chains for the signing authorities need in this case to be attached to the attribute certificates.

6 Conclusions

The authorization system described here is optimized to work in the environment of ad hoc proximity networks. To do that, it concentrates much of the functionality of the authorization system in the Authorization Engines. These cache security information from the domain, being thus able to take authorization decisions even when not in contact with the rest of the network. The disadvantages of this approach is that propagation of certain types of information (security policies, certificate revocation lists) may be slower. The use conditions are stored within the objects, so that they are accessible whenever the object is accessible.

Further directions of research are to define the model in more detail, and to design and implement an authorization system based on it.

References

[1] D.B. Johnson. *Routing in Ad Hoc Networks of Mobile Hosts*. In Proc. of the Workshop on Mobile Computing Systems and Applications, pages 158-- 163, 1994.

[2] R. Meier, M.-O. Killijian, R. Cunningham, V. Cahill. *Towards Proximity Group Communication*, Advanced Topic Workshop - Middleware for Mobile Computing, Heidelberg, Germany, November 2001.

[3] G.-C. Roman and P. J. McCann and J. Y. Plun. *Mobile UNITY: Reasoning and Specification in Mobile Computing*. ACM TOSEM, VOL. 6, no. 3, pages 250-282, July 1997.

[4] M. Thompson, W. Johnston, S. Mudumbai, G. Hoo, K. Jackson, A. Essiari. *Certificate-based Access Control for Widely Distributed Resources*. In the Proceedings of the Eighth Usenix Security Symposium, Aug. `99.

[5] P. Nikander, J. Partanen. *Distributed Policy Management for JDK 1.2*. Proceedings of Network and Distributed System Security Symposium, San Diego, CA, Feb 3-5, 1999.

[6] S. Farrell, R. Housley. *An Internet Attribute Certificate Profile for Authorization*. Internet Engineering Task Force Draft, work in progress.

[7] R. Housely, W. Ford, W. Polk, D. Sodo. *Internet X.509 Public Key Infrastructure*. Internet Engineering Task Force Draft, PKIX Working Group, work in progress.

Dynamic Service Composition: Why and How

Sven Lämmermann

Royal Institute of Technology

Department of Microelectronics and Information Technology
Electrum 229
S-16440 Kista, Sweden

laemmi@it.kth.se

Abstract

In this paper we discuss services and automated composition of (distributed) services; in particular we consider properties of services in terms of decomposability. We define components and composition with respect to services and distributed computing, and we introduce the notion of runtime service composition. The questions of why runtime service composition is needed and how it can be provided will be answered. We also present requirements for runtime service composition methods.

1 Introduction

We believe that on-the-fly composition of services from components is becoming attractive, but requires advanced composition methods. Composability, configurability, and extensibility of services provided by interconnected computer systems and their implementations have recently been the targets of active research efforts (Nikolaou, Marazakis, and Papadakis 1998, Beringer, Wiederhold, and Melloul 1999, Casati *et al.* 2000, Ben Shaul, Gidron, and Holder 1998, Mennie and Pagurek 2001, Mennie and Pagurek 2000, Fuentes and Troya 1997). For various reasons, many research contributions are contributing to this topic. On the one hand, results of traditional research communities such as software engineering, and even more importantly automated software engineering, and distributed systems are needed for providing dynamic services composed at service request time. For instance, a service system is software that requires being maintainable and flexible, which is obviously a software engineering problem. On the other hand we currently face a trend away from traditional service systems, which follow the client-server computation model and are not likely to change frequently, to dynamic services that change as systems evolve and requirements change. This trend coincides with research in the area broadly referred to as peer-to-peer computing, or peer-to-peer networking, or simply P2P.

Peer-to-peer is the latest moniker for a research area that tries to solve a number of problems in modern distributed computing. The aim of research in peer-to-peer computing is to greatly increase the utilization of three valuable and fundamental Internet assets: information, bandwidth, and computing resources. All of these are vastly underutilized at this time, partly due to the traditional client-server computing model (Gong 2001). A typical example is popular web sites. While available network bandwidth has tremendously increased, we still feel congestion while accessing popular web sites, because everyone accesses popular sites.

In particular, research in peer-to-peer computing addresses besides others, problems of interoperability of service systems and platform independence. Together with problems addressed by active networks research (Tennenhouse *et al.* 1997), it gave us motivation for conducting research in the area of runtime service composition.

2 Traditional Services

Every service possesses its own properties in terms of communication, security, quality of service, etc. (Hiltunen 1996). Defining a canonical model for all services turns out to be a hard task. However, services can roughly be classified as follows:

- **Request/response**-oriented services involve a single send–receive message at client side and a single receive-send at server side.
- **Session**-oriented services involve exchange of several messages between client and server. They may be stateful and stateless. Session oriented services have longer duration (Christophides *et al.* 2001).
- **Event**-oriented services are services where the client is passive and waits for a message from the server that a certain situation occurred. Typically, a client subscribes for an event.

All classes represent distributed services involving one or more *service points*. We define a service point as a node in a computer network that runs software that implements some service. This software corresponds to

servers in the traditional client-server computing model and corresponds to objects in distributed systems.

Every service (or distributed object) is associated with an interface and location. The interface of a service describes how to access this service. The location specifies at which service point this service can be accessed. Service location information is used for service request routing. In Internet protocol (IP) networks service request routing reduces to IP routing.

A service location is either statically known or dynamically evaluated. Most traditional services require a user to supply the location of a service point, i.e. a service user must be location- and context-aware. This is significantly different in dynamic service environments. There, a user asks for a service by specifying the required service rather than accessing a particular service interface at a supplied service point whose location information must be supplied by the user.

3 Runtime Service Composition: Why?

We define *dynamic service composition* as a method to synthesize service programs from components within the user interaction loop ("on the fly") on the basis of a service request that specifies the requested service. The actual service program does not exist before the service request is issued. The specification of a service is a query to a composition subsystem rather than the access of a particular service interface. Service composition involves component selection, establishing communication channels between components, and generation of interface code that fits selected components in the context of their use. We sometimes refer to dynamic service composition as *just-in-time* or *runtime* service composition.

We believe that the ability to tailor services (not only network services) to application requirements automatically and at runtime will ultimately enhance flexibility and adaptability as well as maintainability of service systems, because it enables us more agile development of services with minimal interruption to running systems. It also enables us to extend protocols in a relatively easy way, even such that backward compatibility is preserved. On the other hand, dynamic adaptation of composite services is needed in order to accommodate changes that cannot be accommodated on lower systems levels, for instance communication and operating system level. Such changes are especially significant in mobile computing systems (Tosic, Patel, and Pagurek 2001). Several research projects in the area of operating systems are based on the premise that traditional operating system structuring limits performance, flexibility, and functionality of applications. For example, (Cao, Felten, and Li 1994) demonstrate that application-level control of file caching reduces application runtimes by 45%. Similarly, application-specific virtual memory policies increase application performance (Harty and Cheriton 1992, Krueger *et al.* 1993), while exception handling is an order of magnitude faster if the signal handling is deferred to applications (Thekkath and Levy 1994). Therefore, configurability and extensibility of both the abstractions provided by the operating systems and their implementations have recently been the targets of active research efforts. The Synthesis system (Pu, Massalin, and Ioannidis 1988) was one of the first projects to explore configurability and extensibility in the context of operating systems. Perhaps the key contribution of Synthesis is its use of optimization techniques from the area of dynamic code generation in its kernel design. Such techniques can produce efficient executable code since they can take advantage of the extra information in the runtime execution context. For example, frequently executed kernel calls can be regenerated at runtime using compiler code optimization ideas such as constant folding and macro expansion. Synthesis, therefore, is an example of adjusting or configuring the implementation to improve performance without modifications to the high-level operating system abstractions provided for the applications.

Although building systems from components has attractions, this approach also has problems. Can we be sure that a certain configuration is correct? Can it perform as well as a monolithic system? These two questions are discussed in (Liu et al. 1999) for the ENSEMBLE communication architecture (Hayden 1998). This work describes how correct protocol stacks can be configured from ENSEMBLE's micro-protocol components, and how the performance of the resulting system can be improved significantly using the formal tool NUPRL (Constable *et al.* 1986). The goal of this work is also to demonstrate that the NUPRL **L**ogical **P**rogramming **E**nvironment (Constable *et al.* 1986) is capable of supporting the formal design and implementation of large-scale, high-performance network systems. The NUPRL LPE has been used in the verification of protocols for the ENSEMBLE group communication toolkit (Kreitz, Hayden, and Hickey 1998, Hickey, Lynch, and van Renesse 1999), in verifiably correct optimizations of ENSEMBLE protocol stacks (Kreitz 1999, Liu *et al.* 1999), and in the formal design and implementation of new adaptive network protocols (Liu *et al.* 2001, Bickford *et al.* 2001a, 2001b). Its has been shown that with help of the NUPRL formal system, configurations of components may be checked against specifications, and how optimized code can be synthesized from these configurations. The performance results show that one can substantially reduce end-to-end latency in the already optimized ENSEMBLE system. Although the domain is rather limited — all components are protocol layers, and all configurations are stacks of these — the researchers believe that this approach could generalize and scale to more general configuration and component types.

A very important aspect of runtime service composition is that new composite services need not be envisioned at design time, which is known as unanticipated dynamic composition (Kniesel 1999). This feature provides considerable flexibility for modifying and extending the operation of software systems during runtime.

3.1 Interoperability

Most service systems are built for delivering a single type of service, e.g., file sharing services, instant message services, and electronic banking service systems. All of them are designed to serve a particular purpose and have therefore varied characteristics in their architecture and properties. They lack a common underlying service infrastructure. Each service software vendor tends to create its own proprietary service system, including service architecture and communication protocol, which makes those service systems incompatible. Efforts in creating software and systems primitives commonly used by all service systems are duplicated and each vendor creates its own service user community. As a result, a service client participating in multiple communities created by different service implementations must support multiple implementations, each for a distinct service system or community, and serve as the aggregation point. As a consequence, users are locked into one community. Service providers have to offer their service in ways that are specific to how each community operates (Gong 2001).

3.2 Platform Independence

In order to make services and their employing communities easily accessible, platform independence becomes a crucial issue. Services of most systems are offered through a set of APIs that are delivered on a particular operating system. Those services use a specific networking protocol. Service client developers are forced to choose which set of APIs to program to, and consequently which set of service customers to target (Gong 2001). If the same service has to be provided to different communities, the same service must be developed for all platforms used by the communities.

From the software (and service) engineering point of view supporting multiple implementations of the same service software is inefficient and makes complex service systems hard to maintain, especially if requirements rapidly change and new services must be introducible without effecting older systems (i.e. keeping a service system backward compatible).

Runtime service composition can help to tackle this problem in the following way. Services should be decomposed into their components. Components developed for a particular platform running on a specific operating system exhibit their interface by describing computability of objects and the respective domains of these objects, which is needed to preserve type compatibility across different platforms. The composition process detects the necessity of usage of a component. Components are not moved to any site. They are executed at their respective provider site (which can be a client or server). In this way we can create distributed implementations of service software that do not depend on any platform or operating system.

3.3 Active Networks

Active networks permit applications to inject programs into the nodes of local area networks and, more importantly, wide area networks. This supports faster service innovation by making it easier to deploy new network services. Traditionally, new network services are deployed at end-systems. For example, a cache of an http-server is deployed as overlay of the server. Recent results of research in active networks show that implementing new services at nodes interior to the network or at the network layer often offers better functionality, performance, and better exploitation of resources (Wetherall, Legedza, and Guttag 1998).

These results are visible in many applications. For instance, multimedia applications benefit from real-time and multicast services where bandwidth reservation is needed to ensure that data is delivered in time. When overlays are used, bandwidth and latency costs increase, which is not desirable. Another example is http-servers that benefit from caching and load balancing. Intercepting repeated http-request packets at routers and then distributing requests across multiple servers minimizes latency and bandwidth usage compared to a proxy agent.

Introducing and changing network protocols requires standardization and backward-compatible deployment. This makes the process of changing network protocols lengthy and difficult. Runtime service composition can support active networks to address the problem of slow network service evolution by introducing fully automated service synthesis into the network service infrastructure itself, thus allowing not only many new services to be introduced much more rapidly, but also supporting a very dynamic evolution of services. Because of the ability to tailor network services automatically to requirements (which may frequently change) imposed by the user or the environment, we are in position to make services to be much more flexible and adaptable.

4 Runtime Service Composition: How?

Here we present our approach to runtime service composition. We describe the main ideas and mention the concepts and techniques used. Although the presentation is rather brief and informal, it should

highlight major features of the method proposed in this thesis. Our research in runtime service composition is focused on two questions:

- In what way can services be composed efficiently and fully automatically without increasing service latency?
- How can we give a composed service a concurrent implementation, i.e. how can we parallelize a composed service efficiently and fully automatically?

There are another questions that must be answered before runtime service composition can be deployed, e.g., the practical question of how to route service requests in very large communication networks. Runtime service composition does not address a particular service point; rather it tries to find out which service points can be used to compose a particular service. In small networks such as local area networks this is not a problem because service information related to service points can simply be maintained in a local database. However, this is impossible for large networks such as the Internet. Recent results of research in peer-to-peer computing propose some possible solutions; see for instance (Gong 2001). This question is still subject of research and goes far beyond the scope of this thesis.

The general idea behind our approach to runtime service composition is that a service request specifies a service, i.e. its input/output relation, rather than accesses a particular service interface. A service specification can be modeled as a hypothetical judgment whose evidence should be established by a proof system. Formal evidence of a hypothetical judgment is presented by a constructive derivation, which is a proof of existence of the specified service. Proof search is logical service composition and the result of proof is a service program if evidence of the hypothetical judgment can be provided. A service request is a service composition problem.

Taking this as a starting point we explain now what runtime service composition involves, what it requires, and how our approach to service composition works in general.

4.1 Service Components

We define a service component (we write component if this cannot cause any confusion) as a pre-programmed function that is a black box that provides some functionality of a service and may require another service at the same time. A service component that requires services is configurable. For instance, a component that provides a printing service may get different file converters as input depending on file type and printer type respectively. Service components are composable, reusable, and replaceable self-contained units. They encapsulate some functionality, business logic, or any kind of computation and communication. Service components can be reused in different service compositions. A service component is relatively independent from other service components in a particular composition; there are only dataflow dependencies among components. A component can be detached from a composed service and replaced with another appropriate component, possibly from a different source.

In (Tosic, Mennie, and Pagurek 2001) it has been pointed out that there is a distinction between a service component and a composed service. A service has some business-oriented or goal-oriented meaning for an end-user, while a service component has value primarily in service compositions and not when used in isolation. This distinction is context-dependent. The same unit of functionality can be a service component in one context and a service in another.

Other approaches to runtime service composition make also a distinction between a service component as a unit and implementation-level concepts such as software components. In (Tosic, Patel, and Pagurek 2001) a service component is viewed as a higher-level concept that can be implemented using software components. We do not make such a distinction since our approach to runtime service composition uses only pre-programmed components where the internal structure (i.e. implementation of a service component) is uninteresting for us. We are interested only in functionality provided by a component and conditions about when a particular component can be used and what conditions hold after component execution. Flexibility is added to components by parameterization, if needed. In this sense, components are black boxes but they are configurable. For example a network communication component may require a security component, where the security component is not statically known at specification time. A proper security component must be found automatically at service composition time.

4.2 Service Component Specification

Our approach to provide runtime service composition uses a specification of a requested service, where the service specification is included into a service request. The specification of a service is a query to a composition subsystem that attempts to synthesize a service program that satisfies the given service specification.

The composition subsystem operates on logical specifications (for the sake of brevity we say specification further on). It is required that a component to be used in service compositions has to be supplied with such a specification. We define the specification of a component as a description that states the conditions under which this component can be used and what conditions hold after its application. Specifications of components should be encoded in logic using an appropriate specification language. We propose a specification language that can be mapped into the logical language of a constructive logic. The logic is

expressive enough for describing, first, the structure of hierarchical configurations appearing in services and, second, data flow between the service components.

Communication and computation requires interpretation of names (Vivas Frontana 2001). We use names in specifications to express input and output conditions. Operationally, names represent data items that take on values of a particular domain. Logically, names occurring in specifications of components are propositions that represent conceptual roles that values of their corresponding data items play in computation. For instance, if we write `hostname` as proposition then values of the concept hostname may be of type string but identify a host in a network in computation. If we would write `string` as proposition then any string would be an appropriate value and the information that a hostname is required would be lost.

In order to compose service components of different sources component providers have to agree on naming conventions (i.e. conceptual roles) that are applied to specifications of service components. In order to be understood precisely one must agree on a common ontology. This requirement is not new to the service community. Typical examples are all string-based protocols like HTTP and SMTP, and network management systems. Network management systems, for instance, use management information bases (MIBs) where every MIB entry has a unique object identifier.

4.3 Service Composition

A service is a more or less complex software system that can be described in two distinct ways, as a running system or as a composition of software components. A running system can be viewed as a collection of interacting objects. The same system can be described as composition of software components at system specification level (Nierstrasz and Meijler 1994). Accordingly, we argue that every service, which possesses an intrinsic service flow from one runtime entity to another, is runtime composable. This requires the decomposition of the service under consideration into components that correspond to runtime entities of the running service. Service flow among runtime entities can be represented as higher-order dataflow.

Dataflow schemas have the valuable property that they can be used as data structure for a time-efficient planning algorithm. The algorithm works in simplest cases in forward-direction, and is suitable for parallel implementation. Planning can be regarded as proof search in intuitionistic propositional logic. A thorough study on this topic can be found in (Tyugu and Uustalu 1994).

We use this algorithm to compose services. A service composed in this way does not have constraints on execution order other than imposed by data dependencies and by logical constraints explicitly expressed in the specifications of components. Moreover, its specification contains explicit and easily usable information about all data dependencies that must be taken into account when synchronizing concurrent execution of its parts (Lämmermann, Tyugu, and Vlassov 2001). We use this information for parallelization of composed services.

Our proposed runtime service composition method works as follows. When the requested service is specified, we run the planning algorithm. From its output, which is a proof of existence of the requested service if a realization for that service exists, we select and configure the components required for realizing the requested service. If composition involves local components we generate the interface code that fits components in the context of their use or, in other words, the program code that glues components to a service program. If components are distributed over several service points then we exercise coordination rather than gluing.

5 Sample Application

Here, we outline a new network service aimed at improving an existing application. It is our intention to show the diversity of services facilitated by runtime service composition in concert with active networks.

Obtaining stock quotes via the Internet is popular. Having fast access to up to date quotes during periods of heavy server load is crucial. Web caches do not help in this context. First, stock quotes are dynamic data that is usually not cached by web caches. Second, the granularity of objects stored in web caches is inappropriate for this application: clients request a web page that contains a short customized list of quotes but considering hundreds of stocks, the number of possible web pages is huge and, hence, the probability of cache hits is very low.

Active networks allow tailoring caching strategies to suit the application needs. Assume we want to implement a customized protocol for accessing stock quotes, which also allows implementing a reasonable caching strategy at an interior node in an active network. Assume some node of our active network implements a web cache, which does not cache dynamic data. We can model this web cache by two components. The main component, which is configured with a cache component, processes the HTTP request and does the cache lookup. On cache hit it sends the response. On cache miss the HTTP request will be forwarded to the corresponding server. The basic idea is that web pages are cached at network nodes as they travel from the server to a client. Figure 1 depicts this scenario, where components are boxes and arrows indicate dataflow between them.

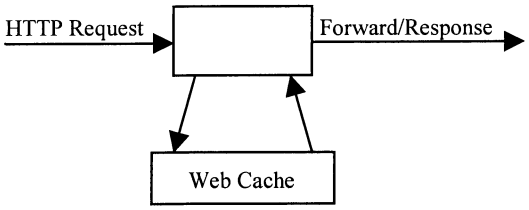


Figure 1 Web cache of active network node

In order to enable an active network node to cache stock quotes we need to upload four more components to this node: one that acts as protocol filter, one that can handle the stock quote protocol, a stock quote cache, and a stock quote formatting component. The protocol filter receives any HTTP and stock quote request. It returns which request it has received. The stock quote protocol handler is configured with the stock quote cache and formatting component. On stock quote request and cache hit the formatting component assembles stock quotes into a viewable format. In Figure 2 we show all components boxes and their respective dataflow dependencies as arrows leading from one box to another.

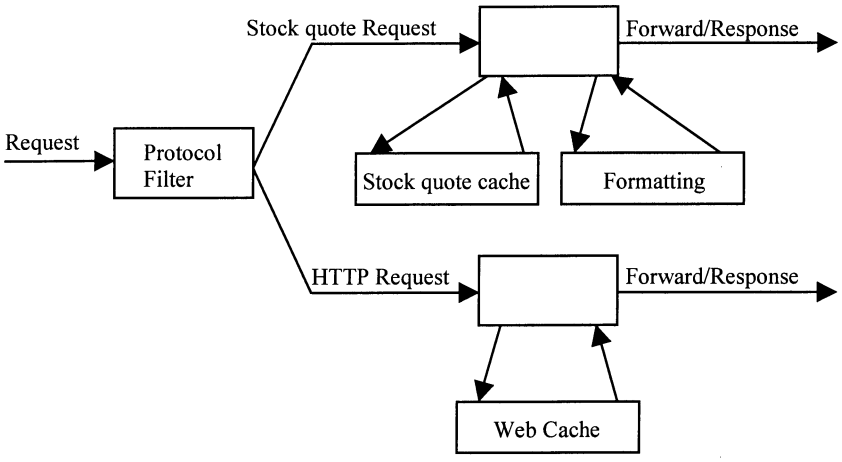


Figure 2 Web and stock quote cache of active network node

The basic idea is that in order to extend the functionality of the original web cache service we simply upload new components supplied with specifications to the network node. The next HTTP or stock quote request will automatically invoke the composition subsystem of this node to reconfigure the service by synthesizing a new service application that can handle both HTTP and stock quote requests.

6 Requirements for Runtime Service Composition

In general, runtime service composition is software composition, but it requires special treatment if the composition process becomes part of the user-service interaction loop. Composing services statically means composition of software at service design time and requires specialized architectures for each kind of service. For instance, in (Fuentes and Troya 1997) a component-oriented service architecture is proposed as flexible service creation framework. It is specialized for multimedia services. We claim that a runtime service composition method should be service domain independent, because a composite service may involve components of different service domains. For instance, a composite online-banking service may require multimedia support as well as transaction service support.

We focus on a composition method for services that are put together from pre-programmed service components. In this context, dynamically composing a service from components means:

1) Finding all appropriate components that are needed for the service program.
2) Adapting the components to the particular use.
3) Synthesizing the interface code that matches the components to the context of their use.

A service is composed on the basis of a service request that specifies the service. For instance, the service to withdraw funds from a bank has the following request:

"Given bank, account and amount that has to be withdrawn, build a service program that receives a bank identifier, an account number, and amount as input and returns funds."

Composing a service on request puts hard requirements on the composition method:

1) The composition method must be efficient, because the composition is performed in a dynamic service environment (in the user interaction loop).

2) The composition method must work autonomously, because the user has no abilities to provide help in the synthesis process.

3) The composition method must guarantee correct deployment of components.

4) Even more importantly, the composition method must also provide a specification language that enables one to specify services and components in a reasonably easy way.

We define a runtime composite service as a service where the actual service program is composed on service request. Hence, the first requirement, the efficiency of the composition method, results from the additional step (i.e. the actual service composition) that must be taken in a dynamic service environment. The service response time in traditional service environments depends on resource latencies (network, database, etc.), server load, and service execution time. In our case the service response time includes additionally the service composition time. Therefore the composition method must be time-efficient in order to provide a service composition with low costs.

In dynamic service composition environments it is not acceptable to involve end users in service composition. Hence, there is no user-service interaction during the service composition process. That gives us the second requirement of the composition method: the composition method must proceed autonomously.

The third requirement, the guarantee of correct deployment of components, is obvious. This requirement can be satisfied by the logical correctness of the composition method and by correctness of specifications of components.

A key requirement to our composition method is to provide a specification logic that enables us to reason automatically about service components. The specification logic allows us to find automatically all appropriate service components that are needed for a service to be composed and to synthesize the interface code that matches the service components (calls methods of classes). An essential feature of our proposed logic is the possibility to express dependencies of service components from different sources. This enables us to specify service component capabilities and the context where a service component can be used during application composition.

7 References

Ben-Shaul, I., Gidron, Y., Holder, O. 1998. A Negotiation Model for Dynamic Composition of Distributed Applications. Proceedings of the Ninth International Workshop on Database and Expert Systems Applications (DEXA '98), Vienna, Austria, pp. 820-825.

Beringer, D., Wiederhold, G., Melloul, L. 1999. A Reuse and Composition Protocol for Services. Proc. Symp. On Software Reusability (SSR'99), ACM SIGSOFT, Los Angeles CA.

Bickford, M., Kreitz, C., van Renesse, R., and Constable, R. 2001a. An experiment in formal design using metaproperties. In Lala, J., Maughan, D., McCollum, C., and Witten, B., eds., DARPA Information Survivability Conference and Exposition II (DISCEX 2001), volume II, pp. 100-107, IEEE Computer Society Press.

Bickford, M., Kreitz, C., van Renesse, R., and Liu, X. 2001b. Proving hybrid protocols correct. In Boulton, R., and Jackson, P., eds., 14th International Conference on Theorem Proving in Higher Order Logics, LNCS 2152, pp. 105-120, Springer Verlag.

Cao, P., Felten, E., Li, K. 1994. Implementation and performance of application-controlled file caching. In Proceedings of the 1st USENIX Symposium on Operating Systems Design and Implementation (OSDI), pp. 165–178.

Casati, F., Ilnicki, S., Jin, L. J., Krishnamoorthy, V., Shan, M.-C. 2000. Adaptive and Dynamic Service Composition in eFlow. HP Labs Technical Reports, HPL-2000-39, 20000406.

Christophides, V., Hull, R., Karvounarakis, G., Kumar, A., Tong, G., Xiong, M. 2001. Beyond Discrete E-Services: Composing Session-Oriented Services in Telecommunications. Proceedings of the Second International Workshop, Rome, Italy, Lecture Notes in Computer Science 2193 Springer 2001, pp. 58 ff.

Constable, R. L., Allen, S. F., Bromley, H. M., Cleaveland, W. R., Cremer, J. F., Harper, R.W., Howe, D. J., Knoblock, T. B., Mendler, N. P., Panangaden, P., Sasaki, J. T., and Smith, S. F. 1986. Implementing Mathematics with the NuPRL proof development system. Prentice Hall.

Fuentes, L., Troya, J. M. 1997. Component Oriented Service Architecture as a Flexible Service Creation Framework. In Proceedings of Communication Networks and Distributed Systems Modeling. CNDS'97, Phoenix, USA, pp. 68-73.

Gong, L. 2001. JXTA: A Network Programming Environment. IEEE Internet Computing, (5)3:88--95.

Harty, K., Cheriton, D. 1992. Application-controlled physical memory using external page-cache management. In Proceedings of the 5th International Conference on Architecture Support for Programming Languages and Operating Systems, pp. 187–199.

Hayden, M. 1998. The Ensemble System. PhD thesis, Cornell University.

Hickey, J., Lynch, N., and van Renesse, R. 1999. Specifications and proofs for Ensemble layers. In Cleaveland, R.,ed., 5th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, LNCS 1579, pp. 119-133, Springer Verlag.

Hiltunen, M. A. 1996. Configurable Fault-Tolerant Distributed Services. Ph.D thesis, Department of Computer Science, University of Arizona Tucson, Arizona 85721, TR96-12.

Kniesel, G. 1999. Type-Safe Delegation for Run-Time Component Adaptation. In Proceedings of ECOOP '99 (LNCS 1628), Springer-Verlag, Lisbon, Portugal, pp. 351-366.

Kreitz, C. 1999. Automated fast-track reconfiguration of group communication systems. In Cleaveland, R., ed., 5th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, LNCS 1579, pp. 104–118, Springer Verlag.

Kreitz, C., Hayden, M., and Hickey, J. 1998. A proof environment for the development of group communication systems. In Kirchner, C., and Kirchner, H., eds., 15th Conference on Automated Deduction, LNAI 1421, pp. 317-331, Springer Verlag.

Lämmermann, S., Tyugu, E., Vlassov, V. 2001. Concurrent Implementation of Structurally Synthesized Programs. In Proceedings of the 6th International Conference on Parallel Computing Technologies, PaCT-2001.

Liu, X., Kreitz, C., van Renesse, R., Hickey, J., Hayden, M., Birman, K., and Constable, R. 1999. Building reliable, high-performance communication systems from components. In 17th ACM Symposium on Operating Systems Principles (SOSP'99), volume 34 of Operating Systems Review, pp. 80–92.

Liu, X., van Renesse, R., Bickford, M., Kreitz, C., and Constable, R. 2001. Protocol switching: Exploiting metaproperties. In Rodrigues, L., and Raynal, M., eds., International Workshop on Applied Reliable Group Communication (WARGC 2001), pp. 37–42, IEEE Computer Society Press.

Mennie, D., Pagurek, B. 2000. An Architecture to Support Dynamic Composition of Service Components. 5th International Workshop on Component-Oriented Programming, WCOP 2000, Cannes, France.

Mennie, D., Pagurek, B. 2001. A Runtime Composite Service Creation and Deployment Infrastructure and Its Applications in Internet Security, E-Commerce, and Software Provisioning. Proceedings of the 25th Annual International Computer Software and Applications Conference, Chicago, Illinois, IEEE Computer Society Press, pp. 371-376.

Nierstrasz, O., Meijler, T. D. 1994. Requirements for a Composition Language. ECOOP 1994 Workshop on Models and Languages for Coordination of Parallelism and Distribution, Springer Verlag, LNCS 924, pp. 147-161.

Nikolaou, C., Marazakis, M., Papadakis, D. 1998. Synthesis of Services in Open Commerce Environments. EUROMED-NET 98 Special Track Electonic Commerce on the Internet, Nicosia, Cyprus.

Pu, C., Massalin, H., Ioannidis, J. 1988. The Synthesis kernel. Computing Systems, 1(1):11–32.

Tennenhouse, D. L., Smith, J. M., Sincoskie, W. D., Wetherall, D. J., Minden, G. J. 1997. A Survey of Active Network Research. IEEE Communications Magazine, vol. 35, no. 1, pp. 80-86.

Thekkath, C., Levy, H. 1994. Hardware and software support for efficient exception handling. In Proceedings of the 6th International Conference on Architecture Support for Programming Languages and Operating Systems, pp. 110–121.

Tosic, V., Mennie, D., Pagurek, B. 2001. On Dynamic Service Composition and Its Applicability to E-Business Software Systems. In Proceedings of the Workshop on Object-Oriented Business Solutions at ECOOP 2001, Budapest, Hungary, pp. 95-108.

Tosic, V., Patel, K., Pagurek, B. 2001. E-Business Service Components with Multiple Classes of Service and Dynamic Adaptability Mechanisms. CITO Knowledge Network Conference 2001, Ottawa.

Tyugu, E., Uustalu, T. 1994. Higher-Order Functional Constraint Networks. In B. Mayoh *et al.*, editors, Constraint Programming, volume 131 of NATA ASI Series F, Springer-Verlag, Berlin, pp. 116-139.

Vivas Frontana, J. L. 2001. Dynamic Binding of Names in Calculi for Mobile Processes. Ph.D thesis, Department of Micoelectronics and Information Technology, Royal Institute of Technology, TRITA-IT AVH 01:02, ISSN 1403-5286, Stockholm.

Thread Migration Protocol in the Mozart System

PCC submission

Dragan Havelka[1], Seif Haridi[2]

[1]IMIT-KTH Isafjordsgatan 22, Electrum 229, 164 40 Kista, Sweden. E-mail: dragan@kth.se.

[2]IMIT-KTH Isafjordsgatan 22, Electrum 229, 164 40 Kista, Sweden. E-mail: seif@it.kth.se.

Abstract

The popularity of applications that exploit mobile and peer-to-peer interactions is growing. These applications are currently being built on platforms, created for the traditional client-server paradigm, that do not fully support the specific requirements of mobile and peer-to-peer interactions.

The Mozart programming system provides support for distributed decentralized and fault tolerant computation. Computation in Mozart is organized in concurrent stationary threads. Introducing the thread mobility in the Mozart system creates a perfect programming platform for development of mobile peer-to-peer applications.

Implementation of mobile threads in Mozart has two major milestones. The first milestone is to extend Mozart distribution subsystem with thread serialization (i.e. with routines that convert memory representation of a thread to a sequence of bits and vice versa). The second milestone is to design a mobility protocol that precisely describes the process of thread migration.

In this paper we present the mobile protocol for threads in the Mozart system.

1 Introduction

The last few years we have witnessed a growing interest in systems with support for decentralized computing and systems that include large number of components that are physically distributed and operate autonomously. The evolution of communication technology (e.g. wireless communication that offers new modes of exploiting the network) makes the situation even more complex. In more radical mobility settings such as *ad hoc networks*[1], users can completely bypass the fixed network structures. In these networks, communication is enabled at all levels by the mobile hosts. These types of environments are known as Peer-to-Peer environments.

The Mozart programming system provides a good programming platform for development of distributed decentralized fault tolerant applications. Transparent distribution (i.e. a program's behavior is independent of how it is partitioned between sites) overcomes the gap between centralized and distributed applications.

Despite of the good distribution platform that Mozart offers, still the system can be improved. One important modification is providing support for thread migration, which can improve the system in many aspects.

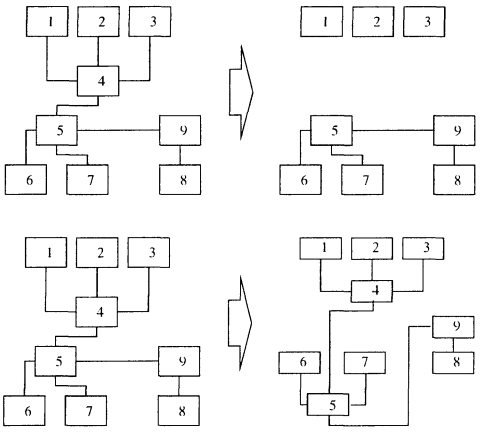


Figure 1: Ad hoc network (fragmentation and latency problems)

In the case of extreme mobile and decentralized systems (e.g. *ad hoc networks*) we can have a number of mobile devices connected to each other without the base stations. Each device (node) in the network can play the role of the base station (i.e. be a link between nodes). In Figure 1, we see that nodes 4 and 5 connect two subnetworks (1, 2, 3, 4) and (5, 6, 7, 8, 9). If node 4, due to battery failure, goes down, *network fragmentation* will occur. Another scenario is that node 5, due to its mobile nature, moves away from node 4. This would make node 5 inconvenient to continue be the link for nodes 6, 7, 8, and 9 which are now much closer to the node 4. A third scenario is when a new node with better resources joins the network. Reassigning the job of the link node, then, can

considerably improve the overall performance. Those exaples illustrate the need for importan

According to the Mozart concurrent programming model, the algorithm for a node acting as a base station in the given examples would be assigned to one thread of execution. By providing mobile threads in the system we introduce solutions for fragmentation and latency problems. In the example with battery failure, we can move *the base station thread* on the link node to the second best node, and in that way avoid *network fragmentation.* In the example where the link node moves out of distance from other nodes, *the base station thread* should move to the node with better network position. Moreover, thread mobility simplifies the process of network reconfiguration. We are aware that this is not the only way to solve the above-mentioned problems, but in this case, the solution is simple.

The thread migration can improve the system by:

- Making the system fault tolerant, by moving execution from nodes that are planed for upgrade or from nodes with bad network connection
- Helping the system to avoid network fragmentation
- Helping the system to perform network reconfiguration
- Moving the execution closer to the resource

Implementation of mobile threads can be divided in two major parts: The first part is to extend the distributed subsystem with serialization (marshaling) routines that transform a thread from memory representation to network representation and back. We have described the serialization model in [2]. The second part is to design a mobility protocol that precisely describes the process of thread migration which we are presenting in this paper.

The rest of the paper is structured as follows. Section 2 presents the mobility control that exists in the Mozart Programming System. It also discusses the issues why the existing model is not well adjusted to thread mobility. Section 3 describes the graph model and how it is used to bridge distributed semantics with the mobile state protocol. Section 4 explains how thread mobility fits in the distribution graph. Section 5 gives a short description of the failure model. Section 6 describes the thread mobile protocol. Section 7 gives some concluding remarks together with a short description of future work.

2 Mobility Control

In the Mozart system there are stateless and stateful entities. In the case of sateless entities mobility is trivially achieved via copying. On the other hand, stateful entities are classified as mobile or stationary where mobile entitites are managed through a mobility protocol.

Any stateful entity in the Mozart Programming System (e.g. cell, object, port, or thread) is characterized by its home site. The home site is the place where a stateful entity holds its state. For stationary entities (e.g. port, thread) the home site never changes, while for mobile entities (e.g. cell, object) the home site can change. It is important to mention at this point that in the presented work we are making threads mobile entities.

Mobility control defines what happens to the home site for each operation on the stateful entity. For example, invoking an assignment operation on a cell implies changing the cell's home site to the invoking site.

2.1 Implicit Mobility

Let us take a closer look on the mobility control of cells. As we previously mentioned, cells are mobile and can be accessed from many sites. Each of these sites knows a Mozart-assigned globally unique cell name, while only the home site contains the cell's state. Invoking the assigment operation from any site causes a synchronous move of the state to that site. This is done by using a mobile state protocol for cells. Invoking a cell-access operation does not move the state, but only the cell's content is copied to the invoking site.

From this model it is apparent that cells are freely mobile (i.e. a cell state will move to each site that updates the cell). The programmer is not aware that migration of the cell's state has occurred. He/she cannot see the difference between the assignment operation invoked on a local cell and a remote cell. This kind of mobility is called the implicit mobility.

2.2 Explicit Mobility

A differnet kind of mobility control is explicit mobility. In this case, an entity stays stationary unless explicitly moved. This means that we provide programmers with operations on entities that will explicitly initiate migration (e.g.{`Pull ThrStub T`}).

A thread as a language entity is special because a thread is a place where execution happens. We want that the programmer has complete control over and awareness of execution. The implicit migration does not fulfill these demands and brings up many issues (e.g. security issues). Hence, we find an explicit mobility protocol a better candidate for controlling thread mobility.

3 The Graph Model and Thread Mobility

To get a better understanding of the mobile state protocol we will use the framework described in [3]. In this framework, a graph model of the distributed execution of the Oz Programming Model (OPM) is used to bridge the distributed semantics with the mobile state protocol (cells, objects) and the implementation architecture. The OPM is a small kernel language [4] used to define Distributed Oz. Detailed information about the OPM and the graph model can be found in [3].

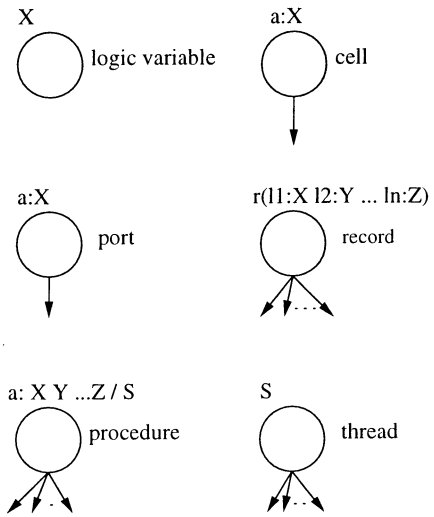


Figure 2: Nodes in the language graph

As described by the framework, an OPM computation can be represented in terms of two graphs:

- The language graph which represents the model of the OPM computation space. Each Oz entity corresponds to one node in the language graph (see Figure 2, 3(a)).
- The distribution graph which is an extension to the language graph where each node is annotated with a site. If a node is referenced by a node on another site, then we map it to a set of nodes. This set is called the access structure of the original node (see Figure 3(b)). Nodes that are only accessed locally (local nodes) do not have an access structure. An access structure consists of a set of nodes: One manager node on the site where the node is located, and one proxy node for each site that has a node reference.

4 Thread Mobility in the Distribution Graph

Now we can explain how thread mobility fits in the distribution graph model. A thread is an entity that could be referenced from remote sites. Passing thread references to remote sites results in the creation of an access structure (i.e. the creation of a manager node on the site where the thread resides and a proxy node on each remote site that has a reference to the thread). The manager has a list of all proxies. The proxies are not aware of each other and they only know the location of the manager node.

Thread migration is triggered by executing a `Pull` operation on a remote site that has a thread reference. Consequently, the proxy on this remote site requests thread migration from the manager. Depending on the manager state, immidiate or deferred migration is executed.

Execution of migration is performed as follows: First, the manager and the proxy of the migration-requesting site exchange roles. Second, all other proxies are updated with the address of the new manager.

All requests that reach the old manager during the migration process are forwarded to the new manager as the last step of the migration process.

Note that we have modified the current mobility protocol in Mozart, by introducing the notion of a mobile manager. In the original protocol the manager node is fixed to the site where the entity was created. This approach simplifies the implementation, but it introduces unwanted dependency on the site where the entity was created. That is something we absolutely want to avoid in the case of threads. The idea behind mobile threads is to build systems in a peer-to-peer fashion while the solution with stationary managers is more appropriate for the client-server model.

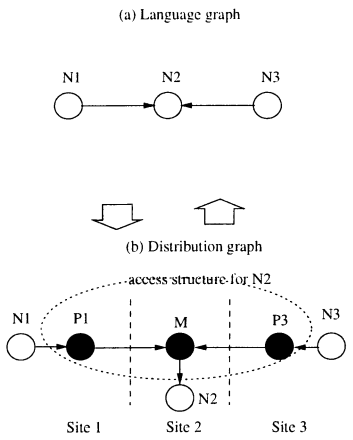


Figure 3: From language graph to distribution graph

5 Failure model

The failure model is still under discussion, but most likely it will have many similarities with failure model for cells and objects. It is important that the failure model reliably inform the programmer about failures and allow him/her to take necessary actions.

6 Description of Thread Mobile Protocol

The protocol is defined with respect to a single thread, that is accessible from a set of sites. Each of these sites that does not have thread's state has a proxy node responsible for part of the protocol for that site. In addition, there is a single manager node responsible for coordinating the proxy nodes, and it resides on the site that contains the state of the thread. These nodes together implement the distributed semantics of a thread. Thread proxies exchange messages with threads in the engine. To initialize migration from one of the proxy sites (explicitly by call to {`Pull ThreadRef Thr`}), an engine thread sends a message to a proxy and blocks waiting for a reply. After executing its protocol, the state of the thread will be on the site that has initialized migration. The proxy on that site will become the manager, and the manager will send a reply giving the thread reference to the engine thread that has called {`Pull ThreadRef Thr`} (i.e. Variable `Thr` will be bound to the reference of the migrated thread). The manager can receive several requests during the migration process. All these requests are saved and forwarded to the new manager.

6.1 Messages

The protocol uses the following nodes and messages. P_i denotes the addresses of proxies and M denotes the address of the manager in the distribution graph corresponding to the thread T. The messages are given in Table 1.

Receivers and messages		
Manager	Proxy	Engine Thread
get(P) **allInformed**()	**request**(E_T, T) **put**(T, LOP) **newMan**(M_n) **ackNewMan**(P) **end**(LOR, NP)	**proceed**(T)

Table 1: Migration protocol messages sorted by receivers

E_T is the reference to the engine thread that has initialized the migration. LOP is the list of all proxies. LOR is the list of queued migration requests and NP is the list of the new proxies that have arrived to old manager during the migration process.

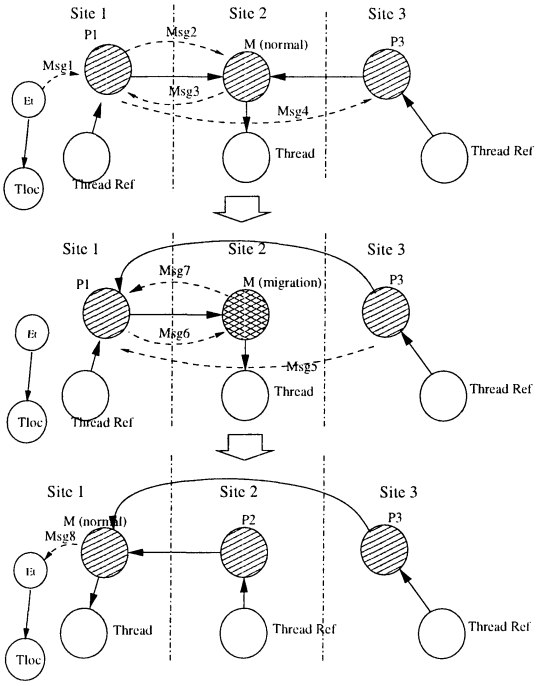


Figure 4: Pull initiates migration of thread

6.2 Outline of the protocol

The Protocol is defined using reduction rules of the form:

Condition
Action

A rule condition consists of a node state and a received message. A rule action consists of a sequence of operations on the node state with optional sending of messages.

See Figure 4 in conjuction with the following rules for a description of the protocol.

Msg1: request(E_T, T) *and* Proxy(*normal*).
This message is sent by one of the engines' threads E_T, as part of executing {`Pull` T T_{loc}}. Thread E_T blocks until a reply arrives. The proxy sends the message **get**(P) to the manager M.

Msg2: get(P) *and* Manager(*normal*) .
The manager changes its state from *normal* to *migration* and suspends the thread. It sends the serialized thread together with the list of all proxies back to the proxy P in the message **put**(T, LOP). If the manager state was *migration* when the message **get**(P) arrives, the message is just saved in the queue of requests that would be forwarded to the new manager.

Msg3: put(T, LOP) *and* Proxy(*normal*).
P has now a complete thread but it still did not become the manager. Proxy P sends to each proxy in the list the message **newMan**(P) to inform all about the address of the new manager, and waits for a reply from them. Proxies that do not reply under a given time are considered dead and are removed from the list of proxies.

Msg4: newMan(P) *and* Proxy(*normal*).
Proxy updates the address of the manager and sends an acknowledgment message **ackNewMan**(P).

Msg5: ackNewMan(P) *and* Proxy(*normal*).
Proxy updates the counter with received replies and if all proxies have replied, sends the message **allInformed**(P) to the manager.

Msg6: allInformed(P) *and* Manager(*migration*).
The manager sends the message **end**(LOR, NP) back to the proxy, invalidates the reference to the local thread and changes its state to a proxy.

Msg7: end(LOR, NP) *and* Proxy(*normal*).
The proxy becomes the manager and the thread received in **Msg3** becomes officially the thread state. The new created manager sends the message **proceed**(T) to the engine thread E_T.

Msg8: proceed(T) *and* Engine thread(*suspended*).
The engine thread invokes the binding of T_{loc} with T.

7 Conclusion and future work

In this paper we presented a mobile protocol for threads in the Mozart system. The protocol ensures that thread migration is done in a consistent way. Implementation work is still in progress and possible refinements of the protocol, such as a definition of the failure model, are under disscussion.

References

[1] Gianpaolo Cugola, Gian Pietro Picco. *PEER WARE: Core Middleware Support for Peer-to-Peer and Mobile Systems.* ESEC/FSE 2001

[2] Dragan Havelka, Seif Haridi 2001. *Thread migration in Mozart.* PCC Workshop Nynashamn, April 2001.

[3] Peter Van Roy, Seif Haridi, Per Brand, Gert Smolka, Michael Mehl, and Ralf Scheidhauer 1997. *Mobile Objects in Distributed Oz.* Available at http://www.mozart-oz.org.

[4] Gert Smolka 1995. *The Oz Programming Model.* In Computer Science Today. Lecture Notes in Computer Science, vol. 1000. Springer-Verlag, Berlin, 324-343.

A Tool for Real Time Simulations of TCP/IP over Wireless Fading Channels

Anna Ewerlid and Nilo Casimiro Ericsson
Signals and Systems Group
Uppsala University, Sweden
{Anna.Ewerlid, Nilo.Casimiro.Ericsson}@signal.uu.se

Abstract

We describe an ongoing effort to create a platform for realistic real-time simulations of wireless communication systems. The purpose of the work is primarily to evaluate the performance of existing and new approaches to provide wide area mobile and wireless Internet. The new approaches comprise of enhancements to transport protocols, and spectrally efficient allocation of the wireless resources.

The simulation tool is, just like the communication protocol stack, divided into several separate layers, each providing its functionality. It is also distributed over several computers that communicate over an ethernet.

The main contribution, that makes the real-time simulation possible, is provided by very heavy calculations and extensive off-line simulations of the lower layers in the communication stack. Through these simulations we obtain large tables of probabilities that are accessed on-line by the simulator running the higher layers.

1 Introduction

We are developing a simulation tool for wireless, mobile, wide area networks. The purpose of the simulation tool is to enable complex simulations of communication systems to be carried out in real-time. Ultimately, we want to investigate and verify the theoretical work and results that we have carried out in the Wireless IP (WIP) project within the Swedish research program for Personal Computing and Communications (PCC) [2].

Our results this far show that a high gain in spectral efficiency can be obtained using channel prediction and scheduling algorithms [7, 14]. Through channel prediction [5], we obtain estimates of future transmission quality over the channels between a base-station and several mobile terminals. These estimates are fundamental in the following step, the media access scheduling. The scheduler's task is to match the offered Internet traffic to the different channels, so that we utilize the scarce bandwidth as efficiently as possible. The scheduler should also take higher layer requirements into account, e.g. priorities for different users, and constraints imposed by the applications.

We would also like to evaluate the implications of our approach for the higher layers in the communications protocol stack: How will different flavors of TCP (Transmission Control Protocol) react on the offered wireless services [3, 11–13, 15]? The media access scheduler and the TCP protocol work on different time-scales, TCP in the scale of Round-Trip Times (RTT) and the scheduler in the scale of channel quality fluctuations, but are not necessarily independent of eachother.

At first, simulating this seems like a straight-forward problem to solve: We could generate error-patterns due to erroneous reception at the mobile client, then simply apply these patterns to the transmitted data. However, our complex link layer ARQ system together with the media access scheduler influenced the higher layers, i.e. the transport layer and the TCP protocol. This interplay between layers that operate on different time scales is crucial to the understanding and optimization of the whole protocol stack. This means that we can't simply do the link and transport layer simulations separately and independently.

In Section 2 we briefly describe the communication system that we primarily want to evaluate. In the following section we describe and discuss in more detail the problem of using layered simulations for this particular type of communications system. Finally, in Section 4, our solution to the simulation problem is proposed, and the paper is concluded in the following section.

2 Simulated System Background

This section briefly describes the system we intend to simulate, in order to evaluate its influence on the applications that generate the data traffic handled by the system. A central term in these simulations and system proposals is the presence of *fast fading channels*. Fast fading occurs when a radio receiver travels through areas with reflecting walls and other obstacles generating multi-path propagation of the transmitted radio signal, which consequently interferes with itself. When the interference is destructive, the radio signal fades out. The fast fading can give rise to fluctuations in the received signal power of more than $10dB$ within time spans of $1ms$, depending on the carrier frequency of the signal and the speed of the mobile terminal.

There is also a *slow fading* present in the mobile wireless channels. This is due to the distance variations between the base-station and the mobile terminals, i.e. the

path loss. In this slower time-scale we expect to see an interaction with the TCP flow control mechanisms.

2.1 Hybrid ARQ and Scheduling

To achieve a high system throughput also over fading channels, adaptive methods for the adjustment of the modulation alphabet and the coding complexity (the transmission rate) can be used.

One approach to achieve adaptation, is the Hybrid type-II ARQ (Automatic Repeat reQuest) scheme combined with an Adaptive Modulation System (AMS) [9], further described below. A rather different approach is based on prediction of the channel quality, thereby allowing for scheduling of the transmission among multiple users, taking the channel variations into account explicitly [6].

The idea behind HARQ-II/AMS is to reduce the transmission rate at each retransmission attempt. Therefore, more immunity against channel impairments can be provided in the following transmission attempts while still having the prospect of a high transmission rate in the first attempt. The drawback of this approach is the delay imposed by the method: The adaptation of coding rate/modulation is only controlled by the ACK/NAK (ACKnowledgement / Negative AcKnowledgement) feedback signal which increases the system delay in poor channel conditions.

In the scheduling approach, prediction of different user channels provide a basis for detailed scheduling of the transmission, by combining time-slot allocation and adaptive modulation. This approach can also take into account the desired error-probability and the priority associated with different users, as well as the current traffic situation. Moreover, the frequency band can be used efficiently, since the different users are allocated time-slots when their transmission conditions are predicted to be favorable, allowing them to use a high modulation level. The resulting constant and low (user/application-specified) error-rate provides the error correcting codes with manageable data, avoiding bandwidth consuming re-transmissions. The main drawback is the sensitivity to channel prediction errors.

An even more promising solution, according to preliminary investigations, would be a combination of the two methods [8]. The combined method successfully reduces the drawbacks of the individual methods described above: Knowledge of the channel conditions indicates the initial coding rate/modulation to use, reducing the delay due to numerous NAKs. The presence of the ARQ-scheme provides robustness against incomplete information in the scheduler, due to channel prediction errors.

Ordinary Internet traffic should be able to run on top of this type of system, since we don't aim at making any large changes to the widely accepted TCP/IP protocol stack. However, we would like to evaluate different flavors of the flow control algorithms within the TCP transport protocol, and their performance over our enhanced wireless links.

In Figure 1 we present an overview of the system we would like to simulate, and also qualitatively evaluate by running it in real-time, with real applications running on the end-points that communicate over the simulated links, using real transport protocols.

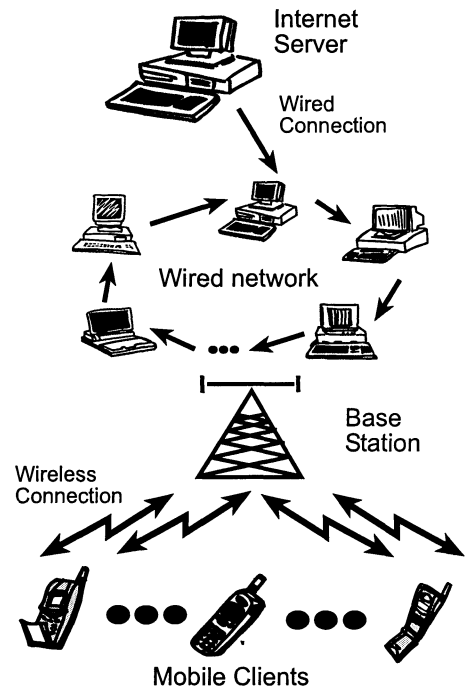


Figure 1. Overview of the communication system that we need to simulate.

3 The Simulation Problem

We need to run a realistic communication system *simulation* in real-time. At the lowest level, we need to simulate mobility of terminals in order to generate realistic fading radio channels. The next level contains the wireless link approach presented in the previous section. It comprises of a predictor for the fading channels, an advanced ARQ system using incremental redundancy and channel coding, and a media access scheduler. In the level above this, we have a number of transport protocols (TCP and UDP in different flavors), some of which have their own ARQ mechanisms that take action when things appear to go wrong. In this level we also have aggregation of traffic from different nodes in the network, destined for one of the mobile terminals under the control of that particular base station. Above that we have applications that utilize these underlying protocols for transmission of application and control information.

3.1 Physical Layer

In the physical layer we have to take into account the varying properties of the physical radio channel, properties that depend on a large amount of parameters, such as the speed at which the mobile terminal is moving, the environment in which it moves, and the interference from other neighboring transmitters. This type of simulation requires excessive calculations when a realistic outcome is desired. We use methods that involve ray-tracing of individual paths of the radio signals, in order to quantify the

interference at specific points in space (the points where the receiving mobile terminals are situated) [5].

We have previously been using very short recordings of channel sounding measurments, that is, a kind of pinging from a base station that is recorded by a mobile receiver. These measurments only allowed us to run the simulations for about $140ms$, which is much too short to see the consequences of a time-varying channel quality at the higher layers. To overcome this problem, channel data are generated by a sophisticated model for the wireless fading channels, a model that allows us to extract arbitrarily long sounding measurements through simulations. The measured data has been used to validate the channel model.

3.2 Link Layer

The link layer contains the functionality for error control and error protection of the data transmitted over the wireless link. It also takes care of the ARQ mechanism, which in our particular case is an incremental redundancy protocol that basically never retransmits the same data in case of a negative acknowledgement (NAK), but instead transmits a data chunk with additional redundant information to help the decoder at the receiver to correctly decode the message. The decoding of these types of codes (convolutional codes) is very processor and memory demanding, since they result in state machines with a very high number of possible states, out of which the most probable should be chosen.

One decoder is furthermore required for each mobile terminal that we wish to simulate. The computational complexity forbids us to run a complete set of decoders in real-time. Thus, the need to find a realistic approximation to the decoding and ARQ system.

3.3 Media Access Scheduler

On top of the link layer we need to run a media access scheduler that not only takes into account the demands from the queues in the receiving buffers on the wired network side, but also is aware of the channel conditions for each active link and makes efficient time-slot allocations based on this information. The scheduler runs a number of different algorithms for evaluation [7].

The decision of the scheduler is highly dependent of the instantaneous channel quality in a specific link between the base station and a mobile terminal. The scheduler may decide that access should be given to a user in a specific time-slot, based on both the predicted quality of the channel, and the amount of data that is queued up in the buffer for that user.

3.4 Our Layered Simulation

We found it appropriate for our current and future requirements, to divide the simulation process into three steps. Step one and two are carried out once and for all for a given system, whereas the third step, the system simulation, can be done over and over again:

1. Run channel simulator
   - Set up specifications (environment, mobile speed, carrier frequency, carrier bandwidth) for N channels
   - Generate sounding measurements and store in N files
2. Run link layer simulator
   - Set up link layer specifications (ARQ, adaptive modulation, error control coding)
   - Generate statistics for all possible and interesting outcomes from the link layer protocol and store in fast accessible data structures
3. Run overall real-time simulation
   - Set up applications (clients, services, TCP flavors) and scheduling algorithm
   - Use sounding measurements for scheduling
   - Use link layer statistics for each transmitted link layer frame
   - Run the clients over the simulated system

3.4.1 Channel Sounding Simulation

Based on advanced ray-tracing simulations of individual wireless links, where a mobile terminal travels through a "landscape" of electromagnetic wave diffracting objects, we are able to generate arbitrarily long channel sounding measurements. The transmitted signal is subject to self-interference, path loss, and measurement noise, making the "measurement" as realistic as possible. The resulting channel sounding measurements have been validated with respect to all the important statistics, using the real measurements mentioned above.

The sounding data will be used in three ways:

- In the scheduler, where it is used as the predicted channel quality. The predictions can not be assumed to be accurate, so a prediction error is added to the "real" channel quality. The prediction error distribution has recently been found, and will be introduced to the simulation model.

- In the decoder at the receiver, where it is used as an estimation of the instantaneous channel quality. This estimation is also assumed to be inaccurate, so an estimation error is added to the "real" value. This error is smaller than the prediction error.

- In the transmission simulation, where it is used as the "actual" channel quality, that gives rise to a certain distortion of the received data.

3.4.2 Hybrid ARQ Simulation

In Section 2.1 we described the Hybrid ARQ and Scheduling used for the link layer transmissions. Here we describe how we simulate this layer off-line, in order to speed up the overall system simulation.

The procedure is best described by outlining an example from our system:

A link layer transmission frame consists of 48 data-carrying time-slots, each of which can be allocated to a

different user (or mobile terminal). A time-slot can contain two or more link layer packets, depending on the modulation level (numbers of bits/waveform). Each link layer packet is encoded with an error correcting (outer) code and an error detecting (inner) code. The transmission rate (a combination of the modulation level, and the coding rate) is decided by the scheduler, depending on the predicted channel quality. The actual coding rate is decided by the ARQ-system, depending on which of up to three transmission attempts that is performed, since the coding rates need to be taken in a specific order.

Up to three transmission attempts are performed for a link layer packet before giving up, and each transmission attempt will depend on the previous one, since the Hybrid ARQ system uses all the previously received versions of the (erroneous) packet in the decoding process. This will generate a tree of depth, $d = 3$, and width according to all the possible combinations of the parameters *channel quality, coding rate, and modulation level.* The coding rate and modulation level are bounded discrete values, but the channel quality is a bounded continuous value. Therefore we need to discretizise the channel quality into a limited number of intervals.

The statistics from the link layer simulations is generated by simulating a large number of link layer packet transmissions for each of the combinations of the parameters presented above. Most of the combinations will result in a successful transmission for all the packets, thus terminating that branch of the tree at the first level. Some combinations will however cause a number of erroneous transmissions, issuing a second attempt by the ARQ system. For each combination of the transmission parameters, the probability of erroneous reception is approximated by the ratio between the number of erroneous packets and the total number of packets transmitted.

This procedure is repeated until all three transmission attempts have been performed, collecting the probabilities for erroneous reception in each node in the tree. In the case that very few packets are subject to a second or third attempt for a combination of the transmission parameters, the second and third transmission attempts are repeated multiple times with the same incrementally redundant packet that, after transmission, will be combined with the previously erroneously received packets.

This simulation process will result in a large combinatorial tree, requiring a fast lookup procedure for each combination of the transmission parameters and each transmission attempt.

4 Overview of the Simulator

The simulator is written in C++, under the Linux operating system. Large parts of the code consists of a library of routines for mathematics, communications, and signal processing, developed at the Department of Signals and Systems, Chalmers University of Technology [1].

We are aiming at making as realistic simulations as possible of the communication system that we want to investigate, including all the layers in the protocol stack. The simulation setup consists of four different machines, connected via Ethernet, that run different parts of the simulation, see Figure 2. Two of the machines are designated for running multiple client and server applications. They represent the communicating end-points in the simulated network. One machine, called the *Bridge*, takes care of the delaying and dropping of packets, according to the outcome from the fourth computer, the *Simulator*.

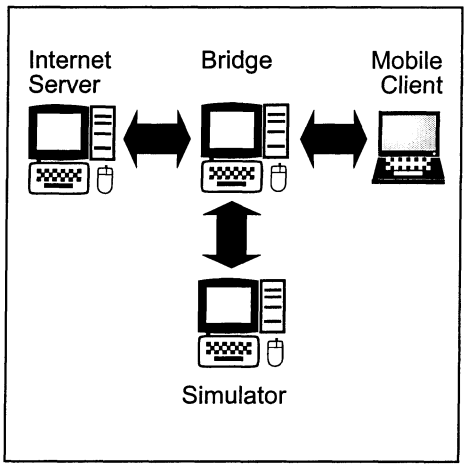


Figure 2. The simulation setup consists of four computers that communicate over an Ethernet.

It is only the higher layers, from the network and upwards, that are simulated in real time. The lower layers, that is the link and physical layers, have been simulated in advance, generating large databases that are used by the simulation of the higher layers. The reason for this partitioning became evident in the light of the computation requirements that we described in the previous section.

The data traffic that arrives at the simulated base station from the wired network is brought up to the network layer (IP) in order to separate the different source-destination pairs. We also to peek into the IP header to extract the type-of-service octet, that tells about different requirements on delay, throughput, and reliability.

The functions described above, are all simulated in the *Simulator* computer in Figure 2. The simulator thus simulates everything that happens from the network layer and down, in the base station, over the wireless link, and in each receiving mobile terminal.

Next, we describe the higher layers in the protocol stack. They run on the end-points of the simulated system, which in this case run on separate machines, and are referred to as *Mobile Client* and *Internet Server* in Figure 2.

4.1 Description of the Simulation Environment

An overview of the simulation environment is presented in Figure 2. The system consists of 4 dual-processor computers that are running Linux RedHat 7.2 with kernel 2.4.13.

One of the important parameters for the scheduling is the type of service. The system is designed to handle packets of different kinds of connections in different ways. To evaluate how they are handled by our system,

we should simulate various kinds of services (telnet, FTP, HTTP, etc.). The connections are initiated by the Mobile Client. It establishes a UDP or TCP connection with the Internet Server. We place different versions of TCP (Westwood, Reno, Vegas, etc.) on them for comparison. Moreover, we can enable various end-to-end improvements, such as SACK (Selective ACKnowledgement), ECN (Explicit Congestion Notification), and ELN (Explicit Loss Notification), etc.

Packets are trapped in the machine referred to as the Bridge, the book-keeping machine, which places them in a holding-queue. The arrival of a packet results in the transmission of a UDP request to the Simulation machine, with information about the arrived packet. After the simulated transmission is finished (succesfully or not), a response will come, telling the Bridge to either release the packet for delivery to the destination, or to drop it as a consequence of an unsuccessful transmission over the wireless link.

The Bridge provides a connection between the Client and the Server in such a way that the two machines are unaware of any special arrangements made between them. TCP and UDP packets are lifted to the application layer and placed in a queue until a response comes from the Simulation machine, and all other kinds of packets (e.g. ARP, ICMP packets) are forwarded directly to the destination machine.

The fourth machine, referred to as the Simulator, simulates the wireless links, according to the description in the previous section. The Simulation software is organized in two processes, one for the actual simulation, and one for maintaining the connection with the Bridge machine. This structure allows us to perform the tasks simultaneously on the separate CPUs.

The Communication process is responsible for the contact with the Bridge machine. Whenever the Bridge machine gets a TCP or UDP packet, it places it into the holding-queue and sends a request to the Simulator, containing information about the packet (packet size, type of service, etc.) The Communication process of the Simulator receives the request and places a dummy packet into the appropriate queue in the network layer buffer.

Based on the contents of the queues and the predicted channel quality, the scheduler in the Computation process, will generate decisions that will be executed by the link layer. The link layer will follow the scheduler's decision, but give precedence to link layer packets that were NAKed in the previous transmission attempt. In the case that all the transmission attempts failed, the Simulator will tell the Bridge to drop the specific packet. In the case that an IP packet is successfully transmitted, the Simulator will tell the Bridge to release the packet.

5 Conclusions

Simulation of wireless Internet connections in real-time is an extremly complex and challenging task. We are creating a system that will allow us to simulate the end-to-end connection between the Mobile Client and the Internet Server. The assumed wireless link uses advanced algorithms for channel prediction, scheduling, and error correction. The real-time simulation system described in the article is going to be used for study of the implications of all these algorithms on the TCP/IP traffic. Moreover, we are going to evaluate various proposals, that have come over the years, for the improvment of TCP/IP performance over wireless links [4, 10, 12].

Appropriate performance measures are throughput, latency statistics and average data rates over wireless links. More complete results of the simulations will be presented at the conference.

References

[1] IT++: A Communication Simulation Library for C++. http://www.s2.chalmers.se/software/index.html.

[2] The Wireless IP Project within the SSF PCC (Personal Computing and Commnication) Program. http://www.signal.uu.se/Research/PCCwirelessIP.html.

[3] M. Allman, V. Paxson, and W. Stevens. TCP Congestion Control. RFC2581, Apr. 1999.

[4] E. Amir, H. Balakrishnan, S. Seshan, and R. H. Katz. Efficient TCP over networks with wireless links. In *HoTOS*, pages 35–40, 1995.

[5] T. Ekman. *Prediction of Mobile Radio Channels.* Licentiate Thesis, Uppsala University, Dec. 2000.

[6] N. C. Ericsson. Adaptive Modulation and Scheduling of IP traffic over Fading Channels. In *Proc. IEEE Vehicular Technology Conference*, volume 2, pages 849–853, Amsterdam, the Netherlands, Sept. 1999.

[7] N. C. Ericsson. On Scheduling and Adaptive Modulation in Wireless Communications. Licentiate Thesis, June 2001.

[8] S. Falahati and N. C. Ericsson. Hybrid type-II ARQ/AMS and Scheduling using Channel Prediction for Downlink Packet Transmission on Fading Channels. In *Proc. PCC Workshop*, Nynäshamn, Sweden, Apr. 2001.

[9] S. Falahati and A. Svensson. Hybrid type-II ARQ Schemes with Adaptive Modulation System for Wireless Channels. In *Proc. IEEE Vehicular Technology Conference*, volume 5, pages 2691–2695, Amsterdam, the Netherlands, Sept. 1999.

[10] K. Fall and S. Floyd. Simulation-based Comparisons of Tahoe, Reno, and SACK TCP. Technical report, Lawrence Berkeley National Laboratory, Dec. 1995. ftp://ftp.ee.lbl.gov/papers/sacks.ps.Z.

[11] S. Mascolo. Smith's principle for congestion control in high-speed data networks. *IEEE Transactions on Automatic Control*, 45(2):358–364, Feb. 2000.

[12] S. Mascolo, C. Casetti, M. Gerla, S. S. Lee, and M. Sanadidi. TCP Westwood: congestion control with faster recovery. Technical Report 200017, Computer Science Department, UCLA, Los Angeles, CA, 2000.

[13] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanov. TCP Selective Acknowledgement Options. RFC2018, Oct. 1996.

[14] T. Ottosson, M. Sternad, A. Ahlén, A. Svensson, and A. Brunström. Toward a 4G IP-based Wireless System Protocol. In *RVK 2002*, Kista, Sweden, June 2002.

[15] G. Wu, Y. Bai, J. Lai, and A. Ogielski. Interactions between TCP and RLP in Wireless Internet. In *GLOBECOM*, pages 661–666, Rio de Janeiro, Brazil, December 1999.

Effect of Channel Prediction Errors on Adaptive Modulation Systems for Wireless Channels

Sorour Falahati[1], Arne Svensson[1], Torbjörn Ekman[2] and Mikael Sternad[2]

[1]Department of Signals and Systems, Chalmers University of Technology, SE-412 96 Göteborg, Sweden.

Email: {Sorour.Falahati, Arne.Svensson}@s2.chalmers.se

[2]Signals and Systems, Department of Material Science, Uppsala University, SE-751 20 Uppsala, Sweden.

Email: {Torbjorn.Ekman, Mikael.Sternad}@signal.uu.se

Abstract

We study the optimum design of an adaptive modulation scheme based on M-QAM modulation assisted by channel prediction for the flat Rayleigh fading channel. The data rate and transmit power are adapted to maximize the spectral efficiency, subject to average power and instantaneous BER constraints. Optimum solutions for the rate and transmit power are derived based on the predicted SNR as well as the prediction error variance which are assumed to be available at the transmitter. The analytical results are evaluated and presented.

1 Introduction

Adaptive modulation, or link adaptation, is a powerful technique for improving the spectral efficiency in wireless transmission over fading channels. The modulation parameters, such as signal constellation size, transmitted power level and data rate are here adjusted according to the channel conditions. The adaptation can also take requirements of different traffic classes and services such as required bit error rates, into account. In the case of *fast* link adaptation considered here, we strive to adapt to the small scale fading. The receiver estimates the received power and sends feedback information via a return channel to the transmitter, which adjusts its modulation parameters. Due to the unavoidable delays involved in power estimation, feedback transmission and modulation adjustment, the adaptation needs to be based on *predicted* estimates of the power of the fading communication channel.

The optimum design of adaptive modulation systems is extensively studied and thoroughly covered in the literature (e.g. see [1], [2], [3]). However, in the proposed solutions, perfect knowledge of the channel conditions at the transmitter as well as error free channel estimates at the receiver are common assumptions for the system design and performance evaluation. In real systems, these assumptions are not valid. This leads to performance degradation such as decrease in the throughput and increase in the delay as well as failure in providing the required service such as the expected bit error rate. Therefore, a solution based on a more realistic assumption is of great importance and interest.

In this paper, we intend to design an optimum adaptive modulation system for a given channel prediction error variance with a corresponding statistical model, which maximizes the spectral efficiency while satisfying a BER requirement. The proposed system utilizes a linear filter to predict the channel quality at the receiver [4], [5]. Moreover a statistical model which reasonably well describes the properties of prediction errors is investigated. The model gives a good estimate of the mean, variance and probability density function of the errors. Then, the statistical model is taken into account for performance analysis of the adaptive modulation schemes such as for example, the Bit Error Rate (BER) versus Signal-to-Noise Ratio (SNR) for given prediction error variances.

This paper is organized as follows. Section 2 describes the system model and the notations which are used throughout this study. The channel prediction is explained in Section 3. The BER formula as a function of predicted instantaneous SNR is evaluated in Section 4 and the optimal rate and power adaptation are derived in Section 5. Analytical results are presented in Section 6 and finally, some conclusions are drawn in Section 7.

2 System Model

In the adaptive modulation scheme, M-QAM modulation schemes with different constellation sizes are provided at the transmitter. For each transmission, the modulation scheme and transmit power are chosen to maximize the spectral efficiency for the instantaneous BER and average power constraints based on the instantaneous predicted SNR. The channel is modeled

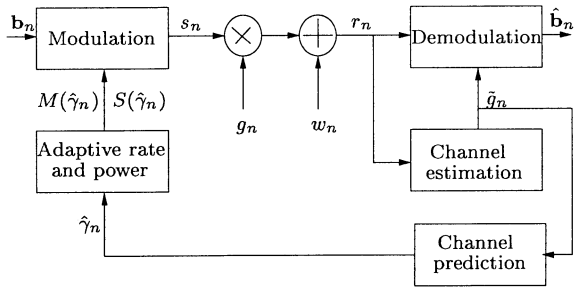


Figure 1: Discrete model of the system.

by a flat Rayleigh fading. At the receiver, demodulation is performed using the channel estimation. The discrete model of the system is depicted in Figure 1. All the signals are sampled at the symbol rate where the index n represents the signal sample at time nT_s where T_s is the symbol period. g_n is the complex channel gain with uniformly distributed phase within $[0, 2\pi)$ and Rayleigh distributed amplitude with pdf (probability density function) given by

$$f_{|g|}(|g_n|) = \frac{|g_n|}{\sigma_g^2/2} \exp\Big(-\frac{|g_n|^2}{\sigma_g^2}\Big) \qquad (1)$$

where σ_g^2 is the average power of the fading process. Moreover, the auto-correlation function of the complex channel gain is given by

$$r_g(m) = E(g_n g_{n-m}^*) = \sigma_g^2 J_0(2\pi f_d T_s m) \qquad (2)$$

where $E(\ )$ is the Expectation function, the asterisk denotes the conjugate operator, $J_0(\ )$ is the Bessel function of order zero and f_d is the maximum Doppler frequency[1]. w_n is a sample of the complex AWGN with zero mean and time-invariant variance σ_w^2. y_n is the noisy observation of g_n at the receiver which is used by an FIR filter at the receiver to predict the instantaneous received SNR denoted by $\hat{\gamma}_n$ where an error free feed-back channel is assumed. Furthermore, the feed-back transmission delay is presumed to be taken up by the predictor since the prediction is performed beyond the corresponding delay.

Based on $\hat{\gamma}_n$, the modulation scheme with constellation size $M(\hat{\gamma}_n)$ (out of N constellations available at the transmitter) which transmits $k(\hat{\gamma}_n) = \log_2 M(\hat{\gamma}_n)$ bits per symbol, and the transmit power $S(\hat{\gamma}_n)$ are selected. Each block of $k(\hat{\gamma}_n)$ data bits denoted by $\mathbf{b}_n$, is Gray encoded and mapped to s_n which is a symbol in the signal constellation and is transmitted over the flat Rayleigh fading channel. The received sample, r_n, is used to estimate the channel gain $\tilde{g}_n$ which is used to demodulate r_n to detect the transmitted bits denoted by $\hat{\mathbf{b}}_n$. Since the estimation error is believed to have a minor effect on the performance compared to the prediction error, perfect channel estimation is assumed in the demodulation here, i.e. $\tilde{g}_n = g_n$. In future extension of this work, we however hope to be able to include estimation error to the model. The SNR prediction $\hat{\gamma}_n$ is here assumed to be based on noisy channel estimates $\tilde{g}_n$ and will therefore have a prediction error, with variance σ_ϵ^2.

In this study, the following notations similar to [1] are used. Let $\bar{S}$ denote the average transmit signal power and $\bar{\gamma} = \bar{S}/\sigma_w^2$ denote the average received SNR for the channel gain with normalized average power ($\sigma_g^2 = 1$). For a constant transmit power $\bar{S}$, the instantaneous received SNR is $\gamma_n = \bar{\gamma} p_n$ where $p_n = |g_n|^2$ is the instantaneous channel power gain. Accordingly, the instantaneous predicted received SNR is $\hat{\gamma}_n = \bar{\gamma}\hat{p}_n$ where $\hat{p}_n = |\hat{g}_n|^2$ is the predicted instantaneous channel power gain. For the transmit power $S(\hat{\gamma}_n)$, the instantaneous received SNR is given by $\gamma_n(S(\hat{\gamma}_n)/\bar{S})$. Also, the rate region boundaries are denoted by $\{\hat{\gamma}_i\}_{i=0}^{N-1}$, defined as the ranges of $\hat{\gamma}_n$ values over which the different constellations are used by the transmiter, where $\hat{\gamma}_N = \infty$. When the predicted instantaneous SNR belongs to a given rate region, i.e. $\hat{\gamma}_n \in [\hat{\gamma}_i, \hat{\gamma}_{i+1})$, the corresponding constellation of size $M(\hat{\gamma}_n) = M_i$ with $k(\hat{\gamma}_n) = k_i$ bits per symbol is transmitted. Finally, there is no transmission if $\hat{\gamma}_n < \hat{\gamma}_0$ meaning that $\hat{\gamma}_0$ is the cutoff SNR.

3 Channel Prediction

The channel gain in a Rayleigh fading mobile radio channel takes values from a complex time series which can be modeled as a correlated complex Gaussian stochastic process. The absolute square, i.e. the power, of the time series is predicted based on linear regression of the observations of the complex time series. The unbiased quadratic predictor that is optimal in the Mean Square Error (MSE) sense is derived in [4] where it is shown that the same prediction coefficients that are optimal for the linear prediction of the complex tap, are optimal for the quadratic unbiased predictor.

The complex channel gain is observed in noise as

$$y_n = g_n + e_n \qquad (3)$$

where e_n is assumed to be a complex Gaussian random variable which is independent from g_n, has zero mean and auto-correlation given by

$$r_e(m) = E(e_n e_{n-m}^*) = \sigma_e^2 \delta(m). \qquad (4)$$

where $\delta(\ )$ is the Kronecker delta function.

From the past and present observations, the power of the signal at time $n + L$, i.e. $p_{n+L} = |g_{n+L}|^2$, is predicted by an unbiased FIR predictor. In the vector

[1] In the following r_g is equivalently used instead of $r_g(0)$.

formulation, the unbiased predicted power based on the past noisy observations is given by

$$\hat{p}_{(n+L|n)} = \theta^H \varphi_n \varphi_n^H \theta + r_g - \theta^H \mathbf{R}_\varphi \theta \quad (5)$$

where

$$\varphi_n = [y_n, y_{n-1}, \cdots, y_{n-(K-1)}]^H, \quad (6)$$

is the regressor, K is the number of filter coefficients, θ is a column vector containing the complex filter coefficients, H is the Hermitian operator and $\mathbf{R}_\varphi$ is the $K \times K$ correlation matrix for the regressors. $\mathbf{R}_\varphi$ is given by $\mathbf{R}_\varphi = \mathbf{R}_g + \mathbf{R}_e$ where

$$\mathbf{R}_g(n,m) = r_g(|n-m|) \quad \text{for } n,m = 1, \cdots, K \quad (7)$$

and

$$\mathbf{R}_e = \sigma_e^2 \mathbf{I}_{K \times K} \quad (8)$$

where $\mathbf{I}$ is the identity matrix. Moreover, (5) corresponds to the absolute square of the complex prediction, but with a bias compensation added to it to provide an unbiased power estimate, i.e.

$$\hat{p}_{(n+L|n)} = |\hat{g}_{(n+L|n)}|^2 + r_g - \theta^H \mathbf{R}_\varphi \theta. \quad (9)$$

The miminum MSEs of the channel gain prediction error (i.e. $\epsilon_{c_n} = g_n - \hat{g}_{(n|n-L)}$) and the power prediction error (i.e. $\epsilon_{p_n} = p_n - \hat{p}_{(n|n-L)}$) are given by

$$\sigma_{\epsilon_c}^2 = r_g - \mathbf{r}_{g\varphi}^H \mathbf{R}_\varphi^{-1} \mathbf{r}_{g\varphi} \quad (10)$$

and

$$\sigma_{\epsilon_p}^2 = r_g^2 - |\mathbf{r}_{g\varphi}^H \mathbf{R}_\varphi^{-1} \mathbf{r}_{g\varphi}|^2, \quad (11)$$

respectively, where $\mathbf{r}_{g\varphi}$ is the cross-correlation between the signal and the regressor given by

$$\begin{aligned} \mathbf{r}_{g\varphi} &= \mathrm{E}\{g_n \varphi_{n-L}\} \\ &= [r_g(L), r_g(L+1), \cdots, r_g(L+(K-1))]^T. \end{aligned} \quad (12)$$

Moreover, the filter coefficient vector θ, that minimizes the corresponding error variances is given by

$$\theta = \mathbf{R}_\varphi^{-1} \mathbf{r}_{g\varphi}. \quad (13)$$

Also (10) and (11) are related through

$$\sigma_{\epsilon_p}^2 = \sigma_{\epsilon_c}^2 (2r_g - \sigma_{\epsilon_c}^2). \quad (14)$$

In order to solve the optimization problem, the pdf of the instantaneous SNR is required. Since g_n is a complex Gaussian random variable, both the true and predicted powers, i.e. $p_n = |g_n|^2$ and $\hat{p}_n = |\hat{g}_n|^2$ are χ^2 distributed. In [5], it is shown that the pdf of γ conditioned on $\hat{\gamma}$ is given by

$$\begin{aligned} f_\gamma(\gamma|\hat{\gamma}) = &\frac{\mathrm{U}(\gamma)\mathrm{U}(\hat{\gamma} - \bar{\gamma}\sigma_{\epsilon_c}^2)}{\bar{\gamma}\sigma_{\epsilon_c}^2} \exp\left(-\frac{\gamma + \hat{\gamma} - \bar{\gamma}\sigma_{\epsilon_c}^2}{\bar{\gamma}\sigma_{\epsilon_c}^2}\right) \times \\ &\mathrm{I}_0\left(\frac{2}{\bar{\gamma}\sigma_{\epsilon_c}^2}\sqrt{\gamma(\hat{\gamma} - \bar{\gamma}\sigma_{\epsilon_c}^2)}\right), \end{aligned} \quad (15)$$

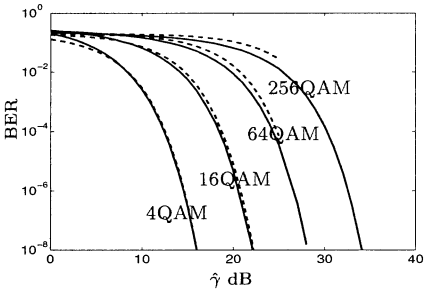


Figure 2: BER versus instantaneous predicted SNR of M-QAM schemes for $\bar{\gamma} = 30$ dB and $\sigma_{\epsilon_p}^2 = 0.001$. The solid lines and dashed lines correspond to (19) based on (17) and (18), respectively.

where U( ) is the Heaviside's step function and I_0( ) is the zeroth order modified Bessel function. The pdf of $\hat{\gamma}$ is given by

$$f_{\hat{\gamma}}(\hat{\gamma}) = \frac{\mathrm{U}(\hat{\gamma} - \bar{\gamma}\sigma_{\epsilon_c}^2)}{\bar{\gamma}(r_g - \sigma_{\epsilon_c}^2)} \exp\left(-\frac{\hat{\gamma} - \bar{\gamma}\sigma_{\epsilon_c}^2}{\bar{\gamma}(r_g - \sigma_{\epsilon_c}^2)}\right). \quad (16)$$

The index n is dropped in the pdf expressions since g_n and $\hat{g}_n$ are both stationary random processes.

4 M-QAM BER Performance

The transmitter adjusts the constellation size and the transmit power based on the instantaneous predicted SNR $\hat{\gamma}$. Evaluation of the optimal power and constellation size (or rate) which maximize the spectral efficieny and satisfy the BER requirement, requires the BER formula as a function of $\hat{\gamma}$. Assuming a square M-QAM with Gray encoded bits, constellation size $M(\hat{\gamma})$, and transmit power $S(\hat{\gamma})$, the instantaneous BER as a function of γ and $\hat{\gamma}$ on an AWGN channel, is approximated by [6]

$$\begin{aligned} \mathrm{BER}(\gamma, \hat{\gamma}) \approx &\frac{2}{\log_2 M(\hat{\gamma})}\left(1 - \frac{1}{\sqrt{M(\hat{\gamma})}}\right) \times \\ &\mathrm{erfc}\left(\sqrt{1.5 \frac{\gamma \frac{S(\hat{\gamma})}{\bar{S}}}{M(\hat{\gamma}) - 1}}\right) \end{aligned} \quad (17)$$

which is tight for high SNRs. In [1], it is shown that (17) can be furhter approximated as

$$\mathrm{BER}(\gamma, \hat{\gamma}) \approx 0.2 \exp\left(\frac{-1.6\gamma}{M(\hat{\gamma}) - 1} \frac{S(\hat{\gamma})}{\bar{S}}\right) \quad (18)$$

which is tight within 1 dB for $M(\hat{\gamma}) \geq 4$ and BER$\leq 10^{-3}$. By averaging (18) over the whole range of the instantaneous true SNR γ, the instantaneous BER as a function of the instantaneous predicted SNR is

obtained as

$$\begin{aligned}\mathrm{BER}(\hat{\gamma}) &= \int_0^{\infty} \mathrm{BER}(\gamma,\hat{\gamma}) f_{\gamma}(\gamma|\hat{\gamma}) d\gamma \\ &\approx \frac{a(\hat{\gamma})}{b(\hat{\gamma})} \exp\left(\frac{c(\hat{\gamma})^2}{4b(\hat{\gamma})}\right), \quad \hat{\gamma} \geq \bar{\gamma}\sigma_{\epsilon_c}^2\end{aligned} \tag{19}$$

where

$$a(\hat{\gamma}) = \frac{0.2}{\bar{\gamma}\sigma_{\epsilon_c}^2} \exp\left(-\frac{\hat{\gamma}-\bar{\gamma}\sigma_{\epsilon_c}^2}{\bar{\gamma}\sigma_{\epsilon_c}^2}\right), \tag{20}$$

$$b(\hat{\gamma}) = \frac{1}{\bar{\gamma}\sigma_{\epsilon_c}^2} + \frac{1.6}{M(\hat{\gamma})-1}\frac{S(\hat{\gamma})}{\bar{S}}, \tag{21}$$

and

$$c(\hat{\gamma}) = \frac{2}{\bar{\gamma}\sigma_{\epsilon_c}^2}\sqrt{\hat{\gamma}-\bar{\gamma}\sigma_{\epsilon_c}^2}. \tag{22}$$

In Figure 2, the instantaneous BER is illustrated for the constant transmit power $S(\hat{\gamma}) = \bar{S}$ and different constellation sizes where both (17) and (18) are used in (19). It is shown that the results based on the approximation given by (18) are tight enough.

5 Optimal Rate and Power Adaptation

The spectral efficiency of a modulation scheme is given by the average data rate per unit bandwidth (R/B) where R is the data rate and B is the transmitted signal bandwidth. When a modulation with constellation size $M(\hat{\gamma})$ is chosen, the instanteneous data rate is $k(\hat{\gamma})/T_s$ (bps). Assuming the Nyquist data pulses $(B = 1/T_s)$, the spectral efficiency is given by

$$\eta_B = \frac{R}{B} = \sum_{i=0}^{N-1} k_i \int_{\hat{\gamma}_i}^{\hat{\gamma}_{i+1}} f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma} \quad \text{bps/Hz} \tag{23}$$

We would like to maximize the spectral efficiency subject to the average transmit power constraint as

$$\int_0^{\infty} S(\hat{\gamma}) f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma} \leq \bar{S} \tag{24}$$

and instantaneous BER constraint given by

$$\mathrm{BER}(\hat{\gamma}) = P_b. \tag{25}$$

Under a BER constraint, $S(\hat{\gamma})$ can be evaluated in terms of $\hat{\gamma}$ and $M(\hat{\gamma})$ as the following: Let us define

$$x = \frac{1}{\bar{\gamma}\sigma_{\epsilon_c}^2 b(\hat{\gamma})}, \quad 0 \leq x \leq 1 \tag{26}$$

and

$$y = \frac{\hat{\gamma}}{\bar{\gamma}\sigma_{\epsilon_c}^2}, \quad y \geq 1 \tag{27}$$

and rewrite (25) as

$$0.2xe^{(1-y)(1-x)} = P_b. \tag{28}$$

Taking the natural logarithm of (28) and then using the Taylor approximation $\ln x \approx x - 1$ about $x = 1$, we obtain

$$x \approx 1 + \frac{1}{y}\ln\left(\frac{P_b}{0.2}\right) \tag{29}$$

and consequently

$$\begin{aligned}S(\hat{\gamma}) &\approx h(\hat{\gamma}, M(\hat{\gamma})) \\ &= \frac{-(M(\hat{\gamma})-1)}{1.6}\frac{\bar{S}\ln\left(\frac{P_b}{0.2}\right)}{\hat{\gamma}+\bar{\gamma}\sigma_{\epsilon_c}^2\ln\left(\frac{P_b}{0.2}\right)}\mathrm{U}\left(\hat{\gamma}-\bar{\gamma}\sigma_{\epsilon_c}^2\right).\end{aligned} \tag{30}$$

As explained in Section 2, a signal constellation and consequently a transmission rate is assigned to each predicted SNR region boundary. Therefore, for a given number of signal constellations, the optimum predicted SNR region boundary for each constellation corresponds to the optimum transmission rate, accordingly. Moreover, the transmit power can be obtained from the instantaneous predicted SNR based on (30). Therefore, the optimization problem can be simplified to finding only the optimal region boundaries. The Lagrangian function is given by

$$\begin{aligned}J(\hat{\gamma}_0,\hat{\gamma}_1,\cdots,\hat{\gamma}_{N-1}) &= \sum_{i=0}^{N-1} k_i \int_{\hat{\gamma}_i}^{\hat{\gamma}_{i+1}} f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma} \\ &+ \lambda\Big(\sum_{i=0}^{N-1}\int_{\hat{\gamma}_i}^{\hat{\gamma}_{i+1}} h(\hat{\gamma},M_i) f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma} - \bar{S}\Big)\end{aligned} \tag{31}$$

where the optimal region boundaries are found by solving

$$\frac{\partial J}{\partial \hat{\gamma}_i} = 0, \quad 0 \leq i \leq N-1 \tag{32}$$

which results in

$$h(\hat{\gamma}_i, M_{i-1}) - h(\hat{\gamma}_i, M_i) = -\frac{k_{i-1}-k_i}{\lambda} \tag{33}$$

where $k_{-1} = 0$ and $M_{-1} = 1$ are assumed. Taking into account (30) and (33), the optimum region boundaries are obtained by

$$\hat{\gamma}_i = \ln\left(\frac{P_b}{0.2}\right)\left(\frac{\bar{S}}{1.6}\frac{M_{i-1}-M_i}{k_{i-1}-k_i}\lambda - \bar{\gamma}\sigma_{\epsilon_c}^2\right) \tag{34}$$

where the Lagrangian multiplier $\lambda \neq 0$ is numerically evaluated based on the power constraint given as

$$\begin{aligned}&\sum_{i=0}^{N-1}\int_{\hat{\gamma}_i}^{\hat{\gamma}_{i+1}} h(\hat{\gamma},M_i) f_{\hat{\gamma}}(\hat{\gamma}) d\hat{\gamma} = \\ &\rho\exp\left(\frac{\sigma_{\epsilon_c}^2}{r_g-\sigma_{\epsilon_c}^2}\left(1+\ln\left(\frac{P_b}{0.2}\right)\right)\right)\times \\ &\sum_{i=0}^{N-2}(M_i-1)\left(\mathrm{Ei}(\rho\lambda P_i) - \mathrm{Ei}(\rho\lambda P_{i+1})\right) + \\ &(M_{N-1}-1)\mathrm{Ei}(\rho\lambda P_{N-1}) \leq \bar{S}\end{aligned} \tag{35}$$

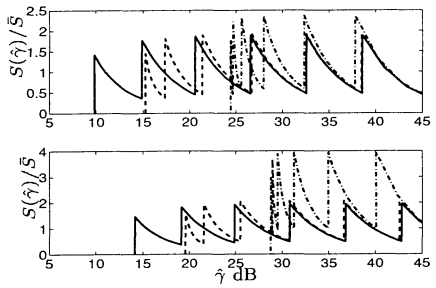


Figure 3: Optimum normalized transmit power versus instantaneous predicted SNR of M-QAM schemes for $\bar{\gamma} = 30$ dB. The upper and lower plots correspond to $P_b = 10^{-3}$ and 10^{-7}, and the solid, dashed and dashed-dotted lines correspond to $\sigma^2_{\epsilon_p} = 0.001$, 0.01, and 0.1, respectively.

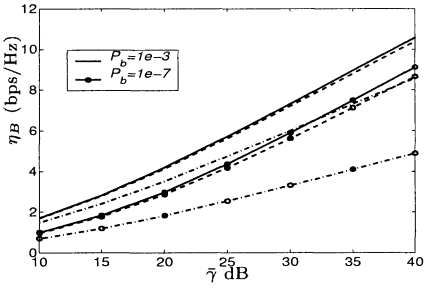


Figure 4: M-QAM Spectral efficiency versus average received SNR for $P_b = 10^{-3}$ and 10^{-7}. The solid, dashed and dashed-dotted lines correspond to $\sigma^2_{\epsilon_p} = 0.001$, 0.01 and 0.1, respectively.

where

$$\rho = \ln\left(\frac{P_b}{0.2}\right)\frac{\bar{S}}{1.6}\frac{1}{\bar{\gamma}(r_g - \sigma^2_{\epsilon_c})}, \quad P_i = \frac{M_{i-1} - M_i}{k_{i-1} - k_i} \tag{36}$$

and Ei( ) is the Exponential Integral given by $\int(e^x/x)dx = \mathrm{Ei}(x)$. When the optimal region boundaries are determined, the maximum spectral efficiency is evaluated through (16) and (23) as

$$\eta_B = \sum_{i=0}^{N-1} k_i\Big[\exp\Big(-\frac{\max(\bar{\gamma}\sigma^2_{\epsilon_c}, \hat{\gamma}_i) - \bar{\gamma}\sigma^2_{\epsilon_c}}{\bar{\gamma}(r_g - \sigma^2_{\epsilon_c})}\Big) - \exp\Big(-\frac{\max(\bar{\gamma}\sigma^2_{\epsilon_c}, \hat{\gamma}_{i+1}) - \bar{\gamma}\sigma^2_{\epsilon_c}}{\bar{\gamma}(r_g - \sigma^2_{\epsilon_c})}\Big)\Big]. \tag{37}$$

6 Results

We assume six different M-QAM signal constellations corresponding to 4-QAM, 16-QAM, 64-QAM, 256-QAM, 1024-QAM and 4096-QAM, are available at the transmitter. Also, a flat Rayleigh fading channel with $r_g = 1$ is presumed. The optimal region boundaries for the cases where the required instantaneous BERs are 10^{-3} and 10^{-7} and the average received SNR $\bar{\gamma} = 30$ dB are evaluated and shown in Figure 3, for the prediction error variances $\sigma^2_{\epsilon_p} = 0.001$, 0.01 and 0.1. It is observed that the optimal region boundaries are increased when the prediction error increases, as expected. The maximum spectral efficiency for $P_b = 10^{-3}$ and 10^{-7} and $\sigma^2_{\epsilon_p} = 0.001$, 0.01 and 0.1 are illustrated in Figure 4. The curves show that the gain in the spectral efficiency using good predictors is considerable as compared to the poor predictors. The reduction in spectral efficiency with an increasing prediction inaccuracy becomes larger when the required BER is increased. Compare $P_b = 10^{-7}$ to $P_b = 10^{-3}$ in Figure 4.

7 Conclusion

The optimum design of an adaptive modulation scheme based on M-QAM modulation is investigated. The transmitter adjusts the transmission rate and power based on the predicted SNR to maximize the spectral efficiency while satisfying the instantaneous BER and average transmit power constraints. The instantaneous BER formula as a function of predicted SNR, average SNR and predicted error variance is evaluated. Optimum solutions for the adaptive rate and transmit power are derived. The analytical results show that the predicted SNR boundaries for a given constellation size are increased when the prediction error increases. Moreover, the spectral efficiency decreases as the predictor error variance increases and (or) the required BER decreases, as expected. However the loss in spectrum efficiency can be significantly reduced by using a good predictor.

References

[1] S. T. Chung and A. J. Goldsmith, "Degrees of Freedom in Adaptive Modulation: A Unified View," *IEEE Transactions on Communications*, vol. 49, no. 9, pp. 1561–1571, Sept. 2001.

[2] A. J. Goldsmith and S. Chua, "Variable-Rate Variable-Power MQAM for Fading Channels," *IEEE Transactions on Communications*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[3] X. Tang, M. Alouini, and A. J. Goldsmith, "Effect of Channel Estimation Error on M-QAM BER Performance in Rayleigh Fading," *IEEE Transactions on Communications*, vol. 47, no. 12, pp. 1856–1864, Dec. 1999.

[4] T. Ekman, M. Sternad, and A. Ahlén, "Unbiased Power Prediction on Broadband Channel," in *Proc. Radio Vetenskap och Kommunikation*, Sweden, Stockholm, June 2002.

[5] T. Ekman, "Power Prediction based on Linear Regression," Technical Report UPTEC 02004R, Signals and Systems, Dept of Material Science, Uppsala University, Sweden.

[6] J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, 4th edition, 2001.

Packet Transmission with Turbo Codes over Fading Channels

Bartosz Mielczarek[1] and Arne Svensson[2]

[1]Chalmers University of Technology, SE-412 96 Göteborg, Sweden. E-mail: bartosz@s2.chalmers.se.

[2]Chalmers University of Technology, SE-412 96 Göteborg, Sweden. E-mail: arne.svensson@s2.chalmers.se.

Abstract

This paper presents the main issues of using turbo coding schemes in wireless packet transmission networks, including rate compatibility, channel estimation and header encapsulation. We claim that incorporating payload information bits, pilot bits and header bits into the code structure yields performance improvements in terms of bit error probability and easiness of implementation. We propose a novel idea of code construction using frequent trellis termination which has interesting properties making it particularly suitable for packet transmission.

1 Introduction

Turbo codes [1, 2] are one of the most promising coding techniques for very noisy channels. Their exceptional performance allows transmitting the signal with very low power which is one of the most crucial demands in future mobile systems. There exist, however, a number of problems with the practical use of turbo codes in systems operating over fading channels.

Wireless packet transmission applications usually employ different block lengths and code structures to adapt to changing parameters of a channel. For example, if the fading channel quality decreases, the problem can be counteracted by decreasing the rate of the code which presents a problem with turbo codes. The natural rate for the typical parallel turbo code is 1/3 and reducing it demands additional hardware blocks. Even if the rate can be decreased, using a fixed interleaver size means sending additional blocks of data over the channel, which causes additional delays and requires buffering at the receiver.

Another inherent feature of turbo codes is the iterative decoding, i.e., the quality of the decoded bits becomes better after each iteration. While this allows for significant reduction of the decoder complexity, it creates problems for certain applications demanding fast, reliable decoding with low bit error rate (BER). One example is multiuser packet transmission. In such systems, each packet is marked with a so-called header which identifies the sender and the receiver of a packet. If the header is turbo encoded, in most cases more than one iteration is needed to decode it reliably. If there are many users in the system, a lot of packets will not be intended for the particular receiver and their decoding will take a lot of power and cause significant delays.

Finally, when the detection of the signal is coherent, the problem of channel estimation using pilot symbols appears. Inserting external pilot bits into the packet reduces the payload and shortens the interleaver length of a turbo code which decreases its performance.

In this paper we address all these issues. We will show that frequent termination of the first turbo code trellis can be used to alleviate the above problems. The rate compatibility can be provided by changing the termination patterns of the trellis and the interleaver length can be kept constant which simplifies the hardware design. Moreover, channel estimation can be achieved by using fixed code sequences instead of external symbols. Finally, by proper design of the code termination, header bits can be decoded with high reliability already after the first iteration. The additional benefit lies in simplification of the turbo decoder hardware.

2 System model

The system analyzed in this paper consists of four blocks: turbo encoder, Rayleigh fading channel, channel estimator and turbo decoder. A brief description of these blocks follows.

2.1 Turbo encoder

The typical parallel turbo encoder consists of two identical recursive systematic convolutional encoders (RSC), connected in a parallel way by a pseudo-random interleaver. The stream of N input bits x_k^s is fed to the first encoder without any modifications and is randomly interleaved for the second encoder. The

streams of systematic bits x_k^s and parity bits x_k^p are then transformed into the antipodal format $\pm\sqrt{E_c}$ (where E_c is the signal energy), punctured, if necessary, to rate r_0 and BPSK-modulated. The N/r_0 code symbols are then channel-interleaved and P pilot symbols used for channel estimation are added. The total length of the frame is therefore $N/r_0 + P$.

2.2 Rayleigh fading channel

The encoded signal is transmitted over a flat Rayleigh fading channel with normalized Doppler frequency $f_D T$ (where f_D is the Doppler frequency and T is the symbol period) and AWGN. After down-converting, matched-filtering and normalizing to unit energy, the received complex baseband signal can be expressed as

$$z_k = a_k x_k + n_k \tag{1}$$

where a_k is the complex fading amplitude and n_k is white, complex Gaussian noise process with variance $E[n_k n_k^*] = N_0/2E_c$ per dimension. The fading amplitude is Rayleigh distributed and the phase error follows the uniform distribution. The correlation between the amplitudes a_k is

$$R_c(n) = E[a_k a_{k+n}^*] = J_0(2\pi f_D n T) \tag{2}$$

where J_0 is the zero-order modified Bessel function of the first kind (note that the fading process is normalized to unit power). The power spectrum of the fading process is the well known U-shaped Jakes spectrum resulting from isotropic scattering.

2.3 Channel estimation

After receiving the signal, the first task of the receiver is the channel estimation where the complex fading amplitudes a_k are estimated as $\hat{a}_k$ prior to decoding. Note, that the channel estimation is performed prior to channel deinterleaving to keep the correlation between samples.

In this paper, we use a coherent, Pilot Symbol Assisted Modulation (PSAM) algorithm which periodically inserts known symbols into the encoded data stream. The classical linear approach to the channel estimation in the presence of noise is the Wiener filtering which minimizes the minimum square error of the estimation defined as

$$\mathbf{w}(k) = \arg \min_{\mathbf{w}(k)\in\mathbb{R}^K} E\left[|a_k - \hat{a}_k|^2\right], \hat{a}_k = \mathbf{w}^T(k)\mathbf{z} \tag{3}$$

where $\mathbf{w}(k)$ are the filter coefficients calculated for every bit position $k \in 1, 2, ...N/r_0 + P$ and $\mathbf{z}$ is the vector of length $N/r_0 + P$ containing the received pilot symbols at P positions and zeros in the positions corresponding to the encoded bits.

To solve (3), the well known Wiener-Hopf equations can be used to calculate the filter coefficients $\mathbf{w}(k)$ as solution to the equation

$$\mathbf{R}\mathbf{w}(k) = \mathbf{r}(k) \tag{4}$$

where autocorrelation matrix $\mathbf{R}$ can be calculated for a given fading rate and noise level as

$$\mathbf{R}_{n,k} = \begin{cases} R_c(0) + N_0/E_c & n = k \\ R_c(n-k) & \text{otherwise} \end{cases}. \tag{5}$$

The vector $\mathbf{r}(k)$ must be calculated for every symbol position k as

$$\mathbf{r}_n(k) = R_c(n-k). \tag{6}$$

After channel estimation, the received complex channel symbols are demodulated and fed to the turbo decoder as

$$y_k = \frac{\Re\{z_k \hat{a}_k^*\}}{|\hat{a}_k|} \tag{7}$$

In the following sections we will use y_k^s and y_k^p for demodulated noisy systematic bits and parity bits, respectively.

2.4 Turbo decoder

In this paper we implement the BCJR algorithm as the component code decoder. The principle of the turbo decoding (see [1]) can be formulated as

$$L(\hat{u}_k) = L_C y_k^s + L_e(\hat{u}_k) + \\ + \log\left(\frac{\sum_{S^+} \alpha_{k-1}(s')\tilde{\gamma}_k^{(e)}(s',s)\beta_k(s)}{\sum_{S^-} \alpha_{k-1}(s')\tilde{\gamma}_k^{(e)}(s',s)\beta_k(s)}\right) \tag{8}$$

where $\tilde{\gamma}_k^{(e)}(s',s)$ is given as

$$\tilde{\gamma}_k^{(e)}(s',s) = \exp\left(\frac{1}{2} L_C y_k^p x_k^p\right) \tag{9}$$

and the $\alpha_{k-1}(s')$ values are the probabilities of arriving at trellis state s' with forward transitions and the $\beta_k(s)$ values are the probabilities of arriving at trellis state s with backward transitions. The summations are performed over S^+ state transitions corresponding to data bit $x_k^s = 1$ and S^- state transitions corresponding to data bit $x_k^s = -1$, respectively.

The first term $L_C y_k^s$ where $L_C = 4E_c/N_0$ is the channel reliability factor, is called the *channel value.* The second term, $L_e(\hat{u}_k)$, is the *a priori* information provided by the second BCJR decoder. Finally, the third term, the *extrinsic* information, is calculated by the first BCJR decoder based on the whole received sequence and the code trellis. This value is then passed to the second decoder in the next iteration.

In this paper, we assume that the turbo decoder does not incorporate the channel model into its structure, i.e., the classical turbo decoder is used for all the simulations.

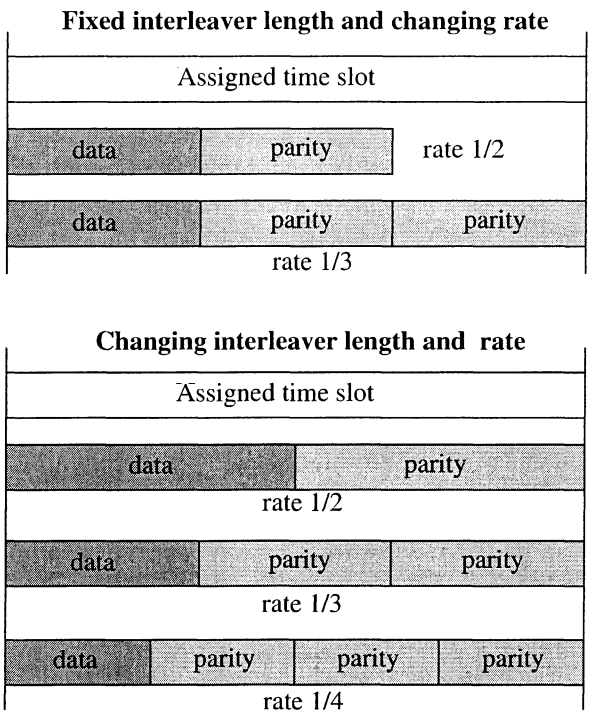


Figure 1: Possible packet organizations

3 Turbo codes in packet systems

3.1 Packet division

Two possible solutions for using turbo codes in packet systems are shown in Figure 1. The simplest implementation employs a constant interleaver length which does not vary with changing rates. This solution, while easy to implement, limits the available rates. For example, if the assigned time slot is equal to the codeword length of rate 1/3 code, decreasing the rate will mean that additional parity bits must be transmitted in a separate packet. This is not always possible due to the delay requirements of the services and/or buffer storage capabilities of the receiver. The second problem arises when the rate is increased. If the codeword gets shorter, a fraction of the time slot will not be used which reduces the bandwidth efficiency.

Another solution is to change the interleaver length for each rate change. If the rate decreases, the interleaver length decreases as well and vice versa. This way the whole packet is always optimally used, i.e., there are no empty spaces for high rates and there is no need to transfer the codeword over more than one packet. Unfortunately, such a solution involves implementing an interleaver for each possible rate and shortens the available interleaver impairing the BER performance.

3.2 Pilot symbols and header bits

In a wireless system, not all the bits in a frame can be used to transmit data. Some of the available bandwidth must be used for pilot symbols and header bits. If the pilot symbols which do not belong to the code structure are inserted into the packet, the effective payload length decreases. This means that the length of the interleaver must be decreased (so the codeword will fit into the remaining slot space) and the resulting bit error rate of the code will decrease as well.

Moreover, in realistic applications, there appears frequently a need for unequal error protection (UEP) which is used to increase reliability of decoding of a selected group of bits (in multiuser scenarios, they can include header or packet identification bits). For the same reasons as above, it is preferable that those bits are encoded using the same turbo code as the other data bits. However, without any modifications to the code, unequal error protection is difficult to achieve for the arbitrary number of header bits.

4 Frequent trellis termination

A possible solution to the above problems is a frequent trellis termination and insertion of the pilot and header symbols into the code structure (see [4]).

Figure 2 shows the principle of the algorithm operation. The data sequence is divided into n sub-blocks of fixed length. Some of the sub-blocks begin with a fixed pilot sequence which is used by a channel estimator. The header and payload data bits follow and the first code trellis is terminated by inserting tail bit sequences. This splits one long trellis of the first encoder into a set of shorter blocks, carrying pilot bits, header bits, data bits and tail bits. Since the second encoder is left unchanged, the principle of a turbo code operation is the same, i.e,. the second trellis is not terminated at all due to the presence of a pseudo-random interleaver (this means that all the classes of bits will be spread randomly in the whole packet).

Employing the above method gives the system two additional levels of adaptivity. Firstly, the number of pilot symbols can be adaptively changed without the need to reconstruct the frame structure and secondly, by changing the number of sub-blocks, it is possible to almost arbitrarily change the rate of the code. This stems from the fact that inserting tail bits which carry no useful information is equivalent to reducing the code rate. The effective rate of such a code is given as

$$r = r_0 \left(\frac{N - nK}{N + r_0 P} \right) \tag{10}$$

where K is the memory of the code.

In general, if the channel quality decreases, more pilot symbols can be inserted in the beginning of each

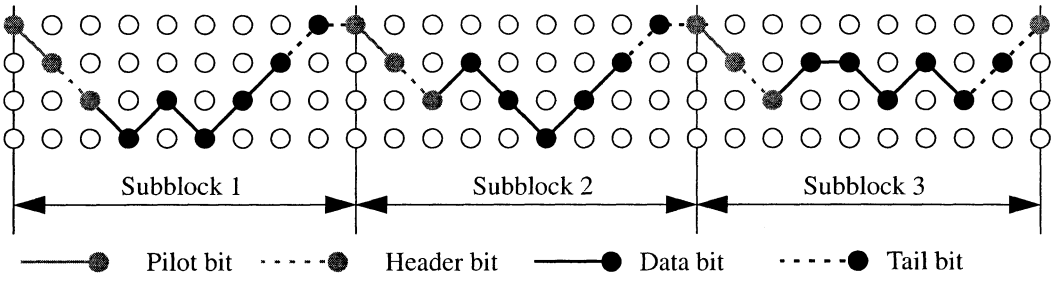


Figure 2: Frequent termination of the first turbo code trellis

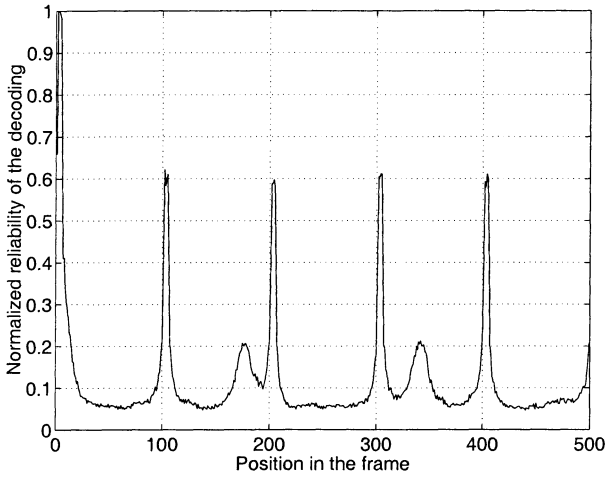


Figure 3: Absolute extrinsic information passed from the first to the second decoder as a function of the position in the code block (N=500, with 10 header bits placed in pairs every 100 bits $E_b/N_0 = 7$ dB).

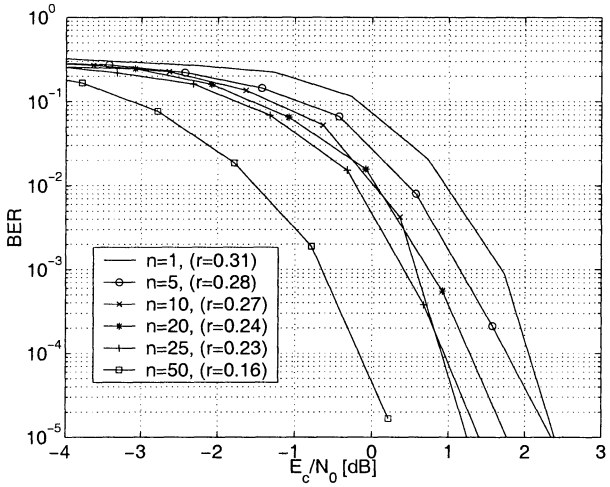


Figure 4: Bit error rates for different trellis termination patterns

sub-block and the termination of the trellis can be done more frequently which reduces the code rate and provides better protection for fewer transmitted bits.

5 Header decoding

Since the reliability of bit detection increases close to the trellis termination points, a straightforward way to decrease the BER of header bits is to insert them between two trellis terminations (see [3]). In this way, the header bits are 'wedged' between the well-defined, all-zero state on the left and the all-zero state on the right. This will lead to a better reliability of decoding as Figure 3 shows. In this figure, we present the average reliability of the extrinsic information generated by the first APP decoder as a function of the position in the code block. One can clearly see that the reliability of bit decoding is much higher for the header bits. Additional high reliability peaks result from uneven spread of channel estimation errors.

6 Simulations

The proposed code modifications were tested by Monte Carlo simulation. We used a parallel rate 1/3 turbo encoder with two RSC (037,033) encoders ($K = 4$) connected by a pseudo-random interleaver of size $N = 500$. Different trellis termination patterns were used, starting with the non-modified code ($n = 1$) and with increasing number of trellis terminations n. Within each frame $P = 150$ symbols were used as pilot symbols. 10000 random data packets were transmitted independently and decoded using our method.

The results of the simulations are shown in Figure 4 with bit error rates for different number of terminations and E_c/N_0, and in Figure 5 where the plots are drawn against normalized SNR defined as $E_b/N_0 = E_c/rN_0$. The first figure shows that by using the frequent termination, it is possible to adjust the code rate according to the channel quality, without changing the length of the transmission. This way the constant bit error rate can be maintained regardless the noise level in the channel. Figure 5 shows the

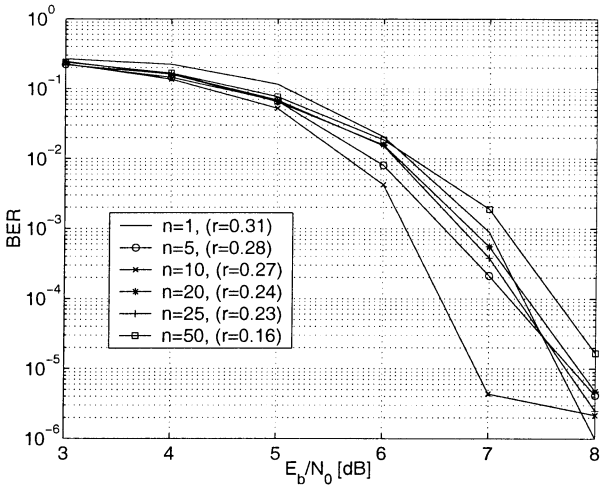


Figure 5: Bit error rates for different trellis termination patterns and normalized SNR=E_b/N_0

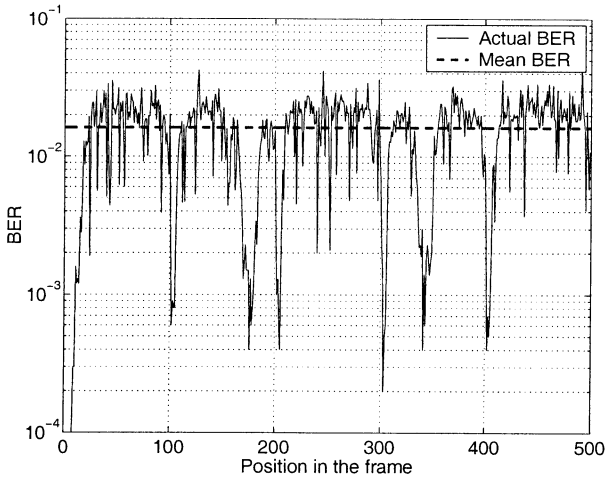


Figure 6: Bit error rates for different bit positions (N=500, with 10 header bits distributed in pairs every 100 bits E_b/N_0 = 7 dB).

performance of the scheme from the information theory point of view. An interesting observation is that, while our codes have worse Hamming distance properties (the frequent trellis termination increases the minimum free distance weight), the suboptimal decoding algorithm actually benefits from faster decoding of certain bits. This is particularly visible when comparing the curve corresponding to 10 terminations with the non-modified code. The gain in the waterfall region is around 1 dB, even though the flattening of the BER curve occurs faster in the first case.

Finally, the header decoding performance of the proposed algorithm is shown in Figure 6, where after only 1 iteration, 10 header bits are decoded with BER $< 10^{-3}$ while the mean BER of other bits is almost two orders of magnitude higher.

7 Conclusions

The practical applications of turbo codes are still few and a lot of research is necessary to gain full insight into the specific problems they present. It is, however, quite clear that a lot can be gained by modifying turbo codes to accommodate specific tasks such as, for example, channel estimation.

We have shown that by modifying the structure of the uninterleaved turbo code component, we can easily adapt the code properties to the changing channel parameters since the rate and the number of pilot symbols do not have to be fixed beforehand. We have also shown that our relatively simple modification of the code can provide very effective and fast header decoding, reducing the delay and power consumption at the decoder. The proposed algorithm can be easily tailored to any protection level of the header bits.

The future work will concentrate on the analysis of the modified turbo decoder properties. The performance loss due to the early trellis termination should be examined in greater detail as well as the convergence properties of the turbo decoder in waterfall region.

8 Acknowledgments

This work was supported by the Swedish Foundation for Strategic Research under the Personal Computing and Communication grant.

References

[1] S. A. Barbulescu. *Iterative decoding of turbo codes and other concatenated codes.* PhD thesis, School of Electrical Engineering, Faculty of Engineering, University of South Australia, Feb. 1996.

[2] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error control-correcting coding and decoding: Turbo-codes (1). In *Proc. IEEE International Conference on Communications*, pages 1064–1070, Geneva, Switzerland, 1993.

[3] B. Mielczarek and A. Svensson. Fast header decoding using turbo codes on Rayleigh fading channels. In *Proc. International Conference on Information, Communication, and Signal Processing*, Singapore, 2001.

[4] B. Mielczarek and A. Svensson. Joint adaptive rate turbo decoding and synchronization on Rayleigh fading channels. In *Proc. IEEE Vehicular Technology Conference*, pages 1342–1346, Rhodes, Greece, 2001.

A new remote sensing tool for forestry applications

Johan E.S. Fransson[1], Per-Olov Frölind[2], Anders Gustavsson[3], Gary Smith[4], Lars M.H. Ulander[5], and Fredrik Walter[6]

[1]Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden, E-mail: Johan.Fransson@resgeom.slu.se
[2]Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping, Sweden, E-mail: perfro@foi.se
[3]Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping, Sweden, E-mail: andgus@foi.se
[4]Chalmers University of Technology, SE-412 96 Göteborg, Sweden, E-mail: smith@rss.chalmers.se
[5]Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping, Sweden, E-mail: ulander@foi.se
[6]Dianthus, Sommarstigen 13, SE-961 32 Boden, Sweden, E-mail: Fredrik.Walter@Dianthus.nu

Abstract

The vision to establish a future sustainable development for various natural resources addresses the importance of access to robust and reliable remote sensing monitoring methods. Recent accomplishments in radar techniques using low frequencies open new opportunities for forest mapping and retrieval of forest parameters. The airborne imaging radar sensor CARABAS-II operates in the lower VHF-band (20-90 MHz). Extensive research efforts conducted with the system have shown promising results in relating VHF SAR data to coniferous forest conditions, e.g. retrieval of stem volume, detection of wind-fellings and thinning cuttings. A fully integrated production line has been developed in a pilot study together with a forest company to evaluate the commercial potential. SAR data covering 25 x 25 km^2 of forested terrain are analyzed in a first test.

1 Introduction

In the order of one fourth of the land surface of the Earth is recognized to be forest covered. However, the amount reported may differ depending on the definition used to characterize an area as a forest. This ecosystem is found within a large variety of regions, ranging from boreal to tropical climate conditions, extending from the interior of continents to mangrove habitats in the littoral zone, and growing on flat to mountainous terrain. Some of the richest examples on biological diversity are encountered in this type of environment, in particular in the tropics. Global estimates on forest cover published by FAO (a United Nations organization) indicate that in 1995 there were 3454 million hectares of forest world-wide, including both natural forest and forest plantations [1]. About 55% are located in developing countries with the remaining part in developed countries. Between 1990 and 1995, there was an estimated global net loss of 56 million hectares of forests. This represented a decrease of 65 million hectares in developing countries and an increase of 9 million hectares in developed countries. Major causes of these forest cover changes include conversion of forests to agricultural land and large infrastructural development in the former case, and forest growth on vacant land in the latter.

Forest ecosystems contain a large part of the carbon stored on land and play a major role in the terrestrial carbon budget. The main open questions concerning the global carbon cycle relate to the lack of knowledge of exchange processes that link the atmosphere and the terrestrial ecosystems. To quantify carbon exchange processes, information is required on the spatial distribution of the carbon stock and carbon fluxes from various ecosystems. Major changes in the carbon stock will influence atmospheric CO_2 levels and therefore also the global climate system. Present methods for measuring and monitoring forest carbon stocks produce estimates of unknown reliability. At a regional scale, methods are based upon measurements at specific locations, whereas global estimates are provided by global inventories and surveys, e.g. the forest resources assessment programmes conducted and regularly issued by FAO. Spatial and temporal simulations on forest growth and productivity are also used. The models, however, generally suffer from unrealistic input variables and a lack of data for validation.

Forestry is also very important to the economies of various countries both in the tropics and in boreal climates. Only about 3% of the world's forests are forest plantations, generally consisting of monocultures [1]. The remaining 97% are semi-natural and natural forests. Commercial activities are found in all three categories. However, the contributions from plantations are increasing as a result of reduction in production from natural forests. The reduction can be explained by deforestation or increased areas being set aside for conservation or other reasons. Forests aimed for commercial use are often managed in a consistent fashion with regular silvicultural treatments, e.g. cleaning and thinning cuttings. During the considerable time periods

This work is financially supported by the Swedish National Space Board, Kempestiftelserna, Holmen Skog AB, the Swedish Council for Forestry and Agricultural Research, Sydved AB, Södra Skogsägarna, and Hildur and Sven Wingquists Foundation.

before harvesting it is of great interest to regularly measure and monitor a number of forest parameters and conditions, e.g. stand-wise stem volume (m^3/ha), tree species composition, and presence of insect pests or other diseases. While wood is the predominant commercial product from most forests, increased attention is being paid to non-wood forest products, particularly in developing countries possessing a rich biodiversity in natural forests. Examples of such valuable products are medicinal plants containing chemical components of interest to the pharmaceutical industry.

Much better information is obviously needed on the quantity, quality, and use of forests. Access to various remote sensing techniques will be decisive to be able to discern the rapid changes of the forest resources presently taking place, both on a global and local scale. Sensor instruments onboard spaceborne and airborne platforms are thus required to fulfil this demand. This paper reviews accomplishments in forestry applications achieved with the airborne sensor CARABAS-II. It is a synthetic aperture radar (SAR) type of sensor but pioneering in the sense that it operates within a different part of the electromagnetic spectrum, i.e. the lower VHF-band. The low frequencies open new opportunities to address some of the needs in forestry management planning and environmental monitoring.

2 VHF SAR imaging

CARABAS-II is the second generation SAR prototype operating with HH-polarized waves in the lower part of the VHF-band, i.e. 20-90 MHz corresponding to wavelengths between 3.3 and 15 m [2]. For several years the proven system concept was the only flying resource available having a capacity to acquire high quality SAR data at frequencies below 100 MHz. The realization of a high performance SAR system in this part of the spectrum is a challenge. Behind the breakthrough for CARABAS-II and its predecessor lies technological advances in antenna and digital receiver design, as well as the ever-growing digital computation power available for image generation. The spatial resolution typically obtained using the full bandwidth and an azimuth integration angle of about 70° is in the order of 2.5 m by 2.5 m, measured in the slant range plane. Major resources have been devoted to implementing efficient signal processing methods, taking into account the much larger amount of contributing data samples from each ground point when illuminating with a wide antenna beam. The radar raw data are presently recorded onboard the Sabreliner aircraft shown in Fig. 1. All of the image generation is carried out off-line on ground, including post-processing steps such as geo-coding and radiometric calibration of the resulting radar images.

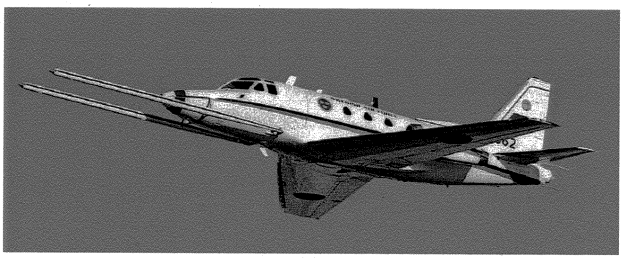


Figure 1: *The FOI developed CARABAS-II VHF SAR system in operation, with the antennas deployed in the front of the Sabreliner aircraft (© Photo: Swedish Defence Images / Peter Liander).*

The CARABAS images have a quite different appearance compared to microwave SAR scenes with similar resolution, i.e. bandwidth. The reason is that the electromagnetic wavelength is about 10-100 times larger than conventional microwave systems, and only meter-scale structures and larger contribute significantly to the image details. For example, grass-covered fields or agricultural areas have a very low backscatter response since their surface roughness is much smaller than the wavelengths of 3.3-15 m. The brightest image structures observed are, in contrast, most often associated with man-made objects, e.g. vehicles, houses, telephone and power cables, and fences. The main VHF radar backscatter from natural objects found in the image is thus related to forested areas as well as topographic features in rugged terrain. These two categories are typically depicted with intermediate to strong pixel intensity levels. The variety of backscattering responses that can be obtained is illustrated in the CARABAS image example given in Fig. 2.

It should be noted that a spaceborne realization of this radar concept might be very difficult. Severe degradations of the radar system performance are expected due to trans-ionospheric propagation effects of the lower VHF-band signals. A future monitoring program on a global scale is thus likely to rely on a stratospheric platform solution, i.e. either a manned/unmanned aircraft or an airship.

3 Stem volume retrieval

An area of about 23 million hectares is covered by forest in Sweden. Presently, it contains around 2850 million m^3 standing volume in growing stock, of which 82% is Scots pine (*Pinus sylvestris*) and Norway spruce (*Picea abies*). The net felling is in the order of 60 million m^3 per year. The slow growth rates imposed by the boreal climate conditions necessitate efficient forestry management planning. Thus, reliable estimates of forest parameters are needed for both short and long term planning, where the primary parameters of interest are stand-wise stem volume and tree species composition.

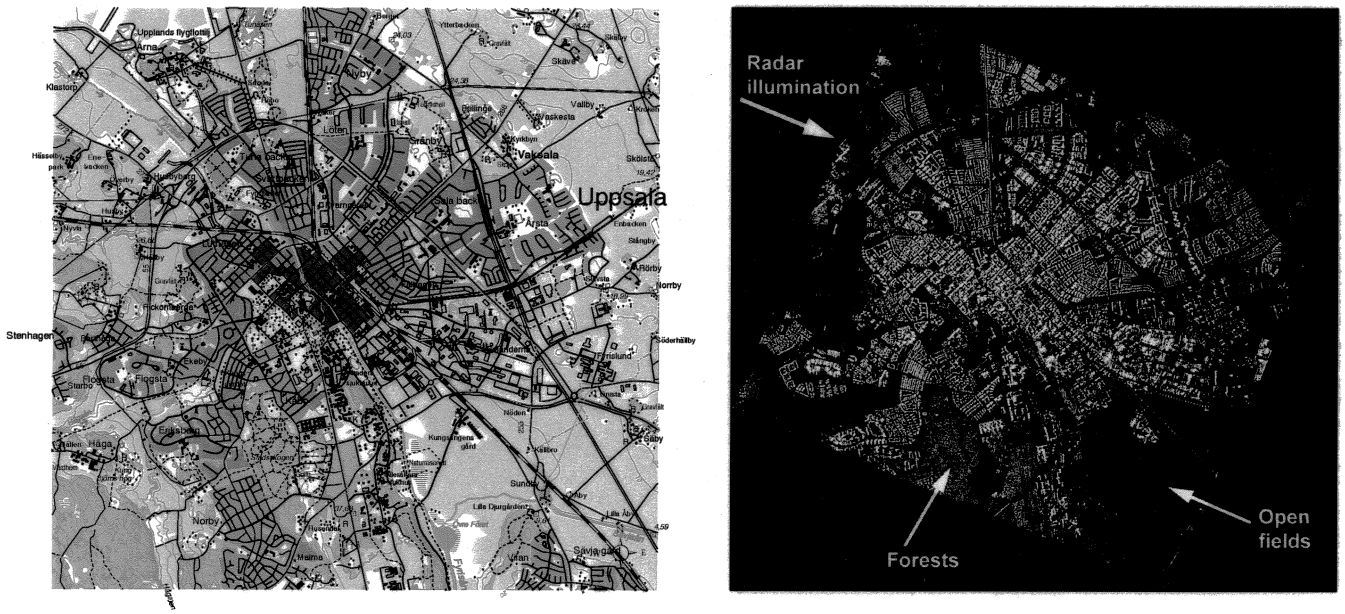


Figure 2: *A geo-coded and radiometrically calibrated CARABAS-II image of the city of Uppsala, covering 8.5 km by 8.5 km. The incidence angle of the radar illumination varies from 55° in near range to 71° in far range. The radar data have been registered from an altitude of 4.7 km. The typical regular pattern in an urban area of building blocks is clearly seen and exhibits the strongest responses. Structures on the annotated open fields are too small relative to wavelength and no significant backscattering can be observed, except for a railway line approaching the city centre and a large farming complex. The low pixel values found here are in the order of the average noise level in the image. The forested areas are depicted with an intermediate brightness of the pixels (Map: Copyright Lantmäteriet, Sweden).*

The meter waves used in CARABAS-II imply that coniferous tree diameters are much smaller than the wavelength. The tree trunk can therefore be regarded as a Rayleigh scatterer in which the image amplitude becomes proportional to trunk volume. Stem volume, i.e. trunk volume per unit area including bark but excluding branches and stumps, can accordingly be extracted by measuring the average pixel amplitude within a specific forest stand. Stem volume relates to biomass but is dependent on the forest type. In Scandinavia, it is typically about 0.6 tons/m^3 for coniferous trees. The low frequencies indicate that attenuation through the canopy layer (needles, leafs, twigs, and branches) is small and that the waves penetrate down to the ground level. The two-way attenuation has been found to be in the range of 1 to 3 dB by examining a comprehensive number of VHF SAR data sets acquired in Sweden and France over coniferous forests with stem volumes up to 1000 m^3/ha. Experiences of stem volume mapping based on the CARABAS-II sensor have been gained from systematic data collection campaigns during the last five years [3-5]. No signs of saturation of the backscattering responses have been observed at the highest stem volume levels analyzed [5]. In Fig. 3, CARABAS-II radar backscattering amplitude is plotted against objectively inventoried stand-wise forest stem volume from two different Swedish test sites, located in northern and southern Sweden, respectively. The relative errors in terms of *RMSE* for the stem volume regression functions developed were 21% in both studies, derived from data samples found in the interval of 80-317 m^3/ha and 80-625 m^3/ha, respectively.

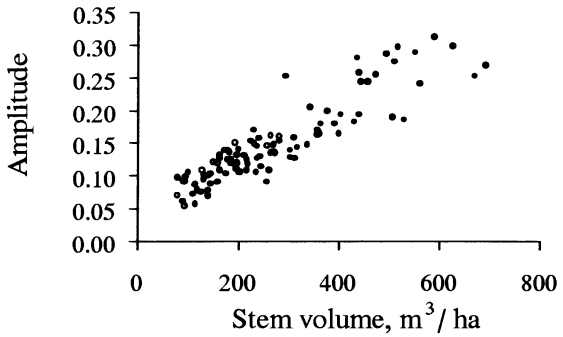


Figure 3: *CARABAS-II VHF SAR backscattering amplitude plotted against objectively inventoried stand-wise forest stem volume. The non-filled and filled circles represent observations from two test sites in northern and southern Sweden, respectively.*

The linear relationship has also been shown through physical modelling of coniferous tree architectures growing on a flat ground [4]. All tree structures except for the tree trunk are neglected as a first order approximation. The total forest backscatter is described as the coherent interaction between the contributions

originating from the direct trunk backscatter, the trunk-ground and ground-trunk dihedral scattering, and finally the ground-trunk-ground scattering. On sloping terrain, however, the dihedral geometry will be affected and the maximum backscattering from the trunk is directed away from the sensor. The amount depends on the local ground slope and its orientation relative to the overall imaging geometry, but the phenomenon is also a function of the tree height relative to the wavelength. The impact from sloping terrain can be seen in Fig. 4, where an area of dense forest has been imaged from two orthogonal directions. For parts of the forest stands, the variations in backscatter are reversed from one image to another. This is due to the effects of topography, and the direction of slope relative to the look direction. Hence, the variations in backscatter induced from sloping terrain represent a problem for retrieving stem volume from a single CARABAS image. Ambiguities in extracted stem volumes may occur even at fairly shallow slopes, i.e. about 4°. A methodology has been identified from modelling to counter this effect. It is based on having access to a number of SAR registrations of an area, acquired from different look directions. The resulting backscattering amplitudes for each forest stand are determined by combining all the images in such a way that the topography modulations are minimized. Each of the backscattering amplitudes plotted in Fig. 3 is based on analyses of four different look directions at both test sites. A refined scattering model, in which the sloping effects are taken into account, is found in [6].

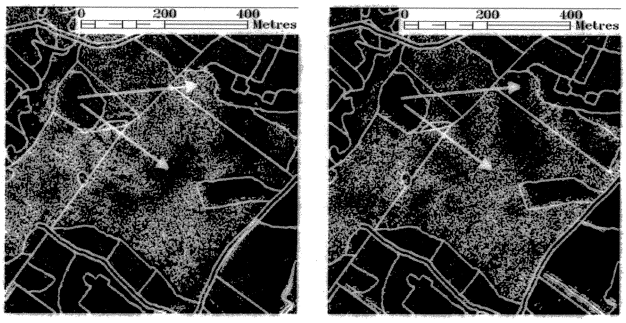


Figure 4: *CARABAS-II image examples for a dense forested area (high backscatter), located in the forestry research park Tönnersjöheden in southern Sweden. Forest stand borders are superimposed. A comparison of the two images shows that apparent within-stand variations of high and low backscatter are reversed when imaged from an orthogonal direction.*

The long-term goal is to establish the VHF SAR technique on a commercial basis for forestry management planning and environmental monitoring. Based on the promising research results, a pilot study in cooperation with the Swedish forest company Holmen Skog AB has been defined and is now running. The objective is to develop a fully integrated and cost-effective production line for mapping of forest parameters over large areas [7]. The input data to the software package are radiometrically calibrated VHF SAR images, information describing the corresponding image generation and imaging geometry, and a digital elevation model. The production line encompasses a number of different processing steps. It starts with a geo-coding process of all individual images and a formation of an image mosaic for each look direction, followed by a semi-automatic data quality check. The latter step includes identification and positioning of disturbing objects within a forested area, e.g. buildings and power lines, and the complete distribution of the objects is described through an image raster layer. At this stage, the data set is prepared for various retrieval algorithms depending on the forest application. Finally, the extracted forest parameters are represented in the given end-user format.

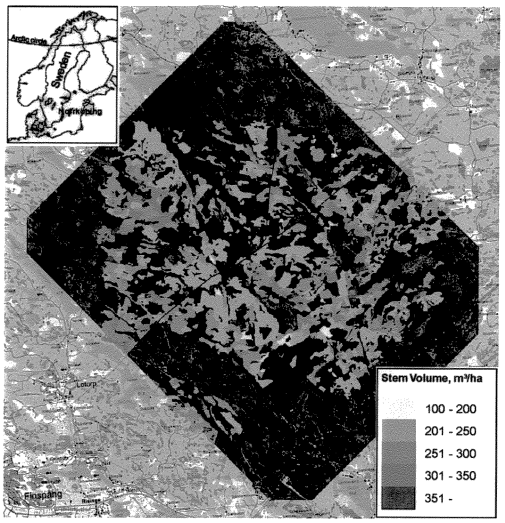


Figure 5: *Visualization example from the developed production line showing the retrieved forest stem volume within a section of the pilot study test area (Map: Copyright Lantmäteriet, Sweden).*

The production line is built on top of a geographical information system (GIS), i.e. software for storage, analysis, and presentation of geographical data. The commercial ESRI ArcGIS™ 8.1 software has been chosen and gives the operator access to all of the GIS-functionality in addition to the production line components found as extension modules. The major software development was completed during the year 2001. The first trials are focusing on stem volume retrieval. An example is shown in Fig. 5. The input SAR data volume consists of about 25 Gbytes, acquired from four different look directions over a 25 x 25 km^2 forested area, located northwest of the city of Norrköping in southern Sweden. The production line is now in a phase of validation and performance optimization.

4 Other forestry applications

Heavy storms cause severe damage world-wide to forested land every year. In order to avoid large economic losses, the wind-thrown timber has to reach the pulp industry rapidly, as timber affected by fungi is not accepted in the pulping process. Therefore, tools for efficient mapping of damaged forested areas are of major importance for assessment of short-term actions, but also for long-term reforestation purposes. In December 1999, several million cubic metres of trees were wind-thrown when devastating storms struck large parts of Europe. In southern Sweden alone, approximately one million cubic metres were damaged. CARABAS-II collected data over two test sites a few weeks after the storm to investigate the possibility to detect wind-thrown forest efficiently, see Fig. 6.

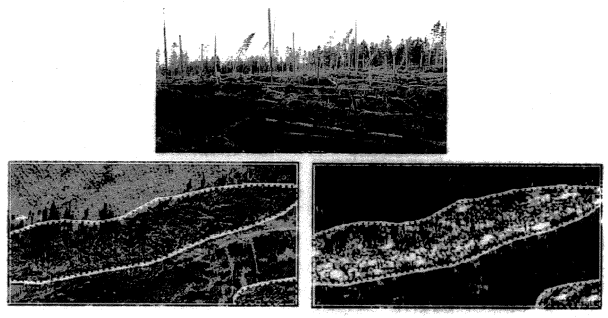


Figure 6: *Comparison between an aerial photo (bottom left) and a CARABAS-II SAR image (bottom right) from the trials of mapping wind-thrown forests. The photo at the top reflects the corresponding situation on ground.*

The results clearly show that the backscattering amplitude, at a given stem volume, is considerably higher for wind-thrown forests than for unaffected forests [8]. Furthermore, the backscattering from fully harvested storm damaged areas was, as expected, significantly lower than from unaffected stands. However, to prevent ambiguities in increased backscattering caused by normal stem volume growth or wind-fellings, multi-temporal change detection techniques using radar images acquired prior to and after wind-fellings would be preferable.

Access to multi-temporal data is also of great interest to obtain better statistics of the thinning cuttings that take place each year. In Sweden, 30% of the average annual harvested stem volume is estimated to originate from thinning cuttings. A similar application is to reveal illegal selective cuttings in regions where bans on logging activities have been imposed. In a project recently initiated, detection of thinning cuttings using CARABAS data is studied. The test site is the forestry research park Remningstorp in southern Sweden, which has been mapped several times prior to and after the thinning activities. A total of 25 thinning stands were thinned between 10-30% of the stem volume with an average of about 20%. The collected data are now in the phase of evaluation [9]. The test site is also used for trials on combining CARABAS-II VHF SAR data with other remotely sensed data sets. Other sensor sources include optical satellite images and high-resolution lidar data. The first results of the benefits obtained are found in [10-11]. The good foliage penetration capability inherent to meter waves enable topographic mapping of the true ground surface beneath the canopy layer by adopting interferometric techniques [12]. An interferomteric registration with microwaves over the same area will instead depict the upper canopy, and by subtracting the two surfaces the tree heights can be estimated.

References

[1] *State of the World's Forests 1999*, FAO, Rome, Italy, 1999.

[2] H. Hellsten, et al., "Development of VHF CARABAS II SAR," in *Proc. Radar Sensor Technology*, SPIE vol. 2747, Orlando, FL, 8-9 April 1996, pp. 48-60.

[3] J.E.S. Fransson, F. Walter, and L.M.H. Ulander, "Estimation of forest parameters using CARABAS-II VHF SAR data," *IEEE Trans. Geosci. Remote Sensing*, vol. 38, pp. 720-727, Mar. 2000.

[4] G. Smith and L.M.H. Ulander, "A model relating VHF-band backscatter to stem volume of coniferous boreal forest," *IEEE Trans. Geosci. Remote Sensing*, vol. 38, pp. 728-740, Mar. 2000.

[5] P. Melon, et al., "On the retrieving of forest stem volume from VHF SAR data: Observation and modeling," *IEEE Trans. Geosci. Remote Sensing*, vol. 39, pp. 2364-2372, Nov. 2001.

[6] G. Smith, et al., "Forest stem volume retrieval with VHF-band SAR," in *Proc. 3rd International Symposium on Retrieval of Bio- and Geophysical Parameters from SAR Data for Land Applications*, Sheffield, UK, 11-14 September 2001, pp. 13-18.

[7] F. Walter, et al., "Development of a fully integrated system for retrieval of forest parameters from CARABAS-II VHF SAR images," in *Proc. SSAB'02, Symposium on Image Analysis*, Lund, Sweden, 7-8 March 2002, pp. 169-172.

[8] J.E.S. Fransson, et al., "Mapping of wind-thrown forests using CARABAS-II VHF SAR image data," in *Proc. IGARSS'01, IEEE International Geoscience and Remote Sensing Symposium*, Sydney, Australia, 9-13 July 2001, pp. 2737-2739.

[9] J.E.S. Fransson, et al., "Detection of thinning cuttings using CARABAS-II VHF SAR data," in *Proc. IGARSS'02, IEEE International Geoscience and Remote Sensing Symposium*, Toronto, Canada, 24-28 June 2002, (in press).

[10] G. Smith, et al., "Forest stem volume estimation using high-resolution lidar and SAR data," in *Proc. IGARSS'02, IEEE International Geoscience and Remote Sensing Symposium*, Toronto, Canada, 24-28 June 2002, (in press).

[11] H. Holmström and J.E.S. Fransson, "Combining remotely sensed optical and radar data in *k*NN estimation of forest variables," Accepted for publication in *Forest Science*.

[12] P.-O. Frölind, A. Gustavsson, and L.M.H. Ulander, "Verification of ultra-wideband VHF SAR data DEM," in *Proc. IGARSS'00, IEEE International Geoscience and Remote Sensing Symposium*, Honolulu, HI, 24-28 July 2000, pp. 3196-3398.

A Study of FMCW-radar for High Accuracy Measurements

MSci Rahman Aljasmi[1] and MSci Jan Kjellgren[2]

[1]Defence Research Establishment, P. O. Box 1165, S-58111 Linköping . E-mail: rahman@foi.se.

[2]Defence Research Establishment, P. O. Box 1165, S-58111 Linköping . E-mail: jankje@foi.se.

Abstract

This study concerns the possibility to use the FMCW-waveform for unambiguous and accurate measurements of target range and target range speed. The study is limited to single target measurements within short ranges. The work is done by two-dimensional signal processing including the estimation of a power spectrum peak and the calculation of the range and the range rate. The work is supported by computer simulation of the signal processing after calculating the received and mixed radar signals.

1. The radar system

The simulated radar system is an FMCW-radar with a coherent transmitter generating an ideal phase modulation, a circulator, an antenna, a mixer and a signal-processor. The signal from the mixer to the signal processor includes both I- and Q-channels.

We study a point target in free space. We have accounted for the thermal noise of the system but neglected e.g. target glint and multi-path phenomena occurring in real situations [1].

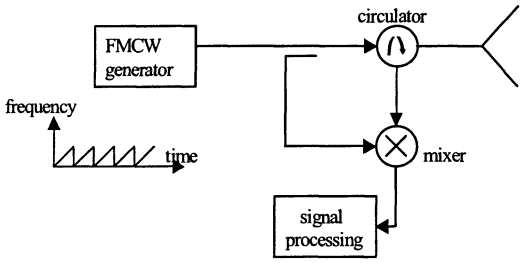


Fig. 1 Block diagram of a FMCW radar.

2. The FMCW-signal

We consider a waveform with a saw-tooth frequency shift [1] with parameters such that no ambiguities will result in either range or velocity and we have also limited the coherent observation time so that a measurable target remains within the same range resolution cell.

The transmitted waveform can be expressed as

$$S_t(t) = \sqrt{2P}\,\mathrm{Re}\left(e^{j2\pi\left((f_c - \frac{B}{2})t_n + \frac{B}{2T_m}t_n^2 \right)} \right) \quad (1)$$

where P is transmitted power, f_c is carrier frequency, B is bandwidth, T_m is modulation time and t_n is mod(t, T_m).

The received FMCW-signal is delayed by the transit time $\tau = 2R/c$, where R is the range and c is the speed of light. The received signal can be expressed as follows [1]

$$S_r = \sqrt{P_r} e^{j2\pi\left((f_c - \frac{B}{2})(t_n - \tau) + \frac{B}{2T_m}(t_n - \tau)^2 \right)} + \omega(t) \quad (2)$$

where P_r is the received power and $\omega(t) \in N(1,0)$ is thermal complex white additive noise.

The received signal is mixed with the transmitted signal, fed through a low-pass filter and sampled according to the Nyquist theorem. The result is a low-pass filtered signal in the base band, which has two frequencies in each period.

A target is supposed to be a point target, which is characterized by its radar cross section, range and range speed.

The range resolution and the range speed are generally obtained by the equations (3) and (4), c.f. [1].

$$\Delta R = \frac{c}{2B} \quad (3)$$

$$\Delta \dot{R} = \frac{\lambda}{2T_{obs}} = \frac{c}{2T_{obs} f_c} \quad (4)$$

where B is the bandwidth, λ is the wave length, T_{obs} is the observation time, f_c is the carrier frequency and c is the speed of light.

The observation time is limited due to the target motion through different range resolution cells, by equation (5).

$$T_{obs} < \frac{\Delta R}{\dot{R}_{max}} \quad (5)$$

where $\dot{R}_{max}$ is the maximum range speed of the targets of interest.

From the equation (3) and equation (5) we have the following condition for the product of the range and the range speed resolutions.

$$\Delta R \cdot \Delta \dot{R} > \frac{c}{2f_c}\dot{R}_{max} = \frac{\lambda}{2}\dot{R}_{max} \qquad (6)$$

We have chosen the signal parameters according to the condition (5).

3. Signal processing

The signal processor is supposed to work in a batch processing mode delivering target range, target velocity and a time stamp for each coherent time interval. The received and mixed signals are calculated according to the radar equation taking into account the change of the measurement geometry. The signal is sampled according to the bandwidth of the signal. We then add noise to the digitized data corresponding to the radar system noise figure. We arrange the digitized data in a 2-dimensional array where one dimension is the sample number within one period of the saw-tooth frequency sweep and the

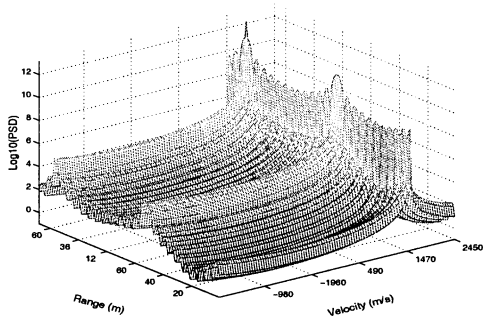


Fig. 2 10-logarithm of absolute value of 2D-FFT of the mixed signal from a target of 60m range and a range speed of 1000m/s (noise free signal).

other is the sweep number within the coherent observation time. We apply a 2-dimensional FFT and a target in general shows up as 2 peaks (see Fig. 2). Both peaks indicate approximately the range and the range speed. The two peaks correspond to the two different frequencies resulting from the mixing process between the frequency modulated transmit signal and a delayed replica.

In this case we only study the use of the larger peak for an accurate estimation of the target parameters. We roughly know the time delay to the target from the approximate position of the maximum peak within the 2D-FFT and by tailoring a filter applied to the original time data only the part corresponding to the most durable frequency will remain. We apply the periodogram method for an accurate estimation of the position of the 2-dimensional frequency peak (see Fig. 3).

The periodogram spectral estimator [2] is

$$P_{per}(f_1, f_2) = \frac{1}{MN}\left|\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x[m, n]e^{-j2\pi(f_1 m + f_2 n)}\right|^2 \qquad (7)$$

We wish to determine the frequency variables f_1 and f_2, which gives the maximum power spectral density at the frequencies f_{1max} and f_{2max}.

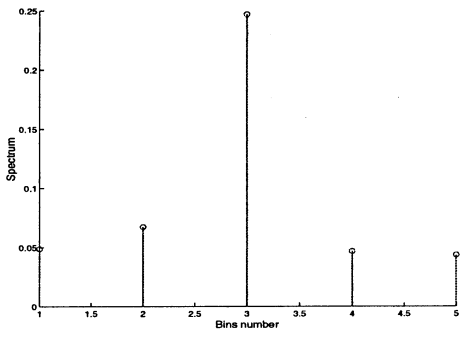


Fig. 3 Truncated spectral-data in one dimension. A principle sketch illustrating a spectral maximum between bin nr 2 and 3.

4. Target parameter estimation

We have chosen the waveform parameters for unambiguous measurements of range and range speed. The range dominates the frequencies within the waveform period and the range speed dominates the phase shift between different periods or sample points separated with the period. When we require an accurate translation from the 2-dimensional frequency position to range and range speed, we must take into account the slight coupling that remains between range and range speed. We then apply an equation system to solve for range and range speed. In our case we use an approximate solution resulting in a small bias error.

Without any noise we aim at a numerical error of the order of 10^{-3} of the resolution cell within the range and range speed dimensions, respectively.

4.1 Range and Range speed measurement

The transmitter frequency of the FMCW radar is changed as a function of time in a known manner. If there is a reflecting object at a distance R, an echo signal will return after the round trip propagation time delay τ = 2R/c. If there is no doppler frequency shift, the beat frequency (difference frequency) measured is due only to the range of the target.

In any practical CW radar, the frequency cannot be continuously changed in only one direction. Periodicity in modulation is necessary and as mentioned earlier we apply a sawtooth modulation, c.f. (1). If the frequency is linearly modulated at a rate $f_m = 1/T_m$ with the bandwidth B, the beat frequency [1] is

$$f_{b1} = B - f_{b2} \tag{8}$$

$$f_{b2} = \frac{2R}{c} B f_m \tag{9}$$

The measurement of the beat frequency determines the range R, c.f. Fig. 4.

In the discussion above, the target is assumed to be stationary. If this assumption does not apply, a doppler frequency shift will be added to the beat frequency and an erroneous range measurement results. The doppler frequency shift causes the frequency of the echo signal to be shifted up or down and the beat frequency will be

$$f_{b2} = \frac{2R}{c} B f_m + f_d \tag{10}$$

with

$$f_d = \frac{2f_c}{c} \dot{R} \tag{11}$$

where $\dot{R}$ is the target velocity relative to the radar and f_c is the carrier frequency.

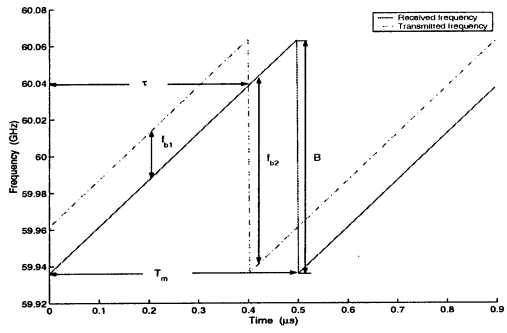


Fig. 4 Sawtooth frequency-time relationship in a linearly modulated FMCW-radar.

From the signal processing we use the estimations of the frequencies f_{1max} and f_{2max} for the range and range speed calculations.

The doppler shift f_d due to the range speed affects the beat frequency f_b. We use one of two different signal filtering alternatives and depending on this choice we apply one of two different pair of equations. We solve (12) for signal delay times $0 \le \tau < T_m/2$

$$\begin{cases} \frac{2B}{T_m c} R + \frac{2f_c}{c} \dot{R} = f_{1\max} \\ \frac{2f_c n}{c} \dot{R} = f_{2\max} \end{cases} \tag{12}$$

and (13) for the Interval $T_m/2 < \tau \le T_m$

$$\begin{cases} B - \left(\frac{2B}{T_m c} R + \frac{2f_c}{c} \dot{R} \right) = f_{1\max} \\ \frac{2f_c n}{c} \dot{R} = f_{2\max} \end{cases} \tag{13}$$

where B is bandwidth, T_m is modulation time, c is the speed of light, f_c is carrier frequency, n is the number of periods, or sweeps, f_{1max} and f_{2max} are the estimated frequencies for the range and radial velocity.

5. Simulation result

The radar parameters are chosen according to Tab. 1.

Notation	Explanation	Value
f_c	Carrier frequency	60 GHz
B	Bandwidth	128 MHz
T_m	Modulation time	0.5 μs
N	Number of period	100
R_{max}	Maximum unambiguous range	~75m
R_{max}	Maximum unambiguous range speed	~2500m/s
ΔR	Range resolution	~1.2m
ΔR	Range speed resolution	~50m/s

Tab. 1 Radar parameters.

To control the numerical procedure we have chosen test cases according to Tab. 2. We suppose that the target moves in a linear path through the radar position. The test results of the range and range speed calculations are presented in Fig. 5 and Fig. 6 respectively.

	Case 1	Case 2	Case 3
Range speed(m/s)	2241.79	1513.79	301.41
Range (m)	70 - 10	70 - 10	70 - 10

Tab. 2 The parameters of the numerical test cases.

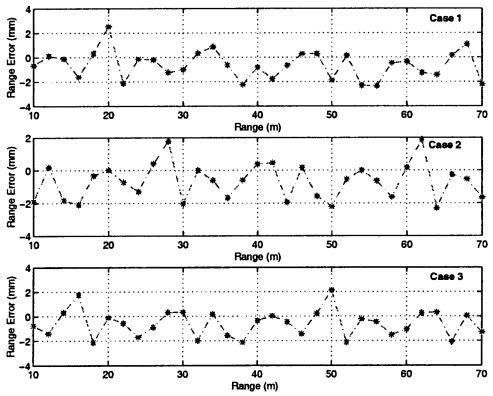


Fig. 5 Numerical range errors for different test cases.

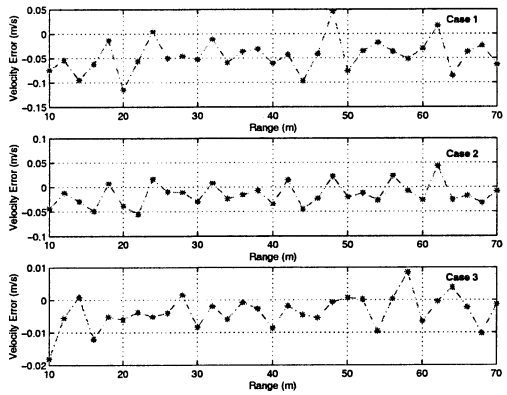


Fig. 6 Numerical range speed errors for different test cases.

The simulations without noise show small numerical errors. The numerical errors of the range measurement show no dependence of the target range or the target range speed. The errors of the range speed measurement display a clear dependence of the target range speed. The maximum numerical errors are all about $2*10^{-3}$ of the corresponding resolution cell.

The errors of the numerical procedure also show a small average bias of about or less than 10^{-3} of the resolution cell.

Supposing normal distributed noise the standard deviation of a radar measurement with noise is

$$\sigma = \pm \frac{\Delta M}{\sqrt{2 \cdot SNR}} \qquad (14)$$

where ΔM is the nominal resolution, ΔR for the range and ΔR for the range speed [1].

To estimate the standard deviation of a noisy signal we have chosen two different distances, 40 m and 60 m for a target with a range speed of 2000 m/s moving towards the radar. We simulate 200 range and range speed measurements in each point and the results are presented in Fig. 7 to Fig. 10.

The means of the range measurements at 40 m and 60 m are 40.01 m and 60.006 m respectively. The means of the range speed measurements at a true speed of 2000 m/s and at the range 40 m and 60 m are 2000.59 m/s and 1999.95 m/s respectively. The standard deviations estimated for the range measurements at true ranges 40 m and 60 m are 4.2 cm and 4.6 cm respectively. The standard deviations for range speed at 40 m and 60 m are 1.2 m/s and 1.7 m/s. The transmitted power is equal in all cases and corresponds to the SNR 37.8 and 30.8 dB at 40 and 60 m respectively. The application of a t-test indicates that the range speed measurements at 40 m are biased [3].

The mixed FMCW-signal during a frequency sweep consists of two parts with two different frequencies. Only the most extended part of the signal is used for the signal processing and the calculation of the range and range speed. In the most unfavorable case we only use 50% of the received power, which increases the measurement noise and the standard deviations. The standard deviations according to (14) are for the range measurements 1.0 cm and 2.4 cm at 40 m and 60 m respectively. The corresponding standard deviations for the range speed measurements are 0.46 and 1.02 m/s.

An obvious improvement is to use both of the target frequencies from the mixer resulting in two different estimations. A properly weighted combined estimation will further reduce the standard deviation.

To achieve the theoretical limit, the estimation has to include a coherent procedure during the whole observation time.

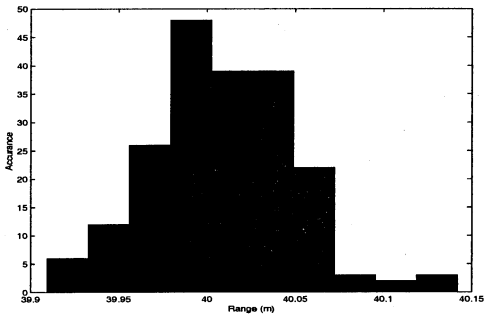


Fig. 7 The distribution of 200 range measurements at 40 m range for the range speed 2000 m/s.

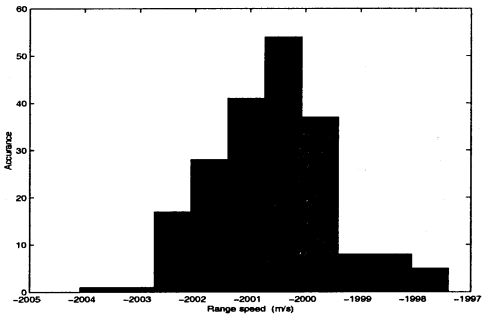


Fig. 8 The distribution of 200 range speed measurements at 40 m range for the range speed 2000 m/s.

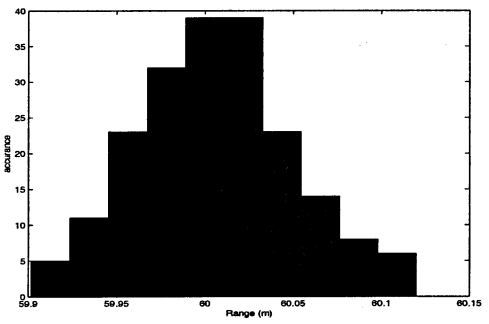


Fig. 9 The distribution of 200 range measurements at 60 m range for the range speed 2000 m/s.

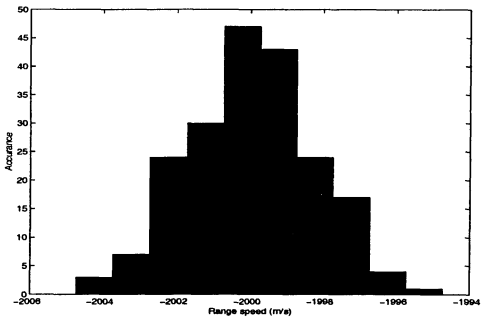


Fig. 10 The distribution of 200 range speed measurements at 60 m range for the range speed 2000 m/s.

6. Conclusion

We have demonstrated a signal processing procedure for range and range speed calculation for an FMCW-radar. The achieved accuracy for both the range and range speed shows a standard deviation of less than 5% of the resolution cell. The standard deviation is between 1.5-4 times the theoretical values.

References

[1] Merill Skolnik. "Radar Handbook". New York: McGraw-Hill Publishing Company (1990), ISBN 0-07-057913-X.

[2] Steven M. Kay. "Modern Spectral Estimation: Theory & Application". Englewood Cliffs : Prentice Hall (1988), ISBN 0-13-598582-X.

[3] A. Hald. "Statistical Theory with Engineering Applications". New York: John Wiley & Sons, Inc (1952).

Autofocusing SAR imagery – some simulation results

Sune R.J. Axelsson

Saab Bofors Dynamics
Swedish Defence Research Agency
S-581 88 Linköping, Sweden
Email: Sune.Axelsson@dynamics.saab.se

Abstract

In this paper, the Phase Gradient Algorithm (PGA) and the Phase Difference Algorithm (PDA) for autofocusing are studied for some various scenes. Sampling requirements and the influence of additive noise are discussed and simulated as well.

1 Introduction

In processing SAR-data into well-focused imagery, the path of the antenna aperture centre (APC) should be estimated with an accuracy of a few percent of the wavelength (See Fig.1). The path information can be achieved by using Inertial Measurement Units (IMU). An alternative solution is the use of autofocus methods [1]-[2].

The simplest types of autofocus technique, such as the Mapdrift and Phase Difference algorithms, only correct for the quadratic phase error or require strong point reflectors in the scene (Prominent Point Processing). These methods can be extended to include higher orders of phase error [4] but the accuracy is reduced when many parameters are to be resolved. A more general and non-parametric method is the Phase Gradient Algorithm [6]-[8], which is studied in this paper for some various scenes.

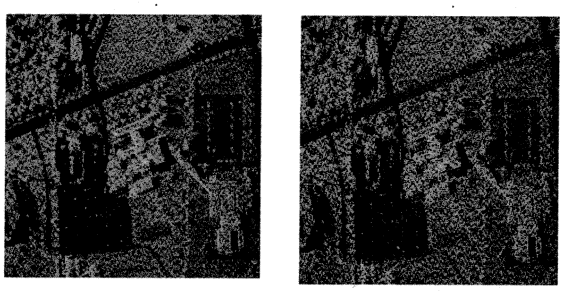


Figure 1: Two simulated images showing the effects of uncompensated sine oscillations in x-, y- and z-directions. The amplitude of the oscillations is 0.03λ (left), 0.05λ (right). Three oscillation periods along the SAR-path.

2 PGA-processing

2.1 General

The input to the PGA-algorithm is SAR imagery data in complex form: C(n,m), where n is the index in azimuth and m is the range index. B(n,m)=|C(n,m)| forms the grey scale image.

The PGA algorithm includes the following main steps:

1. Centre shift of the brightest reflector in azimuth: C2(n,m).
2. Windowing by setting all matrix elements zero for |n| >W/2. The output is C3(n,m).
3. FFT in azimuth of C3(n,m) giving A(n,m). The phase difference is then derived from

$$\Delta\phi_e(n)= \arg\{\Sigma A(n-1,m)A^*(n,m)\} \quad (1)$$

which includes averaging over all range cells (m).

4. The phase error to be compensated is estimated from

$$\phi_e(n) = \sum_{k=1}^{n} \Delta\phi_e(k) \quad (2)$$

from which the linear term is eliminated.

5. Correction of the phase error is carried out:

$$A1(n,m)=A(n,m)\exp[-j\phi_e(n)] \quad (3)$$

6. Inverse FFT back to the image domain gives the corrected SAR-image data: C1(n,m).
7. Steps 1-6 are repeated until the convergence is not improved. This can be measured from the estimated phase error function.

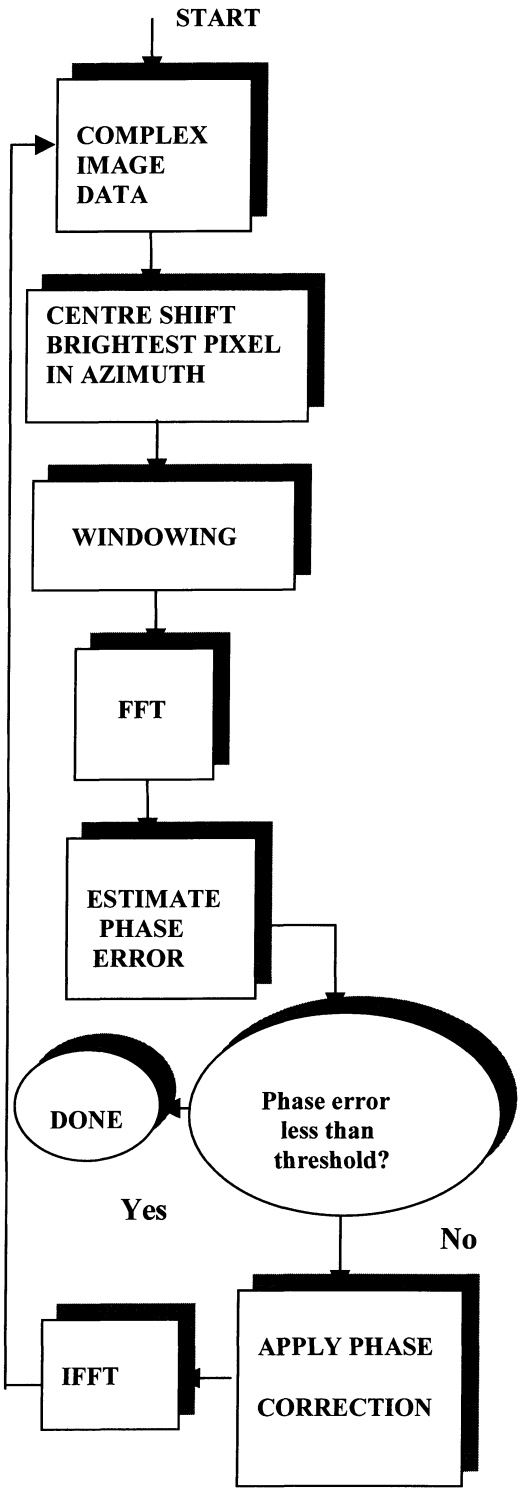


Figure 2: Main steps of the PGA-algorithm.

2.2 Centre shift of the strongest reflectors

The starting operation means a centre shift in azimuth of the strongest reflector at each range. That gives them zero doppler shift. Reflectors positioned offside the centre have a doppler shift which reduces their influence at the following phase estimation over all range cells. The centre shift algorithm includes:

a) Estimate the azimuth position of the strongest reflector in each range cell (m).
b) Estimate it centre shift $\Delta n(m)$ in azimuth.
c) For each range m move all the reflectors in azimuth as $n \rightarrow \text{Modulus}\{n+\Delta n(m),N\}$ giving matrix C2(n,m).

2.3 Selection of window width and threshold

Windowing is then performed, which further enhances the influence of the centre shifted reflectors. The choice of window width is a critical element and various ways of implementation have been discussed.

Obviously, the optimum choice is affected by the power distribution in azimuth. Usually it is recommended that the window width is selected 10 dB below the peak value [2].

For low contrast imagery, one has to start with a wide window that is gradually narrowed for each PGA-iteration. In [8] it was suggested that the threshold should be chosen equal to the average value formed over all resolution elements.

2.4 Phase angle estimation

The phase differences are estimated by forming

$$\Delta\phi_e(n) = \arg\{\sum_m A(n-1,m)A^*(n,m)\} \qquad (4)$$

which means averaging over all range cells (m).

The phase error to be compensated for is derived using $\phi_e(1)=0$:

$$\phi_e(n) = \sum_{k=1}^{n} \Delta\phi_e(k) \qquad (5)$$

from which the linear component is removed.

It can be shown that (4) is a ML-estimate of the phase gradient, and this variance is close to the Cramer-Rao's lower bound [8]

$$\sigma_\phi^2 = (1+2\beta)/(2\beta^2 M) \qquad (6)$$

where β^2 is the signal-to-noise ratio: $\beta^2 = \sigma_a^2/\sigma_\xi^2$.

The analysis is based on the assumption [2] that the doppler signal A(n,m) can be written as

$$A(n,m) = a(m)\exp(j\phi_n) + \xi(n,m) \quad (7)$$

where ϕ_n is the path related phase error and a(m) is the complex reflectance, which is assumed Gaussian with zero mean and variance σ_a^2.

Assuming a(m) constant in azimuth (independent of n) means that the signal model describes a point reflector response (Dirac function in the image plane) surrounded by clutter or noise described by $\xi(n,m)$, which is also assumed gaussian with zero mean and the variance σ_ξ^2.

2.5 How large phase distortions can be accepted?

A. Noise free image

The phase difference is estimated as shown above from (m=1 to M):

$$\Delta\phi_e(n) = \arg\{\Sigma A(n-1,m)A^*(n,m)\} \quad (8)$$

and the phase error from

$$\phi_e(n) = \sum_{k=1}^{n} \Delta\phi_e(k). \quad (9)$$

We realise that the phase of A(n,m) must not jump more than $\pm\pi$ to give a non-ambiguous phase estimation. At presence of noise signals the permitted phase jump interval is further reduced. The tolerance is improved, however, by averaging over many range cells.

B. Presence of noise

Apply the signal model

$$A(n,m) = S(m)\exp[j\phi(n)] + \xi_1(n,m) + j\xi_2(n,m) \quad (10)$$

where $\xi_1(n,m)$ and $\xi_1(n,m)$ are gaussian noise and S(m) is the signal written as

$$S(m) = a(m)\exp[j\theta(m)] \quad (11)$$

The phase error to be eliminated is thus $\phi(n)$. The phase error when we average over M range cells is

$$\Delta\phi_e(n) = \arg\{\Sigma A(n-1,m)A^*(n,m)\} \quad (12)$$

When $SNR=\sigma_s^2/\sigma_\xi^2 >> 1$ the variance of $\Delta\phi$ becomes

$$\sigma_{\Delta\phi}^2 = 1/[M \cdot SNR] \quad (13)$$

which was confirmed by the simulation results of Table I.

TABLE I: Phase error (rms) versus SNR and M.

SNR [dB]	M	σ_Δ [rad]
30	1	0.03
20	1	0.1
10	1	0.3
0	1	1.6
20	127	0.008
10	127	0.024
0	127	0.11
-10	127	0.73

3. PGA-simulations

3.1 One point reflector at each range

Let us start from a SAR-image having N=128 resolution cells in azimuth and M=64 in range. At each range we introduce a point reflector. Its amplitude and position in azimuth are both chosen randomly from a uniform distribution.

A sine wave phase error was introduced at the sampling along the SAR path with the amplitude $5\lambda/4$. The test showed that the phase error was almost completely eliminated after one single PGA-processing. Se Figure 3.

3.2 Many reflectors at each range

If eight reflectors are introduced at each range with random position in azimuth the convergence rate becomes significantly decreased. Figure 4 shows the SAR-image with and without phase error, and finally the result after four PGA-iterations. Still a significant defocusing remains. The gradual improvement of the estimated phase error is displayed as well.

To improve the convergence rate zero padding was applied in the image space. A similar phase error was introduced as in Figure 4, and after one single PGA-processing, a more well-focused image was obtained (Figure 5).

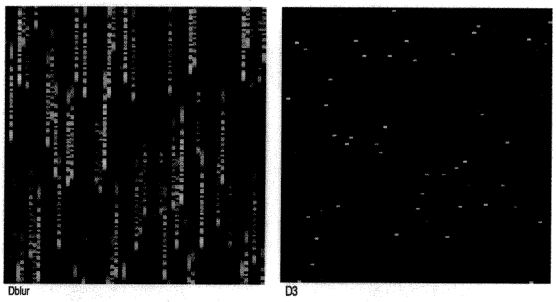


Figure 3. Defocused SAR-image with one reflector at each range (left). After one single PGA-processing (right).

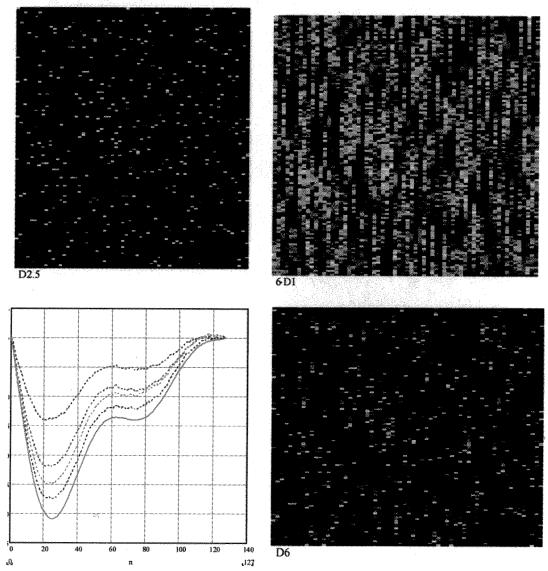


Figure 4: SAR-image with eight reflectors at each range randomly distributed in azimuth (upper left) and after applying phase error function (upper right). Focused SAR-image after four PGA-iterations and the gradual improvement of the phase error estimate below.

3.3 Complex scenes

SAR-imagery was simulated using the method described in [3]. A phase error function was introduced followed by PGA-processing. The resulting imagery after one and two PGA-iterations is shown in Figure 6. A significant sensitivity to window width was noticed and in some cases difficulties to derive full convergence.

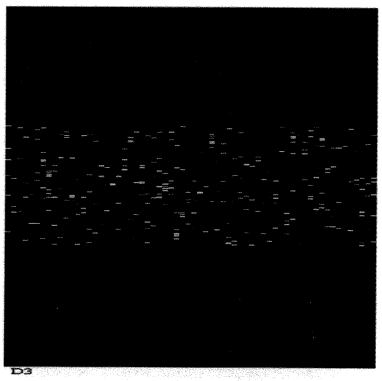


Figure 5: PGA-processing after zero padding and only one PGA-processing.

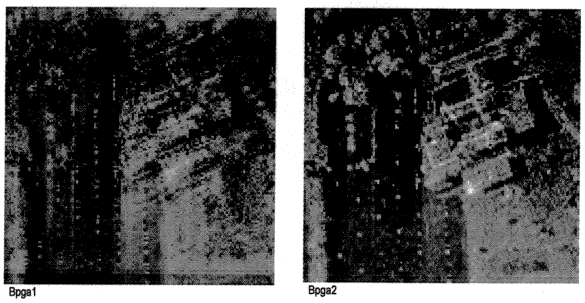


Figure 6. SAR- imagery after one and two PGA-iterations.

3.4 Extended zeropadding

In standard PGA, samples outside the window are replaced by zeros before the FFT and phase estimation occurs. As indicated above the convergence rate increases significantly by increasing the zero padding further to 2N.

A limiting factor is the correlation between adjacent noise samples (ρ) because zero padding also means an interpolation of the FFT-transformed data. It was shown [8] that the correlation gives a biased estimate of the angle difference according to

$$E\{\Delta\phi\} = \arctan\{\sin(\Delta\phi)/[\cos(\Delta\phi)+\rho/SNR]\} \quad (14)$$

where $SNR=\sigma_s^2/\sigma_\xi^2$.

Hicks [8] points out that the bias can be reduced, if instead we perform a DFT over the window followed by an estimate of the phase difference, interpolation and phase estimation.

4. Phase Difference Algorithm

Alternatives to the PGA method discussed above are the parametric methods [1]-[2], such as the Mapdrift Algorithm and the Phase Difference Algorithm. These methods can be extended to include higher orders of phase errors [4] but the accuracy is reduced when many parameters are to be resolved.

Figure 7 shows a simulation of autofocus using the phase difference method [1]. Upper to the left, the SAR-image with no phase error; to the right the image when a quadratic phase error $\phi(t)=a(t/T)^2$ was applied with $a=2\pi*6.92$ radians , where T is the full aperture time.

The first step of the phase difference algorithm divides the azimuth signal history into two subapertures, the two signals are multiplied and the product is Fourier transformed. In the absence of phase error, a maximum for n=0 occurs. When there is a quadratic phase error $a=b\cdot 2\pi$, the maximum occurs for n=b.

In Figure 7, the phase difference method estimated a correlation maximum for n=7, and when that focus correction was applied, the upper right image was obtained. Error functions of higher orders than quadratic have also been simulated but gave less distinct correlation peaks than in Figure 7.

5. Conclusions

In this paper the Phase Gradient Algorithm (PGA) and the Phase Difference Algorithm (PDA) were studied including simulations against both point reflectors and more realistic scenes. The maximum phase jump tolerated between each sample point and the effects of additive noise were analysed. An increased convergence rate was found using zero padding, but a bias effect is then obtained for additive noise. A critical part influencing the convergence of the algorithm is the choice of window width. A large number of iterations are usually required for low-contrast scenes.

References

[1] W.G. Carrara, R.S. Goodman, R.M. Majewski: *Spotlight Synthetic Aperture Radar.* Artech House 1995.

[2] C.V. Jakowatz et al: *Spotlight-Mode Synthetic Aperture radar: A Signal Processing Approach.* Kluwer Academic Publishers, 1996.

[3] Sune Axelsson: Synthetic Generation of SAR Raw Data For Complex Scenes, European Conference on Synthetic Aperture Radar: EUSAR2000, Munich, Germany, 23 - 25 May 2000.

[4] C.E. Mancill and J.M. Swiger: A Map Drift Autofocus Technique for Correcting Higher Order SAR Phase Errors (U), 27:th Annual Tri-Service Radar Symposium Record, Monterey, CA, June 23-25 1981, pp. 391-400.

[5] J.C. Curlander, C. Wu and A. Pang: Automated processing of Spaceborne SAR data, Proceedings of IGARS'88, Munich, Germany, 1982, pp:FA-1:3.1-3.6.

[6] P.H. Eichel, D.C. Ghiglia and C.V. Jakowatz Jr: Speckle processing Method for Synthetic Aperture Phase Correction, Optics letters, vol. 14, No.1, 1989.

[7] P.H. Eichel and C.V. Jakowatz Jr: Phase-gradient Algorithm as an Optimal estimator of the Phase derivative, Optics letters, vol. 14, No.28, 1989, pp. 1101-1109.

[8] M.J. Hicks: On the Convergens of the Phase gradient Autofocus Algorithm for Synthetic Aperture Radar Imaging. Sandia Report SAND 95-2364, January 1996.

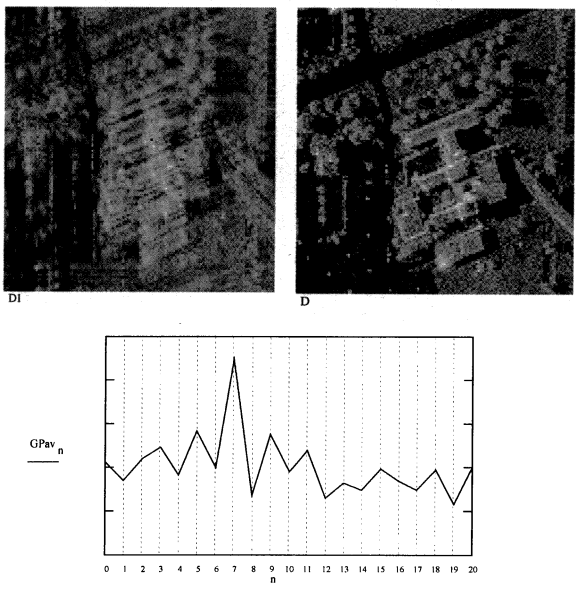


Figure 7: Simulation of auto-focusing using the Phase Difference Algorithm. Upper to the left, when a quadratic phase error was applied with a maximum error of $a=2\pi*6.92$ radians. After PDA-processing to the right. The phase difference algorithm (PDA) estimated a correlation maximum at n=7 corresponding to $a=2\pi*7$ (below).

Deconvolution methods in radar array processing

Anders Nelander[1]

[1]Swedish Defence Research Agency (FOI), Box 1165, 581 11, Linköping, Sweden, E-mail: andnel@foi.se

Abstract

Terrain scattered interference or hot clutter is a recent problem for airborne radar with low sidelobe antennas and conventional adaptive sidelobe cancellation. In this paper a deconvolution approach is proposed to mitigate terrain scattered interference. This approach is based on obtaining an estimate of the complex multipath impulse response from a short time interval in the received signals. The impulse response estimate is then convolved with a direct path reference signal to generate an estimate of the received terrain scattered interference signal. This interference signal estimate is then subtracted from the received main beam signal to generate a main beam signal with mitigated terrain scattered interference.

1 Introduction

Terrain scattered interference or hot clutter is a recent problem in radar electronic counter-countermeasures, especially for airborne radar with low sidelobe antennas and conventional adaptive sidelobe cancellation. The radar jammer can transmit both a direct path signal to a radar antenna sidelobe and multipath terrain scattered signals via both the radar antenna sidelobes and the radar antenna main beam. Low antenna sidelobes and sidelobe cancellation can suppress the direct path signal and the terrain scattered signals from the sidelobe region of the antenna. The remaining problem is the terrain scattered signals into the radar antenna main beam where a high antenna gain must be maintained in the directions to possible targets. These signals are not easily suppressed without also suppressing the target signals.

The scenario is depicted in Fig. 1 where the view is from the receiver antenna in $sin\theta$-space (uv-coordinates). This projection maps the half-space from the antenna aperture normal in which the main beam can be scanned for a planar array antenna. The direct path jammer signal enters the antenna sidelobes from a certain direction and the specular reflection jammer signal can appear from another direction corresponding to the shortest reflection path. Otherwise more diffuse multipath terrain scattering enters from the directions to ground or sea surface areas out to the horizon. The main beam reflection area can give strong terrain scattered jamming signals that compete with the target signals. This is especially serious for low-level targets where there is always a terrain background in the main beam direction. For targets well above the horizon the main beam can be scanned clear of the terrain background. The main beam contribution to the terrain scattered interference is considered to be the most important problem in this paper.

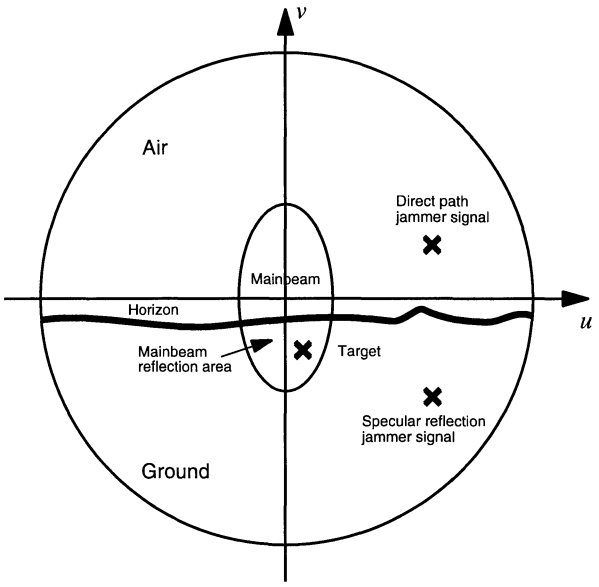


Figure 1: *Terrain scattered jamming scenario projected into uv-coordinates. The main beam is scanned to the antenna broadside and the –3 dB beamwidth covers both clear air and ground areas.*

Radar array processing is usually developed using a covariance-based approach for optimal filtering to suppress interference signals from desired signals. The interference signal covariance matrix is estimated from the received signals and it is used together with the desired signal steering vector to generate the adaptive filter weight vector. This type of processing is used in space-time adaptive processing (STAP) to suppress clutter and jamming in airborne phased array radar with digital beamforming. Various methods to mitigate terrain scattered interference have been proposed, see for example [1]. Two difficulties in these approaches are the nonstationary nature of the interference and the required averaging to estimate the covariance matrix.

In this paper a deconvolution approach is proposed to mitigate terrain scattered interference. This approach is based on obtaining an estimate of the complex multipath impulse response from a short time interval in the received signals. The impulse response estimate is then convolved with a direct path reference signal to generate an estimate of the received terrain scattered interference signal. The interference signal estimate is then subtracted from the received main beam signal to generate a main beam signal with mitigated terrain scattered interference.

The deconvolution approach is in general dependent on a number of limiting conditions and it is sensitive to model errors since it involves inverse filtering with possible singularities and numerical instabilities. However, the interference nonstationarity can be handled by choosing sufficiently short deconvolution processing time intervals for the received signals.

2 Radar signal model

The receiver antenna is assumed to be a planar array antenna with digital beamforming. The radar signals are generally modeled as signals in an array antenna but the signals can also be modeled as signals in a classical adaptive sidelobe cancellation antenna with a main antenna and auxiliary antennas. An airborne pulsed Doppler surveillance radar with lookdown capacity is assumed together with airborne stand-off jammers. The jammer signals are primarily continuous wideband noise signals transmitted from wide beam antennas that generate correlated direct path signals and multipath signals. It is also possible that the jammer signals are pulsed noise or repeater signals.

In the following all signals are assumed to be bandpass filtered to the transmitted signal bandwidth but not pulse compressed, down converted to baseband and sampled with a sufficient sampling frequency. This condition also means that the sampled interference and noise signals are required to fill the sampled signal spectrum. This requirement is necessary to avoid singularities in the deconvolution processing.

The array antenna is used with adaptive beamforming to extract the direct path auxiliary signals from jammers in the sidelobe region. The direct path signals may also include specular or near specular terrain scattered interference signals in the sidelobe region. The goal is to obtain accurate estimates of the jammer signal waveforms as reference signals in the deconvolution processing. The array antenna is also used with adaptive beamforming to generate a receive main beam in the transmit beam direction with suppression of both the direct path jammer signals and the terrain scattered interference signals in the sidelobe region. The resulting main beam signal contains both the target signals and the terrain scattered interference signals from the main beam region. It is assumed that the sidelobe cancellation is sufficient to suppress the direct path jammer signals to about the receiver noise level.

The terrain scattered interference signals are modeled as the result of a convolution of the jammer direct path reference signals with the complex multipath impulse responses that describe the terrain scattering, wave propagation, antenna patterns etc. The impulse responses are time varying and contain delay dependent phase shifts due to Doppler effects in the bistatic airborne geometry. Thus the impulse responses are only valid over short time intervals for the deconvolution processing.

Typically the short time intervals are of the order of the pulse repetition interval of the radar. The time intervals must be chosen longer than the delay extent of the impulse responses to perform the deconvolution processing. But the time intervals must also be chosen sufficiently short to have approximately stationary terrain scattering conditions. The surface clutter signals of the considered airborne radar can often be larger than the terrain scattered interference signals into the main beam. It is assumed here that the main beam clutter signals can be suppressed to give clutter levels well below the terrain scattered interference level.

3 Spatial beamforming

Spatial beamforming is performed by adaptive processing of the received signal samples from the array antenna during the deconvolution processing time interval. The continuous jammer signals are received from fixed directions and the target and clutter signals are received from the fixed transmit beam direction during the short processing time interval. The antenna array covariance matrix is estimated using averaging over range samples, and together with steering vectors to the source directions the adaptive weight vectors can be defined. The spatial beamforming uses the sample covariance matrix estimate defined by

$$\hat{\boldsymbol{R}} = \frac{1}{T}\sum_{t=0}^{T} \boldsymbol{x}(t)\boldsymbol{x}^{H}(t)\,, \tag{1}$$

where $\boldsymbol{x}(t)$ is the received antenna array signal vector over the processing time interval $[0,T]$ in units of the sampling interval. The minimum variance beamforming weight vector is given by

$$\boldsymbol{w}(\theta) = \alpha\hat{\boldsymbol{R}}^{-1}\boldsymbol{v}(\theta)\,, \tag{2}$$

where $\boldsymbol{v}(\theta)$ is the steering vector to direction θ and α is a scaling factor for constant gain beamforming.

The covariance matrix estimate may be modified by diagonal loading and tapering to improve the sidelobe performance. The weight vector is applied to form the signal

$$y(\theta,t) = \boldsymbol{w}^H(\theta)\boldsymbol{x}(t) \tag{3}$$

for different directions to jammers and targets. It is assumed in the following that the sidelobe suppression is sufficient to suppress the direct path jammer signals and sidelobe terrain scattered signals to levels well below the main beam terrain scattered interference signals. Also, the sidelobe suppression is sufficient to give accurate direct path jammer signals.

4 Convolution framework

The convolution of the jammer signals with the complex multipath impulse responses must be developed for sampled continuous signals and time varying impulse responses. The processing interval is chosen to give approximately time invariant impulse responses and small Doppler induced phase shifts. The spatial beamforming generates one main beam signal and an auxiliary direct path signal for each jammer. The main beam signal samples are arranged in a main beam signal vector over the time interval $[0,T]$ as

$$\boldsymbol{m} = \boldsymbol{s} + \boldsymbol{r} + \boldsymbol{n}, \tag{4}$$

with the target signal vector $\boldsymbol{s} = (s_0,\ldots,s_T)^T$, the main beam terrain scattered interference signal vector $\boldsymbol{r} = (r_0,\ldots,r_T)^T$, and the receiver noise vector $\boldsymbol{n} = (n_0,\ldots,n_T)^T$. Using this notation the elements of the main beam signal vector are

$$\{m_i\}_0^T = \{y(\theta_m,i)\}_0^T, \tag{5}$$

for the main beam direction θ_m and time sample number i. Clutter signals are included in the target signal vector and residual direct path signals are included in the terrain scattered interference signal vector.

The direct path signal samples from a jammer are arranged in an auxiliary signal vector over the time interval $[-T,T]$ as

$$\boldsymbol{a} = \boldsymbol{d} + \boldsymbol{n}', \tag{6}$$

with the direct path signal vector $\boldsymbol{d} = (d_{-T},\ldots,d_0,\ldots,d_T)^T$ and the receiver noise vector $\boldsymbol{n}' = (n'_{-T},\ldots,n'_0,\ldots,n'_T)^T$. The elements of the auxiliary signal vector are

$$\{a_i\}_{-T}^T = \{y(\theta_a,i)\}_{-T}^T, \tag{7}$$

for the jammer direction θ_a and time sample number i.

The specular or near specular terrain scattered interference signals in the sidelobe region are not included in this formulation. Also, other residual jammer signals or main beam signals are not included. There may be some correlation between the main beam signal vector and the auxiliary signal vector due to nonorthogonal spatial beamforming weight vectors.

The convolution of the jammer signals with the complex multipath impulse responses is characterized by time causal and time invariant impulse responses. The impulse responses are referenced to a time origin at the start of the processing interval. The time delay extent of the impulse responses sets the required number of negative time samples in the direct path signal vector. This also sets the required number of positive time samples in the main beam signal vector.

The number of positive time samples must be equal to or larger than the number of negative time samples to avoid underdetermined deconvolution processing. In the following the positive and the negative time intervals are chosen equal as in the auxiliary signal vector definition (7). The convolution is computed using a Toeplitz matrix formulation to accommodate the continuous jammer signal. The direct path signal Toeplitz matrix is defined by

$$\boldsymbol{D} = \begin{bmatrix} d_0 & & d_{-T} \\ & \ddots & \\ d_T & & d_0 \end{bmatrix}, \tag{8}$$

and the multipath terrain scattered interference signal vector is given by the matrix-vector convolution product

$$\boldsymbol{r} = \boldsymbol{D}\boldsymbol{h}, \tag{9}$$

with the multipath impulse response vector $\boldsymbol{h} = (h_0,\ldots,h_T)^T$. The multipath impulse response includes effects from terrain scattering, wave propagation, antenna patterns, radar and jammer platform motions and receiver bandwidth.

The multipath impulse response vector is different for each jammer and for several jammers there will be a sum of terrain scattered interference signal vectors. It is assumed that most of the terrain scattering is from the main beam reflection area such that possible Doppler shifts are coupled to the multipath time delay. It is also assumed that in typical scenarios, the impulse response samples are only significantly different from zero for a limited number of samples.

5 Deconvolution processing

The deconvolution processing is developed for the case of a single jammer and a given main beam direction during a short time interval. The signals are described in vector-matrix notation of time samples over the time intervals $[-T,T]$ and $[0,T]$ as in the previous section. The main beam signal vector (4) is formed by adaptive

beamforming over the time interval $[0,T]$ and the auxiliary signal vector (6) is formed by adaptive beamforming over the time interval $[-T,T]$. The auxiliary signal Toeplitz matrix is defined by

$$\boldsymbol{A} = \begin{bmatrix} a_0 & & a_{-T} \\ & \ddots & \\ a_T & & a_0 \end{bmatrix}, \tag{10}$$

and the time origin is chosen such that the matrix is invertible and nonsingular. This means that the time origin auxiliary sample on the main diagonal must not be close to zero and the other auxiliary signal samples must not be too correlated to have parts of the signal spectrum close to zero.

The deconvolution processing is described using the Toeplitz matrix formulation. Formally, the auxiliary signal Toeplitz matrix is inverted and an estimate of the multipath impulse response vector is defined by

$$\hat{\boldsymbol{h}} = \boldsymbol{A}^{-1}\boldsymbol{m}, \tag{11}$$

but in practice the corresponding matrix-vector equation is solved using a suitable linear equations algorithm. Recent structured estimation algorithms that might give smaller errors have not been pursued here, see for example [2].

The estimated impulse response vector is then modified by a threshold operation to keep only the larger samples of the impulse response and setting the rest to zero as

$$\tilde{\boldsymbol{h}} = \hat{\boldsymbol{h}} \odot \boldsymbol{t} \tag{12}$$

with the threshold vector $\boldsymbol{t} = (t_0,\ldots,t_T)^T$ and element wise multiplication. The threshold vector is defined by

$$\{t_i\}_0^T = \left\{ \left|\hat{h}_i\right| > \eta_i \right\}_0^T \in \{0,1\}, \tag{13}$$

where the threshold η_i can be constant or adaptive over the processing time interval. The modified impulse response vector is then convolved with the auxiliary signal to give a mitigated main beam signal vector as

$$\tilde{\boldsymbol{m}} = \boldsymbol{m} - \boldsymbol{A}\tilde{\boldsymbol{h}}, \tag{14}$$

which is the end result of the deconvolution processing.

The impulse response vector estimate is modified by the threshold operation to minimize the residual multipath interference signal without too much signal cancellation. Usually, the impulse response vector has only a few large elements over the processed time interval and ideally most of these are extracted in the modified impulse response. The remaining number of elements are set to zero and this number sets the signal cancellation loss. The deconvolution can also be done iteratively by choosing only a few of the largest elements of the impulse response estimate in each iteration.

The threshold setting must be optimised to maximize the target signal to noise ratio after the deconvolution processing. The target signal vector and the noise vectors will contribute to the estimation errors in (11). For white noise jammer waveforms the inverse of the auxiliary matrix (10) will differ from the inverse of the direct path matrix (8) to give an error noise level in (11) corresponding to the direct path signal to noise ratio in (6). The target signal and noise in (4) will give an error noise level corresponding to multipath signal to target signal and noise ratio. The total error signal can be approximated by a white noise signal vector.

The threshold setting in (13) can be determined by the error noise level estimate or an adaptive threshold. The adaptive threshold can be computed from impulse response samples at large time delays that are free of multipath. The adaptive threshold can also be adapted as the inverse power level of the estimated impulse response vector. The deconvolution processing involves a Toeplitz matrix-vector equation and a Toeplitz matrix multiply that requires fast algorithms for real-time operation. This can be performed by Fourier transform methods for Toeplitz systems [3]. Doppler shifts can be accommodated by element wise multiplication of the Toeplitz matrices by Vandermonde matrices describing the phase progression for different elements in the multipath impulse response. Compensation for Doppler shifts might be used if the jammer platform can be tracked to estimate its velocity vector. Estimation of multipath impulse responses has also been studied in other applications, see for example [4].

6 Limitations

Several limitations of the proposed method can be seen. The method requires coherent waveforms from the jammer antenna in the directions to the radar and to the terrain scattering areas. The auxiliary direct path signal must not contain too much strong diffuse terrain scattering or large time delays of strong specular scattering. The multipath impulse must not span too large time delays compared to the processed time interval. Only the terrain scattering interference from the largest impulse response elements above the threshold are mitigated. Small relative Doppler shifts or known relative Doppler shifts are required to keep the phase errors limited in the processed time interval. Several equally strong terrain scattering interference contributions from different jammers cannot be mitigated. Strong clutter signals comparable to the terrain scattering interference limit the performance of the

impulse response estimation. Pulsed jammer signals can give near singular Toeplitz matrices and inaccurate impulse response estimates. Receiver blanking on transmit gives missing samples in the auxiliary direct path signal and inaccurate impulse response estimates.

7 Simulations

Simulations of the terrain scattered interference mitigation method are shown in following figures. All figures show power in dB for 100 element vectors. The receiver noise level is 0 dB and the direct path jammer power is 60 dB. A target of 30 dB power is at time sample 50. The impulse response is a simple delay of 10 time samples with 30 dB attenuation. This gives a main beam multipath power of 30 dB. Fig. 2 shows the direct path signal and the terrain scattered signal. Fig. 3 shows the estimated impulse response. Fig. 4 shows the main beam signal before and after deconvolution processing. The terrain scattered interference has been suppressed by about 30 dB.

8 Conclusions

A terrain scattered interference mitigation method has been proposed that can obtain 20-30 dB jammer suppression in typical scenarios. The method is based on deconvolution processing and requires accurate estimation of the multipath impulse response.

References

[1] D.J. Rabideau. Clutter and jammer multipath cancellation in airborne adaptive radar. *IEEE Transactions on Aerospace & Electronic Systems*, vol. 36, no. 2, pp. 565-583, April 2000.

[2] S. Van Huffel, H. Park, J.B. Rosen. Formulation and solution of structured total least norm problems for parameter estimation. *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2464-2474, October 1996.

[3] A.E. Yagle. A fast algorithm for Toeplitz-block-Toeplitz linear systems. *Proceedings of the IEEE 2001 International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. IMDSP-P5.4.

[4] M.D. Hahm, Z.I. Mitrovski, E.L. Titlebaum. Deconvolution in the presence of Doppler with application to specular multipath parameter estimation. *IEEE Transactions on Signal Processing*, vol. 45, no. 9, pp. 2203-2219, September 1997.

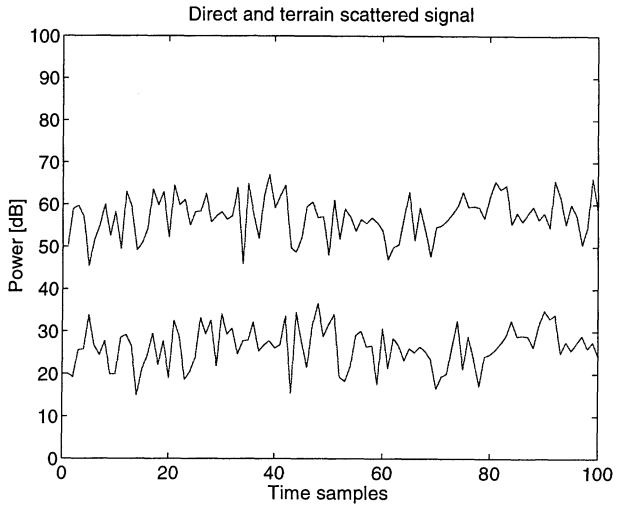


Figure 2: *Direct path and multipath signals.*

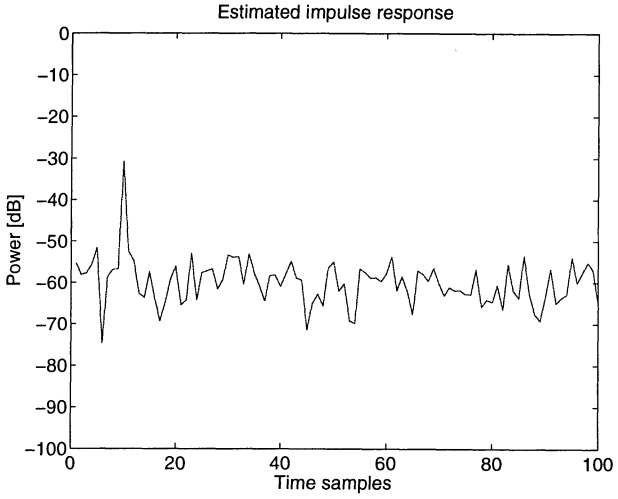


Figure 3: *Estimated impulse response.*

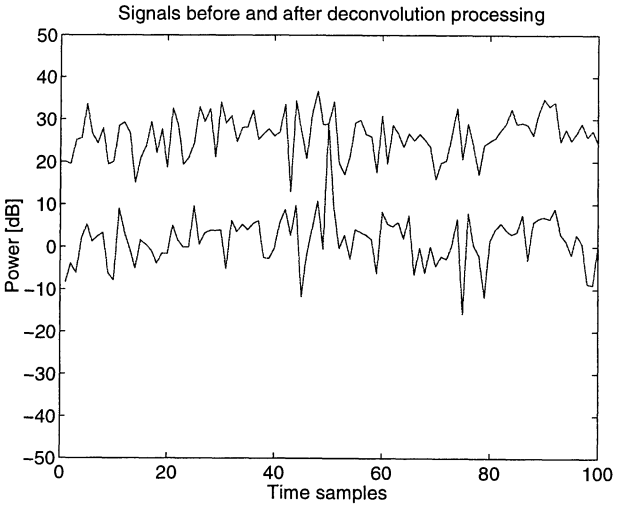


Figure 4: *Main beam signals before and after deconvolution processing.*

Direction Finding for Electronic Warfare Systems Using the Phase of the Cross Spectral Density

Johan Falk[1,2,3], Peter Händel[1,2] and Magnus Jansson[2]

[1]Department of Electronic Warfare Systems, Swedish Defence Research Agency, Linköping, Sweden

[2]Department of Signals, Sensors and Systems, Royal Institute of Technology, Stockholm, Sweden

[3]E-mail: johan.falk@foi.se

Abstract

In modern electronic warfare systems there is a need for direction-finding of transmitters using waveforms for military stealth communication. In this paper, a correlation-based method is investigated utilizing the phase of the cross spectral density to estimate the time-difference-of-arrival from a two-channel digital receiver. A least squares method is reviewed, and its performance is investigated by theoretical analysis and by Monte-Carlo simulations. Proper Cramér-Rao bounds are derived. It is shown that the method is statistically efficient for flat spectrum signals. The method is found to be a promising method for use against military communication in an electronic warfare direction-finding system.

1 Introduction

Electronic warfare (EW) systems for use against military communication sources include direction-finding (DF). New military tactical communication systems use spread-spectrum with filtering to achieve stealth performance, that is, low probability of intercept (LPI). Traditional DF-systems have in general poor performance against these types of signals while correlation-based time-difference-of-arrival (TDOA) DF-systems perform well on wideband signals with low SNR.

In traditional DF-systems based on Watson-Watt, or subspace techniques, antenna arrays with at least 4 elements are used. To obtain smaller, cheaper and mobile EW units, TDOA-based DF-systems are considered. Early work on TDOA-estimation was focused on submarine EW, that is passive sonar array systems [1, 2]. The two main differences between sonar and radio communication TDOA-estimation are the propagation speed (approximately $1.1 \cdot 10^3$ m/s versus $3 \cdot 10^8$ m/s) and the signal characteristics in terms of bandwidth.

An introduction to correlation-based TDOA-estimation can be found in [3]. Two main approaches for correlation-based TDOA-estimation are proposed, time- and frequency-domain based estimators. Most papers published consider estimators in time-domain where the main idea is to find the maximum of the cross-correlation function (CCF) or a weighted version of the CCF [4]. For EW applications, in a multiple signal environment, this is not a tractable method due to the need for frequency filtering of the received signals before TDOA-estimation. When using a frequency-domain based estimator simple frequency filtering and direction filtering [5] can be done to increase the output SNR by supressing signals in all but one angular sector. The beamwidth may be very narrow for large signal bandwidths despite the fact that we only have two receiver sensors. The beamwidth is proportional to the inverse of the signal bandwidth. This is a highly tractable feature in military EW systems due to the high probability of interfering signals and close-by high-power jammers.

To summarize, modern military stealth communication systems utilize short duration LPI waveforms and energy directed transmissions. That is, an EW DF-system needs to perform well on wideband, low SNR, short duration signals with unknown characteristics and origin, while being small, cheap and mobile.

The method presented in this paper is a frequency-domain-based method to estimate the TDOA based on correlation and the phase of the cross-spectral-density.

2 Data model

The transmitted signal is received by two sensors whose outputs, after quadrature mixing, contain two noisy and differently delayed versions of the complex-valued baseband signal $s(t)$. The transmitted signal is assumed to be unknown and accordingly $s(t)$ is modelled as a zero-mean wide-sense stationary process, characterized by its autocorrelation function $\phi_s(\tau)$. Furthermore, it is assumed that $s(t)$ is strictly band-limited into the frequency range $(-B, B)$ Hz and that its power spectral density is continuous in frequency f, that is the transmitted signal is assumed to be strictly band-limited, but broadband. The two chan-

nels in the digital receivers are sampled with sampling frequency f_s Hz, such that $f_s \geq 2B$. Without loss of generality, we let $f_s = 1$ Hz in the forthcoming discussions.

Under an assumption of perfect receiver synchronization, the complex-valued output from two synchronous digital receivers are modelled as

$$\begin{aligned} z_1[k] &= s(t)|_{t=k} + n_1[k] \\ z_2[k] &= s(t)|_{t=k-\Delta} + n_2[k] \end{aligned} \tag{1}$$

where Δ denotes the unknown normalized delay and the actual delay is given by Δ/f_s. The noise terms $n_p[k]$ $(p = 1, 2)$ are assumed zero-mean white (temporal and spatial) circular complex Gaussian with variance σ_p^2, respectively.

Defining the CCF for the received signals as

$$\phi[m] = \mathrm{E}(z_1[k+m] z_2^*[k]) \tag{2}$$

where * denotes complex-conjugate. Then, since $n_1[k]$ and $n_2[k]$ are uncorrelated and $s(t)$ is broadband

$$\phi[m] = \phi_s(\tau)|_{\tau=m+\Delta} \tag{3}$$

where $\phi_s(\tau) = \mathrm{E}[s(t+\tau) s^*(t)]$ is the autocorrelation function of the baseband source signal.

An established method to estimate the unknown delay Δ is to find the argument that maximizes the estimated CCF, say $\hat{\phi}[m]$, followed by some interpolation to find the delay with sub-bin accuracy. In particular, triple parabolic interpolation for time-delay estimation is considered in [6]. Alternatively, the delay can be estimated from the spectral representation of the CCF. The cross spectral density (CSD) is defined by the time discrete Fourier transform of the CCF

$$\Phi(\omega) = \mathcal{F}_d\{\phi[m]\} = \mathcal{F}_d\{\phi_s(\tau)|_{\tau=m+\Delta}\} \tag{4}$$

where $\mathcal{F}_d\{\cdot\}$ denotes the time discrete Fourier transform. A translation in the time domain corresponds to a rotation in the frequency domain, that is

$$\Phi(\omega) = e^{j\omega\Delta}\mathcal{F}_d\{\phi_s(\tau)|_{\tau=m}\} = e^{j\omega\Delta}\Phi_s(\omega) \tag{5}$$

where $\Phi_s(\omega)$ is the power spectral density of the sampled version of the baseband signal $s(t)$. We note that the model (5) is valid for a non-dispersive propagation model.

The direct correlator estimator by peak picking the CCF followed by parabolic interpolation is not statistically efficient at high SNR [6]. Accordingly, alternative methods for the sub-bin search is required. In this paper, we rely on time delay estimation using phase data [2]. We note from (5) that

$$\Gamma(\omega) = \arg\Phi(\omega) = \omega\Delta \tag{6}$$

Now, estimating the unknown Δ in the time domain by maximization of the magnitude of the CCF has been transformed into fitting a straight line to the argument of the CSD. In Figure 1, the linear portion of the phase is used to estimate Δ.

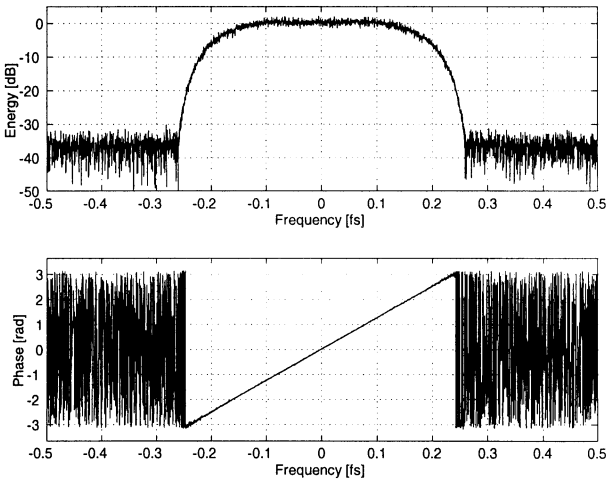


Figure 1: Typical energy-phase representation of the CSD obtained in an EW system.

3 Properties of the sample CSD

Consider the finite length discrete signals $z_1[k]$ and $z_2[k]$ as

$$\{z_1[0], ..., z_1[N-1]\} \tag{7}$$

and

$$\{z_2[0], ..., z_2[N-1]\} \tag{8}$$

where $z_p[k] = 0$ for $k < 0$ or $k \geq N$ $(p = 1, 2)$. Replace the CCF in (2) with the sample expression

$$\hat{\phi}[m] = \begin{cases} \frac{1}{N}\sum_{k=0}^{N-1} z_1[k+m] z_2^*[k] & m = 1-N, ..., N-1 \\ 0 & |m| > N-1 \end{cases} \tag{9}$$

The estimated CSD is now given by the time discrete Fourier transform of $\hat{\phi}[m]$. We have

$$\begin{aligned} \hat{\Phi}(\omega) &= \mathcal{F}_d\left\{\hat{\phi}[m]\right\} \\ &= \sum_{m=1-N}^{N-1} \hat{\phi}[m] e^{-j\omega m} \\ &= \frac{1}{N}\sum_{m=1-N}^{N-1}\sum_{k=1-N}^{N-1} z_1[k+m] z_2^*[k] e^{-j\omega m} \end{aligned} \tag{10}$$

where (9) with zero padding was used in the last equality. Let $p = k + m$, then

$$\begin{aligned} \hat{\Phi}(\omega) &= \frac{1}{N}\left(\sum_{k=1-N}^{N-1} z_2^*[k] e^{j\omega k} \sum_{p=k+1-N}^{k+N-1} z_1[p] e^{-j\omega p}\right) \\ &= \frac{1}{N}\left(\sum_{k=1-N}^{N-1} z_2^*[k] e^{j\omega k}\right)\left(\sum_{p=1-N}^{N-1} z_1[p] e^{-j\omega p}\right) \end{aligned} \tag{11}$$

Accordingly,

$$\begin{aligned}\hat{\Phi}(\omega) &= \frac{1}{N}\mathcal{F}_d\{z_1[m]\}\mathcal{F}_d^*\{z_2[m]\} \\ &= e^{j\omega\Delta}\Phi_s(\omega) + V(\omega) \end{aligned} \quad (12)$$

where $V(\omega)$ is the error term describing imperfections due to finite sample effects and noise. Note the similarity between (5) and (12).

Utilizing the discrete Fourier transform on a regular grid in place of the $\mathcal{F}_d$-operator, we obtain

$$\hat{\Gamma}[k] = \frac{2\pi k\Delta}{M} + v[k] \quad (13)$$

where $M \geq 2N-1$ denotes the number of used frequency bins, and $v[k]$ is the error term describing imperfections due to finite samples effects and noise. With $M = 2N-1$, the frequency index k spans the interval $k = 1-N, ..., N-1$, where N is the length of the two sample records.

4 TDOA estimation

Due to the EW scenario, that is the assumption on lack of knowledge about the signal, we choose an estimator that does not require any probabilistic assumption on the signals. It is shown in [2, 4], that a linear least-squares estimator (LLSE) is statistically efficient for real-valued flat spectrum signals, but additional weighting is needed for non flat-spectrum signals. The proper frequency weighting function for signals with arbitrary spectrum is derived in [7].

From (13), we may constrain a first order polynomial model to have a zero bias term. The LLSE criterion becomes

$$J(\Delta) = \sum_{k=1-N}^{N-1}\left(\hat{\Gamma}[k] - \frac{2\pi k\Delta}{2N-1}\right)^2 \quad (14)$$

which is readily shown to be minimized by

$$\hat{\Delta}_{LLS} = \frac{3}{2\pi N(N-1)}\sum_{k=1-N}^{N-1} k\hat{\Gamma}[k] \quad (15)$$

A thorough derivation of (15) can be found in [8]. Of course the bandwidth of the signal, $(-B, B)$ Hz, must be known and here, $k = 1-N, \ldots, N-1$ denotes the number of frequency-bins allocated by the signal.

The stationary assumption on the signal yields phase values independent of one another [2, 9]. Hence the $\hat{\Gamma}[k]$'s are independent of one another. The variance of $\hat{\Gamma}[k]$ is [2]

$$\sigma^2\left(\hat{\Gamma}[k]\right) \approx \frac{1-C[k]}{C[k]} \quad (16)$$

where $C[k]$ is the discrete coherence function [10] and the approximation comes from a small errors assumption. The result slightly differs from the result in [2] due to the employed complex-valued data model. Now, the variance of (15) follows

$$\sigma^2\left(\hat{\Delta}_{LLS}\right) \approx \frac{9}{4\pi^2N^2(N-1)^2}\sum_{k=1-N}^{N-1}\frac{k^2(1-C[k])}{C[k]} \quad (17)$$

For flat spectrum signals and equal channel noise powers σ_n^2, we obtain

$$C[k] = \frac{\mathrm{SNR}^2}{1+2\mathrm{SNR}+\mathrm{SNR}^2} \quad (18)$$

where $\mathrm{SNR} = \sigma_s^2/\sigma_n^2$ with σ_s^2 being the signal power. Evaluating (17) yields

$$\sigma^2\left(\hat{\Delta}_{LLS}\right) \approx \frac{3(2N-1)}{4\pi^2N(N-1)}\frac{2\mathrm{SNR}+1}{\mathrm{SNR}^2} \quad (19)$$

5 Cramér-Rao bound

The Cramér-Rao bound (CRB) is given by [11]

$$\mathrm{CRB}[\Delta/f_s] = \frac{1}{8\pi^2T}\left[\int_{-\infty}^{\infty}\frac{f^2C(f)}{1-C(f)}df\right]^{-1} \quad (20)$$

where $T = N/f_s$ is the absolute observation time, and $C(f)$ is the coherence function. From [11],

$$C(f) = \frac{\Phi_s(f)^2}{\Phi_1(f)\Phi_2(f)} \quad (21)$$

In (21), $\Phi_s(f)$, $\Phi_1(f)$ and $\Phi_2(f)$ are the power spectral densities of the the continuous time signals $s(t)$, $z_1(t)$ and $z_2(t)$, respectively. We assume that $s(t)$, $n_1(t)$ and $n_2(t)$ are all strictly band-limited to the frequency range $(-B, B)$ Hz. Without loss of generality we may assume Nyquist sampling, that is $B = f_s/2$. The result (20) differs by a factor $1/2$ compared to the expression found in [11] due to the employed complex-valued signal model.

5.1 CRB for flat spectrum signals

Assuming flat spectrum signals, that is

$$C(f) = \frac{\mathrm{SNR}^2}{1+2\mathrm{SNR}+\mathrm{SNR}^2} \quad (22)$$

Accordingly,

$$\begin{aligned}\mathrm{CRB}[\Delta/f_s] &= \frac{f_s}{8\pi^2N}\frac{2\mathrm{SNR}+1}{\mathrm{SNR}^2}\left[\int_{-B}^{B}f^2df\right]^{-1} \\ &= \frac{3f_s}{16\pi^2NB^3}\frac{2\mathrm{SNR}+1}{\mathrm{SNR}^2} \\ &= \frac{3}{8\pi^2NB^2}\frac{2\mathrm{SNR}+1}{\mathrm{SNR}^2} \end{aligned} \quad (23)$$

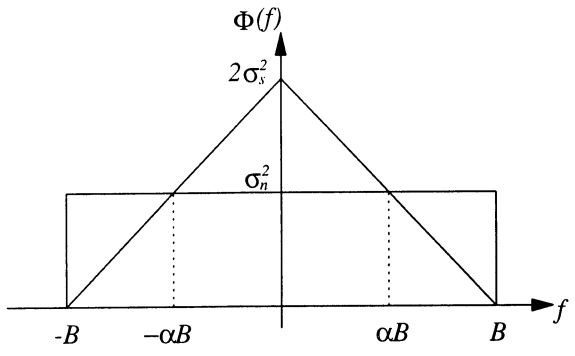


Figure 2: Power spectral density of source and noise.

where $B = f_s/2$ was used in the last equality. Proper scaling gives

$$\mathrm{CRB}\,[\Delta] = \frac{3}{2\pi^2 N}\frac{2\mathrm{SNR}+1}{\mathrm{SNR}^2} \tag{24}$$

The result (24) forms a lower bound on the performance of any unbiased time delay estimator for flat spectrum signals. Note that for large N, (19) equals (24), and thus the LLSE is asymtotically efficient for flat spectrum signals.

5.2 CRB for triangular spectrum

Considering equal channel noise with a flat power spectral density σ_n^2, we define a frequency dependent SNR according to

$$\mathrm{SNR}(f) = \frac{\Phi_s(f)}{\sigma_n^2} \quad |f| \le B \tag{25}$$

Inserting (25) into (20) yields, for $T = N/f_s$

$$\mathrm{CRB}\,[\Delta/f_s] = \frac{f_s}{8\pi^2 N}\left[\int_{-B}^{B}\frac{\mathrm{SNR}^2(f)}{1+2\mathrm{SNR}(f)}f^2 df\right]^{-1} \tag{26}$$

The major contribution to the CRB is given by the frequency regions with high SNR, that is for $\mathrm{SNR}(f) \gg 1$. Thus, the CRB in (26) may be approximated by

$$\mathrm{CRB}\,[\Delta/f_s] \approx \frac{f_s}{4\pi^2 N}\left[\int_W \mathrm{SNR}(f) f^2 df\right]^{-1} \tag{27}$$

where W denotes the high-SNR frequency regions. Consider a signal with power spectral density according to Figure 2, where $\mathrm{SNR} = \sigma_s^2/\sigma_n^2$ is the signal to noise ratio within the the full bandwidth B. Then

$$\mathrm{SNR}(f) = 2\mathrm{SNR}\left(1 - \frac{|f|}{B}\right) \quad |f| \le B \tag{28}$$

The high-SNR frequency region is characterized by $W \in (-\alpha B, \alpha B)$ for some α in the interval $0 < \alpha < 1$.

Inserting (28) into (27) gives

$$\begin{aligned}\mathrm{CRB}\,[\Delta/f_s] &\approx \frac{f_s}{16\pi^2 N\mathrm{SNR}}\left[\int_0^{\alpha B} f^2\left(1-\frac{f}{B}\right)df\right]^{-1} \\ &= \frac{3}{2\pi^2 N\mathrm{SNR}\alpha^3 B^2(4-3\alpha)}\end{aligned} \tag{29}$$

where $B = f_s/2$ was used in the last equality. Proper scaling gives the CRB for the normalized delay Δ

$$\mathrm{CRB}\,[\Delta] \approx \frac{6}{\pi^2 N\mathrm{SNR}\alpha^3(4-3\alpha)} \tag{30}$$

Comparing (24) with (30), we note that the latter result is at least twice as big as the first one. We may define the high-SNR region by α given by the line crossing $\Phi_s(\alpha B) = \sigma_n^2$, that is $\alpha = 1 - 1/2\mathrm{SNR}$. The approximation (30) is quite accurate, as illustrated in Figure 3.

6 Practical considerations

For long delays, the phase unwrapping is problematic and leads to loss in performance due to erroneous estimates. One natural approach is to perform a first initial estimate by peak-picking the magnitude of the CCF and precompensate the data according to this crude estimate. Then, a delay estimate with sub-bin accuracy is obtained by applying the method described above on the precompensated data. The final time delay estimate is obtained as the sum of the crude estimate from the CCF method and the correction obtained from the least-squares fit of the phase of the CSD.

The amount of erroneous estimates can be reduced by averaging the CSD. Here, for simplicity, we average the CSD without overlap, that is based on N samples the CSD is calculated in p disjoint intervals based on N/p measurements. Increasing p lowers the SNR threshold, that is the threshold that appears in non-linear estimation problems. On the other hand, an increased p-value reduces the operating range of the estimator, that is doubling p implies that the operating range of the estimator is reduced by half.

7 Numerical example

Flat spectrum signals are considered. Based on 500 independent simulation runs, the performance of the least-squares fit of the phase of the CSD is investigated. Gaussian data were generated with an integer delay of $\Delta = 3$ samples. A block length of $N = 200$ samples was considered. The mean-square-error (MSE) as function of SNR is depicted in Figure 3. An initial integer-delay estimate was obtained by peak-picking the CCF. The correction was obtained

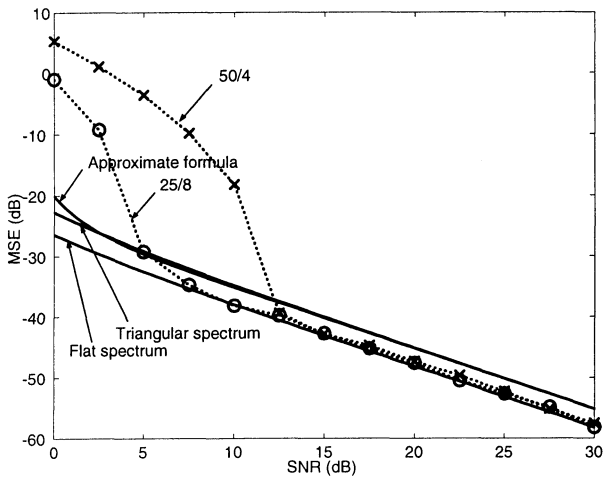


Figure 3: Mean square error as function of SNR for $p = 4$ and $p = 8$, respectively. The Cramér-Rao bound for flat spectrum and triangular spectrum signals are given as reference.

by the least-squares fit (15). A sub-block length of 50 and 25 samples were used, that is $p = 4$ and $p = 8$, respectively. The CSD in each sub-block was calculated using the discrete Fourier transform of the estimated CCF in (9), with some additional zero-padding.

The results are displayed in Figure 3. From the depicted curves we observe that 25/8 (i.e. a sub-block length of 25 samples and eight times averaging of the CSD) has a lower SNR threshold compared to 50/4. The performance above the threshold is similar for both methods, that is they produce estimates with an MSE close to the CRB. The CRB for the considered scenario is also included, given by (24). Further, the triangular spectrum CRB's are included as reference. We note the close correspondence between the exact expression (26) and the approximate (29).

8 Conclusions

Direction-finding in an EW scenario is considered and correlation-based TDOA estimators using the phase of the CSD is analyzed. The results using full-bandwidth flat-spectrum signals show that the considered estimator is asymptotically efficient. An approximate CRB for a triangular spectrum signal is derived. The triangular spectrum is a rough model of an LPI waveform using a roll-off filter. At least a 3 dB loss is encountered compared to full-bandwidth flat-spectrum signals. This method shows promising results for use in an EW system with direction-finding capabilities.

This is a paper based on a preliminary study on correlation-based TDOA in an EW scenario. Future work includes analysis of performance degradation due to the time and frequency synchronization errors between the receiver channels.

References

[1] Carter G.C., TDOA delay estimation for passive sonar signal processing, *IEEE Transactions on Acoustics, Speech, Signal Processing*, June 1981, vol. 29, no. 3, pt. 2, pp. 463-470.

[2] Piersol A. G., Time delay estimation using phase data, *IEEE Transactions on Acoustics, Speech, Signal Processing*, June 1981, vol. 29, no. 3, pt. 2, pp. 471-477.

[3] Carter G.C. [Editor], *Coherence and time delay estimation*, IEEE Press, ISBN: 0-7803-1006-3, 1993.

[4] Knapp C.H., Carter G.C., The generalized correlation method for estimation of time delay, *IEEE Transactions on Acoustics, Speech, Signal Processing*, Aug. 1976, vol. 24, no. 4, pp. 320-327.

[5] Houghton A.W. and Reeve C.D., Direction finding on spread-spectrum signals using the time-domain filtered cross spectral density, *IEE Proc. Radar, Sonar and Navigation*, Dec. 1997, vol. 144, no. 6, pp. 315-320.

[6] Jacovitti G., Scarano G., Discrete time techniques for time delay estimation, *IEEE Transactions on Signal Processing*, Feb. 1993, vol. 41, no. 2, pp. 525-533.

[7] Zhao Zhen, Hou Zi-qiang, The generalized phase spectrum method for time delay estimation, *Proc. ICASSP '84*, March 1984, vol. 3, pp. 46.2/1-4.

[8] Kay S.M., *Fundamentals of Statistical Signal Processing - Estimation Theory*, Prentice Hall, NJ, 1993.

[9] Bendat J.S., Statistical errors in measurement of coherence functions and input/output quantities, *J. Sound and Vibration*, 1978, vol. 59, no. 3, pp. 405-421.

[10] Carter G. C., Knapp C.H., Nuttall A.H., Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing, *IEEE Trans. Audio and Electroacoustics*, Aug. 1983, vol. 21, no. 4, pp. 337-344.

[11] Carter G.C., Coherence and time delay estimation, *Proceedings of the IEEE*, Feb 1987, vol. 75, no. 2, pp. 236-255.

Mapping performance of curved-path SAR

Sune R.J. Axelsson

Saab Bofors Dynamics
Swedish Defence Research Agency
S-581 88 Linköping, Sweden
Email: Sune.Axelsson@dynamics.saab.se

Abstract

In some reconnaissance applications, large areas are to be mapped in short time. The mapping rate can be increased significantly by slowly sweeping the antenna, and furthermore by using a curved SAR-path. In this paper, curved SAR-mapping is analysed in more detail, including highly squinted geometry. Derived relationships show how resolution performance and mapping rate are influenced by the curved SAR-path, and simulation results display how the phase error depends on side acceleration and scene geometry.

1 Introduction

Conventional SAR assumes a straight flight path during the SAR data sampling. At strip mapping, the antenna beam is pointing in a fixed direction. In high-resolution Spot-mode, the main beam direction of the antenna is locked to the scene centre during data collection.

In some reconnaissance applications, the available time for mapping is strongly reduced, or large areas are to be mapped in short time. The mapping rate can be increased significantly by slowly sweeping the antenna. The useful angle interval is limited, however, as a result of the degraded resolution when the antenna beam is approaching the velocity vector. In forward-looking squinted mapping geometry, the useful sweep interval is limited to 20-30 degrees.

In this paper, curved SAR-mapping is analysed in more detail including forward look geometry. Relationships are shown how resolution performance and mapping rate are affected as a result of the curved SAR-path. Simulation results are presented showing how the phase error depends on side acceleration and scene geometry.

2 Scene Centre Rotation

2.1 Straight SAR-path

At SAR-mapping along a straight line, the angular rotation of the scene centre in a platform fixed coordinate system is:

$$\omega_0 = V\sin(\theta)/R \qquad (1)$$

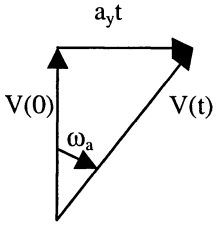
where **V** is the velocity vector and θ is the squint angle of the target measured from **V**.
Example 1: V=300 m/s, R=10 km and θ=30 degrees gives ω=15 mrad/s.

2.2 Curved SAR-path

For curved SAR-mapping an additional rotation is formed which for high side-accelerations becomes dominant.:

$$\omega_a = a_y/V \qquad (2)$$

Example 2: V=300 m/s, a_y=2g give ω_a = 67 mrad/s (4 degrees/s). For a_y=5g, ω_a = 167 mrad/s (10 degrees per second).



3 Resolution for curved SAR

The additional scene rotation reduces the time the reflectors are inside the main beam of the antenna and that put a limit on the maximum possible SAR-aperture.

Assume that the rotation rate of the target reflector in the antenna beam co-ordinate system is ω. The dwell time of the reflector being inside the main beam is then limited by

$$\tau = \theta_A/\omega \qquad (3)$$

The available time of measurement defines a lower limit on the Doppler bandwidth of the SAR-processing, which can be written as

$$B_d = 1/\tau = \omega/\theta_A \qquad (4)$$

The angular dependence of the Doppler frequency shift is

$$f_d = 2V\cos(\theta)/\lambda \qquad (5)$$

where θ is the angle between the velocity vector and the direction to the point reflector.

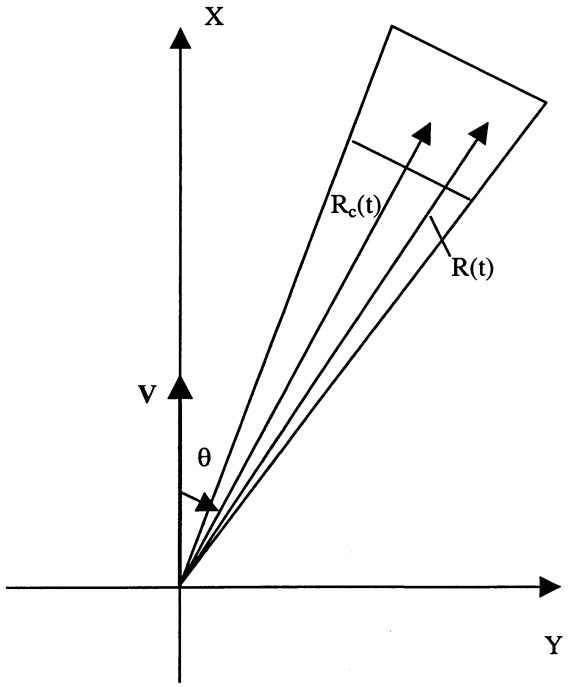


Figure 1: SAR-path geometry.

Differentiation of (5) defines the relationship between synthetic aperture beam width and Doppler resolution

$$\Delta f_d = -[2V\sin(\theta)/\lambda]\Delta\theta \quad (6)$$

Substitution of $\Delta f_d = B_d = \omega/\theta_A$ gives

$$\Delta\theta = \omega\lambda/[2V\theta_A\sin(\theta)] \quad (7)$$

At range R, the resolution limit becomes

$$\Delta R = R\Delta\theta = \omega\lambda R/[2V\theta_A\sin(\theta)] \quad (8)$$

3.1 SAR in straight path

Insertion of the rotation rate of the scene centre

$$\omega_0 = V\sin(\theta)/R \quad (9)$$

into (8) gives the resolution limit

$$\Delta R = \lambda/[2\theta_A] \quad (10)$$

That relationship also shows that ΔR is independent of range and squint angle.
Example 3: For θ_A=1/12 rad (5 degrees) and λ=0.02 m, the range resolution in azimuth becomes ΔR=0.12 m.

3.2 Curved SAR-path and sweeping antenna

Assume that the side acceleration generates the angular rotation ω_a and that the antenna sweeps with ω_A. From (8)

$$\Delta R = \lambda R(\omega_0+\omega_a+\omega_A)/[2V\theta_A\sin(\theta)] \quad (11)$$

Example 4: For a side acceleration of 5g at V=300 m/s), the feasible resolution becomes 0.75 m at 5 km for a Ku-band radar with θ=30 degrees and θ_A=5 degrees. For 10 and 20 km, the corresponding resolution is 1.5 m and 3 m, respectively.

By using a digital array antenna (DGA) and SAR-processing in several antenna channels simultaneously, the observation time can be expanded significantly reducing the above limitations in azimuth resolution for curved SAR-mapping.

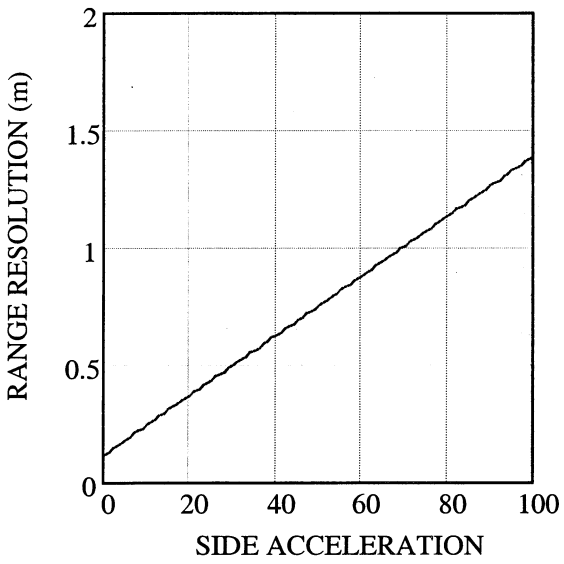


Figure 2: Lower limit of range resolution versus side acceleration (m/s^2) for θ_A=5 degrees, θ=30 degrees, R=5000 m, λ=0.02 m and V=300 m/s.

4 Mapping rate of curved SAR

For conventional strip mapping, the mapping rate is

$$MR0 = VW\sin(\theta) \quad (12)$$

where V is velocity, W is range coverage and θ is the antenna squint angle relative to the velocity vector.

If the angular movement of the reflector is ω (rad/s) due to a curved SAR-path and/or antenna sweeping, the mapping rate is approximately

$$MR = VW\sin(\theta) + \omega WR \qquad (13)$$

As a result the mapping rate increases a factor F given by

$$F = MR/MR0 = 1 + \omega R/[V\sin(\theta)] \qquad (14)$$

Example: For ω=0.17 rad/s (10 grader/sek), R=10 000 m, θ=30 degrees and V=300 m/s, the mapping rate increases about a ten factor compared with strip mapping with fixed antenna beam and straight SAR-path.

5 SAR with scene centre focusing

The Doppler signal from a specific resolution cell can be described as

$$S(t) = \exp[j4\pi R(t)/\lambda] = \exp[j2\pi f_d t + \theta(t)] \qquad (15)$$

where R(t) is the range to the resolution cell measured from the radar antenna, f_d is the Doppler frequency shift and $\theta(t)$ represents the non-linear phase term.

Scene centre focusing is performed by multiplying S(t) by the complex conjugate of the signal from the scene centre

$$S_c(t) = \exp[-j4\pi R_c(t)/\lambda] = \exp[-j2\pi f_{dc} t - \theta_c(t)] \qquad (16)$$

The results is with $\Delta R(t)=R(t)-R_c(t)$ substituted

$$S_1(t) = \exp[j4\pi \Delta R(t)/\lambda] = \exp[j2\pi \Delta f_d t + \Delta\theta(t)] \qquad (17)$$

The linear Doppler term $\Delta f_d = f_d - f_{dc}$ defines the position of the reflector in azimuth and $\Delta\theta(t) = \theta(t) - \theta_c(t)$ is the remaining part of the non-linear phase term, which has been highly reduced as a result of the scene centre correction.

In the scene centre, $\Delta\theta(t)=0$ but increases towards the edges of the SAR-imagery, in particular, when a side acceleration is applied. This is displayed by the simulated SAR-images of Figures 3-4, showing three point reflectors positioned at a distance of 14, 140 and 280 m from the scene centre, which is defined by the lower, left corner (only a quarter of the full 512x512 image is shown). The antenna beam was squinted 30 degrees to the right from the velocity vector.

In Figure 3, the SAR-path is turning right with a_y=20 m/s^2. From (1) and (2), we find ω_a =-67 mrad/s and ω_0 = 30 mrad/s, giving a total scene centre rotation of 37 mrad/s.

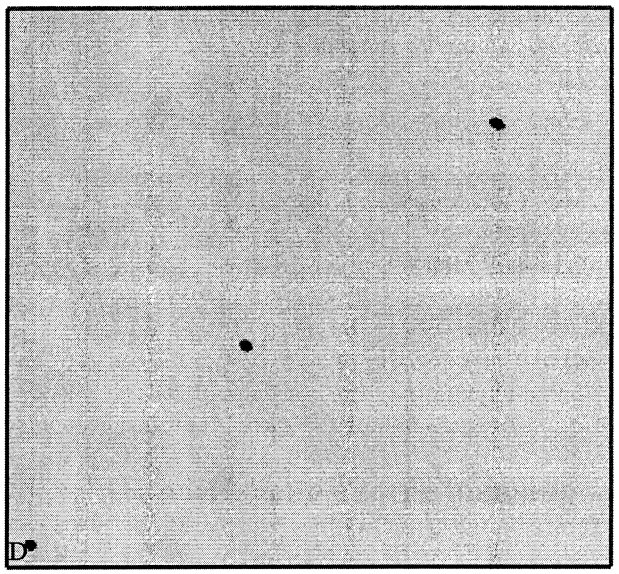


Figure 3: SAR-mapping of three point reflectors to the right with θ=30 degrees, R_c= 5 km, v=300 m/s, ΔR=1 m, λ=0.02 m. SAR-path turning right with a_y= 20 m/s^2.

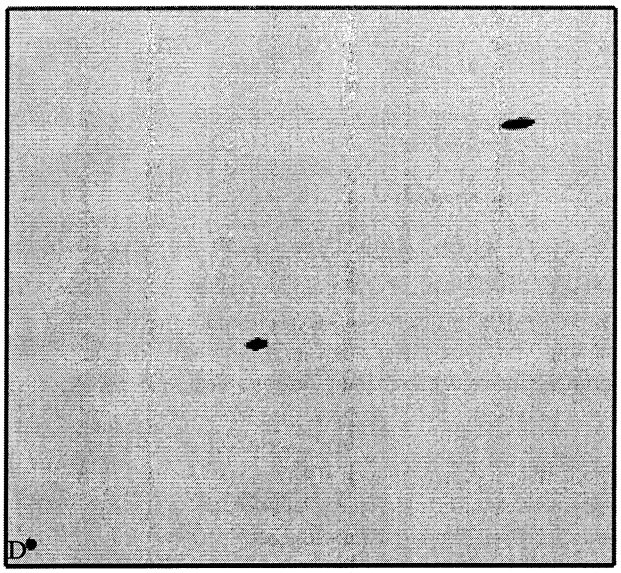


Figure 4: SAR-mapping of three point reflectors to the right with θ=30 degrees, R_c= 5 km, v=300 m/s, ΔR=1 m, λ=0.02 m. SAR-path turning left with a_y= -20 m/s^2.

Figure 4 represents a turn to the left (a_y=-20 m/s^2) giving a scene centre rotation of 97 mrad/s with increased defocusing and mapping rate as a result.

6 Estimated phase error

For perfect scene centre focusing, the non-linear phase error is zeroing in the centre of the image but increases towards the edges, which decreases the image resolution. A phase error of $\pi/2$ gives a slight broadening of the resolution element, and the side-lobe suppression starts degrading [1]. The non-linear phase error was simulated for varying side-acceleration, target range, SAR-image size and resolution.

For the path parameters of Figures 3-4, we find that the non-linear phase error of the top reflector in Figure 3 is about 3 radians (a_y= 20 m/s^2), giving a slight defocusing as a result. But in Figure 4 (a_y= -20 m/s^2), three times higher scene angle rotation is obtained. The non-linear phase variation of the top reflector is then 10 radians over the aperture, explaining the defocusing that occurs. See Fig. 5.

Concluding, a curved path to the left (a_y<0) gives increased image distortion as a result of enhanced image rotation rate and phase distortion, but it highly improves the mapping rate.

A curved path to the right decreases the defocusing and mapping rate. In particular, when the curve radius is close to the target range, a true Spot-SAR geometry occurs.

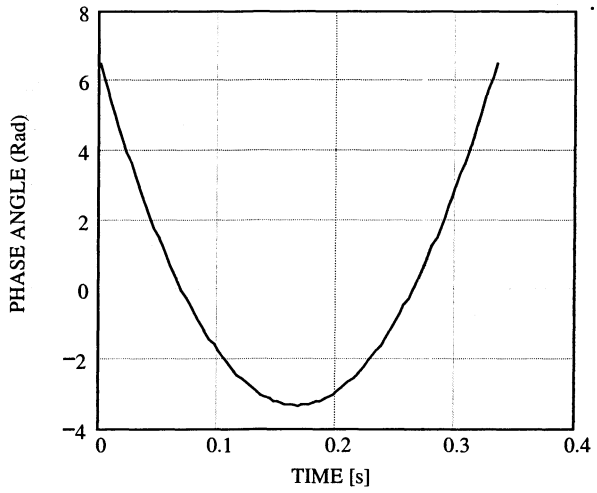


Figure 5: Non-linear phase error $\Delta\theta(t)$ of the defocused top reflector of Figure 4.

7 Concluding remarks

Normally SAR-mapping is performed along a straight line path. Curved SAR-mapping, however, increases the mapping rate significantly. Drawbacks are more complex signal processing and that defocusing may occur.

In this paper, curved SAR-mapping was analysed in more detail including squinted forward look geometry. The relationships of (11)-(14) show how resolution performance and mapping rate are affected as a result of the curved SAR-path geometry. Simulation results were also presented showing how the non-linear phase error depends on side acceleration and scene geometry for a typical example.

A fundamental reason of lower resolution for curved SAR-paths is the reduced reflector time within the antenna beam. By using a digital array antenna (DGA) and SAR-processing in several antenna channels simultaneously, the observation time can be expanded significantly reducing the limitations in azimuth resolution (11) for curved SAR-mapping.

Another source of error is the non-linear phase contribution, which increases when SAR-processing with only IMU-correction of the relative scene centre movement is used.

By applying the back projection algorithm at the SAR processing of the data with correct phase reference for each individual resolution element, the curve effect can be compensated and the non-linear phase error contribution can be minimised. The computation burden increases significantly compared with standard SAR-processing, however. Sub-image processing with individual scene centre correction is one possible way of reducing the computations required.

Acknowledgement

This work was performed at the Sensor Department of Saab Bofors Dynamics in Linköping for the Swedish Defence Materiel Administration (FMV).

References

[1] W.G. Carrara et al: *Spotlight Synthetic Aperture Radar.* Artech House, 1995.

MIMO channel measurements in an urban street microcell

Jonas Medbo, Henrik Asplund, Mattias Törnqvist, David Browne and Jan-Erik Berg

Ericsson Research

Ericsson AB, 164 80 Stockholm, Sweden.

E-mail: {Jonas.Medbo, Henrik.Asplund, Mattias.Tornqvist, David.Browne, Jan-Erik.Berg}@era.ericsson.se

Abstract

Narrowband MIMO channel measurements have been performed at 1947 MHz in an urban street microcell environment in both LOS and NLOS conditions. The measurement data was used to determine the radio channel characteristics as well as the possible MIMO capacity.

The observed capacity was, in average, somewhat less than the ideal. At some LOS locations, a substantial reduction in capacity was observed due to a wave guiding effect of the street canyon. It turned out that the capacity reduction was an effect of both the Ricean K factor and the angular spread.

A maximum likelihood method (ML) was used to accurately estimate the radio channel in terms of plane waves and their corresponding directions at both ends. It is shown that the ML estimated channel can successfully be used for both pathway reconstruction and for capacity determination, thereby indicating that the plane wave approach is appropriate for MIMO channel modeling.

1 Introduction

Multiple Input Multiple Output (MIMO) antenna systems have recently gained a large interest in the research community due to the potential for large capacity increase compared to a single antenna system [1]. In this paper we investigate to what extent this potential can be realized in a street microcell, where the base station antenna is mounted well below roof-top level. We investigate whether the wave-guiding properties of the street canyon lead to poor multipath richness or if additional scattering can provide independent propagation paths that can be utilized by the MIMO system.

2 Channel measurements

The MIMO channel measurements have been performed using a synthetic array technique. A schematic picture of the measurement equipment is shown in Fig. 1.

Two phase-locked signal generators provide CW signals at frequencies f and $f+df$. One of the signals is fed to a reference antenna, while the other is transmitted from the synthetic array antenna, mounted on a linear positioning unit (LPU). At the receiver, the superimposed signals from the two transmitters are measured on a reference antenna and a synthetic array antenna, mounted on a second linear positioning unit. The two antennas are connected to the two ports of a Network Analyzer (HP 8753C), which measures the channel for all combinations of antenna positions on the two linear positioning units. Some key parameters of the equipment are listed in Table 1.

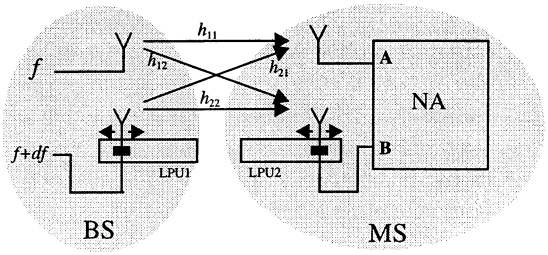


Figure 1: Measurement equipment setup

Table 1: Equipment parameters

Carrier frequency f	1947 MHz
Frequency offset df	100 Hz
Output power	~42 dBm
Antennas	$\lambda/2$ vertical dipoles (2 dBi)
BS array size	7.9λ (219 elements)
MS array size	8λ (41 elements)

The signal h_{11} is used as a phase reference for the signal of interest h_{22}, thereby providing coherency during the synthetic array measurements.

A measurement campaign has been performed in Kista, a part of the city of Stockholm with predominantly office buildings with 5-8 storeys (see map in Fig. 2). The base station (BS) array was mounted on a walkway between two buildings, at a height of 9 m above ground, while the mobile (MS) array was mounted on top of a van at about 3 m over ground. Both arrays were oriented perpendicular to the street, except in two locations where

measurements were performed with the mobile array both perpendicular and parallel to the street, forming a crossed array. To minimize the temporal channel variations, the measurements were performed at night when the traffic intensity was very low.

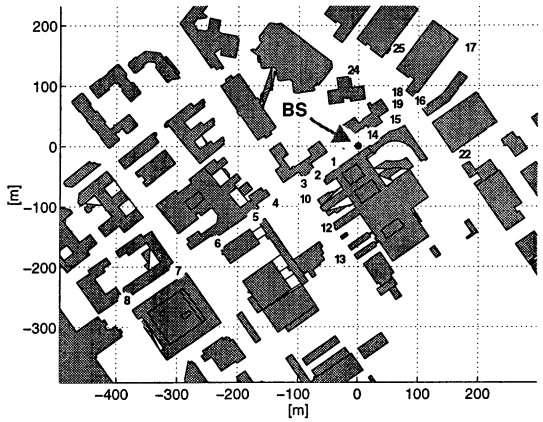


Figure 2: Map over measurement area. The measurement locations are numbered from 1 to 25 and the base station (close to locations 1 and 14) is marked with a circle.

3 Results

In the analysis of the measurement data the focus has been placed on the understanding of those propagation characteristics that have a significant impact on the capacity of a MIMO system. Specifically, the distributions of paths in power, angle of departure and angle of arrival have been studied in some detail.

As a first simple check of the measurement data, the path loss was determined at each measurement location. Measured path loss in a street microcell typically follows the theoretical free space loss in line-of-sight (LOS) conditions while substantially larger losses are observed in non-line-of-sight (NLOS) conditions [2]. This is indeed exactly what can be observed from the present measurements in Fig. 3.

3.1 Spatial channel properties

A high resolution maximum likelihood (ML) based method has been employed to jointly determine the angles of departure (at the transmitter) and the angles of arrival (at the receiver), assuming the channel can be modeled with a finite number of plane waves. This method has previously been demonstrated to produce accurate and reliable results with SIMO data [3]. The plane wave assumption is valid for specular reflections and is reasonable for scattering and diffraction when the scattering point or diffraction edge is not too close to the measurement location (antenna array). Diffuse scattering can also be accounted for using a large number of waves distributed over the corresponding regions in angle.

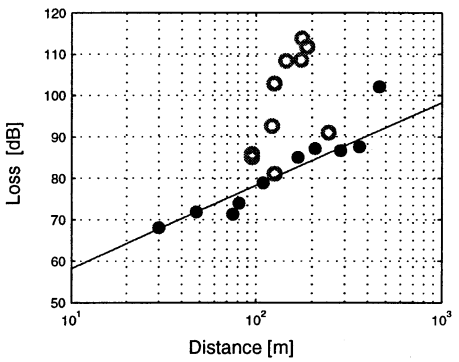


Figure 3: Path loss versus distance for the LOS measurements (filled circles) and the NLOS measurements (open circles). The theoretical free space loss is marked with a solid line.

The measurement data consist of the complex channel transfer matrix $\boldsymbol{H}$. The modeled signal $\tilde{H}_{ij}$ between two array elements i and j is assumed to be the sum of N waves with complex amplitudes s_l, i.e.

$$\tilde{H}_{ij} = n_{ij} + \sum_{l=1}^{N} s_l \exp(-i[\mathbf{r}_i^{BS} \cdot \mathbf{k}_l^{BS} + \mathbf{r}_j^{MS} \cdot \mathbf{k}_l^{MS}]), \quad (1)$$

where n_{ij} is the noise, $\mathbf{r}_i^{BS}$ and $\mathbf{r}_j^{MS}$ are the vectors corresponding to the positions of the array elements, and, $\mathbf{k}_l^{BS}$ and $\mathbf{k}_l^{MS}$ are the wave vectors of wave l at the BS and the MS end respectively. The noise n_{ij} is assumed to be a complex zero mean Gaussian process with the following covariance

$$E\{n_{ij} n_{kl}^*\} = \sigma^2 \delta_{ik} \delta_{jl}, \quad (2)$$

where n^* means n conjugated, i.e. the noise at different elements of $\boldsymbol{H}$ are not correlated.

In a first stage, successive plane waves are identified one-by-one by beamforming. For each wave the likelihood function is maximized with the other wave estimates fixed. This process is stopped when some maximum number of waves is reached. In a second stage the estimates are improved utilizing the fact that two ML estimates are correlated only if they are close in angle at both ends. The maximization has to be done simultaneously only with respect to correlated waves. Therefore, the second stage consists of identification of clusters of correlated waves followed by maximization of the likelihood with respect to each cluster with the other wave estimates fixed.

In the present analysis 50 waves, accounting for about 90% of the total power in most cases, were estimated at each location. The distributions in angle from locations 1 and 22 are shown in Fig. 4. These distributions have been used to reconstruct the three strongest pathways in the maps in Fig. 5. At location 1, the direct path is dominating since there is LOS conditions. The power distribution is more uniform at location 22 where there is no direct path. It seems that in NLOS conditions the main pathways follow the street canyons by diffraction around the corners (Fig. 5).

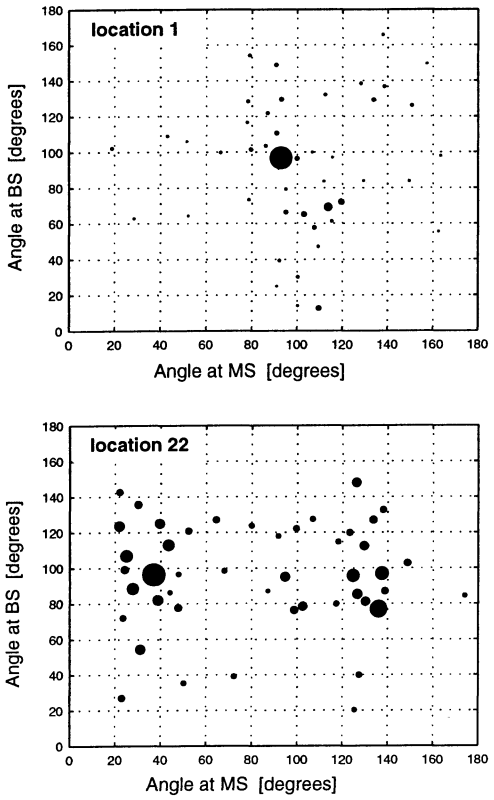


Figure 4: Distribution of estimated plane waves in angle from locations 1 and 22. The area of the circles is proportional to the power.

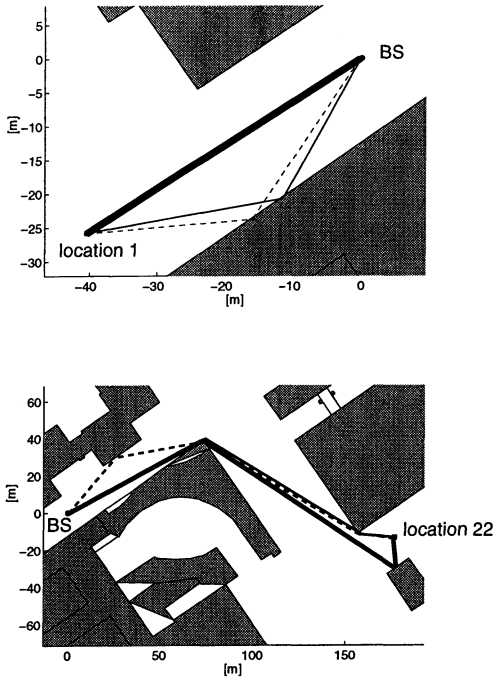


Figure 5: The three strongest paths reconstructed from distribution in Fig. 4 (locations 1 and 22).

3.2 MIMO system capacity

The MIMO system capacity has been investigated using both the measured channel transfer matrix $\boldsymbol{H}$ and the corresponding matrix given by the ML estimated channel. The agreement between the two methods is in general very good showing that the plane wave approximation is accurate enough for MIMO channel modeling.

For a MIMO link, the maximum number of orthogonal channels, N, is the lowest of the number of the transmit and receive antennas [1]. The average signal-to-noise ratio (SNR), P, from one receiver antenna element is the sum from all orthogonal channels i,

$$\sum P_i = P. \tag{3}$$

Each measured transfer matrix $\boldsymbol{H}$ (having M elements) has been normalized such that

$$\sum |H_{ij}|^2 = M. \tag{4}$$

The capacity C is then given by

$$C = \sum_{i=1}^{N} \log_2(1 + \lambda_i P_i) \text{ bits/s/Hz} \tag{5}$$

where the set of λ_i are the eigenvalues of $\boldsymbol{H}\boldsymbol{H}^*$ (where $\boldsymbol{H}^*$ is the conjugate transpose of $\boldsymbol{H}$). The maximum capacity is obtained by water filling i.e. the P_is are adjusted, conditioned by (3), to fulfill

$$\frac{1}{\lambda_1} + P_1 = \frac{1}{\lambda_2} + P_2 = \ldots = D \tag{6}$$

where D is some constant.

The capacity can be calculated for any value of P as long as the SNR of the measured transfer matrix $\boldsymbol{H}$ is substantially higher. To avoid dependence of capacity on absolute signal strength, P has been normalized to 10 dB in the following analysis. Moreover, it is assumed that the theoretical upper and a lower limits of capacity are obtained with a transfer matrix with complex Gaussian distributed elements (Rayleigh fading channel) which are uncorrelated (ideal channel) at the upper limit and fully correlated at the lower limit.

The average capacity, determined directly from the measured $\boldsymbol{H}$ matrix, versus the number of antennas (equal number of antennas at both ends) is shown in Fig. 6. The capacity seems to increase linearly with the number of antennas when the antenna spacing is one wavelength. It is however substantially lower than the ideal capacity. For the smaller antenna separation (0.05 wavelengths) the capacity gain is very close to the lower limit.

In Fig. 7, the capacity for a 8x8 system (normalized to a 1x1 system) with an antenna separation of one wavelength is shown for all locations. The capacity determined from the ML estimated channel is only slightly lower than the capacity determined directly from the measured $\boldsymbol{H}$. The difference disappears when the number of estimated waves is increased to about 200. For channel modeling purpose, however, 50 waves would be sufficient.

At some LOS locations the capacity drops to about 5, which is substantially less than the ideal capacity of about 8. This reduction in capacity can not be explained as an effect of angular spread alone since the rms angular spread values determined from the ML estimated channels (Fig. 8), assuming a uniform distribution in power and Gaussian or Laplacian distributions in angle, results in capacities which are higher than 7.

Another aspect of the radio channel which has impact on the capacity is the distribution in power of the waves.

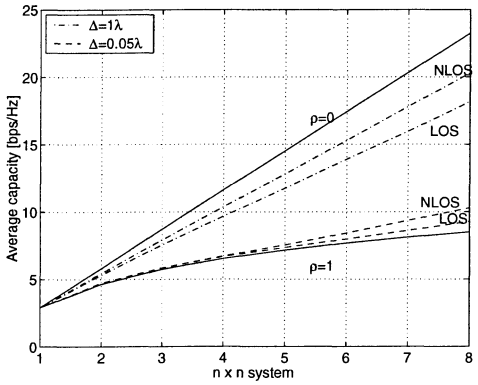


Figure 6: Maximum capacity (Shannon limit at SNR=10dB) versus number of antennas for a nxn system. The average from LOS and NLOS locations are shown for two different antenna separations. The upper (ideal) and lower limits are marked with solid lines.

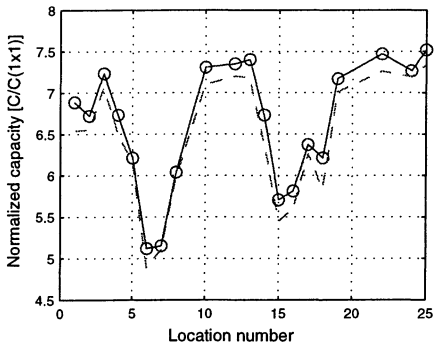


Figure 7: Normalized capacity for a 8x8 system with one wavelength antenna separation. The solid curve with circles correspond to capacity determined directly from H and the dashed curve to capacity determined from the ML estimates.

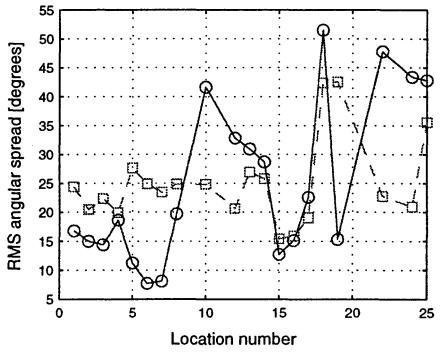


Figure 8: Angular spread (rms) at BS (squares) and MS (circles) determined from the ML estimated channels.

A useful parameter to describe the power distribution is the Ricean K factor, i.e. the ratio between the power of the strongest wave and the rest. The dependence of capacity on the K factor, assuming uniform distribution in angle, is shown in Fig. 9. A comparison with the K factors determined from the ML estimated power distributions (Fig. 9) shows that the impact on capacity is substantial.

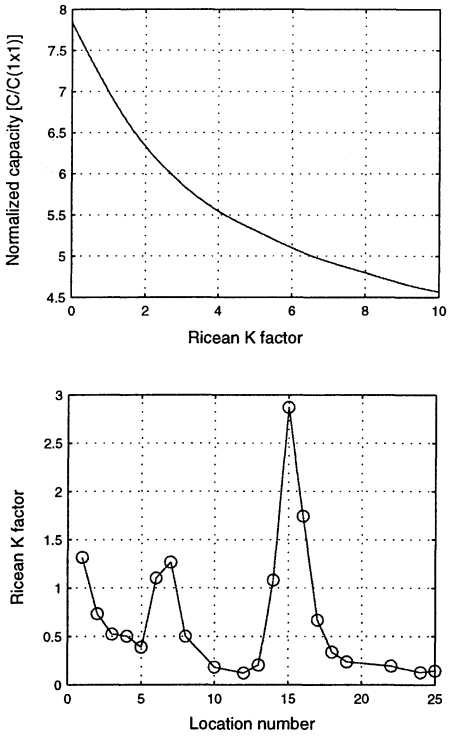


Figure 9: Normalized capacity for a 8x8 system versus K factor (upper plot) and K factors determined from the ML estimated channels (lower plot).

The capacities determined by the K factor alone are shown in Fig. 10 together with capacities determined from both the K factor and the angular distributions. It is obvious that the two curves are strongly correlated. Moreover, it seems that the reduction in capacity due to small angular spread is amplified by a high K factor. It should be noticed here that the upper curve in Fig. 10 is an upper bound which can not be exceeded by increasing the antenna separation.

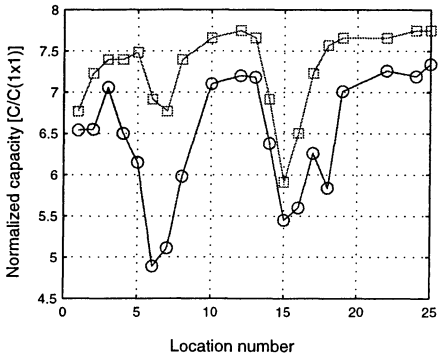


Figure 10: Normalized capacity for a 8x8 system determined from the K factor alone, assuming uniform angular distribution (squares), and from the ML estimated channel (circles).

4 Summary and conclusions

Narrowband MIMO channel measurements have successfully been performed at 1947 MHz in a street microcell environment in both LOS and NLOS conditions. A maximum likelihood method (ML) was used to estimate the channel in terms of plane waves and their corresponding directions at both ends. It was possible to accurately reconstruct the strongest pathways from the estimated directions which indicates that both the measurement and the channel estimation methods are accurate and reliable.

The MIMO capacity was determined both directly from the measurement data and from the ML estimated channel. The agreement between the two methods turned out to be very good, indicating that a plane wave modeling approach is appropriate for MIMO channels. The observed capacity was in average somewhat less than the ideal. At some LOS locations a substantial reduction in capacity was observed. It turned out that this was an effect of relatively large K factors (between 1 and 3) in combination with relatively small angular spreads (7 to 20 degrees). Neither the K factor nor the angular spread alone could account for this capacity reduction. The reason for the increase in K factor as well as for small angular spread values is probably the wave guiding effect of the street canyon.

5 References

[1] J. B. Andersen, "Array gain and capacity for known random channels with multiple element arrays at both ends," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 11, pp. 2172-2178, Nov. 2000.

[2] J.-E. Berg, R. Bownds, F. Lotse, "Path Loss and Fading for Microcells at 900 MHz," in proceedings of the 42nd IEEE Vehicular Tech. conf., Denver, Colorado, pp. 666-671, May 11-13 1992.

[3] J. Medbo, J.-E. Berg, "Spatio-Temporal Channel Characteristics at 5 GHz in a Typical Office Environment", IEEE VTC Fall 2001.

Statistical modelling of the radar background response

Sune R.J. Axelsson

Saab Bofors Dynamics
Swedish Defence Research Agency
S-581 88 Linköping, Sweden
Email: Sune.Axelsson@dynamics.saab.se

Abstract

For an homogeneous scattering ground surface, the amplitude of the received radar signal can be modelled as Rayleigh distributed, which means that it follows an exponential distribution in power. Usually the background is not homogeneous but consists of a variety of surface segments with divergent scattering properties. Various methods of modelling the radar response have been developed and suggested, such as the lognormal, Weibull and Rician distributions. A more recent one is the Rayleigh/Generalized Lognormal Distribution, which gives good agreement with the backscatter statistics from sea for midrange incidence angles. In simulating the radar backscattering from inhomogeneous terrain, the scene can be described as a number of homogeneous sub-surfaces representing N terrain classes with coverage p_1, p_2, ..., p_N, respectively, and with individual scattering characteristics. The geometry of the surface segments and their scattering characteristics can be generated randomly or from a deterministic terrain data base.

1 Introduction

Detection of targets in background clutter is fundamental in most radar applications, and advanced modelling is required to estimate or predict the detection capability and performance.

For a homogeneous scattering ground surface, the amplitude fluctuations of the received radar signal can be modelled as Rayleigh, which means that it follows an exponential distribution in power. Mostly the ground background is not homogeneous but consists of surface segments with different scattering properties. Various methods of modelling the radar response from mixed surfaces have been developed and suggested, such as lognormal, Weibull and Rician distributions. In this paper, some of them are reviewed, and efficient simulation methods are discussed. The detailed modelling of the parameters involved, such as the average backscattering from sea or terrain, is excluded but can be found in a great number of sources.

2 Rayleigh scattering

2.1 General

If the ground surface is rough, the returned wave can be modelled as a sum of contributions from N point reflectors.

$$E_1 = \Sigma E_n \exp[j\omega t - j2kR_n + j\varphi_n] = \Sigma E_n \exp(j\omega t + j\theta_n) \quad (1)$$

in which R_n is the range and $k=2\pi/\lambda$. After demodulation we can rewrite (1) in polar form as

$$E_2 = X + jY = V\exp(j\theta) \quad (2)$$

where

$$X = \Sigma E_n \cos(\theta_n) \quad (3a)$$

$$Y = \Sigma E_n \sin(\theta_n). \quad (3b)$$

For many back-scattering surfaces, such as agricultural fields and forest, the following assumptions can be considered fulfilled: (a) there is a large number of reflectors (N>>1); (b) there is no reflector giving a dominant contribution; (c) the phase angles θ_n (n=1 to N) are assumed to be uniformly distributed between 0 to 2π and (d) are statistically independent. As a result of the Central-Limit Theorem, X and Y follow a Gaussian distribution with zero means and no correlation:

$$p(X,Y) = (2\pi\sigma^2)^{-1}\exp[-(X^2/2\sigma^2) - (Y^2/2\sigma^2)] \quad (4)$$

where $\sigma^2 = E(X^2) = E(Y^2) = (1/2)\,\Sigma E_n^2$.

The probability density function of $V = (X^2 + Y^2)^{1/2}$ and θ, named $q(V,\theta)$, can be computed starting from

$$p(X,Y)dXdY = q(V,\theta)d\theta dV \quad (5)$$

which can be rewritten as

$$q(V,\theta) = p[V\cos(\theta), V\sin(\theta)]\,\left|\delta(X,Y)/\delta(V,\theta)\right| \quad (6)$$

where the Jacobi determinant is $\left|\delta(X,Y)/\delta(V,\theta)\right| = V$.

Insertion into (6) gives $q(V,\theta) = V(2\pi\sigma^2)^{-1}\exp[-V^2/(2\sigma^2)]$. The probability density function of V is then obtained by integration of $q(V,\theta)$ from 0 to 2π

$$p(V) = (V/\sigma^2)\exp[-V^2/(2\sigma^2)] \quad (7)$$

which represents the Rayleigh distribution.

The received power $P=V^2$, which is a measure of the radar cross-section of the ground resolution element, follows a exponential distribution. This is easily shown from $p(V)dV=q(P)dP$ giving

$$q(P) = p(P^{1/2}) \, |dV/dP| . \quad (8)$$

Obviously, the average power is $P_{av} = E(X^2)+E(Y^2) = 2\sigma^2$. From $dV/dP=1/(2P^{1/2})=1/(2V)$ and (8), we derive the probability density function of power P:

$$q(P) = (1/P_{av})\exp(-P/P_{av}) \quad (9)$$

which has the following distribution function

$$F_e(X) = P(P<X) = 1-\exp(-X/P_{av}). \quad (10)$$

2.2 Averaging several samples

At the detection of the radar signal averaging is used to suppress the statistical fluctuations. If B is the doppler bandwidth and T is the time of measurement there are N=BT statistically independent samples $P_n=X_n^2+Y_n^2$ of the instantaneous power. Averaging over N samples

$$P = [P_1 + P_2 + ... + P_N]/N \quad (11)$$

makes P follow the probability density function

$$p_N(P)= \{P^{N-1}N^N/[N-1)!P_{av}^N]\}\exp(-NP/P_{av}). \quad (12)$$

Concluding, (7), (9) and (12) are excellent models for homogeneous surfaces satisfying the conditions a)-d).

3. Dominating reflector

If there is one single reflector giving a dominating response, we can model the received radar signal as sine plus noise [1]. That means that X and Y are still Gaussian but one of them, X for instance, has a non-zero average given by the amplitude response A of the dominating reflector. This gives a modified version of (4)

$$p(X,Y) = (2\pi\sigma^2)^{-1}\exp\{-[(X-A)^2)/2\sigma^2] - (Y^2/2\sigma^2)\} \quad (13)$$

and from that the pdf of V can be determined using similar technique as above

$$p(V) = (V/\sigma^2)\exp[-(V^2 + A^2)/2\sigma^2] \, I_0(VA/\sigma^2) \quad (14)$$

where $I_0(X)$ is the modified Bessel function of order zero.

In (14), $P_{av} = 2\sigma^2$ represents the mean power when the contribution from the dominating reflector is excluded. The pdf corresponding to (9) can be developed from (14) in a straight-forward way using $dP/dV=2V$.

4. Inhomogeneous scene

4.1 General

Let us now discuss a scene, which can be modelled by N areas with internally homogeneous statistics. This can represent a landscape having varying type of vegetation coverage.

The amplitude of the radar signal of area number n is supposed to be controlled by the probability density function $p_n(V)$. What statistics is obtained if we select the resolution element in the scene randomly?

Let P_n denote the probability of sampling area number n, where $\sum P_n=1$. The probability density function of V is then given by (with n=1 to N)

$$p(V) = \sum P_n p_n(V). \quad (15)$$

The relationship is very general and can be applied to varying types of statistics governing the individual sub-areas.

Let us now reformulate the problem and assume that all sub-areas have the same type of statistics $p(x; a_1, a_2, ..., a_M)$, controlled by the parameters $(a_1, a_2, ..., a_M)$, which differ between the different sub-areas, however. In case of Rayleigh/exponential distributions, we have one single parameter: $a_1=\sigma^2$.

If we randomly select a resolution element, the parameters will also be random. Let us denote the pdf: $p_1(a_1, a_2, ..., a_M)$.

If we draw a sample of the radar signal from a randomly chosen resolution element, the probability density function of the signal becomes

$$p(X) = \int..\int p(X/a_1, a_2, ., a_M) \, p_1(a_1, a_2, .., a_M)da_1da_2 ..da_M \quad (16)$$

The relationship (16) can be used to describe the statistics of the amplitude or power.

4.2 Rayleighfading with arbitrary distribution of average power.

Let us assume that the instantaneous power P has an exponential distribution. After averaging over N samples, the probability function is from (12)

$$p_N(P)= \{P^{N-1}N^N/[N-1)!P_{av}^N]\}\exp(-NP/P) \quad (17a)$$

having one single parameter P_{av}. If this parameter varies randomly between different resolution elements governed by the probability density function $p_1(x)$, the pdf of the averaged power P becomes instead

$$p(P) = \int\{P^{N-1}N^N/[N-1)!x^N]\}\exp(-NP/x)p_1(x)dx. \quad (17b)$$

Lewinski [2] used (17b) to compute p(P) for different probability density functions $p_1(x)$ of the background level (σ^0). He also showed that the n:th momentet P^n can be expressed as

$$E(P^n) = E(P_{av}^n)\ (N+n-1)!\ /\ [(N-1)!N^n]. \quad (18)$$

The average of P (n=1) is thus equal to P_{av}. For higher moments (n>1), $E(P^n) > E(P_{av}^n)$, which explains the high tail values, which usually are observed from experimental data of ground clutter.

4.3 Rayleighfading with equal distribution

If P_{av} has a uniform distribution between P_1 and P_2

$$p_1(x) = 1/[\ P_2-P_1]\ \text{ for } P_1<x<P_2 \quad (19)$$

From (17) and (19)

$$p(P) = [(P_2-P_1)(N-1)!]^{-1}[\gamma(N-1,NP/P_1) - \\ - \{\gamma(N-1,NP/P_2)] \quad (20)$$

where γ is the non-complete gammafunction

$$\gamma(a,x) = \int t^{a-1}e^{-t}dt \quad (21)$$

with integration over t from t=0 to t=x.

If we denote $D=P_2/P_1$, $P_m=(P_{u1}+P_2)/2$ and normalize P with respet to the average P_m i.e. $z=P/P_m$ we find

$$p(z)=(1/2)[(D+1)/(D-1)][\ N-1)!]^{-1}\{\gamma[N-1,N(1+D)z/2] - \\ -\gamma[N-1,\ N(1+1/D)z/2]\} \quad (22)$$

Lewinski [2] also showed that the probability of detection Pd(x)=P(z>x) is given by the relationship

$$Pd(x) = (1-a)^{-1}E_2[(1+a)x/(2P_m)] \\ - [a/(1-a)]E_2[(1+a)x/(2aP_m)] \quad (23)$$

where a=1/D and $E_2(x)$ is the integral over $f(x)=e^{-xt}/t^2$ with limits t=1 to infinity.

4.4 Rayleigh fading with gammadistribution

Usually neither σ^0 or P_{av} follow a uniform distribution. An improved approach is to describe the fluctuations of P_{av} using the gamma distribution

$$p_1(x)= m^m x^{m-1}\exp(-mx/P_0)/[\Gamma(m)P_0^m] \quad (24)$$

where $P_0=E(x)$ and $m=P_0^2/E\{(x-P_0)^2\}$.

Insertion of (24) into (17b) yields the pdf of the received power

$$p(z) = [A(N,m,z)/B(N,m,P_0)]\ K_{m-N}[(4mzN/P_0)^{1/2}] \quad (25)$$

where $K_v(u)$ is the modified Bessel funktion of second kind and order v, and

$$A(N,m,z) = 2(Nm)^{(N+m)/2}z^{(N+m-2)/2} \quad (26a)$$

$$B(N,m,P_0) = \Gamma(N)\ \Gamma(m)P_0^{(N+m)/2} \quad (26b)$$

The gamma distribution is often used in modelling the instantaneous power variations for different clutter cases [2], where the parameter m controls the shape and P_0 is the scaling factor.

5. Lognormal and Weibull distributions

The lognormal distribution has been extensively used in the past in modelling the radar clutter from inhomogeneous terrain. If we introduce $P_{db}=10\ ^{10}\log(P)$ and prescribe that P_{db} follows a Gaussian distribution

$$p_1(P_{db}) = (2\pi)^{-1/2}s_{db}^{-1}\exp[-(P_{db}^2-\mu_{db})^2/(2s_{db}^2)] \quad (27)$$

where μ_{db} and s_{db} are mean and standard deviation of P_{db}.

For inhomogeneous background surfaces the Rayleigh distribution underestimates the clutter dynamics, while the lognormal distribution has a tendency of overestimation. The Weibull distribution can be seen as a bridge between the two. By proper parameter selection statistics close to both of them can be defined.

The probability density function of a Weibull distributed variable V is controlled by two parameters: m_v that is the median value of V, and α that influences the asymmetry of the distribution. For the normalised variable $v=V/m_v$, the pdf becomes

$$p(v) = \alpha(\ln 2)v^{\alpha-1}\exp[-(\ln 2)v^\alpha] \quad (28)$$

Mean value and standard deviation are given by [3]

$$\mu_v = m_v(\ln 2)^{-1/\alpha}\ \Gamma(1+1/\alpha) \quad (29a)$$

$$s_v = m_v(\ln 2)^{-1/\alpha}[\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha)]^{1/2} \quad (29b)$$

and the distribution function by

$$F(V<X) = 1-\exp[-(\ln 2)(X/m_v)^{\alpha}]. \quad (30)$$

For $\alpha=2$, in particular, (28) is the pdf of a Rayleigh distributed variable; and for $\alpha=1$ it represents the exponential distribution. The probability of high X-values is further increased when $\alpha<1$.

A recent extension is the Rayleigh/generalized lognormal distribution, which was compared with scatterometer data from a rough sea surface [9]. The backscattering is closely related to the slope distribution of a rough surface showed if the correlation length of the surface undulations is much larger than the electromagnetic wavelength. The proposed σ^0-model in [9] was

$$\sigma^0 = C\exp[\alpha_1 s_x + \alpha_2 s_x^2] \quad (31)$$

where s_x is the wave surface slope.

By assuming a Gaussian surface slope distribution, the $p(\sigma^0)$ can be derived in closed form [9]. In particular for $\alpha_2=0$, $p(\sigma^0)$ equals the lognormal distribution. The distribution of the amplitude V is then derived from

$$p(V)=\int p(V/\sigma^0)\, p(\sigma^0)\, d\sigma^0 \quad (32)$$

where $p(V/\sigma^0)$ can be assumed Rayleigh.

In [9] it was shown that for midrange incidence angle backscatter from sea, the Rayleigh/Generalized Lognormal Distribution gives an excellent agreement with the statistics of scatterometer data at 5 GHz even in the tails, while the best fit Weibull distribution did not model the high amplitude distribution as accurately.

6. Simulation models

6.1 Simulation of the radar signal statistics

Above some statistical distributions of the backscattered radar signal were discussed. In a simulation model the random numbers have to be generated by the computer. Modern programs, such as MathCad and Matlab, have internal procedures generating random numbers for the most common distributions. But some of those derived above are missing. In [4] it is described how to generate random numbers for a given probability density function F(X).

The simplest one is the inverse method. From uniformly distributed random numbers $\{u_n\}$ between 0 and 1, a set of random numbers $\{y_n\}$ is generated from the relationship: $y_n=F^{-1}(u_n)$. A precondition is that the inverse function can be expressed in analytical form or be approximated accurately.

A more general method is the following [4]. Let us assume that the random variable is distributed within a limited interval (a,b), which can be selected so wide that the error becomes negligible. Divide the interval (a,b) in N equal subintervals of length $\Delta=(b-a)/N$ so that $y_n=a+n\Delta$. We now introduce a discrete distribution $z_n=(y_n+y_{n-1})/2$ with point probability $p_n=F(y_n)- F(y_{n-1})$. Finally introduce the random variable w uniformly distributed within $(-\Delta/2, \Delta/2)$, which can be generated from the random variable u uniformly distributed over (0,1) according to $w=\Delta(u-1/2)$. Random variables with the distribution function F(y) can then be generated from $y = z + w = z+\Delta(u-1/2)$. The discrete variable z can be simulated in different ways. The most obvious is to generate a value of the uniformly distributed variable U (0,1) and put $z=z_k$ where k is determined from $p_1+p_2+ \ldots.p_{k-1} <U< p_1+p_2+ \ldots.p_k$. In [4] a procedure developed by Marsaglia is given, which is numerically faster.

A third way of simulating the statistics of the returned radar signal from an inhomogeneous scene is obtained by dividing the generation of random numbers in two steps looking at

$$p(X) = \int\int\ldots\int p(X/a_1, a_2, \ldots, a_M)$$

$$p_1(a_1, a_2, \ldots, a_M)\, da_1 da_2 \ldots da_M \quad (16)$$

Step 1 means a random parameter simulation using the density function $p_1(a_1, a_2, \ldots, a_M)$. With these parameter values inserted into $p(X/a_1, a_2, \ldots, a_M)$ the random variable X is simulated. By this two-steps method simple standard procedures, which are already implemented in computer code, can be applied.

An excellent method of generating the instantaneous radar cross-section from an inhomogeneous background is as follows when Rayleigh fading occurs. For each resolution element the relative coverage p_k and $\sigma_{av}^0(k)$ are defined. The average radar cross-section then is $\sigma_b^{av} = A_0 \sum p_k\sigma_{av}^0(k)$, where A_0 is the area of the resolution element (m^2). The instantaneous radar cross section σ_b is then determined by random sampling from an exponential distribution, or χ^2-distribution when several independent samples are averaged.

6.2 Deterministic scene model built up by homogeneous segments

A deterministic radar model splits up the scene area in homogeneous segments, each one belonging to a certain terrain class, such as roads, forest, harrowed fields, grass, buildings. For each segment, the scattering characteristics are defined versus polarisation, wavelength, observation angle including type of statistical model and parameter set.

The complete model includes as well a geometry and radar model, defining the beam geometry and resolution elements in range and doppler. The instantaneous amplitude and power of the received radar signal are then generated from each resolution element using the statistical method described above.

6.3 Random scene generation

Random scene modelling is an alternative instrument. The average radar cross-section per unit area is described as a two-dimensional stochastic process with the probability density function p(V) and the autocorrelation function R(x,y). If we select a two-parameter model to describe p(V), such as Weibull, and two other parameters (L_x, L_y) to describe R(x,y), the complete model has only four parameters to quantify the scene background. The great advantage is the simplicity, which also means less accuracy, however. Two-dimensional Markov models or fractals can simulate homogeneous surface segments with step transitions at the boundaries. By using brownian fractals, variations within the same class can be modelled as well. See [5] and [6].

6.4 Effects of doppler and signal bandwidths

In modelling the statistical fluctuations of the received signal from a scene, the velocity vector and the antenna beam direction and beam width have great influence upon the doppler bandwidth, and by that the possible averaging over the time of measurement (N= BT).

The doppler frequency in direction θ measured from the velocity vector of the radar **V** is

$$f_d = 2V\cos(\theta)/\lambda \qquad (33)$$

Differentiation of (33) shows that the doppler bandwidth corresponding to a resolution element covering $\Delta\theta$ in θ is

$$B = \Delta\theta(2V/\lambda)\sin(\theta) \qquad (34)$$

For each pixel, a sum signal from the reflectors inside can be defined. After demodulation

$$V = I + jQ = \sum V_n \exp[j2\omega R_n/c + \varphi_n] \qquad (35)$$

Realistic time fluctuations of V are simulated by computing the range values R_n and their changes when the radar is moving. Eq. (35) also describes the decorrelation generated by frequency jumps during measurement i.e. when ω is time varying.

When the ground reflectors are moving as a result of the wind pressure, that effect is also included in (35), because it directly influences the time variation of R_n.

Concluding, the model describes the scene as a set of point reflectors distributed over each resolution element. By modelling the changed geometry when the radar is moving over the surface, realistic time fluctuations of the received signals can be generated.

Acknowledgement

This work on radar background modelling was performed at Saab Dynamics AB for the Swedish Defence Materiel Administration (FMV). Parallel studies supported by FMV focused on the theoretical relationships between σ^0 and the physical characteristics of the backscattering surface.

References

[1] S.O. Rice: Mathematical Analysis of Random Noise, Bell System Tech. J., vol. 23, pp. 282-332, 1944 and vol. 24, pp. 46-156, 1945.

[2] D.J. Lewinski: Nonstationary Probabilistic Target and Clutter Scattering Models, IEEE, vol. AP-31, 490-498, 1983.

[3] F.T. Ulaby and M.C. Dobson: Handbook of Radar Scattering Statistics for Terrain, Norwood, MA; Artech House, 1989.

[4] M. Abramowitz and I.A. Stegun: Handbook of Mathematical Tables, kap. 26.8, 1965.

[5] B.B. Mandelbrot: The fractal geometry of nature. W.H. Freeman, New York, 1982.

[6] Sune Axelsson: IT4 - Irmm. Preliminary modelling of the scattering of mm-waves from vegetation and bare surfaces. Saab RLM-0-91-R94(I), 1991.

[7] Sune Axelsson: Models of sensor applications in the mm-wave chapter 7: Background modelling in active mode. Saab RL-0-82:R30, 1982. In Swedish.

[8] Sune Axelsson: Models of the background signature. VI: The statistical distribution of the radar signal. Saab RD-R97:1003, 1997. In Swedish.

[9] B.E. Barrowes and D.G. Long: Evaluation of compound probability model with tower-mounted scatterometer data. IEEE GRS, vol.40, January 2002, pp. 42-49.

Intentional Electromagnetic Interference (IEMI) Background and Status of the Standardization Work in the International Electrotechnical Commission (IEC)

M. W. Wik[1] and W. A. Radasky[2]

[1] Defence Materiel Administration, SE 115 88 Stockholm, Sweden, mawik@fmv.se
[2] Metatech Corporation, 358 S. Fairview Ave. Suite E, Goleta, CA 93117 , USA, wradasky@aol.com

Abstract

Intentional EMI, or as it is sometimes known, "EM Terrorism," is a new area of concern for public and commercial interests. This paper will review the different categories of EM threats and will also summarize available information concerning the susceptibility levels for commercial equipment. The paper will conclude with a presentation of the standardization work being accomplished by the International Electrotechnical Commission (IEC) in the area of high power transient phenomena. The paper is an updated extract from an article originally published in *The Radio Science Bulletin* No 299 (December, 2001).

1 What is Intentional EMI?

At the URSI XXVIth General Assembly in Toronto, August 1999, a resolution was adopted on "Criminal Activities using Electromagnetic Tools" [1]. According to this, intentional EMI is defined as the "Intentional malicious generation of electromagnetic energy introducing noise or signals into electric and electronic systems, thus disrupting, confusing or damaging these systems for terrorist or criminal purposes." The URSI resolution recommended to provide reasonable data regarding the formulation of standards of protection and support the standardization work.

Over the past two years there has been increasing activity in the area of Intentional Electromagnetic Interference (EMI) which is sometimes referred to as EM Terrorism. In particular at the International Zurich Symposium and Technical Exhibition on EMC in February 1999, there was a well-attended session on EM Terrorism that resulted in the publication of five important papers: dealing with the overall problem [2], presenting an approach to protecting systems from the threat [3], providing HPM test data on automobiles [4], reviewing modeling and simulation methods [5], and describing the use of large-scale simulators to evaluate EM threats [6].

In the summer of 2000 there were two important conferences that provided additional information on this subject. At the EUROEM2000 Conference in Edinburgh in May, there was a special session on Intentional EMI that attracted 15 presented papers although no papers were published. Later in June at Wroclaw, Poland the URSI Commission E session resulted in three published papers dealing with the nature of the threat [7], testing strategies [8], and the approach of the International Electrotechnical Commission (IEC) Subcommittee 77C to develop environment and protection standards to deal with the problem [9].

2 Types of IEMI Threats

In the simplest of terms there are two basic types of IEMI waveforms -- Narrowband and Wideband. However, the meaning of narrowband and wideband concerning pulses needs to be discussed further and this is planned to be included in the IEC work. The narrowband category describes a nearly single frequency that may be transmitted in a pulse on the order of microseconds in length. This type of threat is often referred to as HPM (high-powered microwave). The frequency range for HPM is typically higher than 1 GHz, although from an IEMI point of view, frequencies between 0.2 and 5 GHz seem to be the most important to consider as narrowband threats.

Of course the "single" frequency may vary with time, and the signal can be modulated. It is generally fairly easy to create fairly high power and energy levels with this type of waveform. An example of this type of waveform is one generated in a microwave oven or by radar tubes. The disadvantage of this type of waveform is that systems are affected by the currents and voltages coupled by the incident field, and many system have resonances that create significant susceptibilities to particular frequencies. However, different systems will likely have different resonances, thereby limiting the effectiveness of a single frequency IEMI generator.

For the wideband category, a pulse produces frequency content over a wide range of frequencies. One type of

wideband waveform is UWB (ultra-wideband). These pulses typically have a rise time of less than 100 picoseconds and a pulse width of a few nanoseconds. The main frequency content of the UWB pulse ranges from 0.3 - 3 GHz. It should be noted that the HEMP waveform described by the IEC [10] has a 2.5 ns rise time and a 25 ns pulsewidth. This qualifies a wideband waveform, but it is not considered an ultra-wideband pulse.

The advantage of the wideband pulses is that the resonances of different sized systems can be stimulated simultaneously. The disadvantage is that the energy produced in a single pulse is spread over many frequencies. In addition, the peak electric field value of the pulse is limited by the breakdown field value of air. For these reasons, the wideband pulses are often repetitively pulsed and therefore pose mainly an interference threat to systems as opposed to a permanent damage threat as provided by narrowband waveforms.

In addition to the waveform type, there is the issue of how the disturbing waveform is delivered to the system. The two general categories described in the electromagnetic compatibility discipline are radiated and conducted. In the case of radiated waveforms, an antenna is used to produce electromagnetic fields that propagate to the vicinity of a "target" system. These radiated fields will propagate through the air and through apertures in buildings, automobiles or aircraft. The electromagnetic fields will either couple directly to the cables attached to a piece of equipment or will penetrate directly inside of the equipment, where damage or upset to system operations will occur.

Conducted disturbances can be injected directly, capacitively, or inductively onto power or telecommunication cables. These disturbances will then propagate along the cables until they reach equipment connected to the cables. As in the case of radiated disturbances, both narrowband and wideband waveforms can be injected, however, often the cables transfer functions will limit the frequency content of conducted disturbances to below 1 MHz, particularly on power lines.

3 Susceptibility Levels of Commercial Equipment

3.1 PC Cable Experiments

A recent paper by Radasky, et. al. at the Zurich EMC Conference surveyed equipment failure data from a series of laboratory experiments [11]. Four sets of data -- two conducted and two radiated, provided valuable information.

Power cords and ethernet cables connected to personal computers were tested and surveyed in this paper. Two types of conducted pulse waveforms were used: the CWG (Combination Wave Generator) waveform (1.2/50 microsecond pulse) and the Telecom waveform (10/700 microsecond pulse).

The results indicated that the power ports were not affected for delivered voltages as high as 1.2 kV, although higher levels have been reported to cause damage. For the ethernet cables, the 10Base-2 cards were damaged at 500 volts, which was the lowest test level. The 10Base-T interface was not damaged until a level of 4 kV was reached using the Telecom waveform. It is interesting to note that the 10Base-2 uses a coax cable, however, the shield is not connected at the card, thereby creating an easy damage path.

3.2 Building Power Cable Experiments

An experiment was performed by Fortov et. al. [12] by injecting both narrowband and wideband waveforms onto the power lines entering a five story office building and by making measurements at various power plugs within the building. The measurements indicated that voltages injected on external wiring could propagate well through the internal wiring of a building even when considering multiple switchboards inside the building. It is clear from their work that frequencies less than 1 MHz propagate with low attenuation as do pulses with widths greater than 1 microsecond. Although this study did not address the issue of wiring breakdown directly, it is felt that for the types of pulses considered, that normal building wiring should be able to support peak voltages in the range of 10 kV.

In terms of the vulnerability of computers, both the analyses and limited testing revealed that computer power supplies, and in particular the input filters, appear to be vulnerable to levels of 6 kV for a 50-microsecond pulse. Analyses indicate that levels of 1 - 2 kV would create damage for a 1 ms wide pulse.

By considering both aspects of this work, it appears possible to inject significant levels of voltage into the power wiring system of a building, and that voltage can propagate easily and cause damage to computer power supplies. Of course it is possible that other types of equipment connected to the power system will be vulnerable to injected pulses, although other types of equipment have not yet been considered.

3.3 Radiated Field Effects on PCs

Three types of PCs were tested by LoVetri, et. al. [13] by irradiating them with different frequencies, modulation types and field strength levels. An effort was made to determine the lowest field levels at various frequencies using narrowband waveforms. Angles of incidence and

polarization variations were minimal, so the results should not be considered to be the lowest levels where effects could occur.

The authors concluded in their study that relatively weak fields, as low as 30 V/m, reaching the inside of a personal computer could disrupt its operation. In addition, they noted that the effects were observed only at particular frequencies due to resonances present inside each computer.

3.4 Automobile Test Results and Conclusions

Another experiment was performed in Sweden by Bäckström [4] by irradiating an automobile with narrowband waveforms at high power and field levels (HPM). The frequencies of illumination varied between 1.3 and 15 GHz. Two angles of incidence to the automobile were considered.

The authors of this study noted that effects were more prominent at the lower test frequencies with automobile upsets (including the stoppage of the engine) noted at the lowest test value of 500 V/m. Permanent damage occurred at 15 kV/m at 1.3 GHz and 24 kV/m at 2.86 GHz. They also noted that permanent damage occurred when the automobile was not operating. The types of damage observed included: engine control units, relays, speedometer, revolution counter, burglar alarm, and a video camera.

It is clear that an unprotected (from HPM) electronics system such as an automobile can be vulnerable to HPM fields at relatively low levels. A field level of 500 V/m or lower has the capability to stop the vehicle. Based on this test data and the capability today of building a van with an HPM power source of 10 MW and a directional antenna, it would be possible to damage an automobile at a range of 15 meters and to stop its operation at 500 meters.

3.5 Susceptibility Conclusions

It is apparent from the data summarized here, that both conducted and radiated disturbances are a threat to commercial systems. Clearly it is possible to generate conducted wideband waveforms similar to those used for EMC testing of equipment, but with somewhat higher peak levels than those specified for normal home or commercial usage. For radiated waveforms, it is possible to apply microwave oven parts and those from surplus military radars to generate threatening electromagnetic field levels. Of course generators can be built in laboratories with higher level capabilities, however, source size is an important factor to be considered when creating threat level criteria.

4 Introduction to the IEC

The objective of the International Electrotechnical Commission (IEC) is to produce international standards for electrical and electronic equipment and systems. In particular, the electromagnetic compatibility (EMC) aspects are of interest worldwide for the electronics and power industries in order to harmonize methods of testing and to develop protection methods and devices. The main part of the EMC immunity work is centered in IEC Technical Committee 77 that has three subcommittees dealing with different aspects of EMC (77A, 77B, 77C).

The work described below is part of the approved program of IEC SC77C, "EMC: High Power Transient Phenomena." The intention of the subcommittee is to apply openly published information and to reference existing IEC EMC standards wherever possible, so that there is no duplication of effort. Those working in the SC77C subcommittee and on particular project teams represent their countries and not particular organizations (civilian or military).

Subcommittee 77C currently has sixteen participating member nations: Austria, Bulgaria, China, Finland, France, Germany, Italy, Japan, Mexico, Romania, Russia, Sweden, Switzerland, Turkey, United Kingdom and USA. In addition there are sixteen observing member nations: Belgium, Canada, Croatia, Czech Republic, Denmark, Ireland, Israel, Republic of Korea, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Thailand, and Ukraine. The Chairman is Dr. W. A. Radasky (USA) and the Secretariat has been undertaken by Sweden (Messrs. M. W. Wik, J-O. Sjödin).

The preparation of SC77C standards is consistent with the development of EMC standards within other parts of TC77 (EMC) and is thus structured according to "IEC Publication 61000" which is divided into seven major parts. Only Parts 1, 2, 4, 5 and 6 are utilized by SC77C at present. The current development of standards and reports includes seventeen active projects and/or publications, which are summarized below.

5 Progress of Work and Highlights

Part 1 "*General*" of the IEC 61000 series includes terminology, definitions and other general aspects. "The effects of high-altitude EMP (HEMP) on civil equipment and systems" (61000-1-3, W. A. Radasky, USA) will provide information concerning the effects of high-altitude EMP (HEMP) on electrical and electronic equipment and systems. This information is based on effects observed during high altitude nuclear testing and from tests performed in HEMP simulators in several

countries. The project results in an IEC Technical Report in the beginning of year 2002.

"High power electromagnetic (HPEM) effects on civil systems" (61000-1-5, F. M. Tesche, USA) will result in an IEC Technical Report that discusses the effects of HPEM fields on civilian systems and illustrates the general protection principles that can be applied to protect systems from this newly emerging threat. The present stage is Committee Draft (CD).

Part 2 "*Environment*" gives a description of the high power electromagnetic environment and is divided by SC77C into radiated and conducted parts. "Description of HEMP environment - Radiated disturbance" (61000-2-9, G. Champiot, France) is published as an IEC International Standard. It contains a number of definitions and the radiated parameters for the early-time, intermediate-time and late-time HEMP waveforms. This includes electric and magnetic field time waveforms, HEMP frequency amplitude and energy, weighting of the early, intermediate and late-time HEMP. The standard also deals with reflections and transmission of the HEMP from and through the earth's surface.

"Description of HEMP environment - Conducted disturbance" (61000-2-10, W. A. Radasky, USA) is based on the radiated waveforms in standard 61000-2-9 and is published as an IEC International Standard. It describes the conducted environment applicable for categories of conductors for different positions and illumination cases statistically taken into account. The specified environments are based on extensive theoretical calculations and experimental measurements.

Part 2 also includes "Classification of HEMP environment" (61000-2-11, W. A. Radasky, USA) and groups HEMP environments present at various locations outside and inside of civilian systems. The project is published as an IEC International Standard. One reason for classification is to provide guidance for equipment manufacturers to help them decide on the proper immunity test levels appropriate for their equipment. Another reason is to provide system designers with guidance regarding construction methods and protective measures needed to achieve defined EM classes.

Another new Part 2 project entitled, "High power electromagnetic (HPEM) environments - radiated and conducted" (61000-2-13, D. V. Giri, USA) defines the current and near-future high-power electromagnetic (HPEM) environments that are a potential threat to civilian systems, including both the radiated and conducted environments. It is expected that this standard will become the basis for defining appropriate protection methods in the future. The present stage is Committee Draft (CD).

Part 4 "*Testing and Measurement Techniques*" presently includes five projects. "Test methods for protective devices for HEMP and other radiated disturbances" (61000-4-23, F. M. Tesche, USA) is published as an IEC International Standard. The document contains definitions, shielding effectiveness measurement methods for shielding materials, gaskets and shielded enclosures, and transfer impedance measurement methods for coaxial cables.

"Test methods for protective devices for HEMP conducted disturbance" (61000-4-24, W. K. Büchler, Switzerland) is published as an IEC International Standard. It describes methods to measure the residual voltage on protective devices under HEMP conditions, i. e. for the case of very fast changes of voltage and current as a function of time. This standard complements the standard 61000-5-5, "Specification of protective devices for conducted disturbance."

"HEMP immunity test methods for equipment and systems" (61000-4-25, P. R. Barnes, USA) is published as an IEC International Standard. The document includes a list of immunity tests and environmental conditions, and guidance for the selection of particular immunity tests, test levels, test methods, test equipment, test set-ups, and test procedures. Efforts have been made to employ as many existing EMC test techniques as possible and to avoid more costly special HEMP tests.

"HEMP simulator compendium" (61000-4-32, J. C. Giles, USA) will be published as an IEC Technical Report in year 2002. The report provides information on the various types of existing large HEMP simulators and their uses, performance parameters, limitations, and availability. This will allow all potential simulator users to evaluate the adequacy of available simulators for testing large systems.

"Measurement Methods for High-Power Transients," (61000-4-33, A. Kaelin, Switzerland) is presently in the stage of Committee Draft (CD). This project provides information on the techniques applicable to the measurement of high power transient waveforms. It is important for sensor performance to be standardized so that errors are not made during high field and current testing. Very often normal sensor design and qualification methods are not applicable when high intensity fields and currents are present, often with rise times on the order of one nanosecond. The intention of this work is to identify appropriate sensor calibrations and measurement methods to be used for the measurement of high-power transient electromagnetic disturbances.

Part 5 "*Installation and Mitigation Guidelines*" presently includes five projects. "HEMP protection concepts" (61000-5-3, M. Ianoz, Switzerland) is published as an IEC Technical Report. The concepts presented are based on general principles such as zoning,

grounding, component selection, and circuit and equipment design.

"Specification of protective devices for HEMP radiated disturbance" (61000-5-4, J. Delaballe, France) is published as an IEC Technical Report. This document identifies the parameters required to accurately specify protection from HEMP radiated fields.

"Specification of protective devices for HEMP conducted disturbance" (61000-5-5, W. K. Büchler, Switzerland) is published as an IEC International Standard. This standard identifies the parameters required to accurately specify protection from HEMP conducted environments.

A project originally started by SC77B, "Mitigation of external EM influences" (61000-5-6, W. A. Radasky, USA) was transferred to SC77C several years ago. The project will result in a published standard in the beginning of year 2002. The document contains specific information on installation practices for shielding and other types of electromagnetic protection. As indicated by the title, this publication will cover both HEMP and other external EM disturbances such as lightning and radiated high power EM threats.

"Degrees of protection against electromagnetic disturbances provided by enclosures (EM code)" (61000-5-7, C. W. Jones, USA) describes protection properties offered by different types of equipment enclosures with respect to electromagnetic fields. In addition, standardized test methods are identified along with a codification system that is consistent with the IP code previously established by the IEC. This project has been published as an IEC International Standard.

Part 6 "Generic Standards" includes "HEMP Immunity for Indoor Equipment" (61000-6-6, P. R. Barnes, USA), is presently issued as a Committee Draft for Vote (CDV). This work identifies a generic set of environments and test levels to be applied to electronic equipment located inside of buildings to survive the effects of HEMP.

6 Conclusions

This paper has defined the meaning of Intentional EMI and has briefly described the basic categories of the disturbances. In addition, past experiments have been reviewed to indicate the relatively low levels of radiated and conducted environments that have been shown to affect commercial equipment ranging from PCs to automobiles. Since it is clear that the levels of fields and voltages that can be created using current technology are sufficient to upset and/or damage equipment, there is a standardization effort in the IEC to develop a solution. IEC Subcommittee 77C is in the process of developing a family of HEMP and high power EM transient standards and reports that will help manufacturers and facility owners to protect their equipment from the effects of these disturbances.

References

[1] The Radio Science Bulletin, No 290, September 1999, pp.62-63.

[2] M. W. Wik, R. L. Gardner, W. A. Radasky "Electromagnetic Terrorism and Adverse Effects of High Power Electromagnetic Environments," Workshop W4, *13th International Zurich Symposium and Technical Exhibition on EMC,* February 1999, pp. 181-185.

[3] W. A. Radasky, M. W. Wik, "An Approach to Protect Civil Systems from the Adverse Effects of High Power Electromagnetic (HPE) Environments," Workshop W4, *13th International Zurich Symposium and Technical Exhibition on EMC,* February 1999, p. 187.

[4] M. Bäckström, "HPM Testing of a Car: A Representative Example of the Susceptibility of Civil Systems," W4, *13th International Zurich Symposium and Technical Exhibition on EMC,* February 1999, pp. 189-190.

[5] M. Ianoz, H. Wipf, "Modeling and Simulation Methods to Assess EM Terrorism Effects," Workshop W4, *13th International Zurich Symposium and Technical Exhibition on EMC,* February 1999, pp. 191-194.

[6] J. C. Giles, "Use of Nuclear Electromagnetic Pulse (EMP) Simulators for Protection of Civil systems Against Other Electromagnetic Threats," Workshop W4, *13th International Zurich Symposium and Technical Exhibition on EMC,* February 1999, 195-200.

[7] M. W. Wik, W. A. Radasky, R. L. Gardner, "Intentional Electromagnetic Interference (EMI) -- What is the Threat and What Can We Do About It?" *15th International Wroclaw Symposium and Exhibition on EMC*, June 2000, pp. 896-897.

[8] R. L. Gardner, D. C. Stoudt, C. E. Baum, "Testing Strategies for Susceptibility Testing in High Power Electromagnetics," *15th International Wroclaw Symposium and Exhibition on EMC*, June 2000, pp. 898-901.

[9] W. A. Radasky, M. W. Wik, "The Standardisation of High Power Electromagnetic Transient Phenomena in the IEC," *15th International Wroclaw Symposium and Exhibition on EMC*, June 2000, pp. 893-895.

[10] IEC 61000-2-9, "Electromagnetic Compatibility (EMC) - Part 2: Environment - Section 9: Description of HEMP environment - Radiated disturbance," International Electrotechnical Commission, 1996.

[11] W. Radasky, M. Messier, W. Wik, "Intentional Electromagnetic Interference (EMI) -- Test Data and Implications," *14th International Zurich Symposium and Technical Exhibition on EMC*, February 2001.

[12] V. Fortov, V. Loborev, Yu. Parfenov, V. Sizranov, B. Yankovskii, W. Radasky, "Estimation of Pulse Electromagnetic Disturbances Penetrating into Computers Through Building Power and Earthing Circuits," Metatech Corporation, Meta-R-176, December 2000.

[13] J. LoVetri, A. Wilbers, A. Zwamborn, "Microwave interaction with a personal computer: Experiment and modeling," *13th International Zurich Symposium and Technical Exhibition on EMC,* February 1999, pp. 203-206.

Protection of Aircraft against Electromagnetic Effects

Bo Wahlgren

Saab Avionics AB, 581 88 Linköping, Sweden

Abstract

When designing an advanced and complex system, such as a modern civil or military aircraft, it is absolutly necessary to take into consideration all possible adverse effects caused by electromagnetic fields. This includes the conventional investigation of necessary measures to achive electromagnetic compatibility (EMC) but also design of adequate protection against lightning and radiation from powerful radio and radar transmitters. Future new aircraft will probably be designed to allow for the passangers to use any of their electronic devises during flight.

Very powerful numerical tools are used for prediction of electromagnetic effects. By the use of such tools at the very beginning of a design phase it is possible to avoid expensive modification or redesign at a later phase.

1 Introduction

Electromagnetic Compatibility (EMC) activities have been carried out by Saab in an organized manner from the beginning of development of the 37 Viggen aircraft, i.e. from the middle of the 1960:ies. At that time the complexity of electrical and electronic systems had reached such a level that it was impossible to design a well functioning system without taking EMC into due consideration. The activities have since then been extended to include antenna-to-antenna coupling, Electrostatic Discharge (ESD), lightning, radiation from radio and radar transmitters (HIRF; High Intensity Radiated Fields), High Power Microwaves (HPM), Electromagnetic Pulse (EMP), Personal Electronic Devices (PED) and ionizing radiation.

Continuously increasing system complexity in the development of the 37 Viggen and 39 Gripen military aircraft as well as the Saab 340 and Saab 2000 commuter aircraft has made it necessary to introduce more sophisticated methods for the control of EMC. Specific factors that have made great impact on the work are the introduction of flight safety critical systems, such as electronic flight and electronic engine control systems, and the use of carbon fiber composite materials in structures.

During the 1990:ies the activities have been more internationally oriented. This has resulted in participation by Saab in several EU-projects and in the development of other European aircraft. Presently Saab Avionics is coordinating a large EU-project called EM-Haz. Furthermore Saab Avionics participates in the development of the Airbus 380 aircraft and additional participation is being discussed.

2 Lightning

Lightning is, as we all know from personal experience, a very powerful natural phenomenon. Peak power can be in the order of tens of terawatts corresponding to the power generated by tens of thousands of nuclear generators. However, since every lightning discharge has a very short duration, the released energy is only in the order of that produced by one nuclear generator during some ten seconds.

The probability of an aircraft being subjected to a lightning strike is dependent on such factors as flight duration, altitude and number of starts and landings. For civil aircraft the probability is typically a few times per year. When an aircraft is struck it is subjected to very high electromagnetic strains. Extensive protection measures against the direct and indirect effects are therefore implemented in the design in order to ensure a continued safe flight and landing after such an event.

Direct effects of lightning are experienced when the large current of up to 200 kA passes into and through the structure causing different types of damage mainly through mechanical and thermal action. Modern aircraft are built using more advanced materials than older aircraft, which were almost entirely built using aluminum sheet metal. These new materials are often electrically non-conducting, such as glass fiber composite, semiconducting, such as carbon fiber composite, or consisting of very thin foils and layers of glue, making them less resistant to damage caused by a lightning current and thus requiring additional protection.

Indirect effects are experienced when the electromagnetic fields of lightning induce voltages and currents in the aircraft wiring, which reaches the electrical and electronic equipment causing damage or

disturbing their function. In modern aircraft there are many systems which need to function undisturbed in order to provide a safe flight, such as electronic flight and electronic engine control systems, landing systems and electronic display systems. Protection measures such as cable shields, filtering and transient suppression are therefore implemented in the aircraft installation or in the equipment. In addition such systems normally have redundancy of their flight critical functions.

Verification of the ability of the aircraft to withstand lightning is carried out partly by the use of advanced electromagnetic simulation and computation methods and partly by testing of the hardness against both direct and indirect effects. Connecting a specially designed lightning generator to the aircraft and driving a simulated lightning current through it tests the latter. Induced currents and voltages in the aircraft wiring are measured and compared with the hardness levels, which have also been checked through testing.

3 Radiation from Radio and Radar Transmitters

Radiation from radio and radar transmitters has existed almost during one century and from radar transmitters during half a century. It is, however, not until the last couple of decades that the necessity to protect aircraft against this type of radiation has become evident. The reason for this is of course what has previously been mentioned about introduction of flight safety critical electronic systems. For military aircraft protection against the extension of HIRF into the weapon type of radiation called HPM is necessary.

The most powerful radiation emanates from short and medium wave (SW and MW), TV and radar transmitters. Civil aviation authorities have defined the levels of field strengths that civil aircraft must be designed to withstand. These requirements are up to 100 V/m in the frequency range 10 kHz to 400 MHz, i.e. radiation from radio transmitters, and up to peak field strengths of 3 kV/m in the frequency range 400 MHz to 40 GHz, i.e. radiation from radar transmitters. The field strengths possible to generate by state-of-the-art HPM-weapons and expose a military aircraft to can not be found in open literature, but from nomograms in the open literature they can be estimated to reach levels from 10 kV/m to 100 kV/m.

Verification of the ability of the aircraft to withstand HIRF is, similarly as for lightning, carried out partly by the use of advanced electromagnetic simulation and computation methods and partly by testing. At radio frequencies a test method called Low Level Swept Coupling can be used. The aircraft is then subjected to low level fields from different types of antennas placed close to it while induced currents are measured in the aircraft wiring. The measured currents are then extrapolated to the specified field strengths and compared to the equipment hardness levels. For testing at radar frequencies the aircraft can be subjected to the specified field strengths at discrete frequencies while being in full flight configuration. Any influence on the aircraft systems caused by the radiation can then be identified and evaluated simultaneously.

4 Other Types of Radiation

Radiation from different types of PED:s is a problem today and will continue to be so in the future. It will not be accepted by passengers not to be able to use PED:s onboard and it would be very difficult to guarantee that all such equipment were switched off. Therefore future aircraft will be designed to allow for the use of PED:s. Radiation from PED:s and its coupling to aircraft antennas and wiring are studied and different protection methods are developed in the EU-project EM-Haz.

Another type of radiation, which is and will become an even greater problem for electronics is cosmic radiation. This type of radiation, which is not strictly an electromagnetic but a combination of electromagnetic and particle radiation, emanates from the sun and from cosmos. Its main effect is to cause so called "single event upsets" in electronic components, which can cause permanent changes to information stored in electronic memories. The effects on electronics in space are well known since the beginning of the space era, but for electronics on the earth the effects has only started to show during the last decade, when electronic chips have become sufficiently small to be affected by the lower levels of radiation within the atmosphere. The radiation at aircraft altitudes is about 100 times higher than at ground levels, so the effects are first seen in aircraft electronics. The radiation is also higher at higher latitudes because of the magnetic fields from the earth, the effects of which we can even enjoy visually by observing the Aurora Borealis phenomenon.

5 Modelling

For aircraft projects Saab has adopted a Development Model which starts already at initiation of a project and then runs through a number of phases. In the first phase numerical analysis of the aircraft is carried out using very powerful numerical codes and computers. This results in the definition of requirements to be applied to the complete aircraft including the mechanical structure its installations, wiring, systems and equipment. The model then passes through phases of design and test reviews, testing of equipment and systems, testing and analysis of the complete aircraft, evaluation and finally certification.

Computational Electromagnetics (CEM) uses several numerical methods to solve Maxwell's equations. Some of them run in the time domain and some in the frequency domain. Common for most of them is that they need to be run on very powerful computers. One of the most common methods is called Finite Difference in the Time Domain (FDTD). The result of the calculations is expressed as the electric and magnetic fields at the object (aircraft), which can also be converted to currents, current densities, charge densities or voltages.

Modelling and simulation is continuously growing in electromagnetics as it is in many other disciplines. The reason is that there are many benefits associated with the use of CEM such as early prediction of characteristics, multidisciplinary optimisation, less testing and as a result substantial cost reduction. Therefore a lot of effort is spent on the development of existing and new numerical methods, some of that being carried out in the EU-project EM-Haz. Within Sweden the National Aeronautical Research Programme (NFFP) supports extensive development of CEM carried out by universities, research institutes and industry.

The probability of interference due to coupling from one antenna to another is predicted using different modelling methods. This type of possible interference has been addressed in aircraft development for a long time, but modern aircraft carry more transmitters, receivers and antennas, which increases the risk for interference. In addition, interoperability between different aircraft and between aircraft and other antenna installations has become more important. Antenna-to-antenna modelling and development of more sophisticated modelling tools are therefore of increasing importance.

6 Conclusion

Due to the high level of complexity in aircraft systems concentrated within a small space, the aircraft community has, together with the space community, been in the front as regards the use and development of methods for protection of systems against electromagnetic effects. In the last decade it has become increasingly evident to actors in other civil and military sectors that it is absolutely necessary to obtain knowledge in and to apply this discipline to their projects. Therefore, today we have a large and growing number of interested and knowledgeable individuals who look forward to the further development of this discipline in electromagnetics, which also involves other disciplines such as mechanics, thermal technology, particle physics and microelectronics.

Microwave Breakdown in Slots

D. Anderson[1], U. Jordan[1], A. Kim[3], M. Lisak[1]
O. Lundén[2], M. Bäckström[2], I. Ferrer[2]

[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden.
E-mail: elfda@elmagn.chalmers.se.
[2]Department of Microwave Technology, Swedish Defence Research Agency,
FOI, Box 1165, SE–581 11 Linköping.
[3]Institute of Applied Physics, Russian Academy of Sciences,
46 Ulyanov Street, 603950 Nizhny Novgorod, Russia.

Abstract

In the present work the dynamics of pulsed microwave breakdown in slots at atmospheric air pressure is investigated both experimentally and theoretically, with emphasis on parameters such as microwave intensity breakdown threshold, time-to-breakdown, transmitted energy during the linear breakdown stage and transmitted power during the quasi-stationary discharge stage.

1 Introduction

Microwave breakdown in air is due to the ionising action of free electrons accelerated by the electromagnetic (EM) field. The initial seed electrons are created by background radiation, e. g. cosmic radiation. When the accelerated electrons collide with neutral gas molecules several processes may occur, among them impact ionisation and attachment on the neutral molecules, resulting in gain or loss of free electrons respectively. At subatmospheric pressures or small dimensions, diffusion of free electrons out from the breakdown plasma region to walls or low field regions provides another loss mechanism. The geometry, both physical and electrical, of a region influences both the EM field distribution and the diffusion properties, and thus affects both the ionisation and diffusion rates.

In electronic equipment slots and apertures are encountered frequently, some due to mechanical design and others due to electrical design, e. g. slot antennas. In the case of an EM shielded device, imperfections in the shielding may provide apertures with unwanted transmission properties.

Linear transmission of EM radiation through a slot becomes large at resonance[1, 2]. At the same time the electric field strength becomes large, which may lead to microwave breakdown of the air in the slot, with a discharge sustained by the microwave field. This nonlinear process puts an inherent limit on the maximum power that can be transmitted through a slot of given geometry. This limit may be a problem if the purpose of the slot is to transmit EM radiation, but also an advantage, since it can provide integrated protection against strong external radiation. In the case of imperfections in the EM shield of a device, this latter protection mechanism may be of interest, provided that its parameters can be reliably determined during the design phase of the device.

2 Pulsed Breakdown Theory

When the free electrons present in a gas are agitated by a strong microwave field some of their collisions with the neutral gas molecules cause impact ionisation, thus increasing the number of free electrons. When electrons hit neutral molecules they may instead attach and form negative ions. Since the inertia of the formed negative ions is much larger than that of an electron, these charges are effectively lost from the plasma breakdown point of view. In many cases the process can be modelled by the equation of continuity for the free electron density, cf [3]

$$\frac{\partial n}{\partial t} = \nabla \cdot (D \nabla n) + \nu_i(|E|)n - \nu_a n + S, \qquad (1)$$

where n is the electron density, D is the diffusion coefficient, ν_i is the ionisation frequency accounting for the impact ionisation process, ν_a is the attachment frequency, and S is an independent source of free electrons, e. g. cosmic radiation. Eq. (1) is to be solved with appropriate boundary conditions, i. e. $n = 0$ on walls that absorb electrons and $\nabla n = 0$ on symmetric boundaries or far away from the breakdown region. For the present purpose we use semi-empirical expressions for the model parameters. In the case of air, $D \simeq 10^6/p$ ($\mathrm{cm^2 s^{-1}}$), $\nu_i \simeq 9 \times 10^{-4} p (E_e/p)^\beta$ ($\mathrm{s^{-1}}$), and $\nu_a \simeq 6 \times 10^4 p$ ($\mathrm{s^{-1}}$), where p is the pressure in Torr and $\beta \approx 5.33$. E_e is the reduced electric field, $E_e^2 = E_{\mathrm{rms}}^2/(1 + \omega^2/\nu_c^2)$, where $\nu_c \simeq 5 \times 10^9 p$ ($\mathrm{s^{-1}}$) is the electron-neutral collision frequency, cf [3]. Furthermore, S is taken as a given constant.

Exact analytical solutions can be found in simple geometries, while for more complicated (and realistic) geometries one must resort to approximate analytical or full numerical solutions. If the geomtry is not too complicated, one may simplify the equation by introducing an effective diffusion length, L_{eff}, and a field enhancement factor, α, for the structure. The diffusion losses are approximately represented as $\nabla \cdot (D\nabla n) \approx -(D/L_{\text{eff}}^2)n$, and the enhanced electric field $E = \alpha E_{\text{inc}}$, where E_{inc} is some incident scaling field, is used instead of the real field distribution. The continuity equation simplifies to

$$\frac{dn}{dt} = -\frac{D}{L_{\text{eff}}^2}n + \nu_i(\alpha E_{\text{inc}})n - \nu_a n + S \qquad (2)$$

In the absence of a strong microwave field the stationary solution is $n_0 = S/(\nu_a + D/L_{\text{eff}}^2)$, which determines the background electron density. If the source function S is very weak the average density description provided by n_0 is not very good, and there will be a small fluctuating number of electrons in the system, leading to the problem of stochastically appearing seed electrons for the breakdown.

Breakdown is said to occur for an event that significantly changes the wave propagation properties of the structure. The usual CW-breakdown condition is $dn/dt = 0$, i. e. the boundary between exponentially growing or decaying solutions to Eqs. (1) and (2). If, on the other hand, breakdown must be realised within a finite time, as in the case of microwave pulses, the breakdown condition must be modified. The appropriate condition is that the electron density should exponentiate to a high level n_H, at which the electromagnetic properties of the structure change. Unfortunately, the level n_H is not independent of the structure. Often the level is taken as the critical electron density n_c, at which the electron plasma frequency $\omega_p = (ne^2/\epsilon_0 m)^{1/2}$ equals the wave frequency ω, but this level may be too high in e. g. resonant structures. At the same time, the background level n_0, from which the exponentiation starts, is not very well determined either. One usually expresses the breakdown condition in the logarithmic form

$$\ln\left(\frac{n(t)}{n_0}\right) = \ln\left(\frac{n_H}{n_0}\right) \approx 20, \qquad (3)$$

where the number 20 is somewhat arbitrary (different values are found in the literature, of about the same order).

The solution for the breakdown field strength threshold of Eq. (1) or (2) as a function of gas pressure with the appropriate breakdown condition is called the Paschen curve. To use the simplified model presented in Eq. (2) for a slot the effective diffusion length and field enhancement factor must be determined. The determination of the parameters depends on the

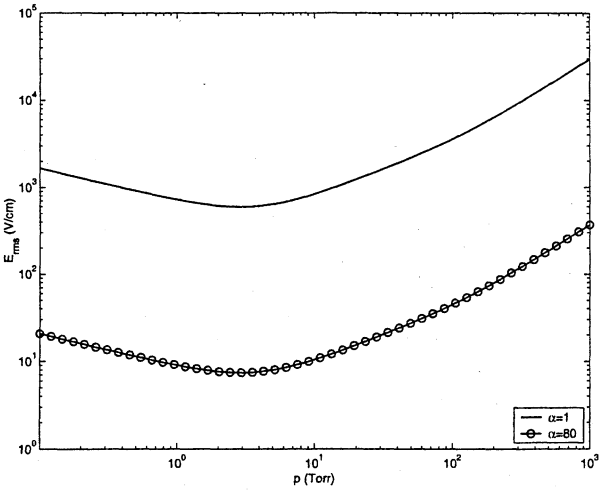


Figure 1: Breakdown field strength vs pressure showing the influence of the field enhancement factor α, for air at $f = 3$ GHz and $L_{\text{eff}} = 0.03$ cm.

local conditions near the slot, since they control whether the breakdown will be in the volume of the slot or near the possibly sharp edges of the surrounding walls. That said, let us consider the simpler case of breakdown of the air in the volume of the slot. For a narrow rectangular slot the field enhancement is strongest when the electric field is polarised perpendicular to the longest transverse slot dimension. In this case the diffusion length can be estimated to $L_{\text{eff}} \approx d/\pi$, where d is the smaller transverse dimension of the slot, since the ionisation frequency is homogeneous in that direction. In Fig. 1 the CW-breakdown thresholds for a geometry with finite diffusion length without field enhancement ($L_{\text{eff}} = 0.03$ cm, $\alpha = 1$) and a hypothetical slot with the same diffusion length ($L_{\text{eff}} = 0.03$ cm, $\alpha = 80$) are compared. The field enhancement factor decreases the threshold by providing a stronger local field.

The effect of finite pulse length can be presented in two ways: the increased breakdown threshold for a given pulse length, i. e. a Paschen curve shifted towards higher thresholds, or the breakdown time t_B vs the incident field. Neglecting the source term S, the solution of Eq. (2) is

$$n(t) = n_0 \exp\left(\int_{t_0}^{t} \nu_{\text{net}}(|E(t')|)\, dt'\right), \qquad (4)$$

where $\nu_{\text{net}} = \nu_i - \nu_a - D/L_{\text{eff}}^2$ is the net ionisation frequency and the exponentiation starts at $t = t_0$ from $n(t_0) = n_0$. With this equation the breakdown condition (3) can be rewritten as $\int_{t_0}^{t} \nu_{\text{net}}(|E(t')|)\, dt' \approx 20$. For a rectangular pulse of duration t_B this reduces to the simple condition

$$\nu_{\text{net}}(|E|)t_B \approx 20 \qquad (5)$$

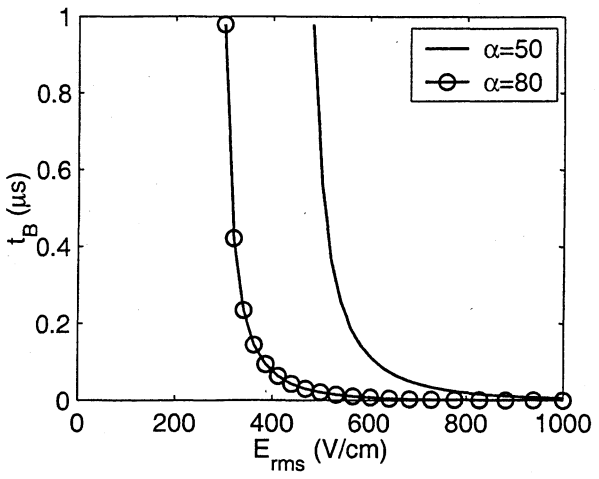


Figure 2: Breakdown time vs incident electric field strength for different field enhancement factors α, in air at $f = 3$ GHz, $L_{\text{eff}} = 0.03$ cm, and $p = 760$ Torr.

In Fig. 2 the latter relation is presented for two different field enhancement factors, $\alpha = 50$ and $\alpha = 80$, with an effective diffusion length of $L_{\text{eff}} = 0.03$ cm, for air of atmospheric pressure and microwave frequency $f = 3$ GHz. The curve has a singularity at the CW-breakdown threshold, and then falls rapidly for increasing field strengths due to the rapid growth of the ionisation frequency with increasing electric field. The effect on a microwave pulse of duration τ_p is called pulse tail erosion, cf Ref. [4]. Short pulses, $\tau_p < t_B$, will be transmitted through the structure unaffected, while for long pulses, $\tau_p > t_B$, the end of the pulse is "eroded" due to reflection and absorption by the breakdown plasma when the structure switches from the linear to the nonlinear stage.

The CW-breakdown threshold puts a limit on the maximum power that can be transmitted through the slot without the signal being affected by breakdown. For pulses of higher power the pulse tail erosion limits the total transmitted energy for a given pulse power. The transmitted energy is

$$\mathcal{E}(t) = \int_{t_0}^{t} P_t(t')\, dt', \tag{6}$$

where P_t is the transmitted power. Before breakdown P_t is the linearly transmitted power, while in the nonlinear phase P_t will typically be a small residual transmitted power. Numerical simulations of the transmitted energy for simplified slot geometries in air were presented by Sieger *et al.* in Ref. [5]. A simple estimate for the transmitted energy in the rectangular pulse approximation is linear growth to the total energy given by

$$\mathcal{E}_{\text{tot}} = P_t \min(\tau_p, t_B). \tag{7}$$

For constant pulse duration t_B, the total transmitted energy scales as $\mathcal{E}_{\text{tot}} \propto E^2$ for field strengths below the pulsed breakdown threshold, while for pulses of higher field strength, the total energy decreases due to the shorter and shorter breakdown times, $\mathcal{E}_{\text{tot}} \propto E^{2-\beta}$.

3 Experiments

A series of experiments in air at atmospheric pressure has been performed, where a slot of dimensions 46×0.2 mm^2 in a metallic plane has been irradiated using a 700 kW magnetron with 1 μs long microwave pulses of frequency $f = 3$ GHz, while varying the incident field strength by varying the distance between the transmitting antenna and the slot. Both the incident field strength and the transmitted power was monitored[6].

In Fig. 3 some typical results are shown for four pulses of different field strength. The weakest pulse propagates unaffected through the slot, while all other show strong pulse tail erosion. Note the small difference in field strength between the weakest and the second weakest pulse, and the great impact it has on the tail erosion, cf the steep decrease in breakdown time for increasing field strength shown in Fig. 2. The strongest pulse causes breakdown early on its leading flank. The picture is confirmed when the transmitted energy is considered. The weakest pulse delivers its energy efficiently, while the two pulses with slightly stronger electric field are able to transmit less than half the energy. The strongest pulse delivers least total energy due to the rapid switch off in transmission caused by the created breakdown plasma. When all different realisations in the experimental series are considered, some variations are of course present, but the general picture remains the same.

4 Conclusions and Outlook

A preliminary study of pulsed microwave breakdown in slots has been presented. Simple theoretical models for microwave breakdown in air around atmospheric pressure have been presented and reviewed, with emphasis on the scaling of breakdown threshold, breakdown time, and transmitted energy with relevant parameters. Furthermore, results from recent experimental work supporting the major features of the theoretical models have been presented.

Future work will be directed towards a more detailed description of both the breakdown and discharge stages in slots of different geometries, with better determination of the electric field enhancement factor and the relative importance of breakdown in the volume of the slot vs breakdown near the sharp edges often present in slot structures.

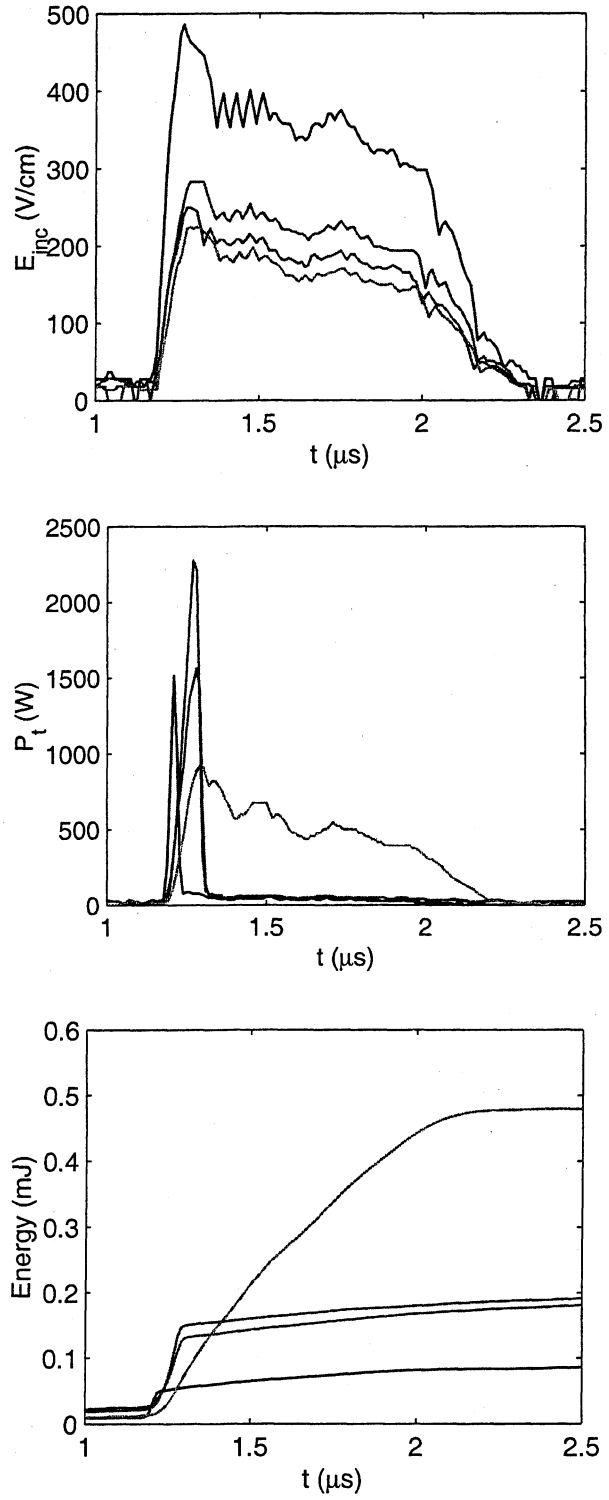


Figure 3: Pulsed breakdown experiment. Four different distances between antenna and slot (different field strengths) are shown. The top panel shows the incident electric field strength, the middle panel shows the power transmitted through the slot, and the bottom panel shows the transmitted energy vs time.

References

[1] R. F. Harrington, "Resonant Behavior of a Small Aperture Backed by a Conducting Body", IEEE Trans. on Antennas and Propagation **30**, 205 (1982).

[2] T. Martin, M. Bäckström, and J. Lorén, "Transmission Cross Section of Apertures Determined by Measurements and FDTD Simulations", 12th International Zürich Symposium and Technical Exhibition on Electromagnetic Compatibility, February 18-20, 1997.

[3] A. D. MacDonald, Microwave Breakdown in Gases, (New York; Wiley) 1966.

[4] S. P. Kuo and Y. S. Zhang, "A theoretical model for intense microwave pulse propagation in an air breakdown environment", Physics of Fluids B **3**, 2906 (1991).

[5] G. E. Sieger, J. H. Yee, and D. J. Mayhall, "Computer Simulation of Nonlinear Coupling of High-Power Microwaves With Slots", IEEE Trans. on Plasma Science **17**, 616 (1989).

[6] M. Bäckström, O. Lundén, P. Andersson, I. Ferrer och J. Lorén, "HPM. Framvägskoppling och skydd. Lägesrapport.", FOI Användarrapport, FOI-R0265SE, november 2001. Totalförsvarets Forskningsinstitut, Sensorteknik, Box 1165, Linköping.

Microwave Coupling into a Generic Object. Angular Pattern of Shielding Effectiveness and at Immunity Testing.

Mats Bäckström, Jörgen Lorén and Magnus Höijer

Swedish Defence Research Agency FOI, Box 1165, SE-581 11 Linköping, Sweden

Abstract

Angular resolved plane wave coupling and coupling in the isotropic environment of a reverberation chamber have been studied for a generic missile. The former yields the angular resolution necessary to include the most susceptible directions in testing. The average coupling, taken over all angles of incidence and polarizations, in the anechoic chamber is shown to be equal to the coupling measured in the reverberation chamber. Typical differences between the maximum and average plane wave coupling are 10 – 15 dB. The results are of crucial importance what concern the relationship between testing in anechoic and reverberation chambers. The results of the low level coupling measurements have been compared to results from a high-level susceptibility testing of the missile. Although the apparent directivity pattern in the RS test differs in its details from the low level coupling measurement, typical lobe widths and the maximum apparent directivity are almost identical.

1 Overview and Objectives

This paper presents results from extensive low-level measurement on the generic missile GENEC, between 0.5 and 18 GHz. Coupling from external field to internal probes has been measured. It also presents results from high level radiated susceptibility (RS) testing at 2.857 GHz. The coupling measurements were performed using plane wave incidence, for a large number of angles of incidence, in an anechoic chamber (AC), as well as using the statistically isotropic environment of a reverberation chamber (RC). The RS test was performed in the RC as well as using plane wave incidence, the latter by use of the Swedish Microwave Test Facility (MTF). There have been several objectives for this study. One is that GENEC provides yet another opportunity to study directivity effects at plane wave illumination, in this case for an object with a well-known external and internal geometry. Earlier coupling studies on various types of real, complex, equipment have shown differences of typically 10 – 15 dB between the maximum and average coupling in the AC, the latter equal to the coupling measured in a RC [1]. This knowledge is of fundamental importance regarding RS testing in these chambers and, in particular, regarding the relation between the environments of the AC and RC.

Our results provide information of the angular resolution that is necessary in order to include the most susceptible directions and polarization in a RS test. Also, it provides information of the magnitude of the expected error bias caused by the fact that the most critical angle of incidence and polarization is not, in general, included in a test. In a test, it is in practice neither possible to predict the worst direction nor to use a sufficiently high angular resolution to include this direction with certainty. The error bias is defined as the ratio between the maximum response at the actual test and the true maximum plane wave response.

The angular patterns from the low level coupling measurements have been compared with theoretical prediction used in spherical near-field testing of antennas [2] and, concerning directivity, with recently developed predictions specially derived for the type of equipment presented here [3]. The measured directivity has been compared with the outcome at the high-level immunity test. A crucial question for this study was how well the angular dependence of the failure level at the RS test could be predicted from low level coupling measurements and theory.

The results of the low level coupling measurements are presented in more detail in [5], and of the high level RS test in [6].

2 Test Object GENEC

The test object, GENEC, is shown in figure 1.The total length is 815.5 mm and the diameter is 105 mm. In the measurements presented in this report all apertures, except the slots at the wing axes, were closed. The size of each wing slot is approximately 5 x 88 mm. The size of each wing is 50 x 120 mm. On the inside of the fuselage the wing axes are connected to the structure. The wings are attached in such a way that they do not have any electric connection to the fuselage.

Two field probes, A and B, were installed inside GENEC, see figure 2.

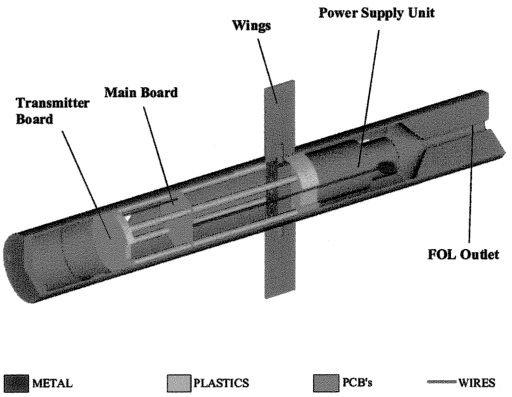


Figure 1. *GENEC, the generic test object.*

Each probe consists of a SMA connector in which a 20 mm centre conductor is inserted. The probes are regarded to be representative for the receiving properties of typical cables and wires.

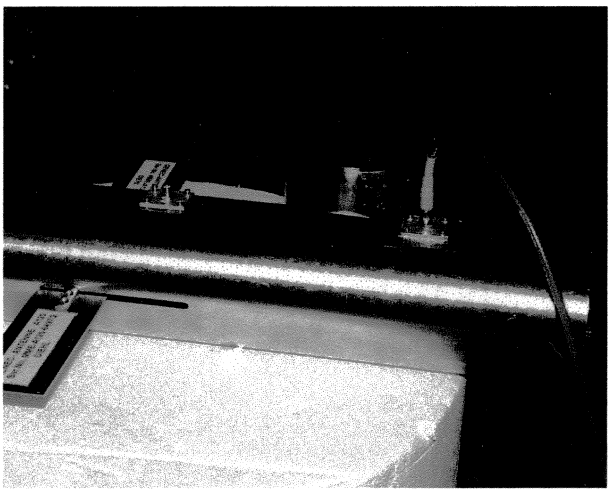


Figure 2. *GENEC, wing slots and location of field probes. Probe position A is close to the wings.*

3 Results

3.1 Low Level Coupling Measurements

The AC measurements were carried out with an angular resolution of one degree, in three perpendicular planes [5]. Two polarizations were used in each plane. Since two probes were used this makes a total of 12 configurations.

Figure 3 shows six examples of angular patterns for the receiving cross section (effective area), defined as the power received by the probe normalised to the power density of the external plane wave. It is clear that the variation vs. angle is "faster" at higher frequencies. This is what to be expected from antenna theory [2].

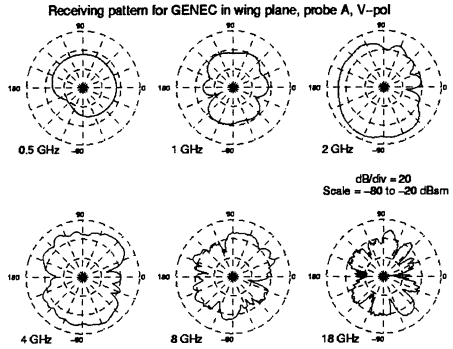


Figure 3. *Angular receiving patterns at low level coupling measurements.*

In [2, pp. 128, pp.190] the maximum angular sampling increment, $\Delta\theta_{\max}$, in order to determine the radiation pattern of an antenna from measured data, is given. It can be approximated by (in degrees):

$$\Delta\theta_{\max} \approx \frac{180}{kr_0 + 10} \qquad (1)$$

where $k = 2\pi/\lambda$ and r_0 is the radius of the smallest possible spherical surface circumscribing the antenna. $\Delta\theta_{\max}$, for r_0 = half the length of *GENEC*, is shown in table 1.

Table 1. Measured and predicted half width, in degrees, of peaks in the receiving pattern of *GENEC*.

Frequency (GHz)	$\Delta\theta_{\max}$ eq.(1)	Most narrow peaks. 3 dB half-width	Maximum peaks. 3dB half width
0.5	13	Min = 33 Ave = 36 ± 3	Min = 41 Ave = 68 ± 5
1	10	Min = 16 Ave = 20 ± 3	Min = 24 Ave = 45 ± 27
2	6.6	Min = 8 Ave = 12 ± 2	Min = 16 Ave = 35 ± 10
4	4.1	Min = 5 Ave = 6 ± 1	Min = 12 Ave = 24 ± 10
8	2.3	Min = 2.5 Ave=3.1 ± 0.6	Min = 3.5 Ave = 15 ± 13
18	1.1	Min = 0.5[1)] Ave < 2.5[1)]	Min = 0.5[1)] Ave = 7 ± 4

1) Estimated. At 18 GHz the angular step of 1 degree is not sufficient to resolve the narrowest peaks.

In table 1 we also show the result of an estimate of the average and the minimum value of the 3 dB half-width of the six narrowest peaks. Also shown is the average and minimum width of the maximum peak of each of the measured configurations. The results in table 1 show that

the sampling criterion given by eq. (1) seems to be well adapted to estimate the angular sampling interval in a radiated susceptibility test.

In figure 4 the receiving cross section is given for the AC as well as for the RC as function of frequency. For the AC the minimum, average and maximum cross sections are given. For the RC, the average cross section, taken over all 200 stirrer positions, is shown.

As expected [1] there are some characteristics to note:

- There is a good match between the average (isotropic) cross-section measured in the RC and the average of the cross-section measured in the AC.
- The maximum values measured in the AC are typically 10 – 15 dB higher than the average cross-section measured in the RC.
- There is a huge difference, around 40 – 50 dB, between the maximum and minimum values measured in the AC. This illustrates the risk for under-testing in an AC if only a few angles of incidence are used.

It can be shown that the ratio between the maximum received power in the AC and the average received power in the RC is equal to two times the maximum directivity, provided the power density in the AC is equal to the scalar power density in the RC, see [7]. The factor of two originates from the effect of polarization. Also, it can be shown that the average of the received power in the AC should be equal to the average received power in the RC. These statements are strictly true only for a full sphere measurement in the AC. Although our measurements were limited to three perpendicular planes we see in figure 4 that the latter statement holds quite well.

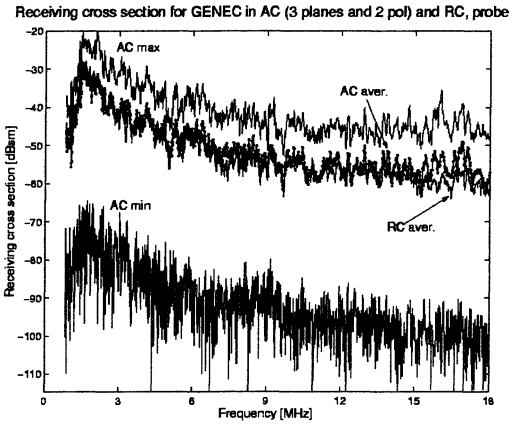


Figure 4. *Impedance matched receiving cross-section in anechoic and reverberation chamber.*

The difference of 10 – 15 dB between the maximum values in the AC and the isotropic value from the RC seen in figure 4, here denoted the *apparent maximum directivity (AMD),* is in accordance with earlier results on other equipment [1]. If a full sphere measurement had been performed in the AC the *apparent maximum directivity* should, as stated above, be equal to two times the maximum directivity of GENEC (since AMD comprise both directivity and polarization). Since our measurements are limited to three planes we expect this to be only approximately true.

As for the angular sampling interval also the maximum directivity can be estimated from antenna theory, see [2, pp. 55]. However, it turns out that this approach severely overestimates the directivity of the type of equipment studied here, by typically 20 - 30 dB, see [3,5,7]. This is not surprising since a shielded enclosure like GENEC is not designed to behave as a highly directive antenna! Recently published results by NIST [3] seem to give a better prediction of measured data. The predictions are also here based on spherical mode theory, but it is assumed that the spherical mode coefficients are independent random variables. For the type of data that are presented in this report, i.e. using three orthogonal cuts with two orthogonal polarizations the results becomes [3]:

$$\langle AMD \rangle = 0.577 + \ln(N_S) + \frac{1}{2N_S} \qquad (2)$$

where < > denotes expectation values and N_s is the number of independent samples, the latter given by:

$$N_S = 12 \cdot (2N - 1) \qquad (3)$$

where $N = kr_0$. Eq. (2) is valid provided $kr_0 > 1$.

In figure 5 the *apparent maximum directivity*, measured in the AC for probe position A, is plotted.

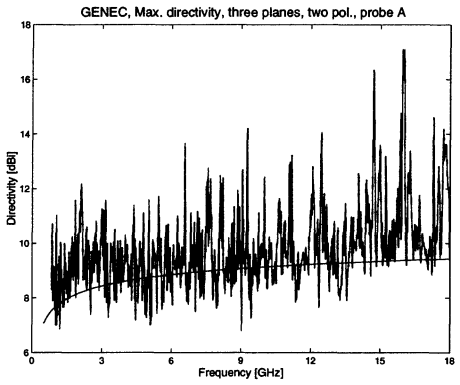


Figure 5. *Measured and calculated (smooth line) apparent maximum directivity.*

In the figure we have also included the expected value according to eq. (2). As can be seen in the figure eq. (2) yields a rather good fit to measured data. It shall be noted, however, that around 5 dB has to be added to the

predicted value to enclose the maximum measured values.

To summarize, the apparent directivity of a test object can be estimated from the size of the object. This is of fundamental importance in relating an RS test in a reverberation chamber to an RS test in an anechoic chamber.

3.2 High-Level Immunity Testing

The high-level immunity investigations were carried out using the Microwave Test Facility (MTF), owned by the Swedish Defence Material Administration (FMV) and operated by Saab Avionics in Linköping, see figure 6. The MTF is described in [4]. High-level tests were also performed in the FOI reverberation chamber. The MTF test was carried out at the frequency 2.857 GHz. The pulse length was 4.7 μs and the pulse repetition frequency was 30 Hz. The field strength (rms) was varied between 250 V/m and 13 kV/m, i.e. between 0.17 kW/m^2 and 0.45 MW/m^2. The irradiation was made in the wing plane of GENEC, cf. [6], with horizontal and vertical polarization respectively.

We early understood that it would take a long time to do a well angular resolved RS test, and that our time at the MTF would be limited. Equation (1) and table 1 was useful to estimate the maximum angular sampling increment that could be accepted. Both eq. (1) and table 1 gives $\Delta\theta_{max} = 5°$. We decided to choose a sampling increment of 3° to have a small security margin.



Figure 6. *GENEC is irradiated from the S-band antenna in the wing plane. All communication with GENEC is done through fibre optic leads to ensure no disturbance from the irradiated HPM-field. The turn table makes it possible to rotate GENEC 360 degrees.*

A lot of work was done to find a good test criterion for when we have a disturbance in the electronics. Finally it was decided to use two digital signals inside GENEC as test criterion. For these two repetitive digital signals a very clear permanent switch from a "1" to a "0" or vice versa could be seen on a large portion of the period of the digital signal, when the irradiating field was increased above a certain level. Eventually, a Bit Error Rate measurement would have given a lower disturbance field.

The result for the horizontal polarization is shown in figure 7. The red solid line curve shows the apparent directivity of the susceptibility of GENEC. The susceptibility is defined as the inverse of the minimum power density, which creates a disturbance in the GENEC-electronics. The directivity is the ratio between the susceptibility and the average susceptibility over all angles. Again, we use the expression apparent directivity to emphasize that it is only the directivity in the plane we have investigated, and it only gives an indication of the whole solid angle directivity. For comparison, the apparent directivity of the two corresponding coupling measurements has also been included in figure 7. The blue dash-dot line curve shows the apparent directivity of probe A, which is positioned close to the wings, where most of the leakage of electromagnetic radiation to the interior of GENEC takes place. The green dashed line curve shows the apparent directivity of probe B, which is positioned closer to the electronics.

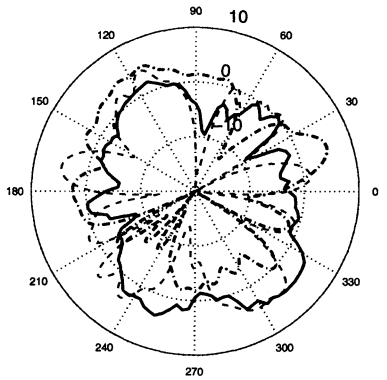


Figure 7. *Apparent directivity (dB) for coupling measurements and RS test of GENEC. The blue dash-dot line curve and the green dashed line curve show the coupling measurement apparent directivity of probe A and probe B, respectively. The red solid line curve shows the apparent directivity for the RS test of GENEC.*

As can be seen in figure 7 the apparent directivity of probe A is different from the apparent directivity of probe B. Hence, it is not possible to find a simple one to one correspondence between the directivity of an arbitrary coupling measurement and the directivity of the susceptibility of a complex object like GENEC. The effect is due to the non-uniform field distribution within

GENEC. However, it is also evident from figure 7, that the apparent directivity of probe B, which is positioned closer to the electronics in GENEC, is more similar to the apparent directivity of the susceptibility of GENEC than the apparent directivity of probe A is.

The apparent maximum directivity of the RS test is 4.7 dB, which is insignificantly smaller than the apparent maximum directivities of probe A and probe B, which are 5.4 dB and 4.9 dB respectively. However, the difference between the apparent maximum and apparent minimum directivity is for probe A, 34 dB and for probe B, 44 dB, but only 14 dB for the RS test. This is, as can be seen in figure 7, a consequence of that for some narrow directions the directivity is very low for probe A and B. (In figure 7 only directivities down to -20 dB can be seen, but the lowest measured directivity is -39dB.)

A possible explanation is that for some angles of incidence the field probe happens to be in an almost complete node. The electronics in GENEC is more complex than a single probe antenna and including non-linear elements, and therefore it is not possible to put the whole GENEC-electronics into a complete node. (The difference between the coupling measurements and RS test is also slightly due the higher resolution in the coupling measurements, but a decrease in the coupling measurements resolution to 3 degrees only reduces the difference between the apparent maximum and apparent minimum directivity to 32 dB for probe A and 37 dB for probe B.)

However, there is generally no practical interest in that for a few narrow angle intervals, the coupling to the interior of GENEC is small[1]. The maximum directivity is therefore a better measure of angular susceptibility and the result, that we have the same maximum apparent directivity in the RS test as in the coupling measurements, is therefore both interesting and important.

In order to study the repeatability, the test was repeated for the horizontal polarization the next day after the set-up had been dismounted and rebuilt. As can be seen in figure 8 the measurement repeatability was good.

4 Acknowledgements

This work was financially supported by the Swedish Armed Forces and the Swedish Defence Material Administration. Our thanks also to the Diehl Company, especially Mr. Frank Sonneman, for all technical information. Finally, we want to thank our colleagues here at FOI, especially Lars Pettersson , Torleif Martin and Olof Lundén, for valuable help and discussions.

[1] However, if coupling measurements are done only for a very few angles there is a non-negligible risk of measuring a far to high shielding effectiveness.

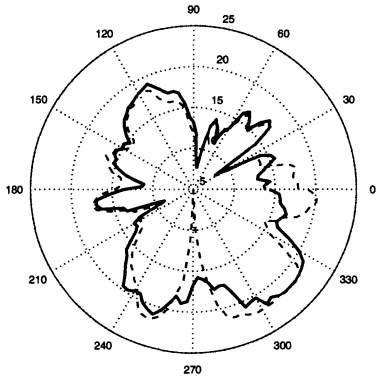


Figure 8. *To test the repeatability of our measurements the same RS test was performed at two different days. The unit dBm^2/W is used for the two susceptibility curves.*

4 References

[1] L. Jansson and M. Bäckström, "Directivity of Equipment and its Effect on Testing in Mode-Stirred and Anechoic Chambers", *Proc. of 1999 IEEE International Symposium on Electromagnetic Compatibility*, Seattle, Washington, USA, 2 – 6 Aug. 1999.

[2] J. E. Hansen (ed.), *Spherical Near-Field Antenna Measurements*, Peter Peregrinus Ltd., London, UK, 1998.

[3] G. Koepke, D. Hill, and J. Ladbury, "Directivity of the Test Device in EMC Measurements", *Proc. of 2000 IEEE International Symposium on Electromagnetic Compatibility*, Washington D C, USA, 21 - 25 Aug. 2000.

[4] *Microwave Test Facility*, Information brochure issued by Saab Avionics AB, Electromagnetic Technology Division, SE-581 88 Linköping, Sweden, 1998.

[5] M. Bäckström and J. Lorén, *Microwave Coupling into a Generic Object. Properties of Measured Angular Receiving Pattern and its Significance for Testing in Anechoic and Reverberation Chambers,* FOI Report FOI –R--0392—SE, February 2002, Swedish Defence Research Agency FOI, Sensor Technology, P.O. Box 1165, SE-581 11 Linköping, Sweden.

[6] M. Höijer, *Comparison between High Level RS Tests and Coupling Measurements*, FOI Report to be published in 2002, Swedish Defence Research Agency FOI, Sensor Technology, P.O. Box 1165, SE-581 11 Linköping, Sweden.

[7] M. Bäckström, J. Lorén, G. Eriksson, and H-J Åsander, "Microwave Coupling into a Generic Object. Properties of Measured angular Receiving Pattern and its Significance for Testing", *Proc. of 2001 IEEE International Symposium on Electromagnetic Compatibility*, Montreal, Canada, 13 - 17 August 2001.

Microwave shielding effectiveness of large mobile military systems and a low cost HPM-indicator

Kurt Eriksson[1], Björn Olsson[1], Karl-Gunnar Lövstrand[2], Barbro Nordström[3] and Mats Bäckström[4]

[1] AerotechTelub AB, Box 360, SE-83125 Östersund, Sweden. E-mail: kurt.g.eriksson@aerotechtelub.se.
[2] Swedish Defence Materiel Administration, SE-115 88 Stockholm, Sweden. E-mail: karl-gunnar.lovstrand@fmv.se.
[3] Swedish Defence Materiel Administration, Box 13400, SE-58013 Linköping. E-mail: barbro.nordstrom@fmv.se.
[4] Swedish Defence Research Agency FOI, Box 1165, SE-581 11 Linköping, Sweden. E-mail: mats@foi.se.

Abstract

The HPM (High Power Microwave) threat has become a burning issue in recent years. As many systems in the past 20 years have been designed with an external protective shield in order to handle the NEMP (Nuclear Electromagnetic Pulse) threat, an interesting issue was to investigate the shielding effectiveness of those systems above 1 GHz. The focus of this paper is to give a summary of the results from these investigations and how they where obtained. The following two questions are also discussed:

Is it possible to build a low cost HPM-indicator in order to see if some equipment has been exposed to HPM?

ESD (Electrostatic Discharge) shielding bags is designed to shield against ESD which have a frequency content above 1 GHz. Would it be possible to use them also as an HPM protection during storage and transportation of spare parts?

1 Introduction

In a project regarding HPM protection measures, financed by the Swedish Defence Materiel Administration, the EMC section at AerotechTelub AB has performed three interesting sub projects:

- Shielding effectiveness measurements on a sample of large mobile military systems with potential shielding at microwave frequencies
- Development, and testing, of a low cost passive HPM indicator
- Shielding effectiveness measurements on ESD shielding bags

2 Shielding effectiveness in large mobile military systems

2.1 Background

Many mobile military systems, such as special vehicles and containers, have often been designed with an external shielding shell in which sensitive equipment, as for example computers, has been installed. There are often several reasons why this solution is chosen, for example:

-mechanical protection and a rugged platform,
-EMI shielding to nearby radio receivers,
-TEMPEST requirements,
-often one important part of the NEMP protection.

The shielding requirement for these systems is typically as SS_C in the EMMA-handbook [1] This requirement specifies a shielding effectiveness of 50 dB in the frequency range 100 kHz to 100 MHz, a reduction with 20 dB / decade from 100 kHz to 10 kHz and from 100 MHz to 1 GHz. This requirement does not specify any attenuation above 1 GHz, but the requirement below 1 GHz forces the designs to be made in sheet metal, quite well jointed together and with EMI shielding gaskets in doors and hatches. These systems are often built in an aluminium sandwich material riveted together at about 50 mm intervals. Cables are brought in and out of the shielding through filters. Due to these design principles, some attenuation above 1 GHz could be expected. An interesting issue was therefore to investigate the shielding effectiveness above 1 GHz.

2.2 Test procedure

The shielding effectiveness was measured typically according to IEEE Std 299-1997 [2].

The measurements were performed with a computer controlled measuring system, consisting of a RF signal generator, power amplifier, transmitting horn antenna, analogue fibre optic link with a broadband monopole

antenna and a spectrum analyser. Due to the nearly isotropic field conditions in the compartments an isotropic or near-isotropic antenna should be used [3].

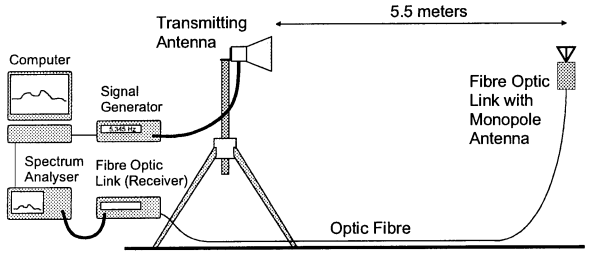


Figure 1: *Measurement set-up.*

Thanks to the very small size of the fibre optic link (47x75x14 mm) it was also possible to measure the shielding inside quite small, shielded cabinets inside the systems.

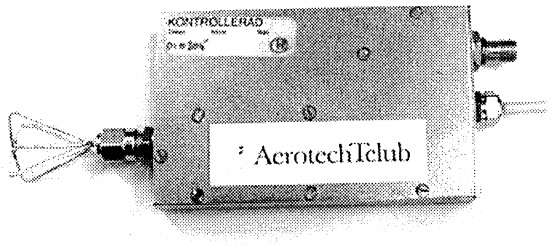


Figure 2: *Fibre optic link with monopole antenna.*

The shielding effectiveness was measured in the frequency range 1- 6 GHz with a step size of 10 MHz.

First, a reference measurement was done in free space with vertical polarisation and a separation distance of 5.5 meters between the transmitting and the receiving antennas. The reference was taken as an average for 12, equally spaced, angles of incidence.

Then the receiving antenna was placed inside the shielded object to be measured. The transmitting antenna was placed outside the object, pointing at the receiving antenna. The distance to the receiving antenna was maintained.

Measurements were performed with both horizontal and vertical polarisation at about 5-10 critical test points on each object such as doors, hatches, joints and cable feedthrough.

2.3 Test samples

Six different test objects with potential shielding were investigated.

- Two different lightweight 20-foot containers
- Two different special purpose vehicles with a shielded compartment
- One armoured wheeled vehicle
- One combat vehicle

The first four objects are designed to have a certain shielding, and they were all built according to the principles described in section 2.1 above. Three of them had one door (about 0.9x2 meter) and one hatch (about 0.5x0.5 meter). The fourth object had two doors (about 0.8x2 meter). All of them were equipped with EMI shielding gaskets of the "contact strips" type in tin-plated steel or beryllium copper.

The two armoured vehicles were both made from welded steel. They did not have EMI shielding gaskets in doors and hatches but the steel overlapped with 30-40 mm.

2.4 Results

The shielding effectiveness of the four, shielded, objects varied in the range of 10-35 dB. The lower attenuation was due to incorrectly installed or adjusted contact strips. Typical faults observed were:

- Paint behind contact strips, or on opposite contact surfaces due to incorrect masking before varnishing.
- Incorrectly mounted strips.
- Doors with improper tolerances or with incorrect pressure in the locked position.

The results showed the importance of quality assurance during manufacturing. One reliably way of doing this is by verifying the shielding effectiveness in one or two positions of each produced unit.

The contact strips in one of the systems, which had been used quite a lot, were in need of cleaning and maintenance. Procedures for this should be included in the normal maintenance routines.

The average shielding effectiveness of the first four objects was about 25 dB. For objects with optimally installed and adjusted contact strips an attenuation of about 30 dB was obtained, and this could probable be taken as a practical maximum value in the frequency range 1-6 GHz for this type of design.

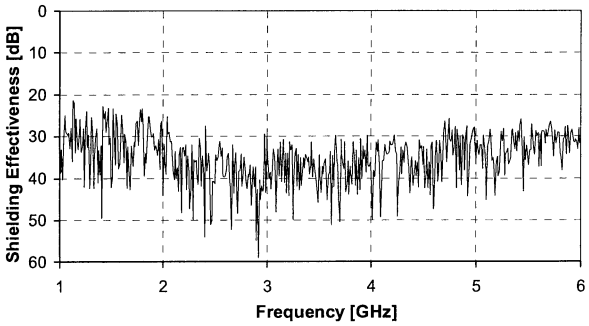


Figure 3: *Typical measured shielding effectiveness of this type of design.*

The large variation in the graph above is caused by standing waves in the test object and the attenuation should therefore be read as the minimum envelope of the curve.

The attenuation from the outside into cabinets inside the outer shield was typically in the range of 40-60 dB.

The shielding effectiveness of the two armoured vehicles was typically 10-20 dB. In directions without doors or hatches it was up to 20 dB and at doors it was about 10 dB. Just in front of the prisms, which are used to allowing the crew seeing out with the hatches closed, there was almost no attenuation as the prism worked as a waveguide.

2.5 Quality factor (Q) of a shielded enclosure

As absorbing material such as wood, plastic materials, equipment, cables and humans in a shielding enclosure also is affecting the shielding effectiveness, it would be interesting to measure this for a typical installation. One way of quantifying the absorption is to measure the Q-factor of a cavity. This was done in one of the special purpose vehicles according to the equation below [4].

$$Q = \frac{16 \cdot \pi^2 \cdot V}{\lambda^3} \cdot \frac{P_{rx}}{P_{tx}} \qquad [4]$$

Both the transmitting and the receiving antenna were placed inside the shielded enclosure. The Q-factor was calculated from the ratio of the transmitted and the received power after taking the reflection coefficients into account. An average of ten measurements, at different antenna positions, is presented in the graph below. This was done without having any personnel inside the enclosure.

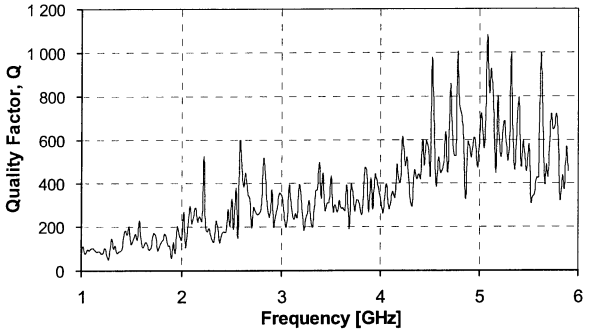


Figure 4: *Measured quality factor of one typical test sample.*

3 A low cost passive HPM indicator

3.1 Background

During trouble-shooting in the field, there is a desire to have a quick way to determine whether the equipment has been exposed to HPM. One way of doing this would be to have a low cost HPM indicator, possible to place on all material and easy to check.

Based on an idea from Karl-Gunnar Lövstrand at Swedish Defence Materiel Administration using an old type of metallized recording paper for electrostatic printers as a passive HPM indicator, AerotechTelub has designed and experimentally tested a prototype indicator.

3.2 The printer paper used

The printer paper used, is first coated with a black contrast layer. An aluminium metal coating is then applied by vacuum evaporation, which makes it conductive. During recording, low voltage pulses are applied to the surface, which evaporates the metal layer, and the black contrast coating becomes visible.

A printer paper of the following manufacturer and type was used for the experiments.

Manufacturer:	Sihl GmbH (Germany)
Type:	AD-paper
Conductivity:	R = 1.5 Ω/sq (measured)

3.3 Design

The basic idea was to create a broadband dipole or a "Maltese Cross", which centre part would be coloured black when exposed to HPM. The main issues were to obtain a good enough sensitivity and a broadband response.

An ESD test generator was first used to see if the material had any potential to be working for low energy pulses such as ESD or HPM. Tests on thin stripes of the paper showed that with a width of about 1 mm, small indications started at about 2-4 kV.

In order to obtain a broadband response and also a lower resistance in the antenna elements, dipoles in the shape of a butterfly was cut in the material. Several different shapes and dimensions were tested. The width of the "waist" determined the sensitivity of the indicator and also the visibility of the indication. The sensitivity was improved with a thinner "waist", but the visibility was instead reduced.

3.4 Results

The indicator was tested with high power pulsed microwaves at the Microwave Test Facility (MTF) in Linköping in Sweden. The indicators were tested at three frequencies 1.3, 2.8 and 5.7 GHz with 1 us pulse length and a pulse repetition frequency (PRF) of 100 Hz. Each exposure lasted for 10 seconds.

Small indications in the form of faint black spots started at field strengths of about 5 kV/m (peak amplitude) for those indicators with a waist of less than 1 mm. At 10 kV/m small black spots were obtained for all the three tested frequency bands.

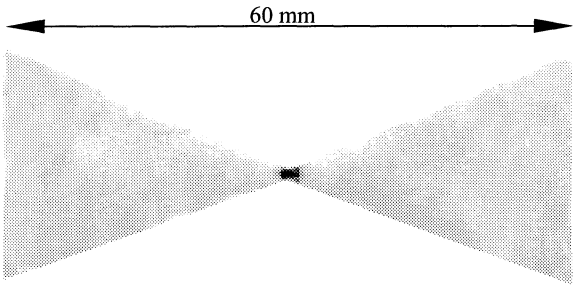


Figure 5: *One indicator tested at 1.3 GHz with 25 kV/m and 100 Hz PRF.*

The metallized paper can be covered with an adhesive plastic film in order to achieve mechanical and weather protection without affecting the performance of the indicator.

By placing one indicator on each side of a paper cube in orthogonal orientation, a multi-angle and -polarisation indicator is created.

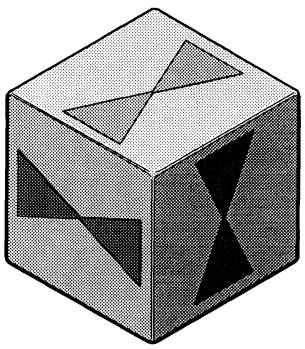


Figure 6: *Six indicators placed on a paper cube in order to obtain "isotropic" response.*

3.5 Conclusions

- It is possible to build HPM-indicator by using metallized recording paper.
- The sensitivity and frequency response of the indicator can be adjusted by changing the width of the waist and the length of the indicator.
- It is possible to build an "isotropic" indicator by placing one indicator on each side of a paper cube.
- It should be possible to mass-produce it at low cost, due to simple technique and low material costs.

4 Shielding effectiveness of ESD shielding bags

4.1 Background

Electronic spare parts can be exposed to HPM during storage and transportation. Most of the electronics, such as printed circuit board assemblies (PCB), are today already stored in some kind of ESD protection packaging. Most of the packaging is of the type ESD shielding bags. These bags offer shielding against ESD, which has a frequency content above 1 GHz. How good is the shielding effectiveness above 1 GHz?

4.2 Test procedure

The same test procedure as in section 2.2 above, was used for the measurement of the shielding effectiveness of ESD bags. The fibre optic link with the broadband monopole antenna was placed inside the ESD bag and the opening of the bag was folded back and taped up in the same way as they are normally used.



Figure 7: *The 3M 210CR ESD bag with the fibre optic link inside it.*

Later, more accurate measurements will be performed in a reverberation chamber.

4.3 Test samples

Initially five different ESD bags, all of the type ESD shielding bags, have been measured.

- 3M 2100R
- 3M 3300
- 3M 210CR
- Eurostat 5547
- MAGNAB Product No 20-003-6002 (Eurostat) "Cycle Shield Bubble"

One of the tested specimen, 3M 3300, has 40 dB EMI shielding in the frequency range 1-10 GHz according to its specification. However, this is according to MIL-PRF-81705D [5], which measure the shielding effectiveness of one piece of the material installed in a hole in a shielded chamber. By doing this, leakage in the seams of a bag is not taken into account and a lower value of a complete bag would be expected.

4.4 Results

The measurements shows that all of the bags except 3M 3300 have a shielding effectiveness of around 20 dB in the measured frequency range 1-6 GHz. The bag 3M 3300 had a somewhat better attenuation, around 25-35 dB. The measured shielding effectiveness may be somewhat overestimated due to the possible change in the reflection coefficient of the monopole antenna, when the antenna and the fibre optic link are placed inside the shielding bags. However, this will probably quite well represent the actual case when the bags are loaded with electronic equipment such as a PCB.

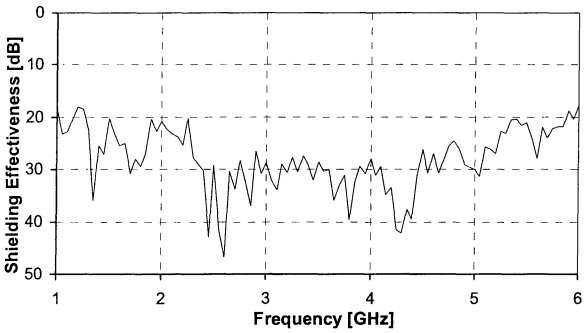


Figure 8: *Measured shielding effectiveness of the 3M 2100R ESD bag.*

4.5 Conclusions

- ESD shielding bags could be used to give some HPM protection.
- The normal way of closing the bags with about 6 cm folded back is adequate.
- Two or more bags can be used outside each other in order to obtain a higher attenuation.

References

[1] FMV, *EMMA, Elektromagnetisk Miljö Användarhanbok,* FMV 27/1 1994.

[2] IEEE Std 299-1997, *IEEE Standard Method for Measuring the Effectiveness of Electromagnetic Shielding Enclosures,* IEEE 9/12 1997.

[3] M. Bäckström, J. Lorén, *Microwave Coupling into a Generic Object. Properties of Angular Receiving Pattern and its Significance for Testing in Anechoic and Reverberation Chambers,* Ch. 3.1.2 and Ch. 4.2.3, FOI Scientific Report, FOI-R—0392—SE, February 2002. Swedish Defence Research Agency FOI, Sensor Technology, P.O. Box 1165, SE-581 11 Linköpng, Sweden.

[4] D. A. Hill, M. T. Ma, A. R. Ondrejka, B. F. Riddle, M. L. Crawford, and R. T. Johnk, *Aperture Excitation of Electrically Large, Lossy Cavities,* IEEE Transactions on Electromagnetic Compatibility, EMC-36, 3, 1994, pp. 169-178.

[5] Department of Defence, MIL-PRF-81705D, *Performance Specification Barrier Materials, Flexible, Electrostatic Protective, Heat-Sealable,* 3/9 1998.

RFI in Operation and Control of Rail Vehicles

Jan-Olof Häggblad[1]

[1]Bombardier Transportation, Dept. ICT/ESV, 721 73 Västerås, Sweden.
E-mail: jan-olof.haggblad@se.transport.bombardier.com

Abstract

RFI (Radio Frequency Interference) has become a more significant factor in vehicle system operation as the use of radio operated telephones transmitting in the 160 to 900 MHz bands becomes commonplace.

There are many examples of accidents, some fatal, that are directly linked to RFI. A year or two ago, a timber processing machine ran amuck while the operator was outside, using a portable radio.

1 Examples out of real life

To illustrate the problems associated with RFI I will try to start the discussion with a few select examples from real life.

1.1 Zapped Wheelchairs

Government safeguards are spotty but concern is growing at the Food and Drug Administration. Last month the agency ordered makers of motorized wheelchairs to shield them from EMI and educate users about the hazard. The FDA said it acted after getting "many reports of erratic unintentional powered-wheelchair movement". In one such incident, according to a recent article by an FDA researcher, radio waves zapped a power wheelchair sending its passenger over a cliff i Colorado causing a broken hip and several other injuries. The victim wasn't identified [1]



Figure 1: Illustration from Dagens Nyheterl [2] showing the effects of RFI on a wheelchair.

1.2 Mobitex data transmitting

Another good example is a tram project, where the vehicle moved up to 40% faster, when the data radio was transmitting.

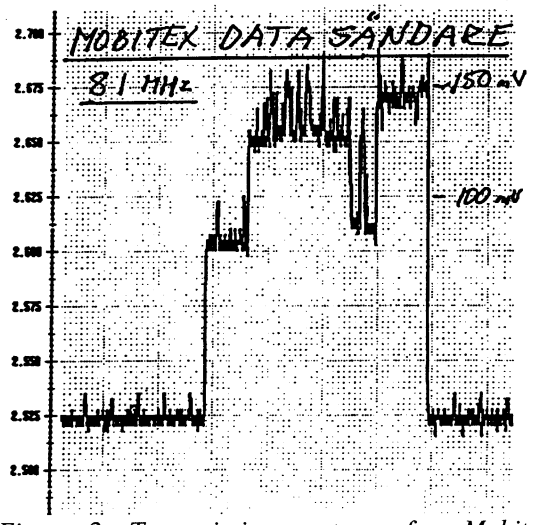


Figure 2: Transmission capture of a Mobitex data transmitter, transmitting at 81 MHz.

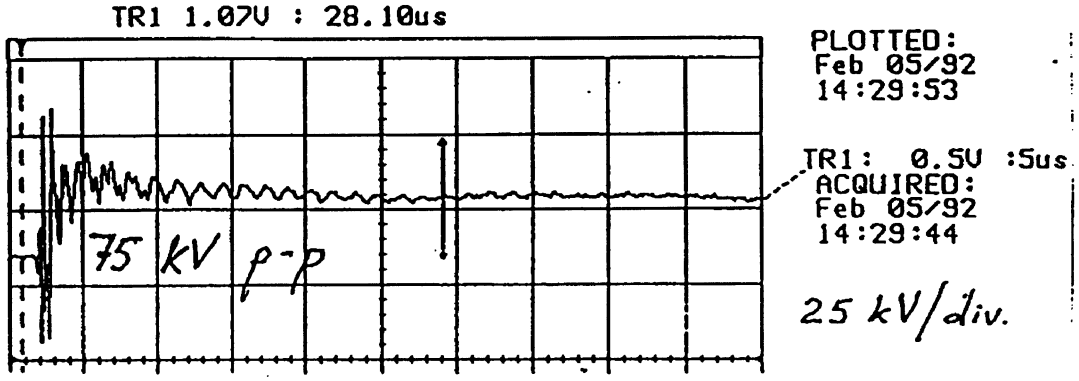


Figure 3: Transient overvoltage caused by the closing of the MCB on a vehicle.

1.3 Closing the MCB (Main Circuit Breaker)

Basically everything that causes switching power supplies is adding to the problem of interference within the vehicle. There is nothing we can do here to temporarily recduce the noise generated by the MCB. This produces a 75-100 kV spark on the roof, giving a field strength of approx. 20 V/m. The oscilloscope plot is shown in figure 3.

1.3.1 Radio development

In 1901, Marconi bridged the Atlantic when the letter s in Morse code traveled from Poldhu, Cornwall, to St. John´s Newfoundland, a distance of nearly 2,000 miles (3,200 kilometres) [3].

For this distance, Marconi replaced the secondary-spark detector with a device known as a coherer, which had been invented by a French electrical engineer, Edouard Branly, in 1890.

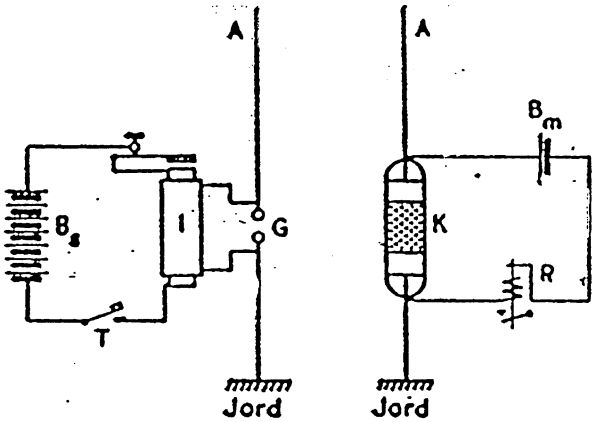


Figure 4: The schematic drawing of the simple transmitter used by Marconi to transmit the first signals across the Atlantic. This transmitter is less power full than our MCB sparks!

1.3.2 X2000 ATC problems

The examples above are well illustrated by the ATC problems which occurred on the X2000 trains at the beginning of revenue service, see figure 5. At neutral sections, when the MCB was operated the transients generated were large enough to interrupt the data transfer between the balis and the ATC system.

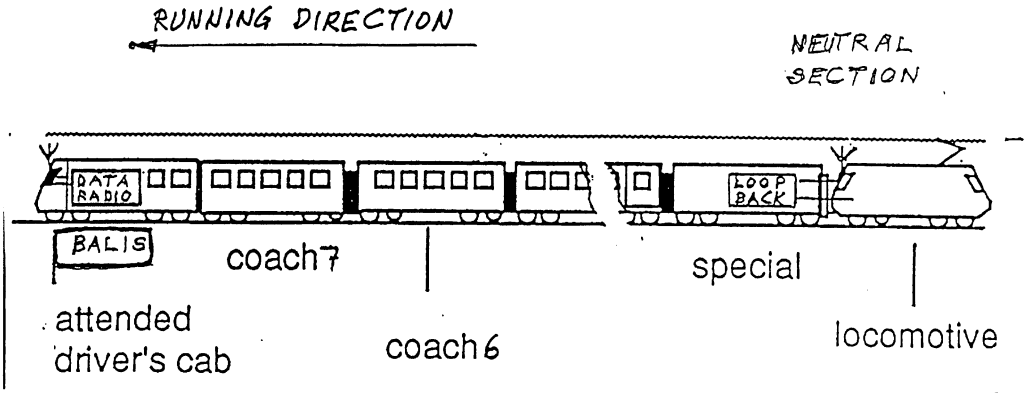


Figure 5: The transients generated by the MCB at neutral sections interrupted the transmission of ATC data.

2 Our present problems

So the question that comes to mind is:
Why do we have so many problems now? Radio transmission itself is not new and FM radio has been used from 1955.

Well part of the answer is that Pulse modulation presents more severe RFI. At the same time higher frequencies are now more common, for instance:

1. Radio communication frequencies in the range of 200 MHz and above are quite common today.
2. Radar Frequencies in the GHz range, 1900 MHz.
3. Amplifiers used over all frequencies.

This is very well illustrated by the problems seen in a push button on FPT, where external RFI could interfere with its normal operation.

2.1 AM detectors

The a.m. detector of a transistor receiver is usually a semiconductor diode feeding a potentiometer load. The potentiometer forms the volume control, which feeds the demodulated signal to the audio-frequency stages of the receiver. The potentiometer is usually shunted by a 10nF capacitor, as shown in figure 6 to filter out the i.f. [4]

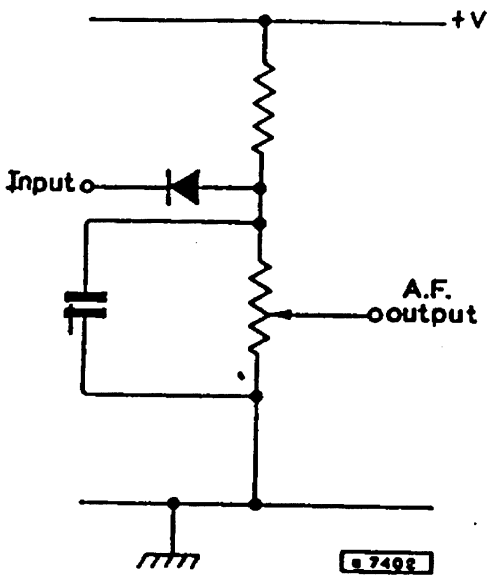


Figure 6: The basic AM detector circuit.

2.1.1 Radio modulation

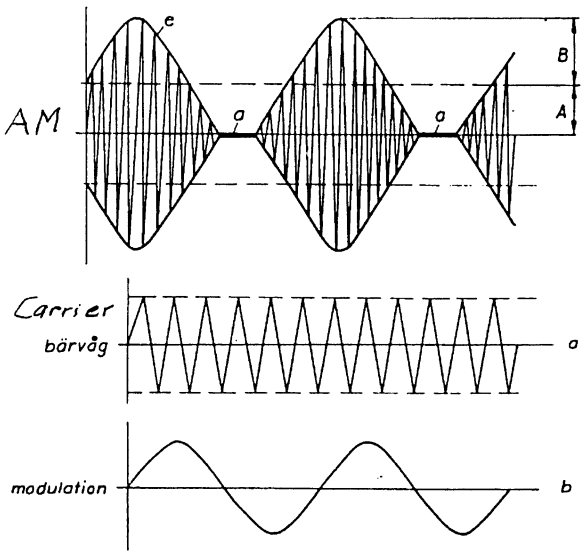


Figure 7: The top graph shows the AM modulated signal, consisting of the carrier wave (a) and the modulation (b).

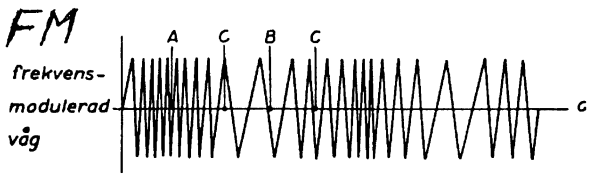


Figure 8: The frequency modulated signal shows an even amplitude and a variation in frequency.

DC= $A_0 = A/2$

$A_1 = 2A/\pi$ ..... $\cos \omega_0 t$ ≈0.64

$A_3 = 2A/3\pi$ ..... $-\cos 3\omega t$

$A_5 = 2A/5\pi$ ..... $+\cos 5\omega t$

$A_7 = 2A/7\pi$ ..... $-\cos 7\omega t$

etc.

$A_2 A_4$ ..... $= 0$

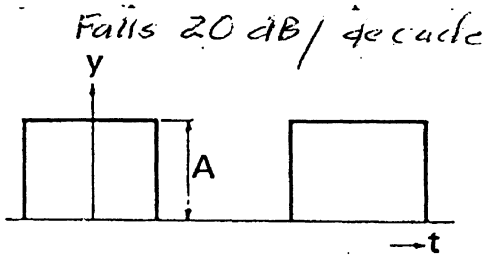


Figure 10: The overtone content of a square wave defined by the constants in the Fourier series.

2.2 Differential-mode radiation, after Ott

Differential-mode radiation can be modeled as occurring from a small loop antenna [5], see figure 9. For a small loop of area A carrying current I, the magnitude of the electric field E measured in free space at a distance r, in the far field, is equal to (Weeks 1964, p. 565, Eq. 91):

$$E = 131{,}6 \cdot 10^{-16} \left(f^2 AI \right) \left(\frac{1}{r} \right) \sin \Theta \qquad \text{Eq. 1}$$

where E is in volts/meter, f is in hertz, A is in square meters, I is in amperes, and r is in meters. Radiation increases with the square of frequency!

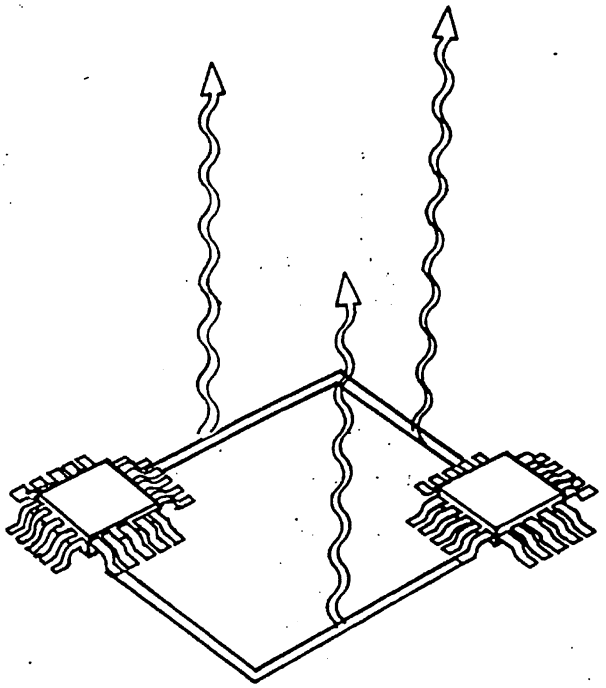


Figure 9: A small on circuit board loop antenna.

2.3 Fourier series

Since digital circuits use square waves, the Fourier series of the currents must be known before calculating the emission. For a symmetrical square wave of a 50% duty cycle (d=0.5), the first harmonic (fundamental) has an amplitude $I_1 = 0.64I$ and only odd harmonics are present. This is for the case where the rise time (t_r) is much less than the period (T). The harmonics fall off with frequency at a 20 dB per decade rate up to a frequency of $1/\pi t_r$; beyond that they fall off at a rate of 40 dB per decade.

If the frequency is doubled, the radiation increases by 6 dB, and if the rise time is halved, the radiation again increases by 6 dB, all other parameters remaining the same. Therefore doubling the frequency and cutting the rise time in half will increase the radiation by 12dB

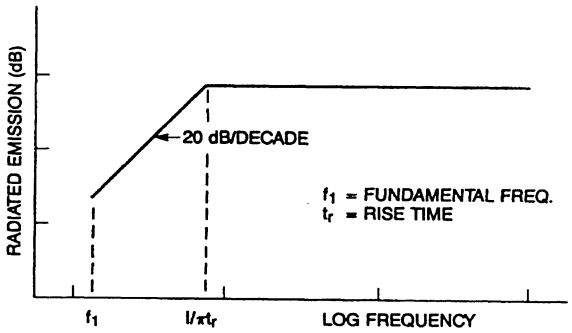


Figure 11: Differential-mode radiated emission envelope versus frequency.

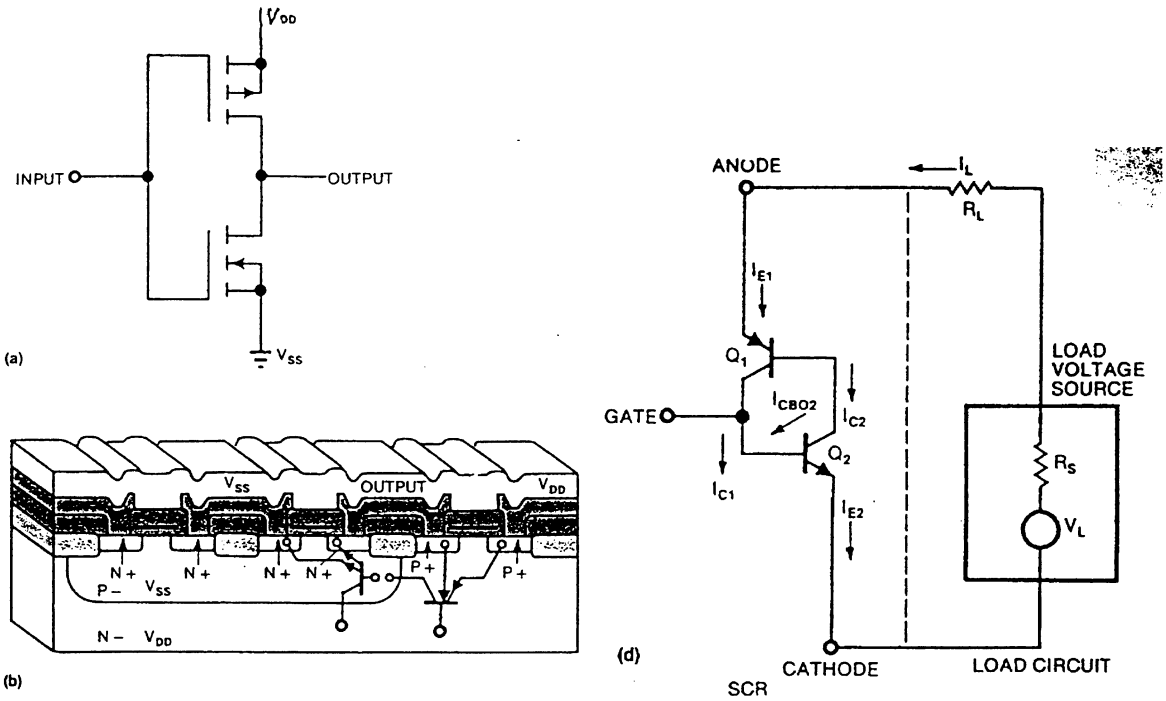


Figure 12: The CMOS latch up is caused by transistor effects in the silicon die.

3 Micro electronics

Where are the SCRs in CMOS devices, and how do they latch up? The answers to these questions could help you prevent latch-up failures in your system. First, consider a CMOS IC´s output driver, shown in figure 12, part (a). It comprises one n-channel MOS-FET with its source tied to VSS and one p-channel MOSFET with its source tied to VDD .

The drains of the two transistors connect to form the output; the gates are tied together to form the input. The fabrication of these transistors in close proximity results in the formation of a parasitic SCR connected directly across the supply rails. When triggered, the SCR presents a low impedance to the supply, allowing excessive current to flow. This situation is damaging to bond wires or to the supply conductors on the die.

How is the SCR formed? Refer to figure 12, parts (b), (c), and (d). The fabrication of the n-channel device results in a vertical npn transistor [6]. The N-substrate, based at VDD, serves as the collector. The P-well forms the base. The source and drain N+ diffusions create the emitters. One emitter is tied to VSS, the other to the output.

3.1 Microcomputer Hardware

Properly designed hardware should have the following circuits:

1. All inputs and outputs must be filtered
2. Power-fail and power-up logic
3. Watchdog logic, checking the program flow of the microcomputer
4. A LED indicator showing program execution
5. Error checking of data and memory (parity check, error correction)
6. False address trapping

2 References

[1] Zapped Wheelchairs, *Wall Street Journal*, Wednesday 15th of June 1994

[2] Mobiltelefon får ej störa rullstol, *Dagens Nyheter*, 15 of August 1995

[3] Marconi: Radio development, *Encyclopedia Britannica*

[4] Mullard, *Transistor audio and radio circuits*, Mullard Ltd 1969

[5] H Ott, *Noise reduction techniques in electronic systems 2 ed*, Wiley 1988.

[6] CMOS Latch up, *Electronic Design News (EDN)*, January 1995

Adaptive Antennas in Wideband Radio Access Networks

Björn Ottersten[1]

[1] Signal Processing/S3, Royal Institute of Technology, Stockholm, Sweden E-mail: bjorn.ottersten@s3.kth.se.

Abstract

Wideband radio access is required to develop future networks providing mobile multimedia services. The design of these systems must be tailored to packet based simplex communication with a wide variety of quality-of-service requirements. Adaptive antennas enable more reliable communication, increased range, and more efficient use of the limited spectral resources when designing the system. This, in turn, is essential to provide adequate quality, coverage, and system capacity at a low cost. This PCC project is divided into several work packages addressing channel and propagation modeling, transmit schemes including space-time coding and transmit diversity, receive strategies, antenna integrated implementation, and ASIC implementation. As part of the project, a test bed has been developed to perform channel sounding and provide experimental data for modeling purposes. Theses work packages are presented in individual papers/posters.

1 Background

Wideband radio access networks intended to support a variety of packet based services for multiple users are currently being envisioned. Typical data rates of 100 Mb/s or more are desirable and services to mobile or portable terminals must be provided. Expected use is indoors as well as outdoors. It is anticipated that there is a large commercial potential if one system can be developed providing both in- and outdoor coverage for a campus-like environment.

There are several scientific challenges associated with realizing a low cost wireless system as described above, some of which this research project addresses. If long range communication is supported, low cost coverage can be provided. Also, seamless coverage of a campus-like area is difficult to obtain at the considered frequency range using traditional techniques. Providing high system capacity is critical and a much more challenging task than supplying a high peak link capacity. Developing radio front ends that can meet requirements on bandwidth, linearity, power, efficiency, and cost are key issues. Furthermore, the air interface of future systems will differ from traditional mobile communication systems in that they are wideband, packet based (carrying data), primarily simplex communication, have to support multi-rate and different quality of service requirements.

Two critical factors in the design of multiple access wireless communication systems are coverage and system capacity. These factors have a direct impact on the cost and quality of the services since the spectral resources are limited and spectral efficiency is necessary to provide high quality services to a large number of users. The spatial dimension is to a large extent unexplored in wireless systems. Traditional telecommunication schemes multiplex channels in frequency and/or time. However, the spatial dimension is in general used in a rudimentary fashion by, for example, using some frequency channels in certain geographical areas (frequency planning) to limit interference. By incorporating adaptive antennas and efficient spatial-temporal processing techniques into future systems, both the capacity and the range may be increased improving service quality and decreasing cost.

This project is carried out by research groups at KTH, CTH, UU, LTH, and Telia Research. The results from individual work packages are presented in individual papers/posters.

A 20 MHz HiperLAN/2 MIMO Channel Model in NLOS Indoor Scenarios

Kai Yu[1], Mats Bengtsson[1], Björn Ottersten[1], Darren McNamara[2], Peter Karlsson[3] and Mark Beach[2]

[1]Department of Signals, Sensors and Systems, Royal Institute of Technology, Stockholm, Sweden
E-mail: kaiyu@s3.kth.se, mats.bengtsson@s3.kth.se, bjorn.ottersten@s3.kth.se
[2]Centre for Communication Research, University of Bristol, United Kingdom
E-mail: Darren.McNamara@bristol.ac.uk, M.A.Beach@bristol.ac.uk
[3]Telia Research AB, Malmö, Sweden
E-mail: Peter.C.Karlsson@telia.se

Abstract

In this paper, we present results of 5.2 GHz wideband indoor multiple input multiple output (MIMO) channel measurements under the EU IST SATURN project . First we show the average power delay profile for the single input single output (SISO) channel. Furthermore, our investigation shows that for non-line-of-sight (NLOS) indoor scenarios, the MIMO channel covariance matrix of each normalized tap of channel impulse response can be be well approximated by the Kronecker product of the covariance matrices seen from the transmitter and receiver respectively. Based on the above results, a wideband statistical model for HiperLAN/2 MIMO channel is presented. Monte-Carlo simulations show reasonably good agreement between the measured data and our model. Finally, we use this model to show the capacity characteristics of the HiperLAN/2 channel in NLOS indoor scenarios.

1 Introduction

Recently, it was reported [1, 2] that channel capacity can be greatly increased by using antenna array at both the transmit and receive side as long as the environment can provide sufficient scattering. A model that can describe the MIMO channel reasonably well is a necessary requirement on designing real MIMO communication systems and therefore attracts great interest. In [3], a wideband model based on channel power correlation coefficients is presented. In [4, 5], a narrowband model has been presented for NLOS indoor scenarios based on the Kronecker structure of the channel covariance matrix. Other models can be found in [6, 7, 8]. However, most models available now are narrowband.

This paper presents the results based on 5.2 GHz measurements conducted by the University of Bristol as part of the SATURN project. We study the statistical characteristics of MIMO propagation channels and propose a wideband model for NLOS Indoor MIMO channels based on these results. This paper is organized as follows. Section II makes a brief description about the measurements in Bristol. The methods used to analyze the measured data are presented in section III. In section IV, we propose a wideband statistical model based on the measurement results and use this model to simulate the HiperLAN/2 channels. Finally, we conclude in section V.

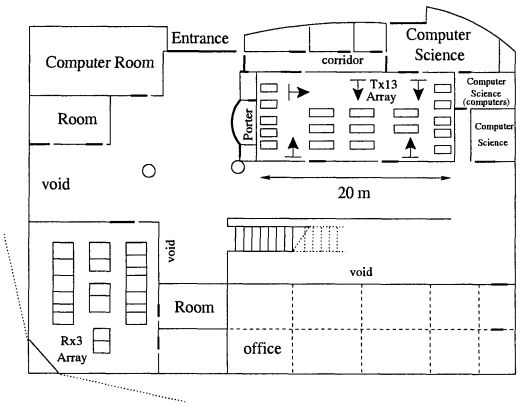


Figure 1: Measurement scenario for NLOS indoor MIMO channel

2 Measurement Setup

The test site was the Merchant Venturers Building of the University of Bristol. The general layout of the test site includes office rooms, computer labs, corridors and open spaces. The entire measurements include 15 transmitter locations and 3 receiver locations. Both line-of-sight (LOS) and NLOS cases were measured. Note that in this paper, all the results are from NLOS cases as shown in Fig. 1. Here, the arrow at each transmitter location indicates the orientation of the transmit array. The transmitter was lo-

cated in a computer lab and the receiver was located in a large modern office with cubicles.

The measurements were carried out using the Medav RUSK BRI vector sounder, which has an 8-element omni-directional uniform linear array (ULA) at the transmitter side and an 8-element ULA with 120° beamwidth at the receiver side (for pictures, see [9]). The distance between two neighboring antenna elements is 0.5λ for both arrays. There is a feedback from the receiver to the transmitter by a cable maintaining the coherence between the transmitter and receiver.

The measurements were centered at 5.2 GHz. A periodic multifrequency signal with 120 MHz bandwidth was sent out by the transmitter and captured by the receiver. The channel response was then estimated and saved in the frequency domain. The maximum expected channel excess delay was $0.8\mu s$, corresponding to 97 frequency subchannels. For each transmit element, one 'vector snapshot' (one measurement from each receive element) is taken by the receiver through switching control circuits. The sampling time for each full MIMO snapshot (8 vector snapshots) is $102.4\mu s$, which is well within the coherence time of this indoor environment. One complete measurement includes 199 blocks with 16 MIMO snapshots within each block, therefore there are 3184 complete MIMO snapshots in total for each frequency subchannel. The time delay between two neighboring blocks is $26.624ms$. This means the total time for one complete measurement is $5.3s$. Finally, it is worth to mention that during the measurements, people were moving around both at the transmitter and receiver side.

3 Analysis Method

3.1 Data Model and Channel Capacity

Assume there are m transmit elements and n receive elements. For a wideband MIMO channel, the input-output relationship can be expressed in the baseband as

$$\mathbf{y}(t) = \mathbf{H}_n^m(t) * \mathbf{s}(t) + \mathbf{n}(t), \quad (1)$$

where $\mathbf{s}(t)$ is the transmitted signal, $\mathbf{y}(t)$ is the received signal, $\mathbf{n}(t)$ is additive white Gaussian noise and '*' denotes convolution. The channel matrix $\mathbf{H}_n^m(t)$ here is an n by m channel impulse response.

It is well known [10] that when the transmitted power is equally allocated to each transmit element and frequency subband, the wideband channel capacity can be expressed as

$$C = \int_W \log_2 \det(\mathbf{I}_n + \frac{\rho}{m}\mathbf{H}(f)\mathbf{H}^H(f))df \quad \text{bits/s}, \quad (2)$$

where W is the overall bandwidth of the MIMO channel, $\mathbf{H}(f)$ is the normalized frequency response of each narrowband subchannel, ρ is the average signal-to-noise-ratio (SNR) at each receiver branch over the entire bandwidth and $(\cdot)^H$ is complex conjugate transpose. Here, we use the same normalization factor for the frequency response of every narrowband subchannel such that

$$\int_W E(\| \mathbf{H}(f) \|_F^2)df = Wnm, \quad (3)$$

where $\| \cdot \|_F$ denotes the Frobenius norm and $E(\cdot)$ denotes expected value.

3.2 HiperLAN/2 Channel Impulse Response

As stated in section II, the measurements are carried out in frequency domain and therefore, we use IFFT to get the MIMO channel impulse response. The Hanning windowing is also used to lower the sidelobe effects.

The HiperLAN/2 standard is originally specified by the ETSI BRAN Project. It can provide high data rate to broadband core networks and mobile terminals. Each HiperLAN/2 channel has 20 MHz bandwidth and therefore 120 MHz measurement bandwidth can provide 6 HiperLAN/2 Channel, i.e. we group 16 successive frequency subchannels in one subset, which corresponding to 20 MHz bandwidth. The IFFT with Hanning windowing is then used to each frequency subset to get 20 MHz channel impulse response.

3.3 Kronecker Structure

It was reported in [4, 5] that for a narrowband NLOS Indoor MIMO channel, the channel covariance matrix can be well approximated by the Kronecker product of the covariance matrices seen from transmit and receive side respectively, i.e.

$$\mathbf{R}_H = \mathbf{R}_H^{Tx} \otimes \mathbf{R}_H^{Rx}, \quad (4)$$

where $\mathbf{R}_H$ is the channel covariance matrix, $\mathbf{R}_H^{Tx}$, $\mathbf{R}_H^{Rx}$ are the covariance matrices at the transmitter and receiver side and '$\otimes$' denotes the Kronecker product. Suppose the channel coefficients are complex Gaussian, it is easy to show from equation (4), as in [7], that

$$\mathbf{H} = (\mathbf{R}_H^{Rx})^{1/2}\mathbf{G}[(\mathbf{R}_H^{Tx})^{1/2}]^T, \quad (5)$$

where $\mathbf{G}$ is a stochastic n by m matrix with independent and identically distributed (IID) $\mathcal{CN}(0,1)$ elements and $(\cdot)^T$ is transpose. Here $(\cdot)^{1/2}$ denotes any matrix square root such that $\mathbf{R}^{1/2}(\mathbf{R}^{1/2})^H = \mathbf{R}$. We will revisit this structure in section 4.

For a wideband MIMO channel, it is also interesting to see whether this structure can be extended to each tap. First we define the transmit, receive and channel covariance matrices from the measured data for the lth tap as

$$\mathbf{R}_{Tx}^l = \mathrm{E}[(\mathbf{h}_i^H(l)\mathbf{h}_i(l))^T] \quad \text{for } i = 1, \ldots, n \quad (6)$$

$$\mathbf{R}_{Rx}^l = \mathrm{E}[\mathbf{h}^j(l)\mathbf{h}^{jH}(l)] \quad \text{for } j = 1, \ldots, m \quad (7)$$

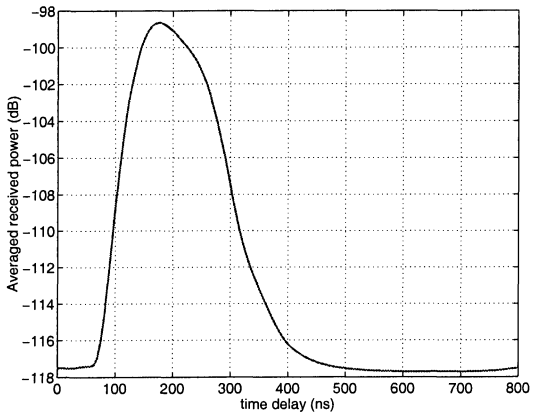


Figure 2: Average power delay profile

$$\mathbf{R}_H^l = \mathrm{E}[\mathrm{vec}(\mathbf{H}_n^m(l))\mathrm{vec}(\mathbf{H}_n^m(l))^H], \quad (8)$$

where $\mathbf{H}_n^m(l)$ is the lth tap of the channel matrix, $\mathbf{h}_i(l)$ is ith row of $\mathbf{H}_n^m(l)$, $\mathbf{h}^j(l)$ is jth columns of $\mathbf{H}_n^m(l)$ and $\mathrm{vec}(\cdot)$ denotes the 'vec' operator.

In [5], an optimal Kronecker factorization method is proposed to factorize the channel covariance matrix $\mathbf{R}_H^l$ into two Hermitian matrices $\mathbf{X}$ and $\mathbf{Y}$. The main idea is to reorder the position of matrix elements so that the original problem boils down to a problem that can be solved by the singular value decomposition (SVD) [11]. More details can be found in [5].

The results of optimal factorization for each tap are then compared with those from the measured data in the sense of model error, which will be defined in section 4.

4 Measurement Results

In this section, the data measured at a single pair of transmit location (Tx13) and receive location (Rx3) are used as an example. Similar results are found for the other four transmit locations. It is interesting to observe that the channel is quite stationary within the $5.3s$ measurement time even though people were moving around, therefore the six 20 MHz MIMO channels are treated as different MIMO channel realizations. Moreover, only pairs of 2 and 3 neighboring elements have been used with different subsets of elements being selected in order to get sufficiently many MIMO channel realizations.

4.1 Average Power Delay Profile

Average power delay profile is used to show how the received power changes according to a fixed time delay reference. It is found by averaging every instantaneous power delay profile over the whole measurement time. Fig. 2 shows the average power delay profile for one SISO channel over $800ns$ excess delay.

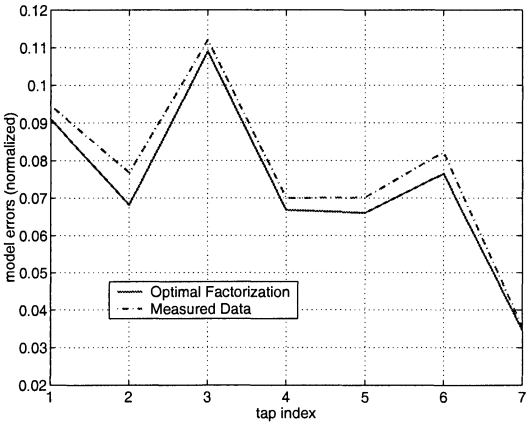


Figure 3: Model errors of Kronecker structure (normalized)

4.2 Kronecker Structure

The channel covariance matrix $\mathbf{R}_H^l$ for lth tap (including the phase) is then calculated and compared with the Kronecker product of the covariance matrices $\mathbf{R}_{Tx}^l$ and $\mathbf{R}_{Rx}^l$ seen from the transmit and the receive side respectively.

To measure the difference between two matrices $\mathbf{A}$ and $\mathbf{B}$, we define the following model error, Ψ as

$$\Psi(\mathbf{A},\mathbf{B}) = \frac{\| \mathbf{A}-\mathbf{B} \|_F}{\| \mathbf{A} \|_F}. \quad (9)$$

For 3x3 setup, the relative model errors for each tap between $\mathbf{R}_H$ (corresponding to $\mathbf{A}$ in (9)) and Kronecker product (corresponding to $\mathbf{B}$) are presented in Fig. 3. For comparison, the model error from the optimal factorization is plotted in the same figure as well. Note that only 7 successive taps that have higher SNR are plotted. It is shown that the mean model error is below 10% (for 2x2 case, the mean model error is below 5%) and the difference between the model error from the optimal factorization and that from the measured data is small. Therefore we conclude that for each tap, the channel covariance matrix can be approximated by the Kronecker structure reasonably well in NLOS indoor scenarios.

4.3 Wideband MIMO Model

Combining the average power delay profile for the SISO channel and the Kronecker structure for each tap, the wideband MIMO channel matrix can be modeled as

$$\mathbf{H}(l) = (\mathbf{R}_{Rx}^l)^{1/2}\mathbf{G}_l(\mathbf{R}_{Tx}^l)^{T/2}, \quad (10)$$

$\mathbf{H}(l)$ is the lth tap of MIMO channel impulse response $\mathbf{H}$, $\mathbf{G}_l$ is the matrix with IID zero mean complex Gaussian elements with power $\bar{p}_l$. The power of lth tap, $\bar{p}_l$ can be determined from the average power delay profile in Fig. 2.

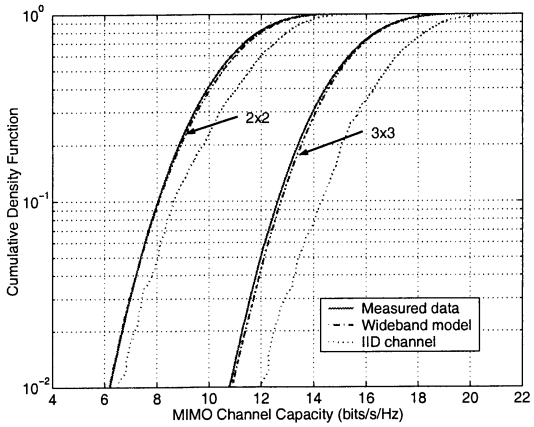


Figure 4: CDF of narrowband channel capacity (normalized) from measured data, wideband model and IID MIMO channel. Power is equally allocated, the SNR at the receiver side is 20dB

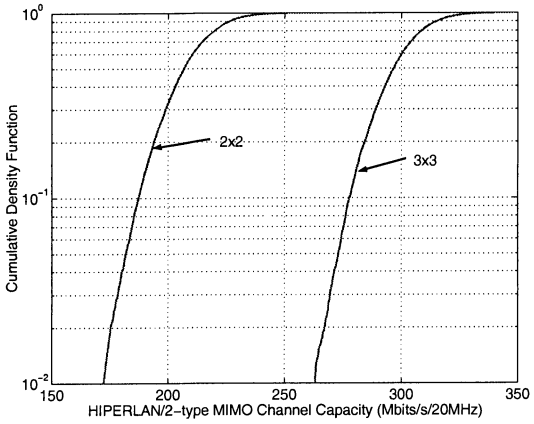


Figure 5: CDF of 2x2 and 3x3 HIPERLAN/2-type MIMO channel capacity. Power is equally allocated, the SNR at the receiver side is 20dB

5000 MIMO channel impulse responses are generated by Monte-Carlo simulations with 2x2 and 3x3 setups. FFT with Hanning windowing is used to transform the simulated impulse response back into the frequency domain and the capacity of the narrowband MIMO channel is calculated and compared with that from the measured data (averaged in frequency domain for both cases and also over spatial domain for the measured data), see Fig. 4. Both curves show reasonably good agreement between the measured data and our wideband model.

4.4 Simulations of HiperLAN/2 Channels

In this section, we will use the above wideband model to generate 1000 realizations of a 20 MHz HiperLAN/2 MIMO channel with 2x2 and 3x3 setup respectively.

Fig.5 shows the CDF of the capacity that one 20 MHz HiperLAN/2 channel can provide. Note that 20 MHz bandwidth is fully used, no bandwidth for guard and pilot tones is considered in this paper. It is shown that with 99% probability, one HiperLAN/2 channel can provide capacity up to 170 Mbits/s (2x2 setup) and 260 Mbits/s (3x3 setup). The simulated capacity curve for the 2x2 setup is similar to that reported in [12].

5 Conclusions

In this paper, we inverstigate the statistical characteristics of NLOS Indoor MIMO channels based on 5.2 GHz measurements. We show the average power delay profile of the SISO channel. Furthermore, it is found that the channel covariance matrix can be well approximated by the Kronecker product of the covariance matrices at both sides respectively. We propose a wideband MIMO channel model based on these results and Monte-Carlo simulations show reasonably good agreement between our model and the measured data. We also use this model to simulate HiperLAN/2 MIMO channels.

Acknowledgment

This work is conducted in part within SATURN (Smart Antenna Technology in Universal bRoadband wireless Networks) funded by the EU IST program.

References

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment," *Wireless Personal Communications*, vol. 6, pp. 311–335, 1998.

[2] I. Emre Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.

[3] K. I. Pedersen, J. B. Andersen, J. P. Kermoal, and P. Mogensen, "A stochastic Multiple-Input-Multiple-Output radio channel model for evaluation of space-time coding algorithms," in *Proceedings IEEE Vehicular Technology Conference*. Fall, 2000, pp. 893–897.

[4] Kai Yu, Mats Bengtsson, Björn Ottersten, Peter Karlsson, Darren McNamara, and Mark Beach, "Measurement analysis of NLOS indoor MIMO channels," in *Proceedings IST Mobile Communications Summit, Barcelona, Spain*, September 2001, pp. 277–282.

[5] Kai Yu, Mats Bengtsson, Björn Ottersten, Darren McNamara, Peter Karlsson, and Mark Beach, "Second order statistics of NLOS indoor MIMO channels based on 5.2 GHz measurements," in *Proceedings*

IEEE Globecom 2001, San Antonio, Texas, USA, November 2001, vol. 1, pp. 156–160.

[6] D. Gesbert, H. Bölcskei, D. Gore, and A. Paulraj, "MIMO wireless channels: capacity and performance," in *Proceedings Global Telecommunications Conference*, November 2000, vol. 2, pp. 1083–1088.

[7] Da-Shan Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, March 2000.

[8] Akbar Sayeed, "Modeling and capacity of realistic spatial MIMO channels," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, vol. 4, pp. 2489 –2492.

[9] D. P. McNamara, M. A. Beach, P. N. Fletcher, and P. Karlsson, "Initial investigation of multiple-input multiple-output channels in indoor enviroments," in *Proceedings IEEE Benelux Chapter Symposium on Communications and Vehicular Technology, Leuven, Belgium*, October 2000.

[10] D. P. Palomar, J. R. Fonollosa, and M. A. Lagunas, "Capacity results of spatially correlated frequency-selective MIMO channels in UMTS," in *Proceedings IEEE Vehicular Technology Conference*. IEEE VTC Fall, 2001, pp. 553 – 557.

[11] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, London, 3rd edition, 1996.

[12] Kai Yu, Mats Bengtsson, Björn Ottersten, Darren McNamara, Peter Karlsson, and Mark Beach, "A wideband statistical model for NLOS indoor MIMO channels," in *Proceedings IEEE Vehicular Technology Conference*. IEEE VTC Spring, 2002.

Active-Reflect Antenna for Power Combining Unequally Spaced Arrays

Dhanesh G.Kurup[1], Anders Rydberg[2] and Mohamed Himdi[3]

[1]Signals and Systems Group, Uppsala University, Box-528, Uppsala,SE-75120, Sweden: dg@signal.uu.se
[2]Signals and Systems Group, Uppsala University, Box-528, Uppsala,SE-75120, Sweden: ar@signal.uu.se
[3]Labarotorie Antennes Radar Télécom, Université de Rennes I, France: Mohamed.Himdi@univ-rennes1.fr

Abstract

In this paper, we present the design of an unequally spaced reflect array using an active reflect antenna. The antenna element in the design uses a novel electromagnetically coupled microstrip antenna element with automatic DC blocking between the orthogonally polarised ports. The behaviour model of the antenna element is first derived using an artificial neural network which is then optimised with the linear model of the transistor to derive the final layout.

1 Introduction

Solid state transistors have the inherent drawback their output power reduces drastically at higher frequencies [1][2]. As new and higher frequencies such as millimeterwave are of interest now a days, mainly due to the scarcity in spectral resources, we need efficient solutions to circumvent this problem. Although device technologies such as high-electron mobility transistors (HEMT) offer the possibility of high gain and superior noise performance compared to MESFETs, in some applications much high power output is required than that can be expected from a single device. The possibilities to increase the power out put which has been explored over these years can be broadly classified into chip level power combining, circuit level power combining and spatial power combining. Chip level power combining which connect devices in series and parallel is limited by the allowed power dissipation in the chip which is often very small. Network or circuit level power combining results in large size of the circuits and the transmission line loss. In this paper we explore the method of spatial combining of power from different devices using an unequally spaced active antenna array [2]. In future, we plan to use a similar active reflect array in a MIMO channel to investigate any gain in capacity compared to the MIMO system with no reflectors in the radio channel [3]. By means of a synthesis based on an differential genetic algorithm we demonstrate that unequally spaced antenna arrays with unequal phases offer superior performance in power combining compared to the conventional equally spaced arrays. Further the spatial power combining topology based on reflectarray principles is explored. The active element for this case uses a novel space efficient single layer active reflect antenna with automatic DC isolation between the transmit and receive ports.

2 Unequally spaced power combining array

The array factor of an antenna array is a non-linear and non-convex function of the element positions. Hence, it is difficult to find a globally optimum solution in the design of an unequally spaced array. In our design, we used a stochastic global optimisation method namely the differential genetic algorithm or differential evolution algorithm [5]. For achieving 100% array efficiency, the unequally spaced array elements are assumed to have uniform amplitudes but the phases can be uniform or non-uniform. For reduced mutual coupling, the lower limit in the design is fixed at $.5\lambda$ where as the upper limit in the distance between the elements is varied from .5 to 1λ. Fig.1 shows the sidelobe level of the array-factor of an unequally spaced 32 element linear array designed using this method. It can be seen from Fig.1 that, by using unequal phases, the sidelobe level in the array is considerably lowered compared to the case with only equal spacing. For instance, when the upper limit in the spacing between the elements is $.7\lambda$, the phase-position synthesis yielded a maximum sidelobe level which is lower by about 4dB compared to unequal spaced array. The unequally spaced active reflect array consists of (i) reflector containing the amplifier integrated active antenna elements and (ii) the feed, which can be a standard high gain horn antenna. For the polarisation of the reflected waves to be perpendicular to the feed radiation, the active antenna element need to be dual polarised. In [6] slot coupled patches are used as the

antenna element in the reflector. The problems with their design is that the structure is multi-layered and large. In the next section we introduces our space efficient design of the active reflect antenna element for the reflector.

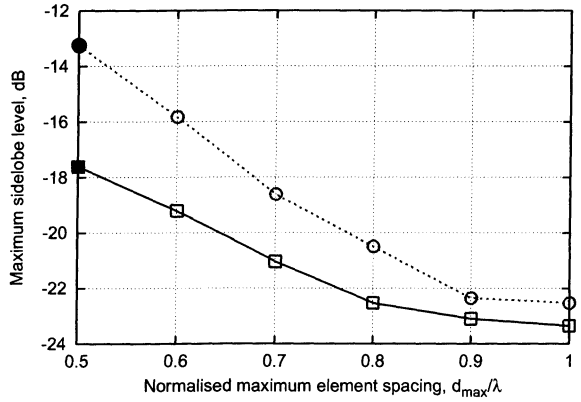


Figure 1: Maximum sidelobe level of an equal amplitude, unequally spaced 32 element array with upper limit in spacing in the DEA synthesis (lower limit is .5λ), ○: position only synthesis.□: position-phase synthesis. ●: uniform phases.■: phase only synthesis.

2.1 Active-reflect antenna

The characteristics of the microstrip patch for the passive part of the active antenna is such that the receive and transmit ports have polarisations orthogonal to each other. The transmit port has a direct contact microstripline feed and the receive port with a novel electromagnetically coupled microstrip-T junction feed, see Fig.2. The usage of the mirostrip T-junction for coupling help us in removing a capacitor (for preventing DC to DC coupling between receive and transmit ports of the active device to be used) and improving the bandwidth. The antenna at the same time is less complex compared to the widely used aperture coupled patch for the same application [6]. The radiation reactance of the antenna at the electromagetically coupled port can be controlled by changing the dimensions 's' and 'L' for a given microstrip T junction, see Fig.2. It is very difficult to have radiation reactance absolutely zero at port 1 (which is often a desirable characteristic as an antenna). However in this case the reactive component is used for the conjugate match of the transistor.The antenna was first designed using MoM software (Agilent-MOMENTUM, [8]) such that radiation reactance at Port 2 is zero at the design frequency of 10GHz. Fourteen different simulations for different length 'L' were carried out. The results were used to train an ANN (Artificial Neural Network, [7]). ANN model was later verified for four other test data, see Fig.3.

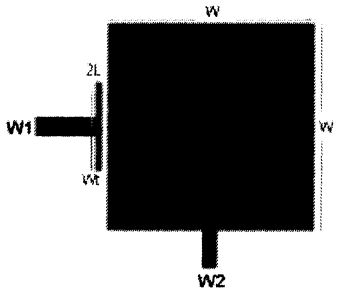


Figure 2: Radiating element for the X-band active reflect antenna.

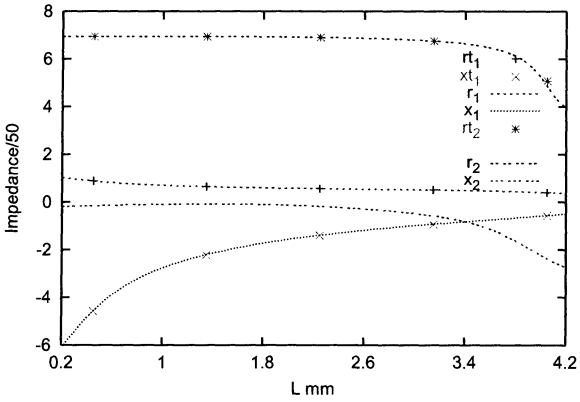


Figure 3: ANN generated output for normalised port 1 impedance, r1+jx1, port 2 impedance, r2+jx2 and the respective test data rt1+jxt1 and rt2+jxt2 for different L, see Fig.2

2.1.1 Behaviour model of the antenna using an Artificial Neural Network

The ANN that we used for training the antenna is a Multilayer Perceptron with 1 input neuron (for L), 6 hidden and 4 output neurons (for R+jX at port1 and 2 respectively) [7]. Fig. 3 show the results of the neural network generated output for the training and test data. It can be seen from Fig.3 that the characteristics of the antenna for different length L of the microstrip-T is now completely described by the neural net. The parameters of the antenna from the MOMENTUM simulation [8] with Port 1 excitation and L=1.4mm are: Directivity= 6.89dB, Gain=6.28 dB, Maximum cross-polarisation level Ecross=-35dB.

With port 2 excitation: Directivity = 6.95dB, Gain= 6.58dB and Maximum Ecross=-41dB. It is noted that Ecross can be controlled by proper selection of L. Isolation between the ports is below -40dB for the stand alone case. The next step is the space-efficient series matching of the ATF-34143 PHEMT for bilateral conjugate match. It is to be noted that most of the previously published active antennas in a similar context either use parallel stub matching which is quite clumsy or no matching at all. Series matching of an impedance Z1 with Z2 consists of series of microstriplines of different widths between Z1 and Z2. It has been noted that a series of microstriplines of non-uniform widths emulates a tapered line and at the same time results in considerably smaller total length. The achieved Gain for the active stage is in the excess of 4dB compared to MAG of the transistor using the low space-efficient series matching, which is noteworthy.

2.1.2 Layout optimisation of the active antenna

The antenna modelled by the ANN is brought into a Genetic Algorithm. The aim was to synthesise a given input impedance by varying the antenna impedance (by changing L of the microstrip-T, see Fig.2) and width and lengths of the 5 section series microstrip lines. The total length of series lines is specified from the desired layout. It has been found that by allowing a width constraint of the series microstriplines the bandwidth of matching can be increased. The final

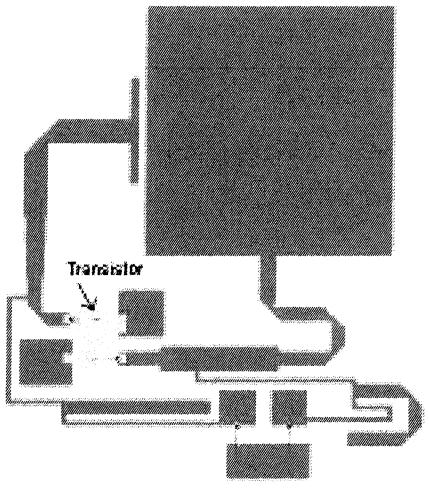


Figure 4: Layout of the X-band active reflect antenna

layout of the active reflect antenna element is shown in Fig.4. The square blocks at the transistor dual source terminals will have via going to ground. Lower arrows shows the positions of 2 capacitors KEMET-case 0603 (b1ypass) to ground. Head of these arrows will have bias and tails will have common ground. The layout size of unit cell shown above along transmit H-plane is ¡.5λ at X-band. The design of the final active reflect array based on the unit cell in Fig.4 is ongoing. The simulated gain of the single cell is around 8dB. Hence for the uniformly illuminated case for 32 element array, the expected total gain is the order (8+G-L)dB, where L is the over all losses in dB for the complete reflect array and G is the antenna array gain. For the 32 element array the directivity is in the order of 20dB. Hence the array gain is also 20dB due to equal amplitudes in the elements.

References

[1] T.Itoh "Recent Progress of Quasi-Optical Integrated Microwave and Millimeterwave Circuits and components", *Alta Frequenza*, Vol.LVIII, No.5-6, pp. 507-515, September-December 1989.

[2] Amir Mortazawi, Tasuo Itoh and James Harvey (ed), "Active antennas and Quasi-Optical Arrays", *IEEE Press*, 1998.

[3] Mattias Wennstrom, Mats Helin, Anders Rydberg and Tommy Oberg, "On the Optimality and Performance of Trnasmit and Recieve Space Diversity in MIMO chaneels", *IEE technical seminar on MIMO communication systems: From concept to implementation*, London, Dec. 12, 2001.

[4] Yahya Rahmat-Samii and Eric Michielssen, "Electromagnetic Optimisation by Genetic Algorithm", *Wiley*, 1999.

[5] Rainer Storn and Kenneth Price "Minimising the real functions of the ICEC'96 contest by Differential Evolution", *Conference Proceedings, IEEE conference on Evolutionary computations*, Nagoya, 1996. pp 842-844, 1996.

[6] A.W. Robinson, M.E. Bialkowsky and H.J. Song "A 137 element active-reflect array with dual feed microstrip patch antenna elements", *Microwave and Optical Techology Letters*, Vol. 26, no. 3, pp. 147-151, August 2000

[7] C.M. Bishop " Neural networks for pattern recognition", *Oxford university press*, 1998.

[8] MOMENTUM: "Method of Moments based planar electromagnetic simulator", *Agilent Technologies*, 1998.

Beamforming-based Estimators for Spatially Distributed Signals

Mikael Tapio*

Chalmers University of Technology, 412 96 Göteborg, Sweden. Email: mito@s2.chalmers.se

Abstract

In wireless communication systems using antenna arrays, it is desirable to estimate the nominal direction and spatial spread of spatially distributed signals, which originate from local scattering around the transmitter. Spatially distributed signals cause conventional high-resolution subspace methods to show degradation in performance, since the underlying data model theoretically is of full rank. Therefore, DOA estimation of spatially distributed signals has recently gained interest in the array processing community, and new high-resolution estimators for distributed signals have been proposed. In this paper two beamforming-based estimators are derived which respectively estimate the nominal DOA and spatial spread of a distributed signal. Finally, numerical examples comparing the performance of two high-resolution methods with the proposed methods, are presented. The non-parametric estimation of the nominal DOA show good performance, while the spatial spread estimate encounters problems with additive noise.

1 Introduction

Many classical Direction Of Arrival (DOA) estimation methods are based on point source models. However in some cases, for example, in indoor radio communications, the point source model assumption is violated, and a distributed source model would be a better approximation [7]. Another example is fast fading in mobile communications, where the elevated base-station antenna, due to local scattering around the mobile, experiences the received signal as distributed in space rather than being emitted from a point source [11]. In conventional high-resolution DOA estimation methods this leads to deterioration in performance, and therefore a number of high-resolution estimators for distributed sources have recently been proposed, e.g. [2, 3, 5, 8, 9]. Many of these new estimators make assumptions on the shape of the signal distribution, assume narrow spatial spreads, and eigen-decompose the full-rank covariance matrix into a pseudo-signal subspace and a pseudo-noise subspace. Most often they render a multi-dimensional optimization problem, implying high computational loads.

In this paper, the possibility of using low-complexity beamforming techniques for localization and spread estimation of distributed signals is investigated. In many applications, e.g. fast fading mobile communications, wireless indoor communications etc., we know little or nothing about the spatial distribution of the signals. Therefore, it would be attractive to make use of a robust and simple beamformer that does not make any assumptions on signal distribution nor does it use any hard-to-choose design parameters.

The paper presents the spatial signal model, the very basics of two parametric high-resolution subspace techniques, and non-parametric estimation using beamforming. Finally, numerical examples are presented comparing the performance of the two high-resolution methods and the proposed beamforming method.

2 Distributed signal model

We use the distributed signal model suggested in [1], in which the distributed signal impinging upon a Uniform Linear Array (ULA) is modeled as being emitted from a tight cluster of L spatially separated point sources (or scatterers), each with random complex gains. Thus, the received signal can be written as

$$\mathbf{x}(t) = s(t)\sum_{\ell=1}^{L}\gamma_\ell(t)\mathbf{a}(\theta_0+\tilde{\theta}_\ell(t)) + \mathbf{n}(t)\,, \quad (1)$$

where $\gamma_\ell(t)$, θ_0, and $\tilde{\theta}_\ell(t)$, are, respectively, the complex gain of ray ℓ, nominal (or mean) DOA, and random spatial (angular) deviation of ray ℓ. The K-sensor array steering vector is denoted by $\mathbf{a}(\theta) = [1, e^{-jkd\sin\theta}, \ldots, e^{-j(K-1)kd\sin\theta}]^T$, where $k = 2\pi/\lambda$ (wavelength – λ) denotes the circular wave number, and d is the element separation. Finally, $s(t)$ denotes the transmitted source signal, and $\mathbf{n}(t)$ the noise, which is modeled as a zero-mean complex circularly sym-

*This work was supported in part by the Personal Computing and Communication Program (PCC) under the project "Adaptive Antennas in Wideband Radio Access Networks". PCC is funded by the Swedish Foundation for Strategic Research.

metric, and spatio-temporally white Gaussian process with variance σ^2.

As in [1, 8], it is also assumed that the scattering environment changes rapidly compared to the mean DOA and spread parameters, i.e. the random complex gains $\gamma_\ell(t)$ are assumed to be temporally white. Further, they are also assumed to be independent from ray to ray, zero-mean, and circularly symmetric

$$E[\gamma_\ell(t)] = 0\,, \qquad E[\gamma_\ell(t)\gamma_k(s)] = 0\,,$$
$$E[\gamma_\ell(t)\gamma_k^*(s)] = \frac{1}{L}\sigma_\gamma^2\delta_{\ell k}\delta_{ts}\,, \quad \forall\,\ell,k,t,s\,, \qquad (2)$$

where $\delta_{\ell k}$ is the Kronecker delta function.

The random spatial deviation $\tilde{\theta}_\ell(t)$ is assumed to be a zero-mean random variable described by a Probability Density Function (PDF) $p(\tilde{\theta};\sigma_\theta)$. The PDF is symmetric in $\tilde{\theta}$ and parameterized by a spread (or standard deviation) parameter σ_θ.

By assuming large L, and using the central limit theorem, it is argued in [1, 8] that $\mathbf{x}(t)$ is a zero-mean complex Gaussian vector with covariance matrix

$$\begin{aligned}\mathbf{R}_x &= E\left[\mathbf{x}(t)\mathbf{x}^H(t)\right] \\ &= E_{\tilde{\theta}}\left[E\left[\mathbf{x}(t)\mathbf{x}^H(t)\,|\,\tilde{\theta}\right]\right] \\ &= P_s\sigma_\gamma^2\int_{\tilde{\theta}} p(\tilde{\theta};\sigma_\theta)\mathbf{a}(\theta_0+\tilde{\theta})\mathbf{a}^H(\theta_0+\tilde{\theta})d\tilde{\theta} + \sigma^2\mathbf{I} \\ &= S\mathbf{R}_v(\theta_0,\sigma_\theta) + \sigma^2\mathbf{I}\,, \qquad (3)\end{aligned}$$

where $S = P_s\sigma_\gamma^2 = E[|s(t)|^2]\sigma_\gamma^2$ is the source signal power including the path gain factor, and $\mathbf{R}_v(\theta_0,\sigma_\theta)$ is the channel covariance matrix (excluding the path gain) of the zero-mean Gaussian channel vectors. In the limit, the cluster of scatterers can be viewed upon as a continuum of scatterers, i.e. the spatio-temporal complex gain can be described by a temporal stochastic process $\gamma(\tilde{\theta};t)$, which has a continuous spatial distribution with covariance kernel $E[\gamma(\tilde{\theta}_1;t)\gamma^*(\tilde{\theta}_2;t)] = \sigma_\gamma^2 p(\tilde{\theta}_1;\sigma_\theta)\delta(\tilde{\theta}_1-\tilde{\theta}_2)$. In this context, $p(\tilde{\theta};\sigma_\theta)$ can be viewed upon as a continuous *Spatial Power Density Function (SPDF).*

The spatial deviation is commonly modeled by a Gaussian or Uniform distribution, i.e. $\tilde{\theta} \sim \mathcal{N}(0,\sigma_\theta)$ or $\tilde{\theta} \sim \mathcal{U}(-\sqrt{3}\sigma_\theta, \sqrt{3}\sigma_\theta)$. For small Gaussian spatial spreads, the $(m,n)^{\text{th}}$ entry of the covariance matrix can be written as [1, 2]

$$\begin{aligned}[\mathbf{R}_v(\omega_0,\sigma_\omega)]_{m,n} &= \exp(j(m-n)\omega_0) \\ &\quad\times\exp(-0.5((m-n)\sigma_\omega)^2), \qquad (4)\end{aligned}$$

where $\omega_0 = kd\sin\theta_0$ denotes the *spatial frequency*, and $\sigma_\omega = kd\sigma_\theta\cos\theta_0$ denotes corresponding standard deviation.

3 Parametric estimation

There are several high-resolution subspace methods that are based on an eigen-decomposition of the covariance matrix. The covariance matrix of a distributed source is theoretically of full rank (equal to the number of sensors), but it is noticed that the signal energy is mainly confined to a lower-dimensional subspace. Hence, the dimension of the signal subspace is approximated by a number, d, much less than the number of sensors. By doing this, a pseudo-signal subspace and a pseudo-noise subspace is obtained. Two methods built on this separation into pseudo-subspaces are briefly described in this section.

3.1 DISPARE

The DISPARE algorithm presented in [5], exploits the approximate orthogonality between the estimated pseudo-noise subspace, from the sample covariance matrix, and the theoretical pseudo-signal subspace. The criterion is formulated as

$$\{\hat{\theta}_0,\hat{\sigma}_\theta\} = \arg\min_{\theta_0,\sigma_\theta} \mathrm{Tr}\left(\widehat{\mathbf{E}}_{pn}^H\mathbf{R}_v^2(\theta_0,\sigma_\theta)\widehat{\mathbf{E}}_{pn}\right)\,, \qquad (5)$$

where $\widehat{\mathbf{E}}_{pn}$ denotes the estimated pseudo-noise subspace. A closely related algorithm is the DSPE algorithm presented in [9].

3.2 WPSF

It is shown in [1, 3] that DISPARE and DSPE fail to give consistent estimates. It is also presented a generalization of Weighted Subspace Fitting (WSF) to handle distributed sources, namely Weighted Pseudo-Subspace Fitting (WPSF). One version that attempts to make the parameterized pseudo-noise subspace as orthogonal as possible to the estimated pseudo-signal subspace, by using an estimate of the optimal weighting matrix, is given as

$$\begin{aligned}\{\hat{\theta}_0,\hat{\sigma}_\theta\} &= \arg\min_{\theta_0,\sigma_\theta} \mathrm{Tr}\Big(\widehat{\mathbf{E}}_{ps}\widehat{\mathbf{\Lambda}}_{ps}^{-1}\widehat{\mathbf{E}}_{ps}^H\mathbf{E}_{pn}\widehat{\mathbf{\Lambda}}_{pn}\mathbf{E}_{pn}^H \\ &\quad+\widehat{\mathbf{E}}_{ps}\widehat{\mathbf{\Lambda}}_{ps}\widehat{\mathbf{E}}_{ps}^H\mathbf{E}_{pn}\widehat{\mathbf{\Lambda}}_{pn}^{-1}\mathbf{E}_{pn}^H \\ &\quad-2\widehat{\mathbf{E}}_{ps}\widehat{\mathbf{E}}_{ps}^H\mathbf{E}_{pn}\mathbf{E}_{pn}^H\Big)\,, \qquad (6)\end{aligned}$$

where $\widehat{\mathbf{\Lambda}}_{ps}$ and $\widehat{\mathbf{\Lambda}}_{pn}$ denote the diagonal eigenvalue matrices for the estimated pseudo-signal and pseudo-noise subspaces, and $\widehat{\mathbf{E}}_{ps}$ and $\mathbf{E}_{pn}$ denote the estimated pseudo-signal and parameterized pseudo-noise subspaces.

4 Non-parametric estimation

The DOA of a single deterministic point source is usually taken as the peak of the Power Azimuth Spectrum (PAS)

$$\hat{\theta} = \arg\max_\theta \hat{P}(\theta)\,, \qquad (7)$$

where $\hat{P}(\theta)$ denotes the estimated PAS, from the sample covariance matrix. Using the true covariance matrix, the PAS is given by

$$P(\theta) = \begin{cases} \frac{\mathbf{a}^H(\theta)\mathbf{R}_x\mathbf{a}(\theta)}{\mathbf{a}^H(\theta)\mathbf{a}(\theta)} & \text{- Conventional} \\ \frac{1}{[\mathbf{a}^H(\theta)\mathbf{R}_x^{-m}\mathbf{a}(\theta)]^{1/m}} & \text{- Mod. Capon,} \end{cases} \quad (8)$$

where $m = 1$ gives the original Capon spectrum [4]. Compared to the Conventional Beamformer (CBF), Capon's beamformer has higher resolution. This is achieved at the expense of reduced white noise suppression, and Capon's beamformer is therefore more suited for high Signal-to-Noise Ratio (SNR). The performance of CBF for distributed sources was analyzed in [6], where the nominal DOA estimate was taken from Eqn. (7).

Assume that the steering vectors corresponding to distinct DOAs are orthogonal for large K, i.e.

$$\lim_{K\to\infty} \frac{1}{K}\mathbf{a}^H(\theta)\mathbf{a}^H(\eta) = \begin{cases} 1, & \theta = \eta \\ 0, & \theta \neq \eta. \end{cases} \quad (9)$$

Hence, for large K we have $\frac{1}{K}\mathbf{A}^H(\theta)\mathbf{A}(\theta) \approx \mathbf{I}$, where $\mathbf{A}(\theta)$ denotes the $K \times L$ steering matrix of the vector $\theta = [\theta_1, \ldots, \theta_L]^T$.

The discrete version of the channel covariance matrix can be written $\mathbf{R}_v = \sum_{\ell=1}^{L} p(\theta_\ell)\mathbf{a}(\theta_\ell)\mathbf{a}^H(\theta_\ell) = \mathbf{A}(\theta)\mathbf{P}(\theta)\mathbf{A}^H(\theta)$, where $\mathbf{P}(\theta)$ denotes a diagonal matrix with the power mass entries $[\mathbf{P}(\Theta)]_{\ell,\ell} = p(\theta_\ell) = p(\theta_\ell; \theta_0, \sigma_\theta)$, i.e. $\mathbf{P}(\theta)$ contains discrete samples of the previously introduced continuous SPDF. Since it is only non-zero values of $p(\theta_\ell)$ that contribute to the channel covariance matrix, we let $p(\theta_\ell) > 0$ to make $\mathbf{P}$ full rank. Note that in this context, θ_ℓ is equal to the nominal DOA θ_0 plus the previously used random deviation $\tilde{\theta}_\ell$ of ray ℓ.

Using the matrix inversion lemma and assuming large K, the inverse covariance matrix becomes $\mathbf{R}_x^{-1} = [S\mathbf{R}_v + \sigma^2\mathbf{I}]^{-1} = [S\mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma^2\mathbf{I}]^{-1} \approx \mathbf{A}\check{\mathbf{P}}\mathbf{A}^H + \sigma^{-2}\mathbf{I}$, where $\check{\mathbf{P}}$ is the diagonal matrix with entries $[\check{\mathbf{P}}]_{\ell,\ell} = -\frac{Sp(\theta_\ell)}{\sigma^4 + \sigma^2 SKp(\theta_\ell)}$. It is now straightforward to insert $\mathbf{R}_x$ and $\mathbf{R}_x^{-1}$ into the expressions for the PAS given by Eqn. (8). Doing so, we find the following relationship

$$P(\theta) \approx \begin{cases} \beta p(\theta) + \varepsilon, & \theta \in \{\theta_1, \ldots, \theta_L\} \\ \varepsilon, & \theta \notin \{\theta_1, \ldots, \theta_L\}, \end{cases} \quad (10)$$

where $\beta = SK$, $\varepsilon = \sigma^2$ for the CBF spectrum, and $\beta = S$, $\varepsilon = \frac{\sigma^2}{K}$ for Capon's spectrum.

As mentioned earlier, when $L \to \infty$, the distributed source can be described by a continuous SPDF. Consequently, if we compute the PAS densely, the SPDF can be acquired from the PAS. Here, ε is taken as the minimum value of the PAS, i.e. $\varepsilon = \min P(\theta)$, which considering Eqn. (10) is reasonable. Hence, we only need to normalize $\check{P}(\theta) = P(\theta) - \varepsilon$ to have unit area over the set $(-\pi/2, \pi/2)$ to acquire an estimate of the parameterized SPDF $p(\theta; \theta_0, \sigma_\theta)$. Using this, the nominal DOA estimate simply becomes the Center of Mass (CM) of $\tilde{P}(\theta)$

$$\hat{\theta}_0 = \frac{\int_\theta \theta\tilde{P}(\theta)\,d\theta}{\int_\theta \tilde{P}(\theta)\,d\theta}, \quad (11)$$

and the spatial spread parameter or standard deviation is acquired as

$$\hat{\sigma}_\theta = \left(\frac{\int_\theta(\theta - \hat{\theta}_0)^2\tilde{P}(\theta)\,d\theta}{\int_\theta \tilde{P}(\theta)\,d\theta}\right)^{1/2}, \quad (12)$$

where $\int_\theta(\cdot)\,d\theta$ in practice can be implemented by a discrete summation (inner product).

Since the true covariance matrix $\mathbf{R}_x$ is unknown, we use the sample covariance matrix which is defined as $\hat{\mathbf{R}}_x = \frac{1}{N}\sum_{t=1}^{N}\mathbf{x}(t)\mathbf{x}^H(t)$ to estimate the PAS. It is well known [10] that $\hat{\mathbf{R}}_x$ converges (with probability 1) to $\mathbf{R}_x$ as $N \to \infty$. Hence, the estimated PAS is a consistent estimate of the true PAS, and our estimate of $p(\theta; \theta_0, \sigma_\theta)$ becomes consistent, thus leading to consistent estimates of the nominal DOA and spread parameters. Note that we have used the assumption of orthogonal steering vectors. Non-orthogonal steering vectors will add bias to our estimates.

Fig. 1 illustrates normalized (to unit area) PAS spectra for three different beamformers; CBF, Capon, and Pisarenko's modified Capon spectrum. The figure suggests that Capon's spectrum has the best fit to the true Gaussian PDF, and Eqns. (11) and (12) give good estimates of the nominal DOA and the spatial spread parameters. It is our experience (which is also hinted at in Fig. 1) that Capon's PAS describes a smooth continuous SPDF well. This SPDF does not necessarily have to be symmetric.

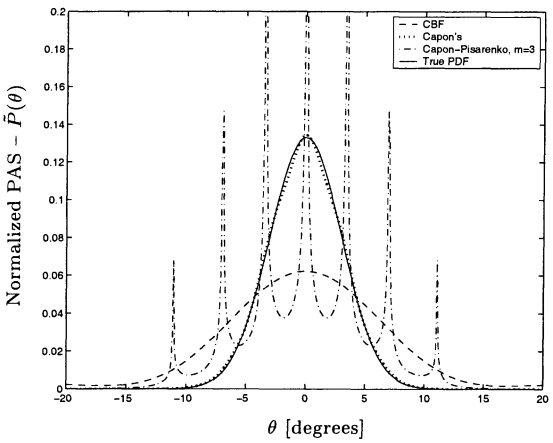


Figure 1: Normalized PAS. Gaussian spread, $\theta_0 = 0°$, $\sigma_\theta = 3°$, eight sensors, true covariance matrix and no noise. Estimated values using Capon's: $\hat{\theta}_0 = 0°$, $\hat{\sigma}_\theta = 3.07°$.

5 Numerical examples

In this section, simulations are presented to illustrate the performance of the different beamforming techniques for a Gaussian distributed source ($\theta_0 = 0°$, $\sigma_\theta = 5°$) versus the more complex DISPARE and WPSF algorithms. The Cramer-Rao bound (CRB) is also included in the figures. Unless otherwise stated, an eight-element ULA with half-wavelength element separation is used. The received signal is generated from the model in Eqn. (1), with $L = 50$ random DOAs and complex Gaussian gains $\gamma_\ell(t)$, and with a new independent realization for each snapshot. The source signal is an equiprobable BPSK sequence. The sample covariance matrix is computed over 100 independent channel realizations (snapshots). The DISPARE and WPSF estimates are acquired by letting the dimension, d, of the pseudo-signal subspace be such that, for an eight-element ULA, $d = 2$ if $\sigma_\theta < 7°$, and $d = 3$ if $\sigma_\theta \geq 7°$. In the examples where the number of sensors is varied from four to twenty, d is varied from two to six, which confines the main part (around 96%) of the signal power to its pseudo-subspace. To locate the solutions, a two-dimensional Nelder-Mead simplex search, initialized at the true parameter values, is used. We also let DISPARE and WPSF use the correct assumption that the deviation is Gaussian distributed.

Figs. 2 and 3 show the Root Mean Square Error (RMSE) versus SNR of the estimated parameters $\hat{\theta}_0$ and $\hat{\sigma}_\theta$, respectively. Figs. 4 and 5 show the RMSE of estimated DOA and spread versus true spread σ_θ for 30 dB SNR. Finally, Figs. 6 and 7 show the RMSE of estimated DOA and spread versus number of sensors. The RMSE is computed over 1000 independent trials for each parameter.

We see that it is better to chose the CM instead of the peak when estimating the nominal DOA. The obvious problem the beamforming techniques encounter is noise. Noise flattens the normalized PAS, hence the beamforming techniques completely fail to estimate the spread for low SNRs (even if we subtracted the minimum value of the PAS prior to normalization). The beamforming techniques also have problems estimating small spreads, which is obvious since the PAS would contain a somewhat narrow lobe even in the case of a single point source. For increasing spread, DISPARE shows deterioration which is probably caused by a bad choice of d, while WPSF and beamforming are less sensitive to an increasing spread. As one might expect, when taking the CM, the performance increases as the number of sensors increases. This is not the case when we take the peak value. Overall, one can conclude that Capon's method shows good performance for high SNRs and spreads that are not too small, and since it does not rely on a choice of d (nor a model), it is more robust.

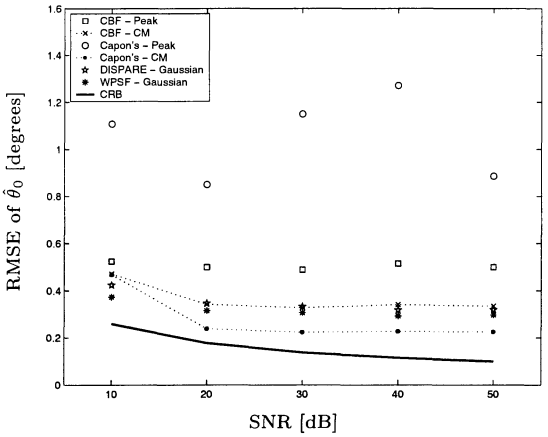


Figure 2: RMSE of $\hat{\theta}_0$ for different SNRs. Gaussian spread, $\theta_0 = 0°$, $\sigma_\theta = 5°$, eight sensors, 50 scatterers.

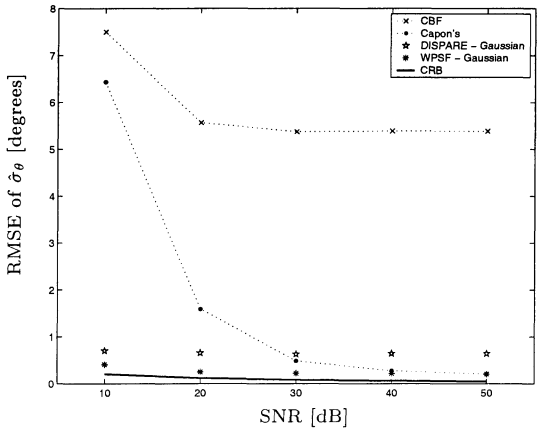


Figure 3: RMSE of $\hat{\sigma}_\theta$ for different SNRs. Gaussian spread, $\theta_0 = 0°$, $\sigma_\theta = 5°$, eight sensors, 50 scatterers.

6 Conclusions

We have investigated the use of simple beamforming techniques to estimate the nominal DOA and spatial spread of a distributed signal. Based on a simple model of the Power Azimuth Spectrum (PAS), it is suggested to estimate the DOA as the Center of Mass (CM) of the PAS, instead of the peak. Capon's spectrum with its high angular resolution capability, showed very good performance for high SNRs. The poor performance for low SNRs is due to noise power in the tails of the PAS. One possible solution could be, instead of subtracting the minimum value of the PAS, to subtract some average power in the tails and then simply "cut out" the main lobe to circumvent the problem with noise power in the tails of the PAS. There is always a trade-off between complexity and performance, and Capon's method has a reasonable balance between the two. Finally as future work, it would be interesting to analyze the performance of the nominal DOA estimate for both of the beamformers.

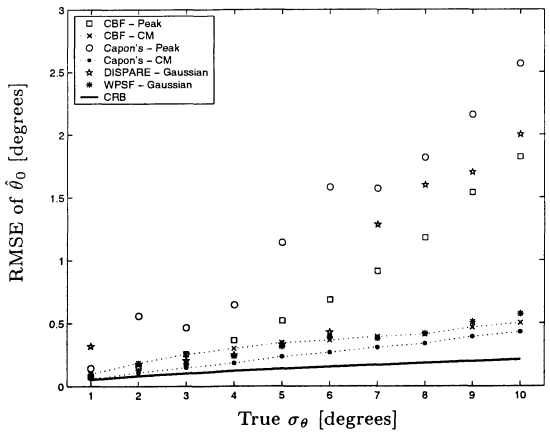


Figure 4: RMSE of $\hat{\theta}_0$ for different spreads σ_θ. Gaussian spread, $\theta_0 = 0°$, eight sensors, 50 scatterers, 30 dB SNR.

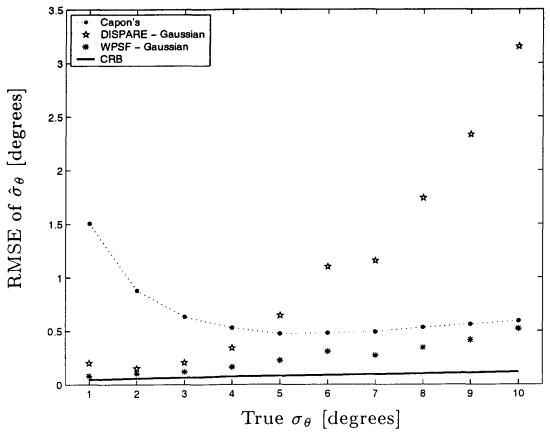


Figure 5: RMSE of $\hat{\sigma}_\theta$ for different spreads σ_θ. Gaussian spread, $\theta_0 = 0°$, eight sensors, 50 scatterers, 30 dB SNR.

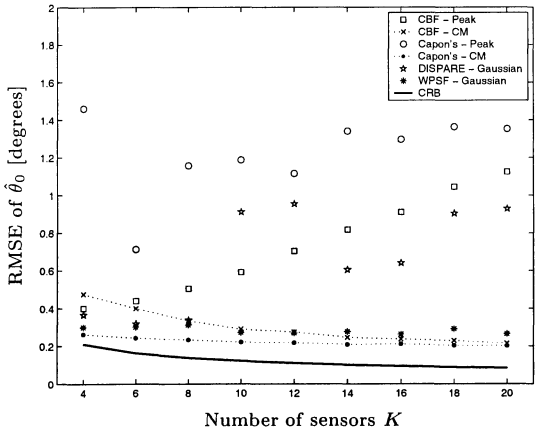


Figure 6: RMSE of $\hat{\theta}_0$ for different numbers of sensors K. Gaussian spread, $\theta_0 = 0°$, $\sigma_\theta = 5°$, 50 scatterers, 30 dB SNR.

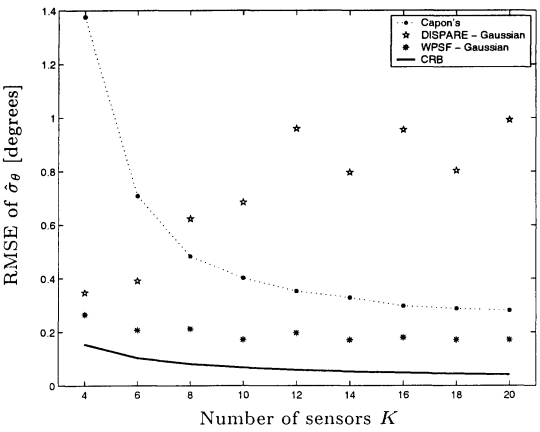


Figure 7: RMSE of $\hat{\sigma}_\theta$ for different numbers of sensors K. Gaussian spread, $\theta_0 = 0°$, $\sigma_\theta = 5°$, 50 scatterers, 30 dB SNR.

References

[1] M. Bengtsson. *"Antenna Array Signal Processing for High Rank Data Models"*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1999.

[2] M. Bengtsson and B. Ottersten. "Low-Complexity Estimators for Distributed Sources". *IEEE Trans. on Signal Processing*, 48(8):2185–2194, August 2000.

[3] M. Bengtsson and B. Ottersten. "A Generalization of Weighted Subspace Fitting to Full-Rank Models". *IEEE Trans. on Signal Processing*, 49(5):1002–1012, May 2001.

[4] J. Capon. "High Resolution Frequency Wave Number Spectrum Analysis". *In Proc. IEEE*, 57:1408–1418, 1969.

[5] Y. Meng, P. Stoica, and K.M. Wong. "Estimation of the Directions of Arrival of Spatially Dispersed Signals in Array Processing". *Proc. IEE Radar, Sonar and Nav.*, 143(1):1–9, Feb. 1996.

[6] J. Goldberg R. Raich and H. Messer. "Bearing Estimation for a Distributed Source via the Conventional Beamformer". *Proc. Stat. Signal Array Process. Workshop*, pages 5–8, Sep. 1998.

[7] D. Tholl and M. Fattouche. "Angle of Arrival Analysis of Indoor Radio Propagation Channel". *Proc. Int. Conf. on Univ. Pers. Comm.*, 1:79–83, Oct. 1993.

[8] T. Trump and B. Ottersten. "Estimation of Nominal Direction of Arrival and Angular Spread using an Array of Sensors". *Signal Processing*, 50(1-2):57–69, Apr. 1996.

[9] S. Valaee, B. Champagne, and P. Kabal. "Parametric Localization of Distributed Sources". *IEEE Trans. on Signal Processing*, 43(9):2144–2153, Sept. 1995.

[10] Q. Wu and D.R. Fuhrmann. "A Parametric Method for Determining the Number of Signals in Narrow-Band Direction Finding". *IEEE Trans. on Signal Processing*, 39(8):1848–1857, Aug. 1991.

[11] P. Zetterberg. *"Mobile Cellular Communications with Base Station Antenna Arrays: Spectrum Efficiency, Algorithms and Propagation Models"*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1997.

Optimised QR-decomposition Core for Adaptive Beamforming

Fredrik Edman and Viktor Öwall

PCC WP2, Dept. of Electroscience, Lund University. E-mail: {Fen,Vikt}@es.lth.se

Abstract

In many applications of modern array signal processing algorithms, the inversion of a data matrix is one of the most computationally demanding task. To be able to perform real-time matrix inversions of large dense data matrices, an optimised ASIC implementation must be considered. The core unit presented in this paper is a QR-processor based on the Squared Givens plane rotation algorithm. The architecture considered is an array structure with locally communicating processor elements. A fixed-point implementation of stable matrix inversion, optimised for numerical stability, power consumption, area, and speed, is the goal.

1 Introduction

Many operations in contemporary signal processing and communication applications require a high throughput rate and superior numerical accuracy. These operations must in many cases be performed in real-time, thus imposing a severe speed constraint. To be able to comply with these demands, special-purpose architectures must be used. Advances in VLSI technology make these architectures technically and economically feasible today. Therefore VLSI design plays an important role in modern real-time signal processing applications.

An integral part of many computationally demanding real-time applications are matrix operations and manipulations. An example of this is communication applications, where solution and analysis of linear systems, parameter optimisation, and matrix manipulation, such as triangularisation and inversion, demand a high throughput. A common way to simplify and reduce the computational complexity of these tasks is to use a factorisation technique such as QR-factorisation. A variety of algorithms on QR-decomposition can be employed [1], including Givens rotations, squared Givens rotations (SGR), Householder and modified Gram-Schmidt orthogonalisation technique. The different techniques are quite similar in performance but when choosing to implement one of these algorithms one must compare such issues as numerical stability and algorithm-to-hardware feasibility.

In several array signal-processing algorithms the inversion of a data matrix is the computational bottleneck. To be able to perform a stable inversion of a large data matrix in real-time an optimised ASIC implementation must be considered. The focus of this project is to optimise the QR-processor based on squared Givens plane rotations, which is the core unit of the matrix inversion problem.

2 The SGR Algorithm

The standard, or generalised, Givens rotations is a well known decomposition technique which was first proposed by Givens [2]. If the matrix $\mathbf{A}$ is an an $n \times p$ matrix the Givens rotations results in the decomposition

$$\mathbf{A} = \mathbf{QR}, \tag{1}$$

where $\mathbf{Q}$ is composed of orthogonalised columns of $\mathbf{A}$ and $\mathbf{R}$ is upper triangular. $\mathbf{R}$ is then a Cholesky factor of $\mathbf{A}^T\mathbf{A}$. Applying Gaussian elimination on $\mathbf{A}^T\mathbf{A}$ will result in an upper triangular $p \times p$ matrix denoted $\mathbf{U}$. Then $\mathbf{U}=D_R\mathbf{R}$ and D_R=Diag($\mathbf{R}$). Then $\mathbf{A}$ may be written as

$$\mathbf{A} = \mathbf{Q}\mathbf{D}_U^{-1}\mathbf{U}, \tag{2}$$

where $Q_A=\mathbf{Q}D_R$ and D_U=Diag($\mathbf{U}$)=D_R^2. The resulting rotation equations from which the updated values are calculated will include operations like square-root and division. These types of operations are not suitable for hardware implementation. Several square-root free methods has been proposed by Gentleman [3], Hammarling [4] and more recently Döhler [5]. The SGR algorithm used in this case was first proposed by Döhler and is a modification of Hammarlings scaled Givens rotation algorithm. This algorithm determine $\mathbf{U}$ directly from $\mathbf{A}$ without calculating $\mathbf{A}^T\mathbf{A}$. This approach only requires about half the number of multiplications needed by the standard Givens to derive $\mathbf{R}$, and is also square-root-free.

Before triangularisation augment the matrix $\mathbf{A}$ to

$$\begin{bmatrix} \mathbf{U} \\ \mathbf{A} \end{bmatrix},$$

where $\mathbf{U}$ is a *zero* $p \times p$ matrix. The scaled rows $\mathbf{u}=r_k\mathbf{r}$ are generated by annihilation of the rows in the sub matrix $\mathbf{A}$ by squared rotations with rows of $\mathbf{U}$. The annihilation process of the kth element of each row in $\mathbf{a}=w^{\frac{1}{2}}\mathbf{v}$ of the partially reduced matrix $\mathbf{A}$, then $\mathbf{a}$ will be paired with the kth row $\mathbf{u}$ in $\mathbf{U}$. The following algorithm describes the complete process of calculating updated rotation values in the SGR algorithm.

```
SGR Algorithm

IF u_k >0
   THEN
        ū=u+wv_k v
        v̄=v-v_k/u_k u
        w̄=wu_k v
   ELSE IF u_k=0
        ū=wv_k v
        w̄=0
   END ELSE
END IF

(where w is a given scalar >0 )
```

After complete annihilation of submatrix $\mathbf{A}$ the matrix $\mathbf{U}$ replaces the upper matrix in the SGR algorithm. The kth rows $\mathbf{u}$ and $\mathbf{r}$ of $\mathbf{U}$ and $\mathbf{R}$ are then connected by $\mathbf{u}=r_k\mathbf{r}$. Hence $\mathbf{U}=\mathrm{diag}(\mathbf{R})\mathbf{R}$ which is the same arithmetically as the the result from Gaussian elimination applied to $\mathbf{A}^T\mathbf{A}$.
Götze has formulated a divide- and root-free Givens rotation algorithm [6]. The divide operation is achieved at a cost of several multiplications, more complex internal cell and greater communication between the cells. For these resons it is most likely not suitible for implementation.

3 Architectural Design

Factorisation algorithms based on Givens rotations are typically implemented in a highly parallel triangular array structure as shown in figure 1.

The triangular array structure consists of a number of communicating processing cells or more often called processing elements (PE). New array data is feed into the top of the array structure every clock cycle. The data then propagates through out the PE structure and exits at the bottom. We have a 100% hardware utilisation when the structure is compleatly filled with data. This modular structure is very suitable for harware VLSI implementation. Since there are two types of PEs in the structure, boundary cells (circular PEs) and internal cells (square PEs). The

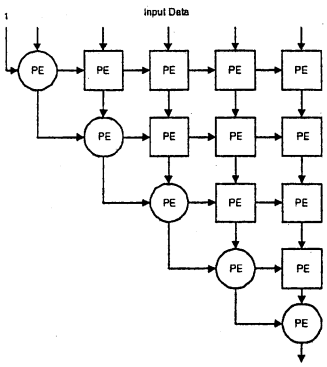


Figure 1: Array structure with communicating PEs.

functionality of the different cells corresponds to the calculation of the updated rotation values in the SGR algorithm. The boundary- and the internal cell formula and function are depicted in figure 2 and figure 3. The variable f is the squared forgett factor β^2, and is simply a constant, with a typical value close to 1 in array processing applications.

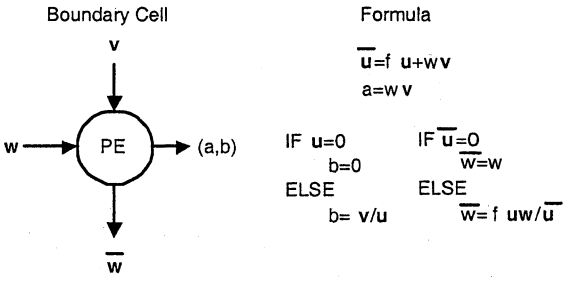


Figure 2: Functional description of boundary cell of the SGR algorithm.

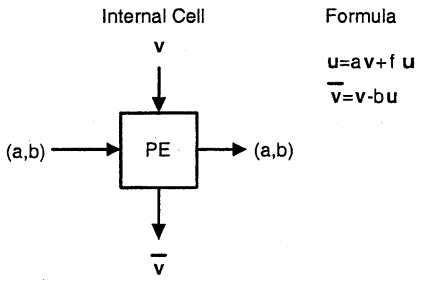


Figure 3: Functional description of internal cell of the SGR algorithm.

As discussed by Walke [7] there are a number of ways to implement the boundary and the internal cell. The decision of which implementation that should be used is based on trade-offs

between numerical precision and stability, latency, and hardware complexity of the design. Despite the reduction of complexity the SGR algorithm offer, the cells are still highly complex and the number of processor elements within the QR array will increase dramatically with the number of inputs thus consuming a large chip area.

4 Hardware Layout Considerations

In modern technologies the interconnections become more and more important for the performance of the circuit. When placing the array structure design on the silicon a preservation of the local connectivity between the cells are required to minimise area and reducing the complexity of the layout. The triangular shape of the array structure depicted in figure 1 is in this sens not optimal for hardware implementation. A more effective solution is a rectangular shaped array [7, 8]. It can be argued that the hardware design tools should take care of this, however presented results show otherwise. The rectangular shape can be obtained by folding the triangular array at cut section A as depicted in figure 4. The new one-time folded array structure depicted in figure 5 is indeed rectangular in shape but parts of the structure have to be globally interconnected which is not optimal. The global connections can be avoided by a second folding at cut section B, shown in figure 5. The resulting two-times folded array structure depicted in figure 6 is much more suitable for hardware implementation than the original triangular shaped array structure. The folded structure will aide the design tools to make a more efficient design on the chip.

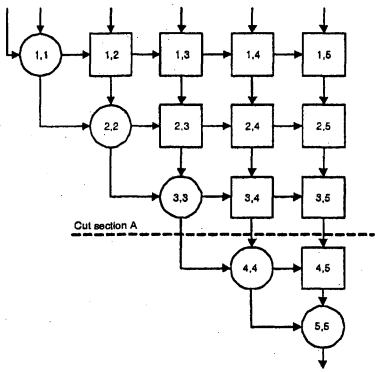


Figure 4: Original array structure.

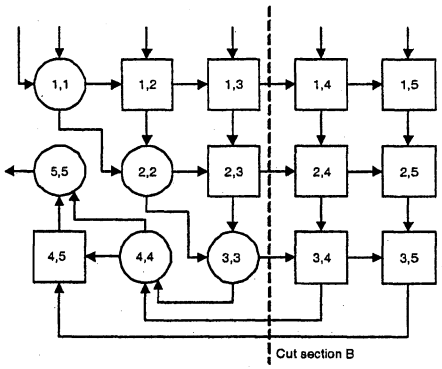


Figure 5: One-time folded array structure.

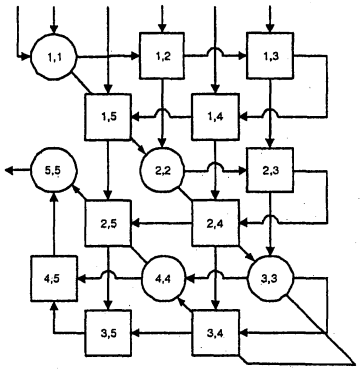


Figure 6: Two-times folded array structure.

5 Optimisation and Simulation

The optimisation of the architecture is divided into two steps.

- The first step is the high-level optimisation. In this step the global architectural structure and interconnection of the communicating PEs are optimised. Global clustering and retiming are investigated to obtain the best possible superstructure and data flow in the structure. Control sequencing of the PEs are also planned in this step.
- The second step is the low-level optimisation. In this step the individual PE is optimised and the realisation of the control circuits is also considered. The PE is optimised for numerical stability depending on the number of bits to use in the representation and calculation of the data.

Simulations are performed at both levels to determine the best course of action.

6 Ongoing and Future Work

Ongoing simulations will deside the best hardware implementation of the processor elements. When the matrix inversion chip has been manufactured it will be put through rigorours testing. The chip will then be integrated into the channel sounding system for multiple input - multiple output (MIMO) channel measurement which is being designed and implemented in PCC WP7 at the Department of Electroscience, Lund University. In parallel the algorithm will be implemented on a standard DSP for performance and cost comparison. Parameters such as power consumption, speed, cost effectiveness and flexibility will be compared. This is important study because it will tell how much one can gain with an ASIC implementation of this kind.

References

[1] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Maryland, ISBN: 0-8018-3011-7.

[2] W. Givens. "Computation of plane unitary rotations transforming a general matrix to trangular form". *J. Soc. Indust. Appl. Math.*, Vol6, No 1, pp.26-50, March 1958.

[3] W. M. Gentleman. "Least squares computations by Givens transformations without square roots". *J. Inst. Maths Applics* 12, 329-336, 1973.

[4] S. Hammarling. "A note on modifications to the Givens plane rotation". *J. Inst. Maths. Applics.* 13, 215-218, 1974.

[5] R. Döhler. "Squared Givens Rotation". *IMA Journal of Numerical Analysis*,11, p.1-5, 1991.

[6] J. Götze and G. J. Hekstra. "An algorithm and architechture based on orthonormal m-rotations for computing the symmetric EVD". *Integration, the VLSI J.*, 20, pp. 21-39, 1995.

[7] R. L. Walke. *High Sample-Rate Givens Rotations for Recursive Least Squares.* Ph.D. thesis Warwick University, 1997.

[8] C. M. Rader. "MUSE: A systolic array for adaptive nulling with 64 degrees of freedom using Givens transformation and wafer scale integrastion". *Proc. of the Inst. Conf. of Application Specific Array Processors*, pp. 277-291, 1992.

Utilizing Channel Information in the Design of Space-Time Block Codes

George Jöngren, Mikael Skoglund and Björn Ottersten *

Abstract

So far, the assumption of no channel knowledge at the transmitter has been inherent in the design of space-time codes for wireless communication. This work, on the other hand, assumes that potentially non-perfect channel information is available at the transmitter and investigates how it can be incorporated into the design of space-time block codes. Efficient codes are found by means of a gradient search over a continuous alphabet. The resulting channel estimate dependent space-time codes can be used, for example, in conjunction with quantized channel information obtained from a feedback link. Simulation results for a Rayleigh fading scenario with such a system using two transmit antennas and one receive antenna show that significant gains are possible.

1 Introduction

The use of space-time codes in wireless communication systems equipped with multiple transmit antennas and one or more receive antennas has recently received considerable attention [1–3].

In general, conventional space-time codes have been designed for situations where the transmitter has no channel knowledge. The present work, on the other hand, considers the design of space-time block codes that may utilize channel information at the transmitter, if available, for increasing the performance. The design is targeted towards a narrowband scenario. Since obtaining accurate channel information is in practice not always possible, a model describing the non-perfect nature of the channel information at the transmitter is exploited. In this way, the resulting codes continue to work well even if the quality of the channel information is poor. The channel knowledge is incorporated into the code construction process by utilizing a design criterion based on an upper bound of the codeword error probability conditioned on the channel information. Efficient codes are found by optimizing the design criterion, with respect to the codewords, using a simple gradient search algorithm. The search is performed over a continuous alphabet. To explicitly illustrate the benefits of using the designed codes, a system with a feedback link, which provides the transmitter with channel information in the form of quantized channel estimates, is evaluated by means of simulations.

Previous and related work include papers about design of conventional space-time codes [1,2,4], design of modulation signal sets [5] and how to utilize non-perfect channel information [6–8], respectively.

The contribution of the present work is the introduction of a general strategy for incorporating non-perfect channel information (at the transmitter) into the design of space-time codes, and the development and study of a method for realizing this strategy in the case of block codes. As will be exemplified herein, an especially attractive field of application for the constructed block codes is in wireless systems where the channel information is highly quantized so that the codewords can be stored in lookup-tables. Even when no channel information is available, the constructed codes perform, for the scenario under consideration, better than the corresponding orthogonal space-time block code (OSTBC). Although the focus is on unstructured continuous alphabet block codes, it is straightforward to modify the construction process to also handle the design of the weighting matrices used in *linear* space-time block codes (also known as linear dispersive codes).

2 System Model

This work considers a multi-input-multi-output (MIMO) wireless communication system employing space-time block codes that utilize partial channel information to enhance the performance. An illustration of a generic model for such a system is found in Figure 1, where the MIMO channel is represented by the matrix $\mathbf{H}$ and the channel information, also referred to as the side information, is represented by the vector ζ. At the transmitter side, a space-time encoder maps the data to be transmitted into a sequence of codeword matrices, which is thereafter split into parallel and generally different symbol sequences, corresponding to the different transmit antennas. To increase the performance of the system, the channel information is used for adapting the codewords to the properties of the transmission medium while taking into account the non-

*This work was supported by the Swedish Foundation for Strategic Research through the project "Adaptive Antennas in Wideband Radio Access Networks" in the Personal Computing and Communication (PCC) program. The authors are with the Department of Sensors, Signals and Systems, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: gjongren@s3.kth.se; skoglund@s3.kth.se; otterste@s3.kth.se).

perfect nature of the former by utilizing a suitable channel information quality measure. The resulting information carrying signals are transmitted over a wireless fading channel. The individual channels between the transmit and receive antennas are modeled as frequency-non-selective. At the receiver, the transmitted data is recovered by means of maximum likelihood (ML) decoding, based on perfect channel state information.

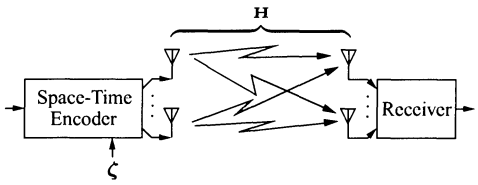


Figure 1: Generic system model.

The signal output from each receive antenna is a weighted superposition of the transmitted signals, corrupted by additive noise. Let M and N denote the number of transmit and receive antennas, respectively. By collecting the filtered and symbol sampled complex baseband equivalent outputs from the receiving antenna array in an $N \times 1$ vector $\mathbf{x}(n)$, the received signal vector can be written as

$$\mathbf{x}(n) = \mathbf{H}^* \mathbf{c}(n) + \mathbf{e}(n)\,,$$

where $(\cdot)^*$ denotes the complex conjugate transpose operator and where the potentially complex valued symbols, transmitted from the M antennas at time instant n, are represented by $\mathbf{c}(n) = \begin{bmatrix} c_1(n) & c_2(n) & \cdots & c_M(n) \end{bmatrix}^{\mathrm{T}}$.The noise term, $\mathbf{e}(n)$, is assumed to be a zero-mean, temporally and spatially white, complex Gaussian random process with covariance matrix $\sigma^2 \mathbf{I}_N$. Furthermore, the MIMO channel is described by the $M \times N$ matrix $\mathbf{H}$, with complex-valued elements h_{ij}. Consequently, the channel between the i^{th} transmit antenna and the j^{th} receive antenna is h_{ij}^*. Alternatively, the MIMO channel is represented by its vectorized counterpart $\mathbf{h} = \mathrm{vec}\,(\mathbf{H})$, obtained by stacking the columns of $\mathbf{H}$ into a vector.

A quasi-static scenario is considered where the channel is assumed to be constant during the transmission of a codeword (or a burst of codewords) but may vary from one burst to another. Assume the codewords are of length L and let K represent the number of codewords. Consequently, mapping the data to be transmitted into codewords results in a code rate of $\log_2(K)/L$ information bits per channel use. The choice of L and K allows an appropriate tradeoff between data rate and time redundancy. Now, without loss of generality, consider the time interval $n = 0, 1, \cdots, L-1$. By grouping L consecutive vectors $\mathbf{x}(n)$ into a matrix $\mathbf{X}$, the received signals corresponding to a transmitted codeword can be written as $\mathbf{X} = \mathbf{H}^* \mathbf{C} + \mathbf{E}$, where $\mathbf{E}$ contains the noise vectors and the transmitted codeword $\mathbf{C} = \begin{bmatrix} \mathbf{c}(0) & \mathbf{c}(1) & \cdots & \mathbf{c}(L-1) \end{bmatrix}$ belongs to the code $\mathcal{C} = \{\mathbf{C}_k\}_{k=1}^{K}$. It is assumed that the transmission is such that all the codewords are equally probable and that the average energy per codeword and information bit is equal to P, i.e., $\frac{\mathrm{E}[\|\mathbf{C}\|_{\mathrm{F}}^2]}{\log_2(K)} = \frac{1}{K \log_2(K)} \sum_{k=1}^{K} \|\mathbf{C}_k\|_{\mathrm{F}}^2 = P$.Here, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

The joint statistics of the channel and the side information is assumed to be such that the channel conditioned on the side information obeys a complex Gaussian distribution with mean vector $\mathbf{m}_{\mathbf{h}|\boldsymbol{\zeta}}$ and covariance matrix $\mathbf{R}_{\mathbf{hh}|\boldsymbol{\zeta}}$. As will be apparent in Section 3, these two quantities are needed in the design of the space-time code. Hence, inherent in the transmission technique, it is assumed that, in addition to the channel information, the transmitter has knowledge about certain scenario specific parameters that allows the conditional mean and covariance to be computed. This is further discussed in connection with the so called example scenario described in Section 5. The example scenario is essentially a special case of the generic model and is used for producing the simulation results of this work.

As previously mentioned, a measure of the quality of the channel information is needed. For this purpose, $\mathbf{R}_{\mathbf{hh}|\boldsymbol{\zeta}}$ is generally used herein, since it describes the remaining uncertainty when the channel information is known. Moreover, the two notions of "perfect channel information" and "no channel information" will be frequently used in the following and can now be given an exact meaning. To be specific, the two extremes of channel knowledge are defined as the following limiting cases:

- "Perfect channel information" $\Leftrightarrow \|\mathbf{R}_{\mathbf{hh}|\boldsymbol{\zeta}}\| \to 0$
- "No channel information" $\Leftrightarrow \|\mathbf{R}_{\mathbf{hh}|\boldsymbol{\zeta}}^{-1}\| \to 0$

Here, $\|\cdot\|$ denotes any norm of a matrix.

3 Design Criterion

The codeword error probability, conditioned on the channel information available at the transmitter, $\Pr[\hat{\mathbf{C}} \neq \mathbf{C}|\boldsymbol{\zeta}]$, serves as an appropriate performance measure for taking the channel information into account. It is clear that minimizing such a criterion for each possible realization of $\boldsymbol{\zeta}$ would also minimize the total codeword error probability, since $\Pr[\hat{\mathbf{C}} \neq \mathbf{C}] = \mathrm{E}\left[\Pr[\hat{\mathbf{C}} \neq \mathbf{C}|\boldsymbol{\zeta}]\right]$. However, in order to get a tractable closed-form expression, the union bound technique is exploited for upper bounding $\Pr[\hat{\mathbf{C}} \neq \mathbf{C}|\boldsymbol{\zeta}]$ so as to obtain

$$\Pr[\hat{\mathbf{C}} \neq \mathbf{C}|\boldsymbol{\zeta}] \leq \frac{1}{K} \sum_{k \neq l} P(\mathbf{C}_k \to \mathbf{C}_l|\boldsymbol{\zeta})\,, \qquad (1)$$

where $P(\mathbf{C}_k \to \mathbf{C}_l|\boldsymbol{\zeta})$ with $k \neq l$ denotes $\Pr\left[\|\mathbf{X} - \mathbf{H}^*\mathbf{C}_k\|_{\mathrm{F}}^2 > \|\mathbf{X} - \mathbf{H}^*\mathbf{C}_l\|_{\mathrm{F}}^2 \,|\, \mathbf{C} = \mathbf{C}_k, \boldsymbol{\zeta}\right]$ which, in turn, represents the aforementioned pair wise error probability. As seen, the above expression is the probability that given a transmitted codeword $\mathbf{C}_k$ and the channel estimate

available at the transmitter, the ML decoding metric corresponding to the codeword $\mathbf{C}_l$ is smaller. By conditioning on the true channel $\mathbf{h}$ and employing a well-known upper bound on the Gaussian tail function an upper bound tight at high SNR:s may be obtained [8]. The final expression for the design criterion is obtained by utilizing this bound to further upperbound the right hand side of (1) and thereafter remove parameter independent terms and factors to arrive at the *design criterion*

$$W(\mathcal{C}) = \sum_{k<l} V(\mathbf{C}_k, \mathbf{C}_l)\,, \tag{2}$$

where $\mathcal{C} = \begin{bmatrix}\mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_K\end{bmatrix}$ is a matrix containing the codewords and where the *codeword pair criterion* is given by

$$V(\mathbf{C}_k, \mathbf{C}_l) = \frac{e^{\mathbf{m}_{\mathbf{h}|\varsigma}^* \mathbf{R}_{\mathbf{hh}|\varsigma}^{-1} \Psi(\mathbf{C}_k, \mathbf{C}_l)^{-1} \mathbf{R}_{\mathbf{hh}|\varsigma}^{-1} \mathbf{m}_{\mathbf{h}|\varsigma}}}{\det(\Psi(\mathbf{C}_k, \mathbf{C}_l))}\,, \tag{3}$$

with $\Psi(\mathbf{C}_k, \mathbf{C}_l) \triangleq \mathbf{I}_N \otimes \left(\frac{(\mathbf{C}_k - \mathbf{C}_l)(\mathbf{C}_k - \mathbf{C}_l)^*}{4\sigma^2}\right) + \mathbf{R}_{\mathbf{hh}|\varsigma}^{-1}$.

It is now possible to construct codes that take the channel knowledge at the transmitter into account by minimizing the design criterion with respect to the codewords in $\mathcal{C}$, while satisfying the output power constraint. In the present work, the focus is on searching for efficient codes over a continuous alphabet, where the individual symbols in the codewords do not belong to a fixed signal constellation. Thus, the code design procedure amounts to solving the following optimization problem

$$\mathcal{C}_{\text{opt}} = \arg \min_{\substack{\mathcal{C} \\ \|\mathcal{C}\|_{\text{F}}^2 = PK\log_2(K)}} W(\mathcal{C})\,. \tag{4}$$

A strategy for solving (4) is discussed in Section 4.

3.1 Asymptotic Properties of the Code

It is in general a difficult task to explicitly characterize the optimal set of codewords. However, in a few special cases it is possible to at least derive some asymptotic properties that the optimal codewords possess. The special cases that are studied in this section are summarized in two theorems concerned with perfect channel knowledge and no channel knowledge, respectively. Detailed proofs are found in [9].

Theorem 1 (Perfect channel knowledge) *Let $\mathbf{M}_{\mathbf{h}|\varsigma}$ be an $M \times N$ matrix, defined through the relation $\mathbf{m}_{\mathbf{h}|\varsigma} = \text{vec}(\mathbf{M}_{\mathbf{h}|\varsigma})$ and let $\|\mathbf{R}_{\mathbf{hh}|\varsigma}\| \to 0$. The range space of each codeword in the asymptotically optimal code then belongs to the range space of $\mathbf{M}_{\mathbf{h}|\varsigma}$.*

The above result is worth elaborating upon. Under conditions of perfect channel information, $\mathbf{M}_{\mathbf{h}|\varsigma}$ is essentially equal to $\mathbf{H}$. The fact that $\mathbf{C}_k^{\perp} = 0$ then implies that in the optimal set of codewords, the range space of each codeword is a subspace of the range space of the channel matrix $\mathbf{H}$. In other words, the energy in each codeword is confined to the space spanned by the left singular eigenvectors corresponding to the non-zero singular values of the channel matrix. Such a property makes sense since there is no point wasting energy in "directions" which do not contribute to the received signal. This means that for the case of one receive antenna, the optimal codewords are of rank one, resulting in a beamforming type of transmission.

Next, the no channel knowledge scenario is considered. Unfortunately, the development is complicated by the large number of terms in the asymptotic criterion function. Therefore, in order to obtain tractable expressions, a special case with only two codewords, $\mathbf{C}_1$ and $\mathbf{C}_2$, is studied. It is reasonable to assume that this provides valuable insight also into the general case of an arbitrary number of codewords.

Theorem 2 (No channel knowledge) *Let $\|\mathbf{R}_{\mathbf{hh}|\varsigma}^{-1}\| \to 0$, $M \leq L$, $\mu > 0$ and assume that the code consists of only $K = 2$ codewords. The two codewords in the asymptotically optimal code then satisfy*

$$\mathbf{C}_1 + \mathbf{C}_2 = 0, \quad \mathbf{C}_1\mathbf{C}_1^* = \mathbf{C}_2\mathbf{C}_2^* = \mu\mathbf{I}_M\,. \tag{5}$$

From (5) it follows that the rows of $\mathbf{C}_1$ and $\mathbf{C}_2$ must be orthogonal and of equal magnitude. Another way of formulating this is to say that the singular values of each codeword are equal. Whether the statement generalizes to any number of codewords is still an open issue. However, numerous experiments where K, M and L were varied and codes constructed as described in Section 4 show that all the resulting codes satisfied the above orthogonality property to a high degree of precision. Information theoretic based arguments in [10] further supports this claim.

4 Numerical Optimization

The optimization problem outlined in the previous section is non-convex and, in general, difficult to solve. This section describes how a numerical gradient search type of technique is used in the present work for finding a locally optimal solution. Although the technique is tailored to the application under consideration, it is similar in structure to the optimization procedure utilized in [5] for the design of efficient signal constellations.

The optimization procedure implements a gradient search which updates $\mathcal{C}$ by taking a step along the direction of the negative gradient. The result is thereafter normalized so as to satisfy the output power constraint of the codeword constellation. Let the gradient $\frac{\partial W}{\partial \mathcal{C}}$ of $W(\mathcal{C})$ be represented by a matrix where the element on the k^{th} row and l^{th} column is equal to the *complex derivative* $\frac{\partial W}{\partial \mathcal{C}_{kl}}$. An expression for the gradient is derived in [9]. Starting with an initial guess $\mathcal{C}_0$ of the solution, the codeword constel-

lation is updated according to

$$\mathcal{C}_{n+1} = \frac{PK\log_2(K)(\mathcal{C}_n - \mu_n \frac{\partial W}{\partial \mathcal{C}})}{\|\mathcal{C}_n - \mu_n \frac{\partial W}{\partial \mathcal{C}}\|}, \quad n = 0, 1, \cdots, \tag{6}$$

where μ_n is the so called step size and $\frac{\partial W}{\partial \mathcal{C}}$ is evaluated at $\mathcal{C}_n$. An appropriate step size is obtained by using successively smaller values of μ_n until the design criterion is found to have decreased. Several alternative methods for choosing the step size exist in the literature. A candidate solution is obtained by iterating (6) until convergence. The gradient search is repeated several times using different $\mathcal{C}_0$ set at random. Finally, the best of the resulting candidate solutions is chosen and designated as the constructed code.

5 An Example Scenario

The generic system model is easily tailored to more specific scenarios. To serve as an example of how the codes in this work may be used, and also for later reference, consider a scenario with a quasi-stationary time-varying channel and a system where channel estimates at the receiver are transported over a feedback link to the transmitter. Vector quantization is employed in the feedback link to keep the bandwidth requirements of the feedback link within reasonable limits. In addition to the resulting quantization errors, a feedback delay further degrades the quality of the channel information.

A possible system setup is depicted in Figure 2. The feedback link consists of an encoder, the actual feedback channel and a decoder. At the encoder, the $MN \times 1$ vector $\tilde{\mathbf{h}}$ is used to represent the *initial channel information* which is fed back to the transmitter. The initial channel information is quantized into a b bit integer i and transported over what is here modeled as an ideal feedback channel, i.e., the channel does not introduce any bit-errors. On the transmitter side, the decoder maps the received integer i into an estimate $\hat{\mathbf{h}}(i)$ of the current channel realization $\mathbf{h}$. A comparison with the generic system model shows that $\hat{\mathbf{h}}(i)$ corresponds to the channel information vector $\boldsymbol{\zeta}$. As previously mentioned, errors in the channel information are not only due to the quantization process. The assumption of a feedback delay in conjunction with a time-varying channel means that the channel information may be outdated by the time it reaches the transmitter side. The latter error source is modeled by assuming that $\tilde{\mathbf{h}}$ is a delayed copy of the current channel realization, correlated, to an arbitrary degree, with $\mathbf{h}$. In the design of the encoder and the decoder, the contributions from these two error sources are reduced by utilizing a modified version of the well-known generalized Lloyd algorithm for minimizing the mean square error $\mathrm{E}[\|\mathbf{h} - \hat{\mathbf{h}}(i)\|^2]$. The design process is straightforward and can be accomplished off-line. The details are, however, beyond the scope of this paper.

A scenario with Rayleigh fading channel coefficients is considered. The channel coefficients are statistically independent and identically distributed with variance equal to σ_h^2. In view of the well-known Jakes model, which describes the time-variations of the channel as a stationary Gaussian process, and since the initial channel information is considered to be a delayed copy of the current channel realization, it is assumed that $\mathbf{h}$ and $\tilde{\mathbf{h}}$ are jointly complex Gaussian and equally distributed. Each initial channel information coefficient $\tilde{h}_{ij}$ is correlated with the corresponding current channel coefficient h_{ij}, and uncorrelated with all others. To describe the degree of correlation, and also to introduce a separate quality measure for the initial channel information, consider the normalized correlation coefficient $\rho = \mathrm{E}[h_{ij}\tilde{h}_{ij}^*]/\sigma_h^2$. Based on this quality measure, perfect initial channel knowledge and no initial channel knowledge corresponds to $\rho \to 1$ and $\rho \to 0$, respectively. The distribution of the current channel and the initial channel information is completely characterized by the covariance matrices $\mathbf{R}_{\mathbf{hh}} = \sigma_h^2\mathbf{I}_{MN}$, $\mathbf{R}_{\tilde{\mathbf{h}}\tilde{\mathbf{h}}} = \sigma_h^2\mathbf{I}_{MN}$, the cross-covariance matrix $\mathbf{R}_{\mathbf{h}\tilde{\mathbf{h}}} = \sigma_h^2\rho\mathbf{I}_{MN}$ and the mean vectors $\mathbf{m}_{\mathbf{h}} = 0$ and $\mathbf{m}_{\tilde{\mathbf{h}}} = 0$.

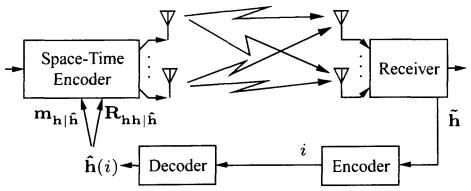


Figure 2: System in the example scenario.

Several parameters are needed in the code design process. For the example scenario, it is assumed that σ^2, ρ and σ_h^2 are known. In practice, these parameters would need to be estimated or taken as design parameters. Based on the assumed parameter knowledge, the conditional mean $\mathbf{m}_{\mathbf{h}|\hat{\mathbf{h}}}$ and covariance $\mathbf{R}_{\mathbf{hh}|\hat{\mathbf{h}}}$, necessary for the construction of the space-time code, can be shown to be given by $\mathbf{m}_{\mathbf{h}|\hat{\mathbf{h}}} = \hat{\mathbf{h}}(i)$, $\mathbf{R}_{\mathbf{hh}|\hat{\mathbf{h}}} = \mathbf{R}_{\mathbf{hh}|\tilde{\mathbf{h}}} - \hat{\mathbf{h}}(i)\hat{\mathbf{h}}(i)^* + \mathbf{R}_{\mathbf{h}\tilde{\mathbf{h}}}\mathbf{R}_{\tilde{\mathbf{h}}\tilde{\mathbf{h}}}^{-1}\mathrm{E}[\tilde{\mathbf{h}}\tilde{\mathbf{h}}^*|i]\mathbf{R}_{\tilde{\mathbf{h}}\tilde{\mathbf{h}}}^{-1}\mathbf{R}_{\mathbf{h}\tilde{\mathbf{h}}}^*$, where $\mathbf{R}_{\mathbf{hh}|\tilde{\mathbf{h}}} = \mathbf{R}_{\mathbf{hh}} - \mathbf{R}_{\mathbf{h}\tilde{\mathbf{h}}}\mathbf{R}_{\tilde{\mathbf{h}}\tilde{\mathbf{h}}}^{-1}\mathbf{R}_{\mathbf{h}\tilde{\mathbf{h}}}^*$ is the covariance matrix of the channel $\mathbf{h}$ conditioned on the initial channel information $\tilde{\mathbf{h}}$. Data from the training of the vector quantizer can be used for computing $\mathrm{E}[\tilde{\mathbf{h}}\tilde{\mathbf{h}}^*|i]$.

Contrary to one of the assumptions in the generic system model, $\mathbf{h}$ conditioned on $\hat{\mathbf{h}}(i)$ is clearly not complex Gaussian distributed, since the feedback link employs quantization. This means that the generic system model used to derive the code construction process is not entirely accurate for the present case, unless the quantization can be neglected such as when an infinite number of bits are used. Nevertheless, as simulation results will show, significant gains resulting from the use of the constructed space-time codes can still be achieved.

Note that several different channel estimate dependent codes can be constructed off-line and stored in a lookup

table, which is indexed by i and appropriately quantized values of σ^2, ρ and σ_h^2. Hence, in such a case, $\hat{\mathbf{h}}(i)$ as well as $\mathbf{m}_{\mathbf{h}|\hat{\mathbf{h}}}$ and $\mathbf{R}_{\mathbf{hh}|\hat{\mathbf{h}}}$ only serve as intermediate quantities in the off-line design process and are not needed on-line.

6 Simulation Results

To illustrate the performance of the constructed space-time codes, simulations of the example scenario described in Section 5 were conducted. The communication system was equipped with $M = 2$ transmit and $N = 1$ receive antennas. By considering two different cases, one with $\rho \to 0$, $b = 0$ (no channel knowledge) and the other with $\rho = 0.9$, $b = 8$ (partial channel knowledge), the impact of the channel information quality on the performance was studied. Note that the former case also corresponds to an open-loop system for which conventional space-time codes are appropriate. Therefore, an OSTBC system utilizing the well-known Alamouti code was used as a benchmark. A comparison with an ideal beamforming system was also performed. For the latter system, the transmission method relied on perfect channel knowledge to multiply each data symbol by a vector matched to the channel, i.e., the transmitted vector was determined as $\mathbf{c}(n) = \mathbf{h}s(n)$, where the vector $\mathbf{h}$ contains the complex conjugate of the true channel coefficients and $s(n)$ represents the n^{th} data symbol.

The assumptions in the simulations were as follows. For all the examined cases, equally probable bits/codewords with an energy of $P = 1$ per information bit were transmitted over the Rayleigh fading channel of the example scenario. The variance of the channel coefficients was set at $\sigma_h^2 = 1$. Furthermore, to make a fair comparison, the effective data rate was fixed at 1 information bit per channel use. Such a rate was achieved by, in the case of the example system, utilizing a constructed code with $L = 2$ and $K = 4$ whereas in the case of the OSTBC system, as well as for the ideal beamforming system, a binary phase shift keying (BPSK) signal constellation was employed.

In Figure 3, the bit error rate (BER) as a function of the SNR is depicted. Clearly, the performance of the two constructed codes, regardless whether the no channel knowledge or $b = 8, \rho = 0.9$ curve is considered, is better than the one for conventional OSTBC. A gain of about 1 dB at a BER of 10^{-2} is obtained for the no channel knowledge case. As can also be verified by asymptotic calculations, the performance of the constructed codes approach each other as the SNR increases. In other words, the usefulness of channel knowledge diminishes as a function of the SNR. On the other hand, if more transmit antennas are used, the gains due to channel knowledge can be substantially larger than for the currently studied scenario.

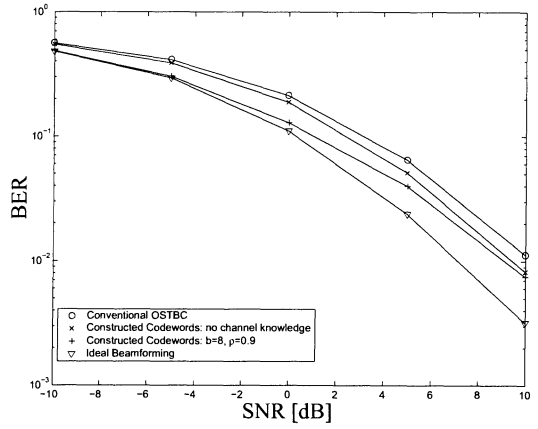


Figure 3: Performance for a two transmit and one receive antenna system.

7 Conclusions

In this work, space-time codes utilizing non-perfect channel information was considered and a design procedure developed. While the focus was on so called continuous alphabet block codes, extensions to finite alphabet and trellis codes are in principle straightforward. A possible application is a wireless system where quantized channel information is available at the transmitter. Even in the case of no channel knowledge, the resulting codes show better performance than conventional space-time codes devised from orthogonal designs, albeit with a higher decoding complexity.

References

[1] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, "Signal Design for Transmitter Diversity Wireless Communication Systems over Rayleigh Fading Channels" *IEEE Transactions on Communications*, vol. 47, pp. 527–537, Apr. 1999.

[2] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction" *IEEE Transactions on Information Theory*, vol. 44, pp. 744–765, Mar. 1998.

[3] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-Time Block Codes from Orthogonal Designs" *IEEE Transactions on Information Theory*, vol. 45, pp. 1456–1467, July 1999.

[4] S. Bäro, G. Bauch, and A. Hansmann, "Improved Codes for Space-Time Trellis-Coded Modulation" *IEEE Communications Letters*, vol. 4, pp. 20–22, Jan. 2000.

[5] G. J. Foschini, R. D. Gitlin, and S. B. Weinstein, "Optimization of Two-Dimensional Signal Constellations in the Presence of Gaussian Noise" *IEEE Transactions on Communications*, vol. 22, pp. 28–37, Jan. 1974.

[6] A. Wittneben, "Optimal Predictive TX Combining Diversity in Correlated Fading for Microcellular Mobile Radio Applications" in *Proc. IEEE Global Telecommunications Conference*, pp. 48–54, 1995.

[7] A. Narula, M. J. Lopez, M. D. Trott, and G. W. Wornell, "Efficient Use of Side Information in Multiple-Antenna Data Transmission over Fading Channels" *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1423–1436, Oct. 1998.

[8] G. Jöngren, M. Skoglund, and B. Ottersten, "Combining Beamforming and Orthogonal Space-Time Block Coding" *IEEE Transactions on Information Theory*, vol. 48, pp. 611–627, Mar. 2002.

[9] G. Jöngren, M. Skoglund, and B. Ottersten, "Design of channel estimate dependent space-time codes." *submitted to IEEE Transactions on Communications*.

[10] T. L. Marzetta and B. M. Hochwald, "Capacity of a Mobile Multiple-Antenna Communication Link in Rayleigh Flat Fading" *IEEE Transactions on Information Theory*, vol. 45, pp. 139–157, Jan. 1999.

Combatting Packet Loss Using Sinusoidal Extrapolation and Interpolation

Jonas Lindblom[1] and Per Hedelin[2]

[1,2] Information Theory Lab, Chalmers University of Technology, SE-412 96 Göteborg, Sweden
E-mail: {jonas.lindblom,per.hedelin}@elmagn.chalmers.se

Abstract

We propose low-delay packet loss concealment schemes for packetized voice communication. The schemes are based on a sinusoidal model of correctly received speech, and the methods are to be implemented in the receivers only. Both *extrapolation*- and *interpolation*-based methods are considered. Initial testing shows promising results; recovered packet losses are often inaudible to a listener.

1 Introduction

Voice communication over packet-switched networks is becoming more and more interesting as the broadband infrastructure is being built-up. Companies and organizations can significantly cut their costs for telephony by utilizing their already existing data networks also for voice traffic. Maintenance is simplified with only one type of network to administrate. Moreover, private users can make international phone-calls basically for "free" over the public Internet. But there are problems related to real-time communication over "best-effort" networks. Real-time applications have very different requirements on the channel compared to traditional data communication applications such as e-mail or file transfers. The traditional applications typically require error-free transmission, which is obtained by the use of *Automatic Repeat Request* (ARQ) schemes built in to the *Transmission Control Protocol* (TCP) [1]. Voice traffic on the other hand, is much more sensitive to delays, but can tolerate a certain degree of packet loss. The *International Telecommunications Union* (ITU) recommends a maximum end-to-end delay of 150 ms for voice communication [2]. This can be a tough limit over a congested network with varying propagation delay. It is therefore of outmost importance to minimize latency at all stages in the system design, and ARQ schemes are not the ideal way to handle packet loss, especially over wireless links.

In this work we focus on low-delay *Packet Loss Concealment* (PLC) schemes that are to be implemented in the receivers only. The objective is to increase the tolerable loss-rate without requiring more bandwidth or delay. Our methods are based on a sinusoidal model of the speech signal from correctly received frames. We are considering both *extrapolation*- and *interpolation*-based methods.

2 PLC based on the Source-Filter Model

The receiver-based PLC scheme is based on loss recovery of the source-filter components of the speech signal which are obtained by means of *Linear Prediction* (LP) analysis [3].

The LP residual and LP coefficients (the spectral part) are treated separately, as indicated in Figure 1.

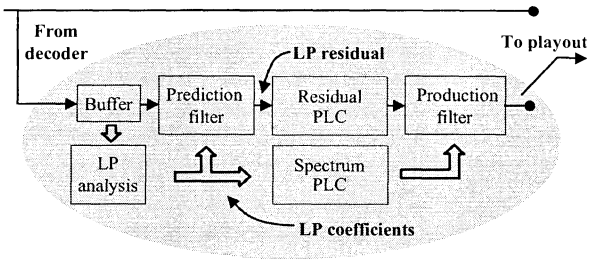


Figure 1: The proposed PLC scheme. The spectrum and residual parts are treated separately.

A fixed frame length of $R = 160$ samples, corresponding to 20 ms of speech at 8 kHz sampling frequency is assumed. The LP analysis is of order $d = 10$, and is performed on all correctly received speech frames to keep the filter states up-to-date. When a speech frame is missing, the switch to the right of Figure 1 selects the output of the PLC unit (the shaded part). The methods to be presented here can be applied in two ways. Either they can be used on their own to recover frame loss in a PCM waveform such as in [4]. Alternatively, they could be incorporated into the receiver of a parametric speech coder that encodes

speech using the source-filter components, such as [5]. Here, we focus on the former.

Previously we have considered PLC based on extrapolation of the source-filter components of the speech signal [6, 7, 8]. Now we extend the scenario from extrapolation to include also interpolation. Obviously this implies an additional delay compared to using the extrapolation scheme only, but it can be motivated by the fact that receivers in packetized speech transmission systems often use *jitter buffers* to combat the varying delays on the network [9]. Our scheme would then work on packets *inside* such a buffer at no extra cost in terms of delay. See Figure 2.

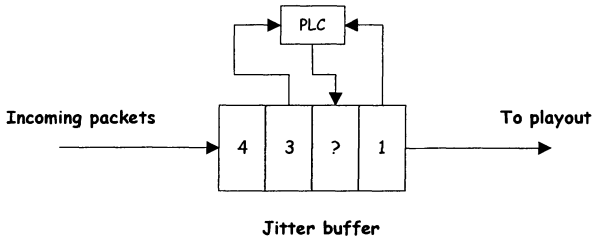


Figure 2: Interpolation based PLC scheme motivated by the use of jitter buffers in VoP systems. Packet 2 of the jitter buffer is concealed by interpolating the parameters of packet 1 and 3.

3 The Sinusoidal Model

The proposed schemes are based on sinusoidal modeling [10] of the LP residual $r(n)$,

$$\hat{r}(n) = \sum_{i=1}^{L} A_i \sin(\omega_i n + \phi_i) \qquad n = 0 \ldots N-1, \quad (1)$$

where L is the number of sine-waves used, $\{A_i\}$ is a set of amplitudes, $\{\omega_i\}$ is a set of frequencies, and $\{\phi_i\}$ is a set of phases. For our experiments on speech sampled at 8 kHz, we use $L = 64$.

In order to find the parameters of the sinusoidal model, a mean-square error criterion is considered,

$$J = \sum_{n=0}^{N-1} \left(r(n) - \sum_{i=1}^{L} A_i \sin(\omega_i n + \phi_i) \right)^2 w(n), \quad (2)$$

where $w(n)$ is a weighting function which turns out to be crucial. The same criterion is used for both extrapolation and interpolation, but the weighting function differs (as will be explained). The criterion is minimized using a recursive method, where one sinusoidal is found at a time. The sinusoidal estimation method resembles the one used in e.g. [11]. For more details, see [8].

4 Extrapolation

4.1 Residual Extrapolation

The main idea with the sinusoidal extrapolation scheme is to model a segment of the last received LP-residual according to (1), and then let the sine-waves continue their oscillations across frame boundaries in order to cover for one or more lost frames. The weighting function $w(n)$ in (2) is important for successful sinusoidal extrapolation. A parabolic-exponential window on the form

$$w(n, N) = \frac{nN - n^2}{N^2} e^{\frac{-\alpha(N-n)}{N}}, \quad 0 \le n \le N-1 \quad (3)$$

is considered. The parameter α determines the shape of the window. We use $\alpha = 5$ in order to emphasize the most recent part of the analysis buffer. The shape of the function in (3) can be found in Figure 3.

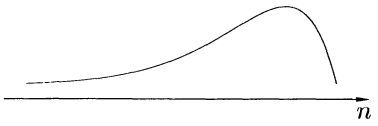


Figure 3: The shape of (3), with $\alpha = 5$.

The choice of the number of residual samples N, to include in the sinusoidal modeling is a key issue. The analysis must be based on a sufficiently large amount of data so that the fundamental frequency (the pitch) is captured. Therefore, we choose N based on a pitch analysis of the residual history. If a reliable pitch estimate is not obtainable (e.g. for unvoiced sounds), then $N = R$. If there is a pitch estimate of p_0 (samples), N is set according to

- If $R/p_0 > 3$ then $N = \lfloor R/p_0 \rfloor p_0$
- If $R/p_0 < 3$ then $N = \lceil R/p_0 \rceil p_0$

if this amount of data is available in the residual history buffer. Frame losses may have occurred and caused the residual history buffer to be at its minimum, which is $R - d$ samples. If this is the case, and the required N is larger than $R - d$, then $N = R - d$ samples will be used for the analysis.

At the end of an error burst, the extrapolated signal is overlap-added with the "true" signal in order to obtain a smooth transition.

4.2 Spectrum Extrapolation

The slowly varying spectral part is recovered by repeating a bandwidth expanded version of the previous LP filter as

$$a'_k = (0.98)^k a_k \qquad k = 1 \ldots d \qquad (4)$$

where $\{a'_k\}$ are the coefficients of the bandwidth expanded production filter. The effect of this is to "soften" the filter. If more than one consecutive frame loss occurs, the bandwidth expansion procedure is employed again. During error frames, the LP production filter is interpolated on a sample-by-sample basis from the filter of the previous frame, to the bandwidth expanded version. For the first correctly received frame after an error burst (from now on denoted "recovery frame"), the LP filter is interpolated from the bandwidth-expanded filter of the last error frame, to the "true" filter, derived from the data of the recovery frame. In this way, smooth transitions are obtained. The interpolation is performed as is described in Section 5.2.

More elaborate spectrum prediction schemes have been suggested [6], but the simple frame repetition scheme performs well in the current setting, and is used successfully in standardized PLC schemes [5, 12]. Initial experiments indicated that the residual part is perceptually more important to recover accurately.

5 Interpolation

5.1 Residual Interpolation

The residual interpolation scheme is also based on fitting a sinusoidal model to correctly decoded data, and then using the model to produce the missing part. For the interpolation case decoded data is assumed to be available on "both sides" of the missing segment. We find the parameters of the sinusoidal model by minimizing the same criterion (2), but with a modified weighting function $w_{\text{IP}}(n)$,

$$\begin{aligned} w_{\text{IP}}(n) &= w(n, N_1), && 0 \le n \le N_1 - 1 \\ w_{\text{IP}}(n) &= 0, && N_1 \le n \le N_1 + R - d - 1 \\ w_{\text{IP}}(n) &= w(N_e - n, N_2), && N_1 + R - d \le n < N_e. \end{aligned} \tag{5}$$

where N_1 and N_2 are the number of samples used for model estimation prior, and after the frame loss, respectively. Furthermore $N_e = N_1 + R + d + N_2 - 1$. N_1 and N_2 are estimated based on pitch analysis, as for the residual extrapolation in Section 4.1. A problem when interpolating in the LP-residual domain is that the frame loss will cause the LP-prediction filter to lose it state, and the first d samples after a loss will be unreliable. With the suggested choice of $w_{\text{IP}}(n)$, we neglect the first d residual samples after a frame-erasure when estimating the model. The shape of (5) is found in Figure 4. By adopting this strategy, we fit the sinusoidal model to data from both sides of the missing segment, emphasizing the parts closest to the "hole". Using (1), it is then possible to repair the waveform using the estimated parameters.



Figure 4: The shape of (5).

If more than one frame is lost, the same technique is employed, but with a larger "hole" in (5).

5.2 Spectrum Interpolation

The spectrum interpolation is performed using the cepstral parameterization [3] of the LP coefficients. The polynomial LP coefficients are first converted to cepstral coefficients, which are expanded to dimension 64, linearly interpolated, converted to reflection coefficients, truncated to dimension 10, and then finally converted back to polynomial coefficients. In this way, smooth transitions are obtained.

6 Experiments and Results

In this section we report on the results from some testing of the proposed PLC system. The ultimate goal of the PLC is to render a repaired waveform that is of *subjectively* good quality to a listener. It is hard to find objective measures that correlate well with subjective performance. Segmental *Signal-to-Noise Ratio* (SNR) for example, is not very useful here, as a small phase error can result in negative SNR, while a listener would not notice the error. And discontinuities (e.g. "clicks") would not affect SNR much, but could be very annoying to a listener. Therefore, we base our evaluation on subjective testing.

In [8], the extrapolation scheme is applied to a multi-sentence recording from 4 different speakers with random frame-erasures ($P_{\text{error}} = 0.15$). The sinusoidal extrapolation scheme, and two benchmark schemes, are then employed to repair the frame-erasures. Benchmark system 1, does not work on the source-filter components but simply replaces lost segments by inserting repetitions of the last correctly received p_0 speech samples. To some extent, benchmark system 1 resembles the techniques in [4]. Benchmark system 2 can be seen as a mix between benchmark system 1 and the proposed sinusoidal extrapolation scheme. It is based on the source-filter components; the spectral part is recovered exactly as in the sinusoidal extrapolation scheme, but the LP residual is recovered using repetitions of the last p_0 residual samples. The repaired waveforms are then presented pairwise (one sentence repaired with one of the methods versus the same sentence, with the same frame-erasures, but repaired with another method) to 10 listeners. The lis-

teners are forced to indicate their preference, but are given the option to repeat a sentence pair an unlimited number of times prior to the forced decision. The listeners' preferences can be found in Table 1.

Method A	Method B	Pref. A	Pref. B
SEP	B1	76	4
SEP	B2	66	14
B1	B2	21	59

Table 1: Results of the subjective testing of the proposed Sinusoidal ExtraPolation (SEP) scheme, Benchmark scheme 1 (B1), and Benchmark scheme 2 (B2). The numbers indicate the number of preferences (out of 80 tests) for a certain method.

According to the subjective testing, the sinusoidal extrapolation scheme is preferred to benchmark system 1 in 95% of the cases and benchmark system 2 is preferred to benchmark system 1 in 74% of the cases. Hence, there is a clear indication that the schemes operating on the source-filter components are preferred by the listeners. Moreover, the sinusoidal extrapolation scheme is preferred to benchmark system 2 in 83% of the cases and thus outperforms both methods based on pitch period repetition. The effect of a frame-loss is illustrated in Figure 5, where it can be seen how the spectral content of the signal is preserved throughout the frame-erasure.

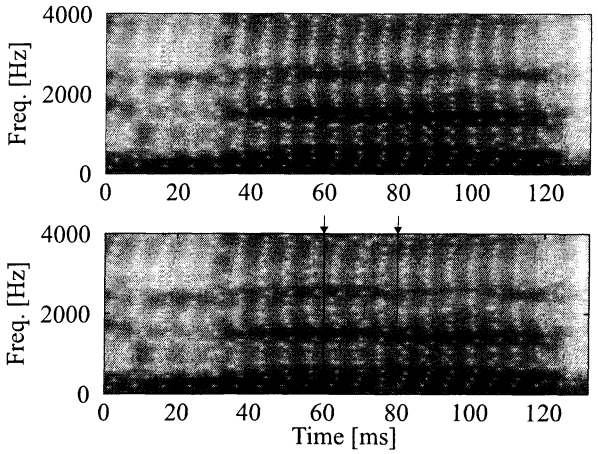


Figure 5: Illustrating the effect of a 20 ms frame loss using spectrograms. Top: the original signal (no loss). Bottom: the arrows and the vertical lines indicate the frame-erasure, which is concealed with the sinusoidal extrapolation scheme.

When it comes to interpolation based PLC, the method is still very much under development. Initial testing seems promising. Most of the time, smooth frame-recovery is obtained. An example can be found in Figure 6 where both interpolation and extrapolation has been applied to the same frame-erasure. As of

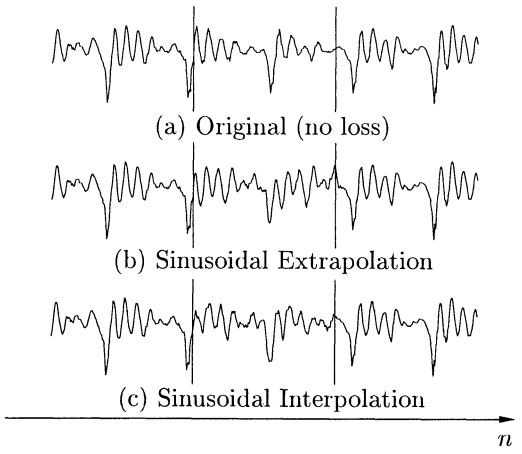


Figure 6: Illustrating the effects of a 20 ms packet loss using the two proposed schemes. The segment between the two vertical lines is recovered in subfigure (b) and (c).

today, the interpolation scheme is however not as robust as the extrapolation scheme. If the speech signal changes too much during an error event, the interpolation based method may produce a slightly distorted result. This is naturally more frequent during long error-bursts. It has also been noted that the interpolation scheme seems to work well on music, which compared to speech, typically can be regarded as a more stationary source

An attractive thought is of course to combine extrapolation and interpolation into one robust PLC scheme. Shorter error-events, say one or two frame-losses, would then be handled by the interpolative scheme, whereas more severe error bursts are handled by means of extrapolation.

References

[1] W.R. Stevens. *TCP/IP Illustrated, Volume 1 - The Protocols.* Addison Wesley, 1994.

[2] ITU-T Recommendation G.114. *One-Way Transmission Time.* ITU-T, 2000.

[3] L.R. Rabiner and R.W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

[4] ITU-T Recommendation G.711 Appendix I. *A high quality low-complexity algorithm for packet loss concealment with G.711.* ITU-T, 1999.

[5] R. Salami, C. Laflamme, J Adoul, A. Kataoka, S. Hayashi, T. Moriya, C Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham. Design and description of CS-ACELP: A toll quality 8

kb/s speech coder. *IEEE Trans. Speech Audio Processing*, 6(2):116–130, 1998.

[6] J. Lindblom, J. Samuelsson, and P. Hedelin. Model based spectrum prediction. In *Proc. IEEE Workshop on Speech Coding*, pages 117–119, 2000.

[7] J. Lindblom. *Packetized Speech Transmission - Combatting the Packet Loss Problem.* Licentiate thesis, technical report no: 403L, Chalmers University of Technology, 2001.

[8] J. Lindblom and P. Hedelin. Packet loss concealment based on sinusoidal extrapolation. Accepted for presentation at IEEE Int. Conf. Acoust., Speech, Signal Processing, 2002.

[9] J.C. De Martin, T. Unno, and V. Viswanathan. Improved frame erasure concealment for CELP-based coders. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 3, pages 1483–1486, 2000.

[10] R.J. McAulay and T.F. Quatieri. Sinusoidal coding. In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pages 121–173. Elsevier Science Publishers, 1995.

[11] E.B. George and M.J.T. Smith. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Trans. Speech Audio Processing*, 5(5):389–406, 1997.

[12] ITU-T Recommendation G.728 Annex I. *Frame or packet loss concealment for the LD-CELP decoder.* ITU-T, 1999.

Unsupervised Detection over Time-Dispersive Vector Channels

Anders Hansson[1] and Tor Aulin[2]

[1]Dept. Computer Eng., Chalmers U. Tech., SE-412 96 Göteborg, Sweden. E-mail: ahansson@ce.chalmers.se
[2]Dept. Computer Eng., Chalmers U. Tech., SE-412 96 Göteborg, Sweden. E-mail: aulin@ce.chalmers.se

Abstract

The method of transmitting equidistant pulses in time to carry discrete information is basic. When the pulses overlap, intersymbol interference is present. Optimal maximum likelihood sequence detection of such signals in additive white Gaussian noise (AWGN) was not known until 1972 [1]. It is less artificial to assume that the pulse shape is not perfectly known to the receiver designer. An unknown pulse shape, which is due to distortion or interference from multiple propagation paths, occurs in most practical transmission media. Our aim is to develop highly efficient detection algorithms for this problem. More specifically, an *a posteriori* probability algorithm is presented.

1 Model

Consider a data sequence $\{d_k\}_{k=0}^{K-1}$ that is linearly modulated and conveyed at the transmission rate $1/T_s$. Let the received signal $r(t)$ be modeled as the transmitted waveform sequence corrupted by a time-dispersive distortion, plus AWGN,

$$r(t) = \sum_{k=0}^{K-1} d_k\, h(t - kT_s) + w(t), \tag{1}$$

where the (noiseless) channel response $h(t)$ to the modulation pulse is compactly supported on $[0, LT_s)$, while $\mathrm{E}\{w(t)w^*(t+\tau)\} = N_0\delta(\tau)$. If the receiver front-end feeds $r(t)$ to a lowpass filter with cutoff rate W (and ideal characteristic), and then uniformly samples the filter output at the Nyquist rate, the nth sample, $n = 0, \ldots, N-1$, in the kth symbol interval is subject to intersymbol interference,

$$\begin{aligned} r_{k,n} &= \sum_{\ell=0}^{L-1} d_{k-\ell}\, h(\ell T_s + nT_s/N) + w_{k,n} \\ &= \sum_{\ell=0}^{L-1} d_{k-\ell}\, h_{\ell,n} + w_{k,n}, \end{aligned} \tag{2}$$

with $\mathrm{E}\{w_{i,m}w_{k,n}^*\} = 2WN_0\delta_{ik}\delta_{mn}$. For simplicity, we restrict our attention to a fractionally spaced sampler that operates at an integer multiple of the baud rate. Equation (2) is more conveniently expressed using matrix algebra,

$$r_{k,n} = \mathbf{d}_k^T\, \mathbf{h}_n + w_{k,n}, \tag{3}$$

with

$$\begin{aligned} \mathbf{d}_k^T &= \begin{bmatrix} d_{k-(L-1)} & d_{k-(L-2)} & \cdots & d_k \end{bmatrix}, & (4) \\ \mathbf{h}_n &= \begin{bmatrix} h_{L-1,n} & h_{L-2,n} & \cdots & h_{0,n} \end{bmatrix}^T. & (5) \end{aligned}$$

Likewise, the observation vector extracted in symbol interval k is

$$\mathbf{r}_k = \begin{bmatrix} r_{k,0} & r_{k,1} & \cdots & r_{k,N-1} \end{bmatrix}^T = \mathbf{D}_k\, \mathbf{h} + \mathbf{w}_k, \tag{6}$$

where

$$\begin{aligned} \mathbf{D}_k &= \mathrm{diag}\left\{\mathbf{d}_k^T, \mathbf{d}_k^T, \ldots, \mathbf{d}_k^T\right\}, & (7) \\ \mathbf{h} &= \begin{bmatrix} \mathbf{h}_0^T & \mathbf{h}_1^T & \cdots & \mathbf{h}_{N-1}^T \end{bmatrix}^T, & (8) \\ \mathbf{w}_k &= \begin{bmatrix} w_{k,0} & w_{k,1} & \cdots & w_{k,N-1} \end{bmatrix}^T, & (9) \end{aligned}$$

and the total observation vector has a similar compact form,

$$\mathbf{r}_0^{K-1} = \begin{bmatrix} \mathbf{r}_0^T & \mathbf{r}_1^T & \cdots & \mathbf{r}_{K-1}^T \end{bmatrix}^T = \mathbf{D}_0^{K-1}\, \mathbf{h} + \mathbf{w}_0^{K-1}, \tag{10}$$

if we let

$$\begin{aligned} \mathbf{D}_0^{K-1} &= \begin{bmatrix} \mathbf{D}_0^T & \mathbf{D}_1^T & \cdots & \mathbf{D}_{K-1}^T \end{bmatrix}^T, & (11) \\ \mathbf{w}_0^{K-1} &= \begin{bmatrix} \mathbf{w}_0^T & \mathbf{w}_1^T & \cdots & \mathbf{w}_{K-1}^T \end{bmatrix}^T. & (12) \end{aligned}$$

A more thorough treatment of optimality considerations for the front-end is found in [2, 3].

2 The APP Algorithm

The state graph of the (digital) encoder/modulator describes how the set of hypothetical data sequences can be arranged in a tree.[1] Each node at depth $k \in \{0, \ldots, K-1\}$ is labeled by the hypothetical data matrix $\widetilde{\mathbf{D}}_0^k$ associated with the path leading to the node.

[1]Uncoded memoryless modulation can be visualized by a trivial state graph having only one state.

Let edge number i at time index k be denoted $e_{k,i}$, and consider its *A Posteriori* Probability (APP) by averaging over all consistent hypothetical sequences,[2]

$$\begin{aligned}\mathrm{APP}(e_{k,i}) &\triangleq \Pr\left(e_{k,i}|\mathbf{r}_0^{K-1}\right) \cong p(\mathbf{r}_0^{K-1}|e_{k,i})\Pr(e_{k,i}) \\ &= \sum_{\widetilde{\mathbf{D}}_0^{K-1}:e_{k,i}} \Pr\left(\widetilde{\mathbf{D}}_0^{K-1}\right) p(\mathbf{r}_0^{K-1}|\widetilde{\mathbf{D}}_0^{K-1}),\end{aligned} \tag{13}$$

where $\cong$ means that the left and right members of the relation are equivalent in an information theoretical sense.[3] Note that the likelihood function, i.e. $p(\mathbf{r}_0^{K-1}|\widetilde{\mathbf{D}}_0^{K-1})$, is undefined due to the composite nature of our problem: both the channel parameter vector $\mathbf{h}$ and the data matrix $\mathbf{D}_0^{K-1}$ are unknown. The problem could have been reduced into a simple hypothesis-testing problem provided that the receiver designer had knowledge of the probability density function (pdf) governing the mapping from the source to the observation space. A solution is then obtained by averaging over the channel parameter vector [4]. We are facing another case of interest, which is perhaps less artificial, since the (deterministic) channel parameter vector has no pdf over which to average. The standard solution is to employ a *generalized likelihood function* [4],

$$\begin{aligned}&\max_{\mathbf{h}\in\mathbb{C}^{NL}} p(\mathbf{r}_0^{K-1}|\widetilde{\mathbf{D}}_0^{K-1},\mathbf{h}) \\ &\cong \exp\left\{-\frac{1}{2WN_0}\left\|\mathbf{r}_0^{K-1}-\widetilde{\mathbf{D}}_0^{K-1}\tilde{\mathbf{h}}_{\mathrm{ml}}\right\|^2\right\},\end{aligned} \tag{14}$$

which involves the maximum likelihood estimate $\tilde{\mathbf{h}}_{\mathrm{ml}}$ of the unknown parameter vector, assuming that the hypothesized data matrix $\widetilde{\mathbf{D}}_0^{K-1}$ is true,

$$\tilde{\mathbf{h}}_{\mathrm{ml}} = \left(\widetilde{\mathbf{D}}_0^{K-1}\right)^{\dagger}\mathbf{r}_0^{K-1}. \tag{15}$$

The expression in (15) is readily obtained using the method of normal equations, where $\left(\widetilde{\mathbf{D}}_0^{K-1}\right)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of $\widetilde{\mathbf{D}}_0^{K-1}$ [5].

Obviously, the generalized sequence APP

$$\Pr\left(\widetilde{\mathbf{D}}_0^{K-1}\right)\exp\left\{-\frac{1}{2WN_0}\left\|\mathbf{r}_0^{K-1}-\widetilde{\mathbf{D}}_0^{K-1}\tilde{\mathbf{h}}_{\mathrm{ml}}\right\|^2\right\} \tag{16}$$

depends upon the *entire* hypothesized data sequence through the associated channel parameter estimate $\tilde{\mathbf{h}}_{\mathrm{ml}}$. This means that any forced-folding of the exponentially growing tree search into a trellis search is suboptimal, as is any search other than the exhaustive search. The same problem occurs in infinite-memory channels, and a sub-optimal detector is then typically obtained by truncating the channel memory [6]. The truncation depth μ folds the search tree into a trellis, where the states can be labeled with the μ most recent hypothesized symbols. A similar result is also achieved for our problem by pruning the search tree into a trellis. This could for example be done by means of the Per-Survivor Processing (PSP) concept [7] in combination with the Viterbi Algorithm (VA) as suggested in [8]. Given that a tree of depth $k=\mu$ has been grown, incremental metrics are first computed for all branches that emerge from each node. If we assume that each symbol $\{d_k\}_{k=0}^{\mu}$ is drawn from an M-ary alphabet, there are $M^{\mu+1}$ nodes at the depth $k=\mu$ (corresponding to $M^{\mu+1}$ sequences), and the number of incremental metrics that has to be computed is $M^{\mu+2}$. Second, the VA chooses a subset of paths through the familiar add-compare-select operations, such that the paths are guaranteed to have different recent history. The PSP principle is intuitive and simply means that each path keeps and updates its own individual estimated parameter vector. We have thus obtained a trellis structure, where the states are defined as the $\mu+1$ most recent hypothesized symbols. Note that there is a sign ambiguity involved in (16): it is impossible to discriminate two message sequences that only differ in sign (antipodal signals), since they are associated with two channel estimates that only differ in sign. Differential encoding can be employed to solve this problem. Each state can then be defined as the μ most recent (hypothesized) differentially encoded bits.

It is further desirable to find a recursive way of computing (16), such that for time index k, values from epoch $k-1$ can be used. This problem is solved by means of the Recursive Least-Squares (RLS) algorithm [9, 2, 10].

Finally, let us present the algorithm without going into any details of its derivation. As already mentioned, the goal is to express the exponent in (16)

$$\widetilde{\Gamma}_0^{K-1} \triangleq \left\|\mathbf{r}_0^{K-1}-\widetilde{\mathbf{D}}_0^{K-1}\tilde{\mathbf{h}}_{\mathrm{ml}}\right\|^2 = \widetilde{\Gamma}_0^{k}+\widetilde{\Gamma}_{k+1}^{K-1}+\tilde{\Lambda}_k \tag{17}$$

recursively,

$$\widetilde{\Gamma}_0^{k} = \widetilde{\Gamma}_0^{k-1}+\tilde{\gamma}_k^{\mathrm{fwd}}, \tag{18}$$

$$\widetilde{\Gamma}_{k+1}^{K-1} = \widetilde{\Gamma}_{k+2}^{K-1}+\tilde{\gamma}_{k+1}^{\mathrm{bwd}}. \tag{19}$$

Two search trees are thus grown simultaneously: a forward tree whose root originates at time index $k=0$, and a backward tree whose root originates at time index $k=K-1$. The incremental metrics for the paths in the forward tree are given by a forward-directed RLS algorithm,

$$\tilde{\gamma}_k^{\mathrm{fwd}} = \tilde{\mathbf{e}}_{k-1}^H\widetilde{\mathbf{V}}_k\tilde{\mathbf{e}}_{k-1}, \tag{20}$$

$$\tilde{\mathbf{e}}_{k-1} = \mathbf{r}_k-\widetilde{\mathbf{D}}_k\hat{\tilde{\mathbf{h}}}_{k-1}, \tag{21}$$

$$\widetilde{\mathbf{V}}_k = \left(\mathbf{I}+\widetilde{\mathbf{D}}_k\widetilde{\mathbf{P}}_0^{k-1}\widetilde{\mathbf{D}}_k^H\right)^{-1}, \tag{22}$$

[2] We do not bother to make a distinction between a data sequence and the data matrix that it implies, since the mapping is one-to-one.

[3] A normalization constant that does not provide any information has been omitted.

with recursions

$$\tilde{\hat{\mathbf{h}}}_k = \tilde{\hat{\mathbf{h}}}_{k-1} + \tilde{\mathbf{P}}_0^{k-1}\tilde{\mathbf{D}}_k^H\tilde{\mathbf{V}}_k\tilde{\mathbf{e}}_{k-1}, \quad (23)$$

$$\tilde{\mathbf{P}}_0^k = \left(\mathbf{I} - \tilde{\mathbf{P}}_0^{k-1}\tilde{\mathbf{D}}_k^H\tilde{\mathbf{V}}_k\tilde{\mathbf{D}}_k\right)\tilde{\mathbf{P}}_0^{k-1}, \quad (24)$$

and initializations

$$\tilde{\hat{\mathbf{h}}}_\mu = \tilde{\mathbf{P}}_0^\mu\left(\tilde{\mathbf{D}}_0^\mu\right)^H\mathbf{r}_0^\mu, \quad (25)$$

$$\tilde{\mathbf{P}}_0^k = \left(\left(\tilde{\mathbf{D}}_0^\mu\right)^H\tilde{\mathbf{D}}_0^\mu\right)^{-1}. \quad (26)$$

Likewise, the incremental metrics for the paths in the backward tree are given by a backward-directed RLS algorithm,

$$\tilde{\gamma}_k^{\text{bwd}} = \left(\tilde{\mathbf{e}}_{k+2}^{\text{b}}\right)^H\tilde{\mathbf{V}}_{k+1}^{\text{b}}\,\tilde{\mathbf{e}}_{k+2}^{\text{b}}, \quad (27)$$

$$\tilde{\mathbf{e}}_{k+2}^{\text{b}} = \mathbf{r}_{k+1} - \tilde{\mathbf{D}}_{k+1}\tilde{\hat{\mathbf{h}}}_{k+2}^{\text{b}}, \quad (28)$$

$$\tilde{\mathbf{V}}_{k+1}^{\text{b}} = \left(\mathbf{I} + \tilde{\mathbf{D}}_{k+1}\tilde{\mathbf{P}}_{k+2}^{K-1}\tilde{\mathbf{D}}_{k+1}^H\right)^{-1}, \quad (29)$$

with recursions

$$\tilde{\hat{\mathbf{h}}}_{k+1}^{\text{b}} = \tilde{\hat{\mathbf{h}}}_{k+2}^{\text{b}} + \tilde{\mathbf{P}}_{k+2}^{K-1}\tilde{\mathbf{D}}_{k+1}^H\tilde{\mathbf{V}}_{k+1}^{\text{b}}\tilde{\mathbf{e}}_{k+2}^{\text{b}}, \quad (30)$$

$$\tilde{\mathbf{P}}_{k+1}^{K-1} = \left(\mathbf{I} - \tilde{\mathbf{P}}_{k+2}^{K-1}\tilde{\mathbf{D}}_{k+1}^H\tilde{\mathbf{V}}_{k+1}^{\text{b}}\tilde{\mathbf{D}}_{k+1}\right)\tilde{\mathbf{P}}_{k+2}^{K-1}, \quad (31)$$

and initializations

$$\tilde{\hat{\mathbf{h}}}_{K-(\mu+1)}^{\text{b}} = \tilde{\mathbf{P}}_{K-(\mu+1)}^{K-1}\left(\tilde{\mathbf{D}}_{K-(\mu+1)}^{K-1}\right)^H\mathbf{r}_{K-(\mu+1)}^{K-1}, \quad (32)$$

$$\tilde{\mathbf{P}}_{K-(\mu+1)}^{K-1} = \left(\left(\tilde{\mathbf{D}}_{K-(\mu+1)}^{K-1}\right)^H\tilde{\mathbf{D}}_{K-(\mu+1)}^{K-1}\right)^{-1}. \quad (33)$$

Finally, the penalty term in (17),

$$\begin{aligned}\tilde{\Lambda}_k &= \left(\tilde{\hat{\mathbf{h}}}_{k+1}^{\text{b}} + \tilde{\hat{\mathbf{h}}}_{\text{c}}\right)^H\left(\tilde{\mathbf{P}}_0^k\right)^{-1}\left(\tilde{\hat{\mathbf{h}}}_{k+1}^{\text{b}} + \tilde{\hat{\mathbf{h}}}_{\text{c}}\right) \\ &+ \left(\tilde{\hat{\mathbf{h}}}_k + \tilde{\hat{\mathbf{h}}}_{\text{c}}\right)^H\left(\tilde{\mathbf{P}}_{k+1}^{K-1}\right)^{-1}\left(\tilde{\hat{\mathbf{h}}}_k + \tilde{\hat{\mathbf{h}}}_{\text{c}}\right), \quad (34)\end{aligned}$$

can be expressed by means of a joint parameter estimate,

$$\tilde{\hat{\mathbf{h}}}_{\text{c}} = -\tilde{\mathbf{P}}_0^k\,\tilde{\mathbf{Q}}_k\,\tilde{\hat{\mathbf{h}}}_k - \tilde{\mathbf{P}}_{k+1}^{K-1}\tilde{\mathbf{Q}}_k\,\tilde{\hat{\mathbf{h}}}_{k+1}^{\text{b}}, \quad (35)$$

where

$$\tilde{\mathbf{Q}}_k = \left(\tilde{\mathbf{P}}_0^k + \tilde{\mathbf{P}}_{k+1}^{K-1}\right)^{-1}. \quad (36)$$

Note that the RLS recursions can be more efficiently computed if we introduce a few intermediate matrices, something which is omitted here in order to keep the presentation short.

3 Numerical Results

The presented algorithm was applied to APP detection of Minimum Shift Keying (MSK) over a channel (L=4, N = 2) characterized by the parameter vector

$$\begin{aligned}\mathbf{h} = [\ & -0.2350 + 0.1708j \\ & -0.6762 + 0.4913j \\ & -0.4218 + 0.3064j \\ & 0 \\ & -0.0007 + 0.0005j \\ & -0.4924 + 0.3577j \\ & -0.6109 + 0.4438j \\ & -0.1463 + 0.1063j\ ].\end{aligned}$$

It was noticed that the bit error rate is unacceptably high for uncoded transmission. We have therefore initiated an investigation of the performance of serially concatenated coded systems. However, the soundness of the generated APP's can easily be illustrated. Fig. 1 shows APP's for a typical frame with a low bit error rate, whereas Fig. 2 shows APP's for a typical frame with an unacceptably high bit error rate. A frame such as the one in Fig. 2 is termed catastrophic since about half its bits are erroneously decoded.

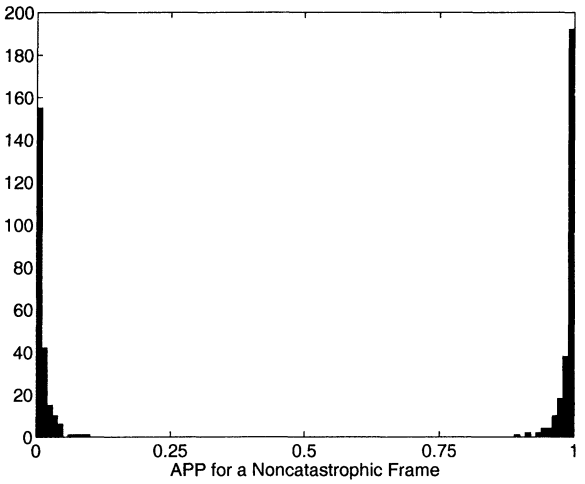


Figure 1: A histogram of the generated APP's in a typical noncatastrophic frame of 512 bits.

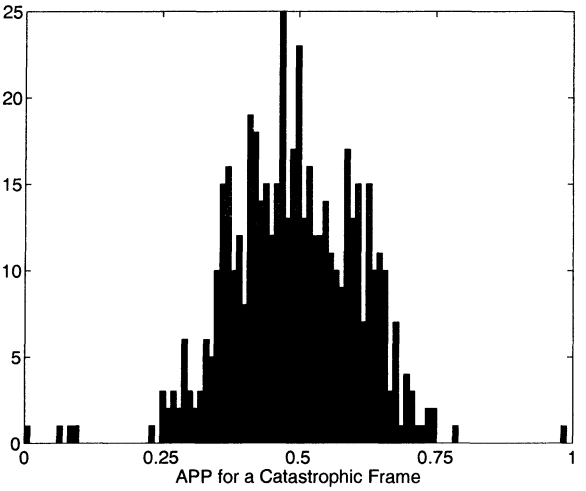


Figure 2: A histogram of the generated APP's in a typical catastrophic frame of 512 bits.

4 Conclusions

An APP algorithm for detection of unknown overlapping pulses has been presented. This algorithm will be used for iterative detection of coded transmission.

References

[1] G. D. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Transactions on Information Theory*, vol. 18, no. 3, pp. 363–378, May 1972.

[2] K. M. Chugg, *Sequence estimation in the presence of parametric uncertainty.* PhD thesis, University of Southern California, Los Angeles, CA, Aug. 1995.

[3] A. Hansson and T. Aulin, "On the discretization of unsupervised digital communication over time-dispersive channels," in *Proc. IEEE International Symposium on Information Theory*, Lausanne, Switzerland, to be published in June 2002.

[4] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* New York, NY: John Wiley & Sons, 1968.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore, MD: The Johns Hopkins University Press, 3rd ed., 1996.

[6] A. Hansson, K. M. Chugg, and T. Aulin, "On forward-adaptive versus forward/backward-adaptive SISO algorithms for Rayleigh fading channels," *IEEE Communications Letters*, vol. 5, no. 12, pp. 477-479, Dec. 2001.

[7] R. Raheli, A. Polydoros, and C. Tzou, "Per-survivor processing: A general approach to MLSE in uncertain environments," *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 354–364, Feb./Mar./Apr. 1995.

[8] A. Anastasopoulos, *Adaptive soft-input soft-output algorithms for iterative detection.* PhD thesis, University of Southern California, Los Angeles, CA, Oct. 1999.

[9] S. Haykin, *Adaptive Filter Theory.* Upper Saddle River, NJ: Prentice Hall, 3rd ed., 1996.

[10] S. Talwar, *Blind space-time algorithms for wireless communication systems.* PhD thesis, Stanford University, Stanford, CA, Jan. 1996.

GMPLS and QoS: New Directions for the Internet

Arne A. Nilsson

Department of Electrical and Computer Engineering
NCSU, Raleigh, NC

Department of Telecommunications and Signal Processing
Blekinge Institute of Technology

Extended Abstract

During the last decade we have seen a tremendous growth in the Internet. The introduction of the World Wide Web coupled with the enormous growth in fiber capacity has meant that the net has virtually exploded in terms of hosts, users, applications, and services. The users of the Internet is requiring more and more services and also guarantees with regards to this service. The requirement to handle specific application traffic with differentiated service quality, is stressing existing Internet and Intranet infrastructure. This includes both hardware and software. The hardware includes routers, switches and cables, and the software is in the form of new traffic engineering and traffic control protocol. The new advances in optical fiber transmission provide the required bandwidth for emerging applications. However, routers and switches that can exploit the optical fiber capabilities are also needed. New IP switching technologies and protocols provide some of the functionality that routers and switches can use to enable optical transport systems to satisfy Internet traffic demands.

As an example the recent development with regards to Multi-Protocol Label Switching (MPLS) is an example of how a new set of protocols for provisioning and managing core networks is growing in popularity. The networks may be data-network like those of ISPs, voice-network as in traditional telecommunications companies, or one of the new networks that combine video, voice and data. MPLS is in effect an addition to an IP network that will permit resources to be reserved and routes pre-determined. MPLS superimposes a connection-oriented framework over the connectionless IP-network. It provides virtual connections through the network to connect nodes that lie at the edge of the network. This is very similar to what is done in ATM networks and there are those that claim that MPLS is just another way of spelling ATM.

The high reliability requirement in telephone networks is such that the subscribers should not have their calls dropped, and should always have access to their service. Downtime must be kept to a minimum, and backup resources must be provided to take over when any component of the network fails. The data world is increasingly demanding similar levels of service to those common in the arena of telephony. Individual customers expect to be able to obtain service at all times and expect reasonable levels of bandwidth. Corporate customers expect the same services, but may also have data streams that are sensitive to delays and disruption.
As a consequence of the convergence of video, voice and data networks modern integrated networks need to be provisioned using protocols, software and hardware that can guarantee high levels of availability. This may be achieved by provisioning redundancy both with regards to hardware and software. When a primary component fails, processing is switched to the stand-by component.

Multi-Protocol Label Switching (MPLS) is rapidly becoming a key technology for use in core networks, including converged video, voice and data networks. MPLS does not replace IP routing, but works alongside existing and future routing technologies to guarantee high-speed data forwarding between Label-Switched Routers (LSRs) together with reservation of bandwidth for traffic flows with differing Quality of Service (QoS) requirements. MPLS enhances the services that can be provided by IP networks, offering scope for Traffic Engineering, guaranteed QoS and Virtual Private Networks (VPNs). MPLS uses a technique known as label switching to forward data through the network. A small, so called MPLS shim label is inserted between the data link layer and the network layer upon entry in the MPLS network. At each hop across the network, the packet is routed based on the value of the incoming interface and label, and sent to an outgoing link with possible a new label value. The path that the packets follow through a network is defined by the swapping of label values. This is really nothing new, the same thing is done in ATM networks and there it is called VPI/VCI. In the traditional TDM network it is called time slot switching. Since the mapping between labels is constant at each LSR, the path is determined by the initial label value. Such a path is called a Label Switched Path (LSP). At the ingress to an MPLS network, each packet is examined to determine which LSP it should use and hence what label to assign to it. This decision is a local matter but is likely to be based on factors including the destination address, the quality of service requirements and the current state of the network. This flexibility is one of the key elements that make MPLS so useful.

The premise of Generalized MPLS is that the idea of a label can be generalized to be anything that is sufficient to identify a traffic flow. For example, in an optical fiber using Dense Wave length Division Multiplexing, DWDM, one complete wavelength could be allocated to a requested flow. The LSRs at either end of the fiber simply have to agree on which wavelength to use. Unlike with non-generalized labels, the data inside the requested flow does not need to be marked at all with a label value; instead, the label value is implicit in the fact that the data is being transported within the agreed wavelength. On the other hand, some representation of the label value is needed in the signaling protocol so that control messages between the LSRs can agree on the value to use.
Generalized MPLS extends the representation of a label from a single 32 bit number to an arbitrary length byte array and introduces the Generalized Label object (in RSVP) and Generalized Label TLV (in CR-LDP) to carry both the label itself and related information. In the talk I will describe in more detail how the switching quantities used in optical networks are represented as GMPLS labels. Additionally I will present how the different QoS classes can be mapped onto the LSPs both in the MPLS and the GMPLS environment. It will be argued that MPLS is a crucial component of many modern long-haul high-speed networks. The long haul carriers are now looking to the next generation of Generalized MPLS services as the next step in delivering high quality, high-speed and reliable network services.

References:

1. White paper from Data Connection: MPLS Virtual Private Networks. A review of the implementation options for MPLS VPNs including the ongoing standardization work in the IETF MPLS Working Group
2. RFC 3031 Multiprotocol Label Switching Architecture
3. Draft-ietf-mpls-generalized-signaling Generalized MPLS. Signaling Functional Description
4. Draft-ietf-mpls-generalized-rsvp-te Generalized MPLS Signaling. RSVP-TE Extensions
5. Draft-ietf-mpls-generalized-cr-ldp Generalized MPLS Signaling. CR-LDP Extensions
6. Draft-ietf-ccamp-lmp Link Management Protocol (LMP)
7. Draft-ietf-mpls-cr-ldp Constraint-Based LSP Setup using LDP RFC 3036 LDP Specification
8. RFC 2205 Resource ReSerVation Protocol (RSVP). Version 1 Functional Specification
9. RFC 2961 RSVP Refresh Overhead Reduction Extensions

Hybrid Modeling of Communication Networks Using Modelica

Daniel Färnqvist*, Katrin Strandemar*, Karl Henrik Johansson*, and João Pedro Hespanha†

* Dept. of Signals, Sensors & Systems, Royal Institute of Technology, Sweden, kallej@s3.kth.se
† Dept. of Electrical & Computer Engineering, University of California, Santa Barbara, USA

Abstract

Modeling and simulation of communication networks using the modeling language Modelica is discussed. Congestion control in packet-switched networks, such as the Internet, is today mainly analyzed through time-consuming simulations of individual packets. We show, by developing a model library based on a recent hybrid systems model, that Modelica provides an efficient platform for the analysis of communication networks. As an example, a comparison between the two congestion control protocols is presented.

1 Introduction

The objective with research on control of networks is often to improve traffic throughput and to better accommodate different service demands. Communication networks experience major problems due to traffic congestion. Today's congestion control is in most networks implemented as end-to-end protocols [1, 2, 3]. The protocols have proved to form the basis of a remarkably robust and scalable system, though the understanding of the basic principles of these complex systems are far from satisfactory [4, 5, 6].

The intention of this paper is to describe initial work on modeling packet-switched communication network using Modelica [7, 8], which is an object-oriented language for modeling physical systems. The standard modeling and simulation environment targeted at network research is the discrete-event simulator ns-2 [9]. Since ns-2 directly implements the Internet protocols and simulates individual packets, it provides on one hand accurate simulation results but on the other hand a rather slow simulation speed. The result of this is that ns-2 is mainly for studying relatively small networks over a short time scale. The other extreme is to use flow models, i.e., to approximate packet transmission with a continuous flow and basically neglect the network protocols. A hybrid systems model, which is based on the average rates but takes packet drops and rate adjustments due to congestion control into account, was recently proposed in the literature [10]. The motivation for this model is to capture the network behavior on a time scale in between packet models and flow models. Studies have shown that the hybrid model is able to model many important network phenomena [10, 11]. In this paper, we show that the hybrid model is suitable for Modelica. Moreover, we show that the model can then be efficiently simulated using Dymola [12], which is a commercial Modelica simulation environment.

The outline of the paper is as follows. In Section 2, a brief introduction to congestion control in communication networks is presented. The hybrid model is described in Section 3 and its Modelica implementation is discussed in Section 4. An example, where two TCP versions are compared for a small wireless network, is given in Section 5. The paper ends in Section 6 with a small discussion on ongoing work.

2 Communication Networks

A packet-switched network can be described by a directed graph, see Figure 1. The nodes represent the routers, which direct the packets from sender to receiver, and the edges correspond to wired or wireless links. The bandwidths of the links are limited, so each router has a buffer where packets are stored if more packets are entering the router than are leaving. In this way, it is possible to deal with minor traffic congestion in the network. If too many packets enter a router in a short amount of time, however, packets will be dropped due to that the router buffer has finite size. The way this congestion problem is handled by the senders on the Internet is through a control mechanism denoted transmission control protocol (TCP). The receiver sends acknowledgments back to the sender, when packets have arrived. In order to efficiently use the network resources, the TCP sender adjusts its sending rate according to a control variable called the window size w. The TCP sender sends w number of packets and waits for their acknowledgments to return. Hence, w corresponds to the number of unacknowledged packets the sender may have in the network. When the sender has received acknowledgments for all w packets, the window size is increased by one: $w := w + 1$. If a packet is dropped (so that no acknowledgment for that packet is received), w is decreased by a factor two. Hence, TCP uses additive increase and multiplicative decrease (AIMD) to regulate the congestion window size based on explicit acknowledgments and implicit negative acknowledgments. Although, the AIMD control strategy

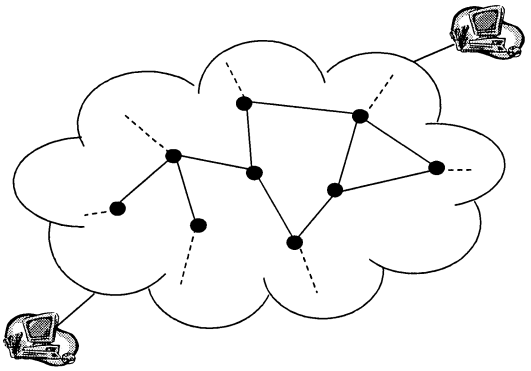


Figure 1: Communication network.

has proved to be efficient, robust and remarkably scalable for the Internet, it is believed that it might be too abrupt for emerging applications such as streaming of audio and video.

3 Hybrid Model

In the hybrid model for communication networks proposed by Hespanha et al. [10, 11], the network dynamics and the TCP dynamics are modeled as hybrid systems (e.g., [13, 14, 15]). A hybrid system is a mathematical model for a dynamic system that has both continuous-time and discrete-event dynamics. The traffic dynamics of a packet-switched communication network is suitable to model as a hybrid system, where the packet flows are approximated by continuous dynamics while packet drops and control protocols are modeled as discrete changes. The motivation for introducing a hybrid model for communication networks is that it captures the behavior of the system on a time scale in between conventional packet models and flow models. Thanks to the model abstraction of packets, the simulation time is in principle not affected by the number of packets transmitted or the queue sizes. Still, the hybrid model shows enough detail to accurately model transmission protocols and packet drops.

3.1 Network Dynamics

Packet transmission rates and queue lengths are continuous-time variables in the hybrid model. The received rate of packets at a router is denoted r and the sent rate is denoted s. The number of packets stored in a router queue is denoted q. The dynamics of the queue is depending on if the queue is full ($q = q_{\max}$) or not ($0 \leq q < q_{\max}$). For each router, we introduce the hybrid system with two discrete states shown in Figure 2. In this model, subscript f refers to flow f, the variables q and s are defined as $q = \sum_f q_f$ and $s = \sum_f s_f$, and the bandwidth of the outgoing link is equal to a constant B. Note that when the queue is full, a drop will be generated as soon as the variable z is equal

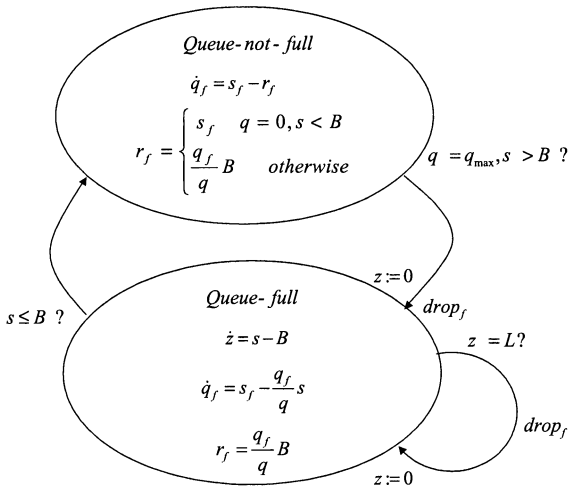


Figure 2: Hybrid queue model.

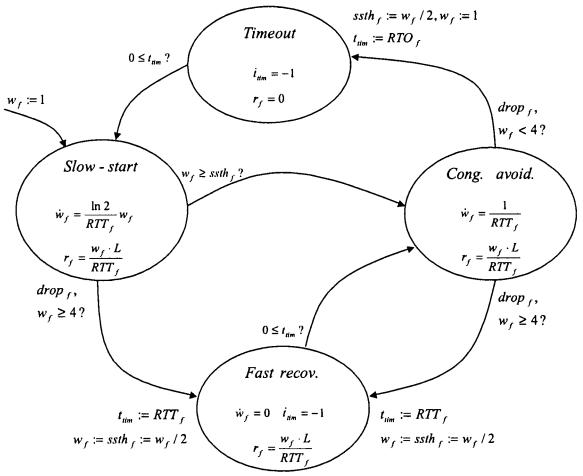


Figure 3: Hybrid TCP model with state transitions.

to the predefined packet size L. Which flow f of the incoming flows that will lose a packet is determined by the distribution of the flows in the queue.

3.2 TCP Dynamics

For TCP, the sending rate r and window size w are conveniently modeled as continuous variables, while the various modes of TCP (such as slow-start, congestion avoidance, and fast recovery) represent discrete states. The hybrid TCP model consists of four discrete states as shown in Figure 3. Here the round-trip time RTT_f for flow f is the time between sending a packet and receiving the corresponding acknowledgment. It is given by the sum of the physical propagation time and the queueing times. The hybrid TCP model uses an estimate of RTT_f to update the window size w_f and to derive the sending rate r_f, which also depend on the discrete state of the TCP.

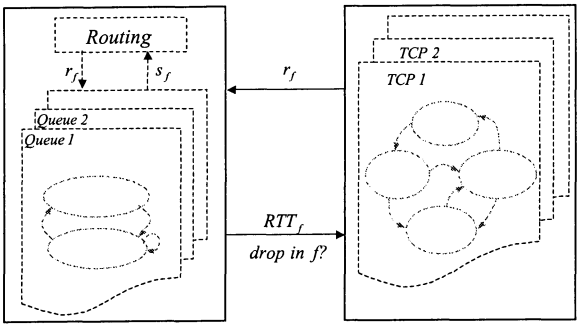


Figure 4: Composition of hybrid communication network model. The network dynamics and the TCP dynamics are separated.

4 Modelica Implementation

Modelica is an object-oriented modeling language designed mainly for large heterogeneous physical systems. It is suitable for the hybrid network model described in previous section. For example, the continuous dynamics of the TCP model in Figure 3 is implemented through the following code:

```
if SlowStart then
   der(w)=log(2)*w/RTT;
   r=w*L/RTT;
elseif CongAvoid then
   der(w)=1/RTT;
   r=w*L/RTT;
elseif FastRecov then
   der(w)=0;
   r=w*L/RTT;
elseif TimeOut then
   der(w)=0;
   r=0;
end if;
```

Discrete events, such as the packet drops in the network model, can be implemented in the Modelica language as the following example:

```
when drop then
   reinit(w,w/2);
end when;
```

A communication network library was developed in Modelica. The library contains standard building blocks for network simulation, such as TCP senders, routing tables, and queues. Figure 4 shows the schematic layout of our communication network model. Note that the network dynamics and the TCP controllers are separated. The only information shared between the two submodels are the sending rates r_f, the round-trip times RTT_f, and the drop events. The modular structure allows an easy testing of for instance different TCP controllers applied to the same network topology. Note that only the queue and the

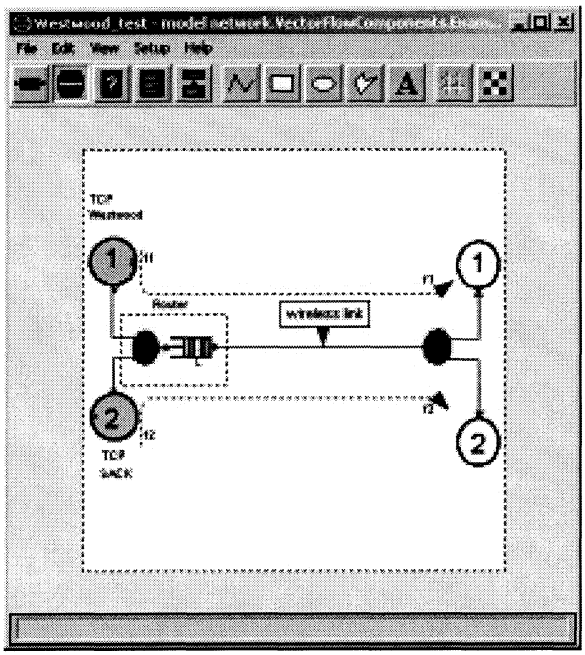


Figure 5: Example of a simple communication network implemented in Modelica.

TCP components contain dynamical equations. The router component implement the network topology.

Appropriate handling of the switching between discrete states is important for accurate and efficient simulation of hybrid systems [16, 17]. Implementations in Simulink showed some problems in this respect. Our implementation in Modelica and simulation in Dymola works well. Note that the simulation time is in general not depending on the queue sizes or sending rates, but instead depends on the number of discrete events generated by packet drops. The simulation time grows considerably when the number of discrete events becomes large. In ns-2, where individual packets are simulated, the simulation time depends on the size of the packet flows, the number of flows, and the queue sizes.

5 Example

An example of a communication network implemented in Modelica is shown in Figure 5, where two senders and two receivers are connected to the network. Let us simulate this simple network. The two flows are sharing the same link capacity. If the sum of the flows at some time instance is larger than the bandwidth of the link, packets will be queued. Sender 1 sends Flow f_1 using a version of TCP called TCP Westwood (TCPW) [18] and Sender 2 sends Flow f_2 using TCP SACK [10]. For a wireless link transmission losses are more likely than for a wired link. For the network in Figure 5, there can be packet losses either due to that the router queue is full or due to that the

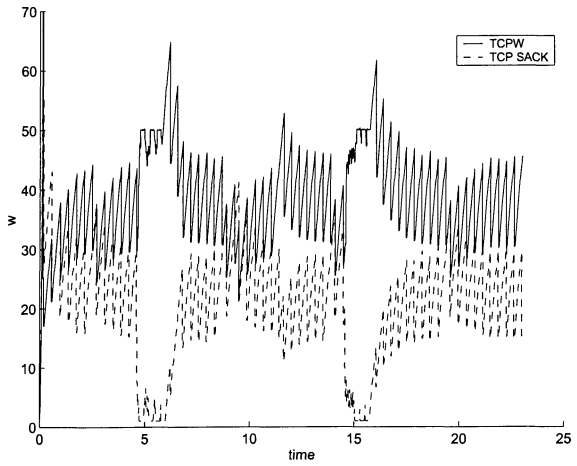


Figure 6: Simulation of the system in Figure 5. The windows size for TCP SACK is considerably reduced when the wireless link is losing a lot of packets (at about $t = 5$ and $t = 15$), while TCPW is able to cope with the losses very well.

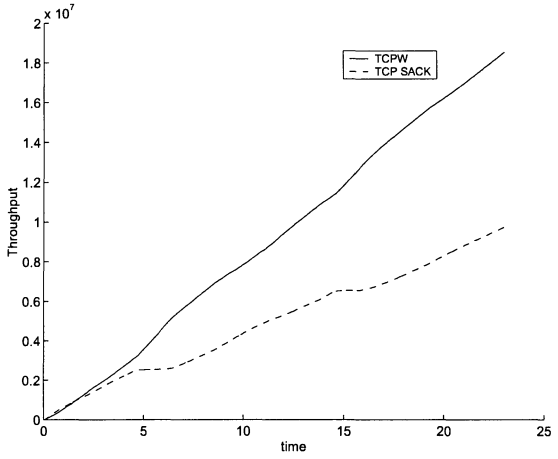


Figure 7: TCPW gives a larger throughput than TCP SACK. The difference is emphasized when there are a large number of packet losses due to the wireless transmission.

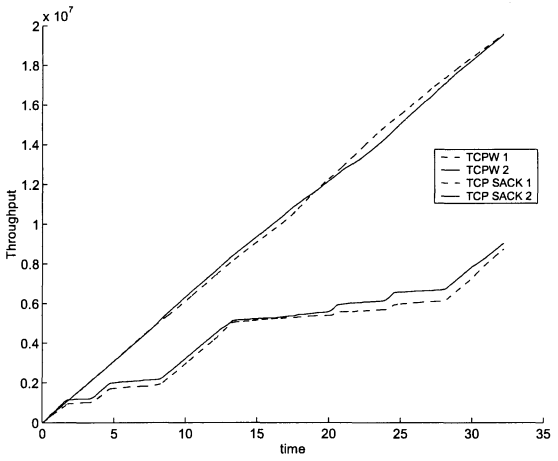


Figure 8: Throughput for two TCP Westwood senders (upper plots) and two TCP SACK senders (lower plots).

wireless transmission loses packets. We model the wireless link as having a good and a bad state. In the good state 0.1% of the transmitted packets are lost, while in the bad state 10% of the packets are lost. The link stays in each state a random amount of time, which is exponentially distributed. TCPW was designed taking wireless links into account, while TCP SACK was designed for wired links. This is illustrated next.

Figure 6 shows the window sizes for a simulation of the network in Figure 5. The solid line corresponds to TCPW and the dashed line to TCP SACK. Note the time intervals at about $t = 5$ and $t = 15$ when the link is in the bad state. The packet losses due to the bad transmission result in a sudden decrease of the window size for TCP SACK, while TCPW are able to compensate for the packet losses of the wireless link. The window size for TCPW is in general larger than for TCP SACK. This gives a larger throughput for the connection using TCPW, as is shown in Figure 7.

Figure 8 shows the throughput when two TCPW's are sharing the same wireless link (upper plots) and when two TCP SACK's are sharing the same link (lower plots). From the simulations we see that the major advantage of TCPW is when the link is in the bad state. When the link is in the good state, the performance of both TCP implementations are roughly equal.

Since TCP (SACK) was developed for wired networks, where packet losses arise only due to buffer overflow, there is a problem using it over wireless links. The reason is that the window size is reduced by a factor two every time a drop occurs, regardless if the drop is due to congestion or to transmission loss. TCPW has another way of updating the window size when a drop occurs. The window size is set to a value based on an estimate of the available bandwidth and the current round-trip time RTT. Since RTT is highly dependent on the queue sizes, so that RTT is small if and only if all corresponding queues are small, a small RTT implies that a detected drop must be due to wireless loss. The aim of TCPW is hence to utilize the available bandwidth more efficiently.

6 Discussion

Ongoing work includes the comparison of various TCP variants, such as Reno, SACK, and Westwood, as well as the proposal of improved congestion control mechanisms. The modeling environment that we have developed based on Modelica provides a flexible modular platform for simulating various networks and protocols. Note, however,

that the hybrid model is not only suitable for efficient simulation, but it is also possible to analyze the model using mathematical tools from hybrid systems theory.

Acknowledgment

The authors want to thank Håkan Hjalmarsson and Gunnar Karlsson for helpful discussions.

References

[1] V. Jacobson. Congestion avoidance and control. In *Proc. of SIGCOMM*, pages 314–329, 1988.

[2] J. Walrand and P. Varaiya. *High-performance communication networks*. Morgan Kaufmann, 2nd edition, 2000.

[3] L. L. Peterson and B. S. Davie. *Computer networks: a systems approach*. Morgan Kaufmann, 2nd edition, 2000.

[4] V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[5] V. Paxson and S. Floyd. Why we don't know how to simulate the internet. In *Proc. of the Winter Simulation Conference*, 1997.

[6] F. Paganini, J. Doyle, and S. Low. Scalable laws for stable network congestion control. In *IEEE CDC*, Orlando, FL, 2001.

[7] *Modelica*. `http://www.modelica.org`.

[8] S. E. Mattsson, H. Elmqvist, and M. Otter. Physical system modeling with modelica. *Control Engineering Practice*, 6:501–510, 1998.

[9] *The Network Simulator ns-2*. Information Sciences Institute, University of Southern California. `http://www.isi.edu/nsnam/ns`.

[10] J. P. Hespanha, S. Bohacek, K. Obraczka, and J. Lee. Hybrid modeling of TCP congestion control. In M. Di Benedetto and A. Sangiovanni-Vincentelli, editors, *Hybrid Systems: Computation and Control*, volume 2034 of *Lecture Notes in Computer Science*, pages 291–304. Springer-Verlag, Berlin, Germany, 2001.

[11] S. Bohacek, J. P. Hespanha, J. Lee, and K. Obraczka. Analysis of a TCP hybrid model. In *Proc. of the 39th Annual Allerton Conference on Communication, Control, and Computing*, 2001.

[12] H. Elmqvist, D. Brück, and M. Otter. *Dymola—User's Manual*. Dynasim AB, Research Park Ideon, Lund, Sweden. `http://www.dymola.com`.

[13] H. S. Witsenhausen. A class of hybrid-state continuous time dynamic systems. *IEEE Trans. Automatic Control*, 11(2):161–167, 1966.

[14] R. W. Brockett. Hybrid models for motion control systems. In H. Trentelman and J. Willems, editors, *Essays in Control: Perspectives in the Theory and Its Applications*, pages 29–53. Birkhäuser, Boston, 1993.

[15] A. van der Schaft and J. M. Schumacher. *An Introduction to Hybrid Dynamical Systems*. Number 251 in Lecture Notes in Control and Information Sciences. Springer-Verlag, London, 2000.

[16] K. H. Johansson, A. Rantzer, and K. J. Åström. Fast switches in relay feedback systems. *Automatica*, 35(4):539–552, April 1999.

[17] K. H. Johansson, M. Egerstedt, J. Lygeros, and S. Sastry. On the regularization of Zeno hybrid automata. *System & Control Letters*, 38:141–150, 1999.

[18] S. Mascolo, C. Casetti, M. Gerla, M. Y. Sanadidi, and R. Wang. TCP Westwood: bandwidth estimation for enhanced transport over wireless links. In *MobiCom*, Rome, Italy, 2001.

Education in Internet Techniques

A project financed by the Knowledge foundation

Arne A Nilsson[1], Lars Magnus Lahne[2],

[1]Blekinge Institute of Technology, 371 79 Karlskrona, Sweden. E-mail: ano@its.bth.se
[2]Högskolan på Gotland, 621 67 Visby, Sweden. E-mail: lml@hgo.se

Abstract

The competence to develop, build and operate large and complex digital networks is a scarce resource in Sweden.

The aim of the project is to establish education that uses limited resources in the form of qualified teachers in an effective way and at same time is available everywhere.

One important objective is to develop a model for co-operation between universities, and between higher education and the business world. Other important objectives are to offer education, which is adapted to the needs of both the academic and the commercial world, and at the same time, will attract new groups of young students.

The main idea is to create a network of institutions for higher education that will develop a virtual education in Internet techniques. This education consists of a number of small modules. The total educational package can be described as a collection of courses/units that can be combined in different ways.

1 Background

The competence to develop, build and operate big and complex digital networks is a scarce resource in Sweden. A growing interest for the use of IP-based networks (Internet) within business and society as a whole has led to a lack of skilled personnel.

Today almost everybody agrees that Sweden should invest considerable resources to build a digital infrastructure that covers the whole country. To make this possible it is essential to educate people and give them the right qualifications to develop, build and run this infrastructure.

There is a lack of teachers who can manage this kind of education. The problem to recruit enough qualified teacher constitutes a serious obstacle in the struggle to supply all interested parties with the competence necessary to build, run and support large and complex digital networks. Another problem is the difficulty to recruit students to courses in technical and natural science.

2 Goals

The purpose of this project is to present recommendations that take their staring-point from the goal to:

- create a model for co-operation between higher education and industry with the purpose to develop and carry through advanced education.
- use available resources in the form of teachers in an effective way.
- help to make it possible for teachers at different levels to increase their competence.
- offer education which is available everywhere in the country.
- attract new groups of young students.
- use the possibilities that are offered by the new information techniques.

3 A virtual education

The main idea is to create a network of institutions for higher education that will develop a virtual education in networking. This education consists of a number of small units. The total educational package can be described as a collection of courses/units that can be combined in different ways as undergraduate studies, further education or as part of different types of courses designed for trade and industry.

Target groups are:

- young students in the "gymnasium" or who just finished their time of study at the "gymnasium".
- students attending undergraduates studies
- people from trade and industry who want to keep up to date and/or enter deeply into different aspects of networking.
- people who want to obtain expert competence in networking.

The different courses are developed by the institution/institutions that have access to necessary competence. One should not forget the possibility that some units could be developed by or in co-operation with

industry that often has access to expertise and other scarce resources. One example is the courses, which have been developed by Cisco and today are used by several universities, colleges and other institutions in Sweden and all over the world.

The courses/units can be grouped in different levels from basic courses, which belong in the borderland between the gymnasium and undergraduate studies to courses at master level. The large technical universities probably have means to develop courses at all levels within their fields of specialization. On the other hand, the small university colleges as for example Gotland concentrate on courses at a lower level.

In industry there is a need for short or very short courses, which convey summaries within different fields of interest. There is consequently reason to add a complementary offer of small units that have the range between 2 hours and two days. This type of short courses is commercially very interesting and can provide the project and involved parties' with fresh resources for development.

Thanks to modern information technology the courses will be available for all members of the network. As a consequence one can offer advanced courses in more than one or two places in the country.

The aim is that one will be able to combine the courses/units in different ways and use them in different situations as undergraduate studies, further education or as part of courses designed for trade and industry. An important goal is to find ways to attract youth to studies in Internet techniques. The "gymnasiet" and other institutions outside higher education can for that reason be important consumers for some of the courses.

Arrangements like this demand that the relation between different courses and the students' previous knowledge is carefully specified. This is especially important when recruiting students in connection with further education and custom-made courses for trade and industry.

The industry mostly wants shorter courses than what is usually the case in higher education. It is normally difficult to reserve several weeks' work in order to take part in a course. Against this background one should divide each course in several units that can be used separately.

4. Research

Close connections with the frontline of research are a necessary condition for the courses in question to be able to meet with the demands from surrounding society. The network shall not administer research, but will obviously have a strong interest that the research institutions can recruit qualified researchers. A possible solution is that the network and the participating institutions make a joint effort to approach trade and industry in order to obtain financial support which will make it possible to expand research.

5. Examination

From a practical point of view, it is good if the institution arranging a course also gives the exam. This does not mean that the institution that developed the course not should have a considerable influence on required standards and how the examinations are carried out.

If the institution that arranges a course is not formally entitled to conduct the examination, the problem can be solved in two different ways. One way is to make an agreement with an institution that has those rights. Another solution is that the students are registered at an institution with necessary rights.

6 Co-operation with trade and industry

A strengthened supply of competence in internet technique is of great interest for trade and industry. Everything indicates that the project must co-operate with private companies. Some examples on fields suitable for co-operation are:

- a survey of the need for further education.
- development of courses that meet with the demands from trade and industry.
- recruitment of teachers and researchers.
- financing of necessary development.
- financing and recruitment of international expertise.

There are also reasons to look into the possibilities to co-operate with a commercial partner in order to expand the market for courses in Internet techniques. The most obvious solution is to use the university systems own organizations for co-operation with the surrounding society as a channel to sell courses.

7 The project

The project is financed by the Knowledge Foundation and administrated from Gotland University College. The goals for the project overlap to great extent those of the Swedish Center for Internet Technique (SCINT). The united efforts are coordinated by a working team consisting of Professor Arne A Nilsson from Blekinge Institute of Technology, Bengt Källbäck, from SCINT and Lars Magnus Lahne from Gotland University College.

Issues on Performance Measurements of TCP

Frida Gunnarsson, Fredrik Gunnarsson, Fredrik Gustafsson

Dept. of EE, Linköpings universitet, SE-582 82 Linköping.
E-mail: {frida,fred,fredrik}@isy.liu.se.

Abstract

Due to the rapid growth of the Internet and the use of packet data traffic, the communication world has to find ways of improving performance of this traffic. New areas of use include, e.g., radio transmissions and real time data. This work discusses some issues of performance when using TCP and also present some of the measures that are used today. Furthermore we discuss this matter from the perspective of a bottleneck queue and present a measure that can be used from that viewpoint in the network. This measure is illustrated by comparing two versions of TCP and form the basis for a discussion of what models will be needed to actually achieve the goals that we have. A control structure is presented and using simulations and a control interpretation we note that RED introduces an implicit wanted steady state value of the queue length. We use the dropping of packets as a control signal and the output signal is the queue length.

1 Introduction

When new application areas for packet data traffic emerge, new difficulties also appear. This has produced a lot of results on how to improve performance when communicating Internet traffic in the new frameworks. E.g., the increased use of mobile services implies that packet data must work well over a radio channel as well as for a wired link. One question that remains is how the performance measure should be chosen. What do we really want to achieve and how can we measure it?

The aim of this work is to illuminate the difficulties with measuring performance and to develop an easy-to-use methodology. We will focus on the bottleneck queue, as this gives a good overview of the network and for some scenarios it is fairly easy to find, e.g., the node between the wired and the wireless network when using packet data over radio. The bottleneck queue measure was introduced in [1]. This will also lead to a need for a mathematical description and ways to calculate performance improvements. The current work will show the new control perspective that dropping of packets gives us and also use simulations to reveal certain properties of Random Early Detection, RED, [2].

In this paper we will illustrate our discussion by comparing two TCP versions, namely TCP Reno and TCP Vegas, in Sec. 5. Before that we discuss what the considerations are, Sec. 2, and show some measurements that are used in current research, Sec. 3. We also show our way of measuring performance in Sec. 4 and finally, Sec. 6, discuss what kind of models that are needed to use this measure in order to improve performance. Furthermore we present a control theory interpretation that will be used as the basis for the final mathematical model and also used to give a intuitive interpretation of the system behavior.

2 Performance Issues

To compare the two TCP versions, we need to discuss what we want to achieve and what is good performance. Good performance means different things depending on who we are asking, e.g., the network provider might have a different measure than the user. How can we find a good common way of measuring performance? First let us discuss some of the different issues we can encounter.

For each user, transmission time and reliability are important. Considering TCP connections this means time, since TCP makes sure that everything goes through no matter what it takes. One issue is thus the time to transmit a certain number of bytes, measured on a per flow basis.

The network provider wants to fully utilize the available capacity and divide it among the users. This gives us the requirement that there should always be packets processed by the slowest node. The issue of dividing the resources in an efficient way is not so easy to resolve. Efficiency can be measured from both the overall network and from the single user and optimizing the efficiency of the network might not be fair among the users and might cause users to change network provider. The two extreme cases are "total fairness" and "best effort" and the matter has been discussed in various scheduling contexts. We will not discuss this matter more here.

These two actors, the user and the network provider, demand things that can very well counteract each other. E.g., if one user gets access to a certain amount of the bandwidth all the time he will be very happy because he does not have to compete with anyone, but the network will not be used very efficiently, since it is unlikely that the user is going to send at full speed day and night. On the other hand, the network provider would like to average

the use of the network over time, which could mean that users would experience large delays, since most traffic occurs at daytime and some of it would be delayed and even sent during the night.

We will now discuss the matter of measuring performance on a deeper level, taking a look at single routers and senders in the network.

3 Measurements Used in Current Research

With all the research efforts put into the area of improving TCP, there are of course a lot of performance evaluations going on. We will here briefly describe some of the measures that are used in different contexts.

References [3, 4, 5]
use the time to transmit a certain number of bytes or the packet bit rate to evaluate mobile Internet access, to analyze TCP when used in GPRS and to analyze file transfers over UMTS, respectively.

Reference [6]
develops a model of the distribution of the queue length and the congestion window and uses these to investigate TCP's performance.

References [7, 8]
consider throughput (bytes/sec) and Jain's fairness index, [9], as performance measures to analyze the improvements of TCP when using TCP Vegas instead of TCP Reno and to evaluate the improvements achieved by a new version of TCP, respectively.

Reference [10]
models RED queue behavior using a probabilistic view and uses the model to calculate RED parameters that fulfill stability and other control criteria.

Considering the discussion in Sec. 2 we will continue with a slightly different view of performance measurements.

4 TCP Performance Using the Bottleneck Queue

We will take on the eyes of a network provider who wants satisfied customers. We will focus on a specific part of the network, namely the bottleneck queue. There can of course be many bottlenecks in one network, but only one per flow. The bottleneck of a flow is the queue that is experienced as the slowest one and consequently the only queue where that flow has more than one packet at a time. When investigating the performance from the bottleneck, we can look at, e.g.,

The time the queue is unused.
This shows how efficiently the network is used. An idle queue is waste of capacity.

The fluctuations of the queue length.
This shows the ability to find a steady state of the network. The more stable the queue length is, the more stable is also the round trip time experienced by each connection. This will also make it easier for new connections if the total network is experienced as being in steady state, and thus scalability is improved.

The fraction of dropped packets.
This shows how fast the algorithms notice changes of the network. A larger number of dropped packets indicates that the algorithm takes longer time to respond to overflows.

The throughput.
This indicates if the queue is often empty. It is also interesting for, e.g., the owner of this bottleneck queue if he is getting payed per delivered byte or packet.

The average queue length
A long average queue means large round trip delays and a high probability to drop packets, while a short average queue means high probability for a unused queue.

Goals: The conclusion is that we want to control the tail of the queue distribution to keep it small and we also want to make the distribution narrow around a not too high and not too low mean value.

5 Comparison of two TCP versions

To give a feeling of the distribution measure, we will compare TCP Reno and TCP Vegas by looking at both queue length histograms and the queue length over time.

5.1 Reno vs. Vegas

We will here briefly describe the differences in the two TCP versions. The main difference lies in the estimation of the available network capacity. Reno is designed to increase the estimate whenever a packet goes through and to decrease it on timeouts or drops. In Vegas a measure has been introduced and the estimate is increased if this measure is above some threshold and decreased if it is below another threshold. Decreasing and increasing are also much more aggressive in Reno than in Vegas. The purpose of this is that Vegas should discover difficulties before they result in a drop and respond accordingly. Reno will not discover problems with congestion until a packet (or more) is dropped.

5.2 Simulations

We have simulated different scenarios using *ns-2* [11], see Fig. 3. The network in the simulation is a number of flows sharing a bottleneck queue and they carry either FTP or HTTP traffic. Fig. 1 shows the simulated network and Tab. 1 shows the used parameter values[1]. We do not aim to decide which of the two versions are the better one but only to discuss the matter of measuring performance using the bottleneck queue characteristics. We have simulated 36 different scenarios and we will look at a few of them in this context.

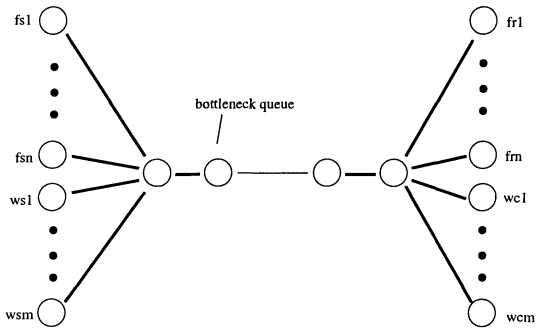


Figure 1: The simulation setup in *ns-2*. n file senders, fsi, with corresponding file receivers, fri, and m web servers, wsi, with corresponding web clients, wci.

Table 1: Parameter settings for the simulations in *ns-2*.

Description	**Setting**
Fixed	
Number of flows, $n + m$	27
Simulation length, T [s]	100.0
Start and end times of FTP flows	uniformly distributed over $[0, aT]$ and $[(1 \quad a)T, T]$, respectively
Changing	
TCP version	Vegas or Reno
Number of web traffic flows, m	0, 9 or 18
a	0.1, 0.3 or 0.5
Bottleneck queue, q	20 or 50 packets

Figure 2 shows the connection between the queue length over time and its histogram. The queue length oscillates around virtually three different levels and each level gives rise to a bump in the histogram. Here the histogram is rotated to indicate the shared axis of the two diagrams. We can view the histogram as the projection of the time plot on the y-axis.

[1]The different setups will from now on be labeled with (ver, m, a, q), e.g., $(R, 9, 0.3, 20)$ means Reno with 9 web flows, start times during the first 30% of the simulation and a queue length of 20 packets.

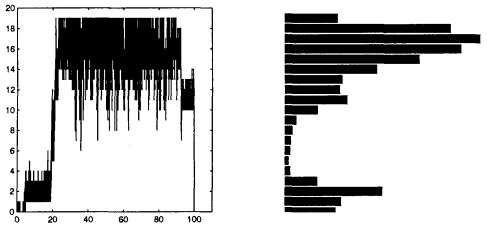


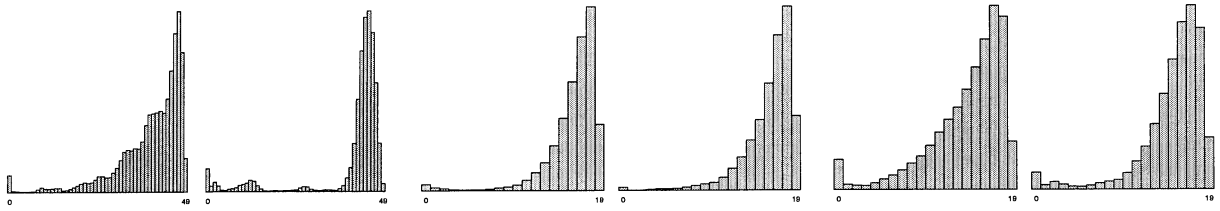
Figure 2: The time profile of the queue length with corresponding histogram. The setting is [V,9,0.5,20]. The shared axis shows the number of packets in the queue. The histogram shows intensity to the right and the left plot has time on the x-axis.

In Figs. 3(a) and 3(b) we see a significant difference in the width of the peaks. Reno has given rise to more oscillations. The peak in the Reno histogram seems to continue beyond the queue length, while the peak in the Vegas histogram is centered somewhat to the left of the maximum queue length. Vegas has thus, in this scenario, been more successful with accomplishing the goals from Sec. 4.

The histograms in Figs. 3(c) and 3(d) are very similar and no significant difference can be found. This setting is also very hard, with 27 file flows sharing a link with the queue length being only 20 packets. In Figs. 3(e) and 3(f) we have the same setting as previously but only 9 file flows and 18 disturbing web transfers and we see that Vegas handles this better with a slightly more narrow peak. Here the center of the peak seems to be more to the left than for the Reno case. We see that it is quite easy in these cases to see what the histograms mean. Of course in reality the number of flows through the link will vary and it will be hard to see where the peak of the distribution will be when looking only at a small sample. We will now discuss the use of the queue occupation histogram measure.

6 Using the measurements

The aim of all this is to reach a mathematical description where we can use common knowledge to tune the performance of a network. As we have concentrated on the bottleneck queue, this is also where we want to find the mathematical representation. Looking at the goals in Sec. 4 we see that they are easily translated into common tasks of a controller. In control theory the specifications for a strategy are often made in a so called step response, where you study rise time and overshoot when the system is subject to an abrupt change. These two correspond to how fast the system reacts to changes and how much it oscillates, respectively. Translating into the histogram framework these will both correspond to the width of the peak of the distribution and the mean value would correspond to a wanted



(a) $(R, 18, 0.1, 50)$ (b) $(V, 18, 0.1, 50)$ (c) $(R, 0, 0.1, 20)$ (d) $(V, 0, 0.1, 20)$ (e) $(R, 18, 0.1, 20)$ (f) $(V, 18, 0.1, 20)$

Figure 3: Results from three different simulation setups for using Reno or Vegas.

steady state value for the queue length, the so called reference value of our control loop. Thus we would like to find a mathematical description of the system that fits into the control theory framework.

One could ask what the controller would be. When we look at the system from the bottleneck queue one answer could be to drop packets as has been proposed in, e.g., papers about RED and ECN [2, 12]. If we keep that thought we can find a control structure to work with. We realize that it is the congestion window of the TCP senders that implicitly decides the size of the queue and that the windows are changed when packets are dropped. There is also a delay in the system, which will be related to the experienced round trip times. The model will of course be highly non-linear and also dependent on the number of flows using the bottleneck queue.

Our aim is to use the probabilistic view introduced in [10] to find a time discrete model with the dropping of packets as a control signal and the queue length as an output, see the upper part of Fig. 4. The model presented in [10] gives us a way of studying approximate behavior of the queues. We will now look at a simple simulation example. The discrete time version of the proposed model is shown in (1).

$$\begin{array}{ll} W_{k+1} & W_k = \dfrac{1}{R_k} + \dfrac{W_k W_{(k\ R_k)}}{2R_{(k\ R_k)}} p(q_{(k\ R_k)}) \\ q_{k+1} & q_k = C + N_k \dfrac{W_k}{R_k} \quad p(q_k) \end{array} \tag{1}$$

W is the average congestion window, q is the queue length, R is the average round trip time, N is the number of connections, p is the probability of dropping a packet, C is the service rate of the queue and k is the time index. In Fig. 4 $u_k = p(q_k)$ and the delay is R_k. The model only takes into account the congestion avoidance phase and does not consider timeouts. For our ordinary FIFO queue $p(q)$ is a step at $q = q_{max}$, and for a queue deploying RED

$$p(q) = \begin{cases} 0 & q < q_m \\ 1 & q > q_M \\ p_M \frac{q\ q_m}{q_M\ q_m} & \text{otherwise} \end{cases}$$

for some parameters q_m, q_M and $p_M \in [0, 1]$. This means that we drop some packets before the queue is actually full.

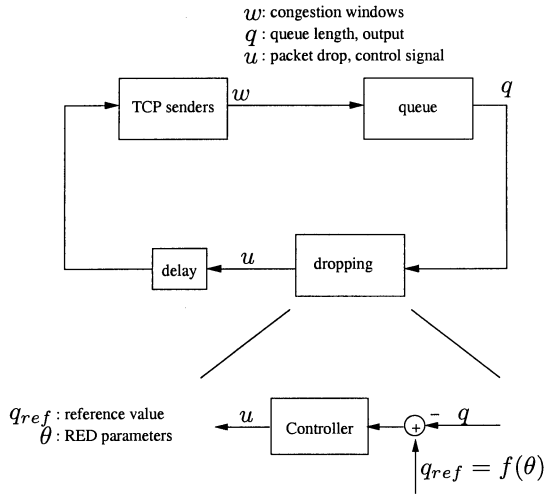


Figure 4: A control model of the bottleneck queue, with the interpretation of RED depicted in the lower part.

Our simulations using the model (1) shows that, for a family of loads, RED will correspond to setting a reference value for the queue length. Let us call the RED parameters $\theta = [\ q_m \quad q_M \quad p_M\ ]$. We assume now that we can find RED parameters that gives us good performance for a certain network setting. Then there is a set

$$L(\theta) = \{N | \theta \text{ gives good performance}\}.$$

In $L(\theta)$ the reference value that the RED queue imposes will be constant. Thus we can vary the load in the network and the steady state value imposed by RED will be the same if RED manages to give good performance at all. This is seen in simulation plots of the queue histograms. Fig. 5 shows two results for two subsets of L. We see that the mean values are nearly the same (the difference is not statistically certain). Thus, RED introduces a controller

with an implicit reference value depending on the RED parameters but not on the load. However the parameters must be adapted to large variations in the load.

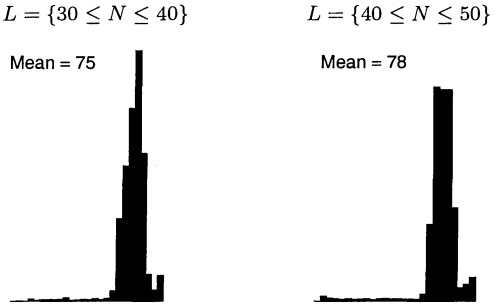


Figure 5: The queue histogram using RED for two subsets of L. The mean are more or less the same.

7 Conclusions

We have discussed the matter of measuring performance of a network and have presented the queue occupation as an appropriate measure. We showed that the aim is to find a system description that would fit in a control theoretic framework and we have in this context presented the goals that we want to achieve:

- Stable system,
- Small overshoot,
- Fast step response and
- Small tails of the queue distribution.

Furthermore we have discussed what models that will be needed to achieve the goals of good performance. We showed a control structure of the system and a nonlinear model. This control structure gives an intuitive way of describing system characteristics. With simulations of the model we also showed that RED introduces a reference value that does not depend on the load in certain valid regions, and also that the parameters of RED must be adapted to large variations of the network load.

Future work includes to incorporate RED in the control structure and translate the RED parameters into reference value and control parameters. We will also further develop the mathematical model and investigate stability and robustness for RED using this model.

References

[1] F. Gunnarsson, F. Gunnarsson, and F. Gustafsson. TCP Performance based on Queue Occupation. In *Towards 4G and Future Mobile Internet, Proceedings for Nordic Radio Symposium 2001*, Mar. 2001.

[2] S. Floyd and V. Jacobson. Random Early Detection gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, 1(4):397–413, Aug. 1993.

[3] J. Sachs and M. Meyer. Mobile Internet - Performance Issues Beyond the Radio Interface. In *Aachen Symposium on Signal Theory 2001*, 2001.

[4] M. Meyer. TCP Performance over GPRS. In *Wireless Communications and Networking Conference, 1999. IEEE*, Sep. 1999.

[5] J. Peisa and M. Meyer. Analytical Model for TCP File Transfers over UMTS. In *3G Wireless 2001*, 2001.

[6] A. Misra, J.S. Baras, and T. Ott. Window Distribution of Multiple TCP's with Random Loss Queues. In *Proceedings of Globecom'99*, Dec. 1999.

[7] L.S. Brakmo and L.L. Peterson. TCP Vegas: end to end congestion avoidance on a global Internet. *Selected Areas in Communications, IEEE Journal on*, 13:1465–1480, Oct 1995.

[8] S. Mascolo, C. Casetti, M. Gerla, M.Y. Sanadidi, and R. Wang. TCP Westwood: Bandwidth Estimation for Enhanced Transport over Wireless Links. In *Proceedings of MobiCom 2001*, 2001.

[9] R. Jain. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling. *SIGMETRICS Performance Evaluation Review*, 19(2):5–11, 1991.

[10] C.V. Hollot, V. Misra, D. Towsley, and W.B. Gong. A Control Theoretic Analysis of RED. In *INFOCOM 2001. Proceedings. IEEE*, 2001.

[11] S. McCanne and S. Floyd. ns Network Simulator. http://www.isi.edu/nsnam/ns/.

[12] S. Floyd. TCP and Explicit Congestion Notification. *ACM Computer Communication Review*, 24(5):10–23, Oct 1994.

Matching Multi-Fractal Process Parameters Against Real Data Traffic

Patrik Carlsson[1], Markus Fiedler[2] and Arne A. Nilsson[3]

[1]Blekinge Institute of Technology, 371 79 Karlskrona, Sweden. E-mail: Patrik.Carlsson@bth.se
[2]Blekinge Institute of Technology, 371 79 Karlskrona, Sweden. E-mail: Markus.Fiedler@bth.se
[3]Blekinge Institute of Technology, 371 79 Karlskrona, Sweden. E-mail: Arne.Nilsson@bth.se

Abstract

Recent analyses of real data/Internet traffic indicate that data traffic exhibits long-range dependence as well as self-similar or multi-fractal properties. By using mathematical models of Internet traffic that share these properties we can perform analytical studies of network traffic. This gives us an opportunity to analyse potential bottlenecks and estimate delays in the networks.

Processes with multi-fractal properties can be modeled by multiplying the output of Markov Modulated Rate Processes (MMRP) each defined by four parameters. The MMRP are easily used in stochastic fluid flow modeling. This model is also suited for analysis of other traffic types e.g. VoIP and thus, it allows for integration of different traffic types, i.e. time-sensitive voice traffic with best-effort data traffic. Using this model we can calculate performance parameters for each individual stream that enters the system/model.

In this paper we show how to construct a multi-fractal process that is matched to measured data from MMRP sub processes.

Keywords: Multi-fractal traffic, fluid flow model, numerical analysis, performance evaluation

1 Introduction

The reason for this research is that we want to develop a mathematical model for Internet traffic, that can be used in a fluid flow analysis. In earlier papers [1] and [2] we demonstrated that the fluid flow model is a usable tool for analytical performance studies. We have used data processes that had multi-fractal properties (*MFP*), and we also expressed desires to do similar studies based on processes that have been modeled based on measured network traffic. In this paper we show how a certain class of multi-fractal processes can be modeled such that they match up well against real network traffic. The tool we now have permits us to do better capacity management instead of today's *rules-of-thumb*.

Since we are to match a mathematical model against measured data, we will start by describing the mathematical model and its basic properties. Following that we will verify this model against simulations that have been performed. The actual matching of the parameters based on real measured traffic will then be described, and finally some conclusions will be drawn.

2 Processes description

The mathematical model that is used to generate a process with multi-fractal properties is formed by multiplying the output of independent 2-state Markov modulated rate processes (called sub-processes). The process will be formed as

$$R(t) = \prod_{i=0}^{n-1} R_i(t)\,. \tag{1}$$

Here $R_i(t)$ is the output rate from sub-process i at time t (data-unit/time-unit). The 2-state Markov-modulated rate process is depicted in Figure 1. This is a well known and often used model. The model was originally suggested in [4].

2.1 Description of a Sub-Process

Each sub-process is a simple low-high process, as described in Figure 1. The sub-processes transition matrix and rate matrix are formed as

$$\mathbf{M}_i = \begin{bmatrix} -\lambda_i & \lambda_i \\ \mu_i & -\mu_i \end{bmatrix} \qquad \mathbf{R}_i = \begin{bmatrix} l_i & 0 \\ 0 & h_i \end{bmatrix}. \tag{2}$$

The first two moments of the MFP rate process are going to be used for the matching of processes to MFP. Thus it is interesting to have analytical expressions for the two first moments.

$$\mathbf{E}[R_i(t)] = \frac{l_i\mu_i + h_i\lambda_i}{\lambda_i + \mu_i} \tag{3}$$

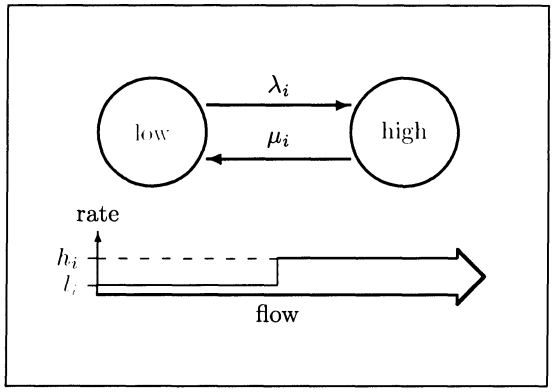


Figure 1: Sub-process i, a 2-state Markov Modulated Rate Process.

$$\mathbf{E}\left[R_i^2(t)\right] = \frac{2\lambda_i\mu_i(l_i-h_i)^2}{t^2(\lambda_i+\mu_i)^3}\left(t - \frac{1}{\lambda_i+\mu_i} + \frac{e^{-(\lambda_i+\mu_i)t}}{\lambda_i+\mu_i}\right) + \frac{(l_i\mu_i+h_i\lambda_i)^2}{(\lambda_i+\mu_i)^2} \quad (4)$$

A more detailed derivation can be found in [3]. It is obvious that the equation for the first moment is not affected by any changes in t. The second moment (cf. Figure 2 for an example) has the following limit properties:

$$\lim_{t\to 0}\mathbf{E}\left[R_i^2(t)\right] = \frac{\lambda_i\mu_i(l_i-h_i)^2}{(\lambda_i+\mu_i)^2} + \mathbf{E}\left[R_i^1(t)\right]^2 \quad (5)$$

$$\lim_{t\to \infty}\mathbf{E}\left[R_i^2(t)\right] = \frac{(l_i\mu_i+h_i\lambda_i)^2}{(\lambda_i+\mu_i)^2} = \mathbf{E}\left[R_i^1(t)\right]^2 \quad (6)$$

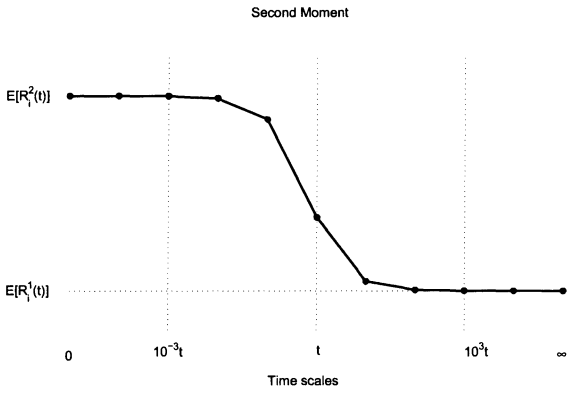


Figure 2: Typical shape of the second moment of a subprocess with $\lambda_i = \mu_i$ and $\mathbf{E}[R_i(t)] = 1$.

2.2 Process

As mentioned before the process is formed by multiplying the output from the sub-processes. The transition matrix is formed using Kronecker addition:

$$\mathbf{M}_\mathrm{D} = \mathbf{M}_0 \oplus \mathbf{M}_1 \oplus \cdots \oplus \mathbf{M}_{n-1} \quad (7)$$

and the rate matrix as

$$\mathbf{R}_\mathrm{D} = \mathbf{R}_0 \odot \mathbf{R}_1 \odot \cdots \odot \mathbf{R}_{n-1}\,, \quad (8)$$

where

$$\mathbf{A}\oplus\mathbf{B} = \mathbf{A}\otimes\mathbf{I_B} + \mathbf{I_A}\otimes\mathbf{B} \quad (9)$$

$$\mathbf{A}\odot\mathbf{B} = \mathbf{A}\otimes\mathbf{I_B}\cdot\mathbf{I_A}\otimes\mathbf{B} \quad (10)$$

and $\mathbf{I_X}$ is a unity matrix with the same size as matrix $\mathbf{X}$. The Kronecker multiplication $\otimes$ is defined as

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,m} \\ \cdots & \cdots & \cdots \\ a_{n,1} & \cdots & a_{n,m} \end{bmatrix}_{n\times m}$$

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & \cdots & b_{1,q} \\ \cdots & \cdots & \cdots \\ b_{n,1} & \cdots & b_{p,q} \end{bmatrix}_{p\times q}$$

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & \cdots & a_{1,m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & \cdots & a_{n,m}\mathbf{B} \end{bmatrix}_{np\times mq} \quad (11)$$

The moments for the process are easily obtained by multiplication of the sub-process moments

$$\mathbf{E}[R(t)] = \prod_{i=0}^{n-1}\mathbf{E}\left[R_i^1(t)\right] \quad (12)$$

$$\mathbf{E}\left[R^2(t)\right] = \prod_{i=0}^{n-1}\mathbf{E}\left[R_i^2(t)\right]\,. \quad (13)$$

3 Simulations

To compare the analytical formulas of the first and second moments, some simulations have been run. In Figure 3, comparisons of first and second moments obtained by simulation and by the analytical method are shown. In Table 1 the process parameters are presented.

Looking at the first row in Figure 3, the match is perfect for both moments. However this process corresponds to a non-varying process, with $l_i = h_i = 2$ hence no variations. Looking at rows two and three, a slight difference between analytical and simulation results is seen. It can probably be accredited to a too small sample space for the simulation. But the match is fairly good at least at higher time scales for the second moment.

4 Matching

Our developed matching method assumes that the first moment has been normalized in such a way that the mean of the first moments at different time scales will be one.

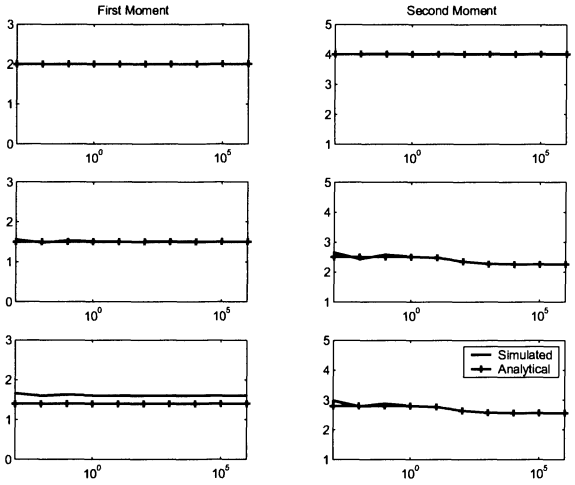


Figure 3: A comparison between moments that where calculated from simulations and moments based on the analytical formulas in Equations 3 and 4.

ID	τ	λ	μ	l_i	h_i	Mean
1	100	0.02	0.02	2	2	2
2	100	0.02	0.02	1	2	1.5
3	83.333	0.03	0.02	1	2	1.4

Table 1: Parameters used in simulation

When matching real-world data there are usually some trade-offs to be made. We have chosen to overestimate the moments. The main reason for this, is that it is better to work against a worst case scenario. Another consideration is the number of processes that are used to match the process, since in an analysis step to follow [1] there is an upper bound on the processing power. That is if we use an infinite number of processes we would probably get a good match, but it would not be usable in future work.

When matching we are trying to match the first two moments. There are a number of methods that could be used to do this. In this paper we have chosen to use a simple and heuristic method. The method works roughly like this:

1. Initialize

   $\vec{D}^2$ The second moment
   $\vec{D}^1$ The first moment
   $\vec{T}$ The time scales

   where D_i^x denotes the i:th element in the vector corresponding to the x:th moment.

2. For each time scale in $\vec{T}$, denoted by T_i, try to find a sub-process that has properties that match it, as if the process were *constant* at time scales smaller than t (cf. Figure 4). In the case $\lambda = \mu = 2/\tau$, which will be used in the sequel, a choice that has proven to be useful is given by $\tau = 10\,T_i$, leading to $\lambda = \mu = 0.2/T_i$.

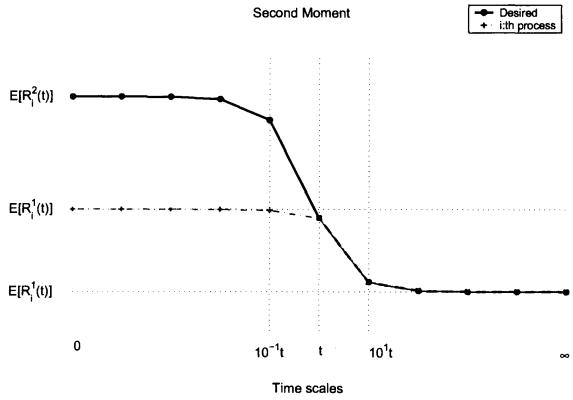


Figure 4: Identification of sub-process i from the desired process.

3. Loop through $\vec{T}$ starting at the largest time scale and going down to the smallest.

   (a) Assign the moments

   $$M^2 = D_i^2 \quad M^1 = D_i^1 . \tag{14}$$

   These will be used as the desired value of the sub-process as $t \to 0$, since the process has been normalized in such a manner that the mean will be one and remembering that the limit value of a sub-process in equation 5. This value can be too large for one single sub-process. The solution is the usage of sub-sub-processes, processes that are identical and together achieve the desired properties on a given time scale.

   (b) Estimate how many sub-sub-processes that are needed to achieve this limit.

   $$k = \left\lceil \frac{\log M^2}{\log 2} \right\rceil \tag{15}$$

   (c) If $k \leq 0$ then this time scale is finished. Continue with the next smaller time scale (step 3a).

   (d) Determine the limit for a sub-sub-process

   $$O^2 = \sqrt[k]{M^2} \quad O^1 = \sqrt[k]{M^1} \tag{16}$$

   (e) Based on the symmetry from $\lambda = \mu = 2/\tau$ an expression for l and h can be obtained

in the following manner:

$$\Delta L = 2\sqrt{O^2 - (O^1)^2} \quad (17)$$

$$l = M_i^1 - \frac{\Delta L}{2} \quad (18)$$

$$h = M_i^1 + \frac{\Delta L}{2} \quad (19)$$

(f) The parameters that identify sub-process i are known (the output rates in the two states, the transition rates and the number of sub-sub-processes that form the sub-process). Create two vectors $\vec{M}_i^2$ and $\vec{M}_i^1$ that span all time scales defined by $\vec{T}$ from these parameters. These vectors specify the moments for sub-process i. Store the parameters needed to create these vectors.

(g) Remove the influence on the desired process that this sub-process has by simply dividing it out, element-by-element. j is the size of the $\vec{T}$.

$$\vec{D}^1 = \left[\frac{D_1^1}{M_1^1}, \ldots, \frac{D_j^1}{M_j^1}\right] \quad (20)$$

$$\vec{D}^2 = \left[\frac{D_1^2}{M_1^2}, \ldots, \frac{D_j^2}{M_j^2}\right] \quad (21)$$

(h) Proceed with the next time scale (step 3a).

4. Now the matched process can be created by combining the sub-processes using element-by-element multiplication using Equations 3 and 4 from the stored parameters. Recreate the moment vectors for each sub-process. Call them $\vec{P}_i^x$ where i is the number of the sub-process i and x indicates the x:th moment. The moments will then be

$$\vec{M}^1 = \begin{bmatrix} P_{1,1}^1 \cdot P_{2,1}^1 \cdot \ldots \cdot P_{j,1}^1 \\ \ldots \\ P_{1,n-1}^1 \cdot P_{2,n-1}^1 \cdot \ldots \cdot P_{j,n-1}^1 \end{bmatrix}^T \quad (22)$$

$$\vec{M}^2 = \begin{bmatrix} P_{1,1}^2 \cdot P_{2,1}^2 \cdot \ldots \cdot P_{j,1}^2 \\ \ldots \\ P_{1,n-1}^2 \cdot P_{2,n-1}^2 \cdot \ldots \cdot P_{j,n-1}^2 \end{bmatrix}^T \quad (23)$$

5. Optional loop: Repeat from step 3 until a desired maximal error or maximal number of processes is reached.

Now it is possible to list the parameters that are needed to match the desired input. When comparing a matched process with a measured process we define the error in the second moment as

$$\overrightarrow{Error} = \left[\left|\frac{D_1^2 - M_1^2}{D_1^2}\right|, \ldots, \left|\frac{D_j^2 - M_j^2}{D_j^2}\right|\right]. \quad (24)$$

ID	λ	μ	l_i	h_i
1	512	512	0.5	1.5
2	128	128	0.5	1.5
3	32	32	0.5	1.5
4	8	8	0.5	1.5
5	2	2	0.5	1.5

Table 2: Parameters used to create the desired process that is matched in Example 1.

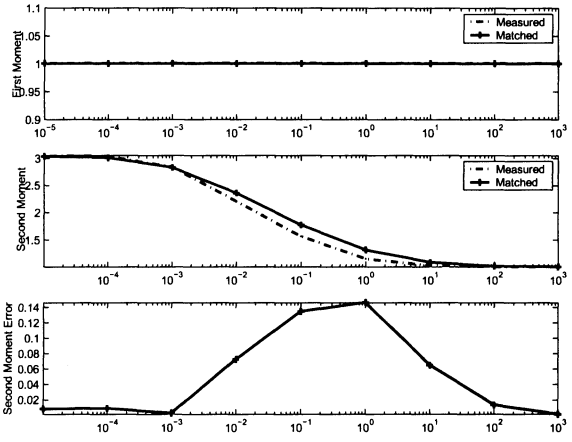


Figure 5: Comparison between a simulated process and a created process (Example 1).

4.1 Example 1: Simulated data

In Figure 5 an example of the matching is shown. Here the desired signal is a simulated one, it was created by five sub-processes listed in Table 2. The first moment is a perfect match, the second moment is over estimated. Looking at the error the over estimation is not that large.

4.2 Example 2: Uplink on small workgroup switch

The data used here came from a mid-size workgroup switch (Cisco 2948G). There were about 36 hosts connected to it, distributed into two categories: department computers and 16 computers accessible to students. The measurements were performed during the summer of 2001, from the 7:th to the 30:th of July, which resulted in more than 2×10^6 samples. The measurement focused on the interface that corresponded to the uplink of the switch. The data was collected using a simple SNMP (Simple Network Management Protocol) polling tool. The tool polled the device roughly once every second, with the odd delay (probably due to processing times in the device and intermediate devices).

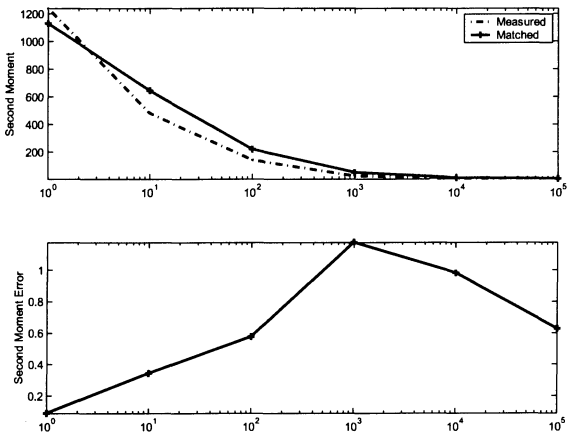


Figure 6: Comparison between a measured process and a created process (Example 2).

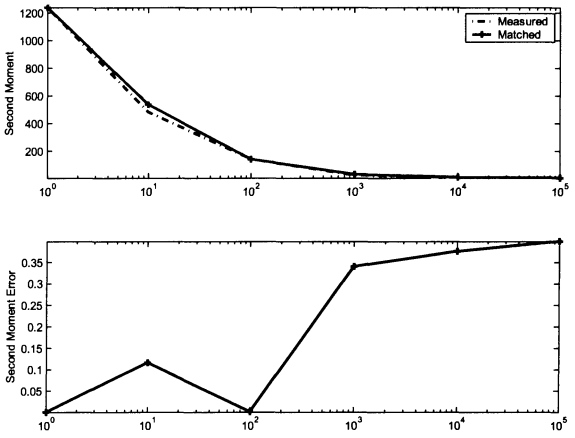


Figure 7: The resulting match after running the method five times with k limited to one.

The comparison of Figure 6 (real data) with Figure 5 (simulated process) reveals that the measurements do not cover the whole set of possible time scales. Unfortunately, SNMP-based tools are basically limited to time scales of one second and more. New tools being able to cope with smaller time scales are currently under development.

The results of the matching shown in Figure 6 are based on the algorithm described in Section 4 without the optional step 5. For most of the time scales, the measured values are over-estimated. However, on the smallest time scale, an under-estimation is perceived. Looking at the error plot in Figure 6, a better match would be desirable. This can be achieved in a simple way by

- using one instead of k homogeneous sub-sub-processes (i.e. with the same transistion and data rates) on a given time scale;
- using option 5, i.e. repeating the matching procedure several times.

Basically, this modification leads to the use of heterogeneous sub-sub-processes (i.e. with different data rates) per time scale under consideration. The improvement of applying these modifications in our example is seen from Figure 7, where step 3 was repeated five times. The match is much better, and there is no underestimation left. However, the number of processes needed increased significantly.

5 Conclusions

In this paper we have presented a method that allows us to match process parameters against real network traffic. It is possible to get a better match using more processes. This, however, will cost in later analysis with regards to processing time. We have found that it is possible to get a fairly good match using only a few processes. We need a measurement tool that allows us to measure at smaller time scales than one second.

References

[1] P. Carlsson and M. Fiedler. Multifractal products of stochastic processes: Fluid flow analysis. Proceedings of 15th Nordic Teletraffic Seminar (NTS-15), Lund, Aug. 22-24, 2000.

[2] P. Carlsson, M. Fiedler, and A. Nilsson. Voice and multifractal data in the Internet. Proceedings of 26th IEEE Conference on Local Computer Networks (LCN2001), Tampa, Nov. 15-17, 2001.

[3] P. Carlsson, M. Fiedler, and A. Nilsson. Moment analysis of sub-processes of multi-fractal sub-processes, Internal Report, 2002.

[4] P. Mannersalo, I. Norros, and R. Riedi. Multifractal products of stochastic processes: A preview.
COST-257 Technical Document 257TD(99)31.
`http://nero.informatik.uni-wuerzburg.de/cost/Final/TDs/257td9931.pdf`

Reducing call setup delays using SIP/SDP compression

Hans Hannu[1], Jan Christoffersson[2], Stefan Forsgren[3], Stefan Wänstedt[4] and Krister Svanbro[5]

[1-5]AWARE at Ericsson Research, Ericsson Erisoft AB, Box 920, 971 28 Luleå, Sweden. E-mail: Hans.Hannu, Jan.Christoffersson, Stefan.Forsgren, Stefan.Wanstedt, Krister.Svanbro { @epl.ericsson.se.

Abstract

This paper introduces a generic compression scheme intended for ASCII based signaling protocols in cellular systems such as UMTS and GSM/GERAN. In this paper, the scheme is applied to SIP/SDP signaling. To increase the compression efficiency, prior knowledge of typical SIP/SDP phrases are used together with previously transmitted messages. The scheme is efficient and robust to message loss and misordering. The compression efficiency on two tested SIP sessions are 8 and 5.7 to 1, respectively. Simulations show that the time used for SIP signaling between the terminal and the first SIP proxy in a call setup is reduced by more than 66%. The simulations further indicate that the system capacity is increased by approximately 17%.

1 Introduction

In the recent years two communication technologies have become commonly used by the general public; cellular telephony and the Internet. These two technologies are now beginning to merge, one could say that the Internet is going mobile. One of the consequences of this merge is that IP [1] protocols will be used in cellular systems all the way to the terminal, for both transport of user data and session control.

The Session Initiation Protocol (SIP) [2] is an application layer protocol for establishing, modifying and terminating multimedia sessions or calls. These sessions include Internet multimedia conferences, Internet telephony and similar applications. The standardization body 3GPP has selected SIP as the session establishment protocol for the IP multimedia subsystem in coming UMTS releases, see [3]. The Session Description Protocol (SDP) [4] is used to advertise multimedia conferences and communicate conference addresses and conference tool specific information. It is also used for general real-time multimedia session description purposes. SDP is carried in the message body of SIP messages. Both SIP and SDP are text based using the ISO 10646 character set in UTF-8 encoding. These protocols have some typical characteristics:

- Requests and reply characteristics. When a sender sends a request, it stays idle until a response is received. Hence, a SIP session takes a number of round trip times.
- The protocols are ASCII based.
- The protocols are generous in size.

In wireless environments and especially in cellular systems, such as GSM/GPRS and UMTS, there is a need to maximize the transport efficiency of data over the radio interface. It is of vital importance to have a sufficient number of users per cell; otherwise the system cost would prohibit deployment. Since it is crucial to use the existing bandwidth carefully, the average user bit rate is typically relatively low compared to the average user bit rate in wired line systems. Cellular links also have characteristics that differ somewhat from a typical fixed link, e.g. copper or fiber. The most important characteristics are the lossy behavior and the relatively large round trip times.

Compressing the IP/UDP headers of the protocols used for carrying user data is one way to make more efficient use of the scarce radio resources, see e.g. [5]. To enhance the radio resource usage even further, compression of the messages from signaling protocols, such as SIP/SDP, should be considered. Compression of SIP/SDP will also improve the service quality, by reducing the user idle time, at e.g. call setup. In [6] there is a comparison between a GERAN call setup using SIP and an ordinary GSM call setup. In this example, a typical GSM call setup takes about 3.6 seconds, and for the corresponding case when using SIP, the call setup is approximately 7.9 seconds. Thus, compression solutions are needed to reduce the signaling delay and the required bandwidth.

Another situation that would benefit from reduced signaling is when carrying signaling messages over

narrow bandwidth links in mid-call. In GERAN, such uncompressed SIP signaling may result in voice frame stealing and degraded speech quality.

The likely introduction of new IP based services with their corresponding signaling protocols makes it reasonable to consider a compression scheme that is generic. The scheme should be generic in the meaning that the scheme efficiently can be applied to arbitrary protocols with certain characteristics, such as the ASCII based protocols SIP, SDP and RTSP [7].

In this paper a compression scheme for compression of SIP/SDP is introduced. The scheme is generic towards ASCII based protocols and resilient against packet loss and misordered packets between the compressor and decompressor. The scheme can be implemented following the standard being developed in IETF, see [8] and [9].

The paper is organized as follows: Chapter 2 describes the compression scheme applied to compression of SIP/SDP. Chapter 3 presents some performance results. Finally, a short summary together with conclusions is given in Chapter 4

2 The Compression Scheme

The compression scheme described in this is an application signaling compression scheme, with the intention to be a multiple protocol compression scheme. The central idea of the scheme is to use previously transmitted messages in the compression process. This gives a significant improvement compared to compressing messages independently. The scheme is based on standard dictionary compression algorithms, e.g. LZSS [10], which are commonly used to compress ASCII (text) files. In general such compression algorithms maintain a byte buffer, which is filled with the text to be compressed. The algorithms scan through the buffer and replace repeated strings with references to the previous occurrence. Thus, the text to be compressed is actually used as the reference dictionary. However, applying these kind of algorithms on relatively small messages (200 - 2000 bytes) with few long repeated text strings, such as is the case for SIP messages, does not result in satisfactory compression. The reason for this is simply that not enough text strings can be replaced by references. This issue has been addressed with the proposed compression scheme.

To increase the compression efficiency, additional dictionaries are used. A static dictionary contains text strings that are typically found in SIP/SDP messages such as header names, etc. A dynamic dictionary contains text strings from already transmitted messages and is updated during the session. These dictionaries together with functionality for message handling form the basis of the scheme.

Since the dynamic dictionary is updated during the session and the fact that messages can be misordered or lost, a mechanism to handle dictionary management is needed. This functionality has been addressed through the use of headers. Each compressed message is supplied with a header that contains:

A sequence number that is used to detect if a compressed message has been lost or misordered. It is always used in the acknowledgement process.

Acknowledgement information used by an entity to inform its opposite entity of which compressed messages that have been received and successfully decompressed.

A CRC calculated over the uncompressed message is carried in the header. The CRC is used to avoid that faulty decompressed messages are passed up to higher layers in the case of link bit errors.

The compression and decompression entities also maintain temporary storage repositories. These are used to store sent and received messages until they are acknowledged and put into the dynamic dictionary. It is of crucial importance that updates to the dynamic dictionaries are done simultaneously at both the compressor and decompressor, since it is not possible to decompress a message using a dictionary different from the one used for the compression. As more and more messages are put into the dynamic dictionary the compression efficiency generally increases. This can be seen when comparing the compression ratio of the first message to compression ratios of later messages compressed using a large dynamic dictionary.

The compression process can be described as follows: The message to be compressed arrives at the compressor. The information from the dictionaries is assembled to form a reference dictionary of the compressor. The compression algorithm processes the message by looking for strings in the message that exist in the reference dictionary, or in the already processed part of the message, and replaces them with references. Before the compressed message is sent, the new information that will be used to compress subsequent packets is stored in the temporary storage repositories. Finally, the header is attached and the compressed message is sent. The reverse process is carried out at the decompressor.

3 Performance Evaluation

The compression scheme has been implemented in C and tested. The results presented in this section are based on compression of the two SIP sessions, A1 and A2, found in [11]. All packets are delivered in order and there is no packet loss. To achieve realistic results, a static dictionary containing most SIP keywords from [2] has been used together with some additions from proposed

extensions of SIP. The compression algorithm used is a slightly modified LZSS compressor with variable length encoding of the references.

3.1 Compression Factor

For the A2 series, the compression scheme achieves an average compression factor of 8, defined as: bytes uncompressed / bytes compressed. The first packet has a compression factor of 2, but as the dynamic dictionary accumulates information the compression factor increases, see Figure 1. The A1 series shows similar results: the factor is 2.3 for the first message and the overall factor is 5.7.

The results illustrate the impact of using a dynamic dictionary that is updated with previously transmitted messages. Even better performance could be achieved by saving parts of the dynamic dictionary from previously compressed sessions. The choice of compression algorithm obviously has an impact on the performance. The maximum possible string length that can be referenced with a single reference is an example of a parameter that will influence the performance.

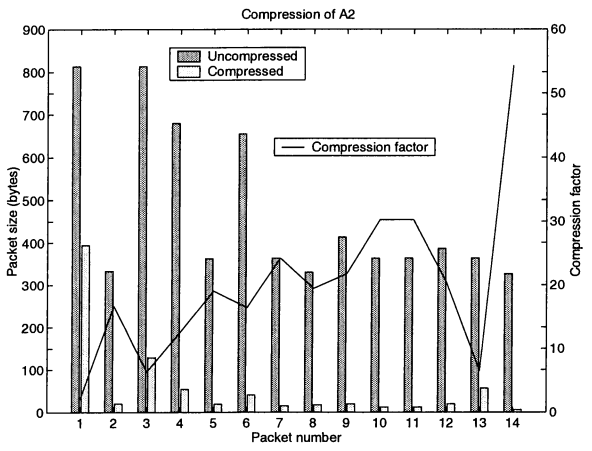


Figure 1. Compression factors of the A2 series and packet sizes before and after compression.

3.2 CPU Consumption

CPU consumption is an indicator of the resources needed to run the algorithm and must be kept at a reasonably low level to make implementations, e.g. in a 3G system feasible. The CPU consumption has been measured on a Pentium III using the assembler instruction *rdtsc*, which reads the number of clock cycles elapsed. The implementation was written in C and was not optimized in any particular way other than that it was compiled with the *gcc* optimization flag -O2. The tests indicate that it takes a total of 100,000 to 350,000 clock cycles to process a typical packet, which on a 600 MHz processor corresponds to approximately 150 to 600 microseconds. These numbers are averages derived from several runs. The first packet in each stream takes roughly twice the time to process as a typical packet for the stream. Compression time does not only depend on packet size, but also on how well the packet can be compressed, which explains why the first packet takes the longest time to process. In general, short and few matches in the dictionary yield longer times, since more dictionary searches have to be performed. The measured time is divided into four parts: protocol actions in the compressor, compressing, protocol actions in the decompressor and decompressing. The protocol actions measured are: processing headers, calculating CRC and maintaining the dictionaries. The tests show that compressing the message requires approximately two thirds of the total processing time and that decompression is by far the fastest part.

3.3 Memory Consumption

The implementation uses RAM for mainly three purposes: the variables, the dictionaries and the buffers. The number of variables is considered small compared to the other memory usage, and is not included in this evaluation. The dynamic dictionary is 4 kilobytes and uses a search structure of 8.5 kilobytes. The temporary storage repository contain at most 12 entire packets each, but in a typical SIP session they contain 1-2 packets each. The test series packet sizes range from approximately 200 to 800 bytes. In addition, input and output buffers are needed to keep the current message. All included, the RAM usage in the implementation is approximately 14 kilobytes.

3.4 System Performance

In addition to the studies of the actual SIP-compression scheme, we have investigated the potential gain in a UMTS system due to compression of SIP-signaling. This investigation was performed using a UMTS system simulator. Speech sessions were initiated by exchanging a series of packets back and forth between the user equipment and the system. All packets have to be acknowledged from the receiving side before a new packet is sent. In our SIP traffic model the average SIP packet sizes are collected in an array while a second array controls the directions in which the packet travels. During simulation average sized packets are sent one at a time, upon acknowledgement of the previous package. When the 200 OK SIP accepts the call, the radio access bearer (RAB) or channel changes from signaling (acknowledged) to a speech RAB (unacknowledged). Channel switching in the simulation is assumed to occur instantly. After a conversation with a mean length of 90s the call is released, exchanging and acknowledging the final BYE and 200 OK SIPs using the signaling RAB.

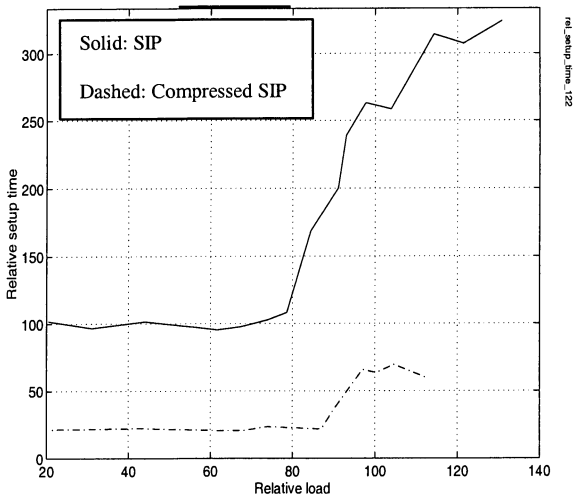


Figure 2. Results of system simulations of SIP signaling and impact on setup time due to SIP signaling.

From Figure 2 the SIP portion of the setup time of a conversation is reduced several magnitudes. Figure 2 shows the relative call setup time due to SIP signaling, where the baseline is the call setup time of uncompressed SIP in a system with low load. The baseline for the relative load is the number of users a corresponding circuit switched system can support. The capacity is measured by the relative load that the system can support before the 95th percentile of the down link block error rate (DLBLER) increases sharply. From Figure 3 the increase is approximately 17 %. System capacity with signaling is low in this study compared to other similar studies. The main reason for this is that as the system load increases to a certain level, retransmission of SIP-packets commences and increases the load even more.

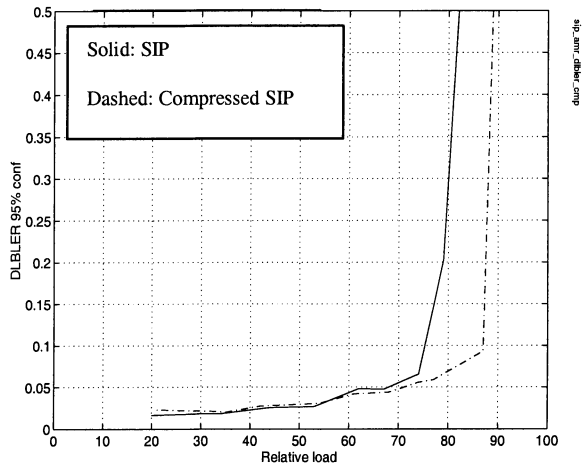


Figure 3. Results of system simulations of SIP signaling and impact on DLBLER.

4 Summary and Conclusions

This paper has motivated the use of SIP compression in 3G networks to reduce call setup delays and bandwidth usage. The possibility to use previous messages in the compression process has been shown to improve the performance of the compression to a great extent. However, use of previous messages also place demands on the robustness of the process. To handle this, the compressed messages have headers attached to ensure that the same messages are used in the dictionary at both the compressor and decompressor. The information in the headers are used to make the compression scheme robust to message loss and misordered messages.

The compression factor obtained for the two test sequences are 8 and 5.7. These results depend on the particular sequences but can be improved by using a more sophisticated dictionary at the start of the compression. The time to compress a typical message is 150 to 600 microseconds on a 600 MHz processor, depending on the size of the message and number and length of the matches found in the message. The memory consumption is small, typically in the vicinity of 14 kilobytes of RAM. Setup delays and system capacity in a UMTS system with SIP-initiated (voice) services are also significantly enhanced by the use of the compression scheme.

5 References

[1] J. Postel, *Internet Protocol*, RFC 791, September 1981.

[2] M. Handley, H. Schulzrinne, E. Schooler and J. Rosenberg, *SIP: Session Initiation Protocol*, RFC 2543, August 2000.

[3] IP Multimedia (IM) Subsystem - Stage 2, 3GPP TS 23.228, v.5.1.0. 2001-06.

[4] M. Handley and V. Jacobson, *SDP: Session Description Protocol*, RFC 2327, April 1998.

[5] C. Bormann, *et. al., RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed*, RFC 3095, July 2001.

[6] Nortel Networks, A Comparison Between Geran Packet-Switched Call Setup Using SIP and GSM Circuit-Switched Call Setup Using RIL3-CC, RIL3-MM, RIL3-RR and DTAP, 3GPP TSG GERAN #2, GP-000508, 6-10 November 2000.

[7] H. Schulzrinne, A. Rao and R. Lanphier, *Real Time Streaming Protocol (RTSP)*, RFC 2326, April 1998.

[8] R. Price *et. al. Signaling Compresssion (SigComp)* Internet draft (work in progress), March 2002. <draft-ietf-rohc-sigcomp-05.txt>.

[9] H. Hannu *et. al. SigComp - Extended Operations.* Internet draft (work in progress), March 2002. <draft-ietf-rohc-sigcomp-extended-02.txt>.

[10] J.A. Storer and T.G. Szimanski, *Data Compression via Textual Substitutions*, Journal of the ACM 29, 1982.

[11] H. Hannu, *Signaling Compression Requirements & Assumptions,* Internet draft (work in progress), December 2001. <draft-ietf-rohc-signaling-req-assump-04.txt>

Scalability limitations of self-healing optical networks due to reliability constraints.

Lena Wosinska

Royal Institute of Technology (KTH), Department of Microelectronics and Information Technology, Electrum 229, S-164 40 Kista, Sweden.E-mail: e-adress1.

Abstract

In this paper we study the limitations that reliability requirements impose on network scaling and design in order to determine the limiting factors for the scalability of optical networks. We analyze reliability performance of networks based on optical fibers and focus on self-healing networks in ring and mesh topologies. We consider connection availability in a mesh network and network availability for a dual ring and study the influence of cable mean time to repair (MTTR) on network availability and scalability.

We present the general availability models for mesh and ring networks and use these models to calculate limitations for network size for given reliability data of the network elements. The goal of this work is not to achieve absolute availability figures. The results focus on the limiting factors for scalability of optical networks due to reliability constraints and guidelines for improvement of the network scalability.

Our results show that allowable network size for a given network availability depends strongly on cable MTTR. This implies that even for networks with built-in redundancy, cable MTTR is of utmost importance.

1 Introduction

With global deregulation in telecommunications, there has been an explosive increase in the deployment of communication technology, products and services. Regardless of the market, communication companies are challenged by increased competition, evolving technologies and expanding customer expectations. As information becomes the world's most valuable resource, the integrity of the infrastructure becomes paramount. To survive in today's competitive marketplace, each operator must operate an efficient, yet reliable network that provides leading edge services with minimal downtime. With the growing competitiveness, carriers have begun to invest heavily in their infrastructure with a significant focus on automated management, maintenance and monitoring solutions for their networks. Traditionally, automated management was considered a means to achieve greater internal efficiencies. Today, network management is seen as a vital business strategy necessary to achieve rapid deployment of new services and to providing the level of service required to grow in the face of intense competition.

Greater competitiveness among operators has led to an increased focus on Quality of Service (QoS) both as a sales tool and as a means of retaining customers. Service contracts promising 99.999% availability are becoming commonplace. With emerging network architectures it is a concern that network resilience and survivability is kept to at least the same standard as traditional SONET/SDH ring architectures provide [3]. It should be noted that even with protection schemes such as 1+1 within SONET, delivering service commitments is a difficult challenge when the outside plant is not managed appropriately.

In this paper we study the limitations that reliability requirements pose on network design and network size in order to determine the limiting factors for scalability of optical networks. A significant improvement in the scalability of networks can be obtained by improving the reparation process. Therefore, we study the impact of cable mean-time-to-repair on network reliability performance in order to point out the importance of effective fault management in optical networks. This paper shows that by decreasing cable mean-time-to-repair operators can either offer higher network availability to their customers, or build larger networks while guaranteeing the same availability as they offer today.

2 Abbreviations

A	Asymptotic availability
ADM	Add/drop multiplex
DP	Diverse protection
FIT	Failures in time, per 109 hours
FDDI	Fiber Distributed Data Interface
MDT	Mean down time, min/year
MTTR	Mean time to repair, hours
NE	Network element
NF	Node failure

OADM	Optical add/drop multiplex
OFA	Optical fiber amplifier
OTDR	Optical Time Domain Reflectometer
OXC	Optical cross-connect
RH	Ring hit
RFTS	Remote Fiber Test Systems
SDH	Synchronous digital hierarchy
SH	Span hit
SHR	Self-healing ring
SONET	Synchronous optical network
U	Asymptotic unavailabilitySvenska

3 Network scenarios and survivability mechanisms

In this work we analyze reliability parameters of optical networks. We concentrate on self-healing networks that, in general, are networks that can survive at least a single failure, of either link or node, by using some kind of survivability/protection mechanism. However, our focus is on failures due to fiber break. There are three main categories of topologies for protection architectures [3]: *point-to-point, ring and mesh.*

Point-to-point is based on a route diversity mechanism and corresponds to a pair of nodes that are connected by two physically diverse transmission paths. In the route diversity mechanism, a dedicated standby line is provided for each working line (see figure 1). This way the network is protected on a 1+1 or 1:1 basis and the network survivability is 100% for a single failure. In this scheme, however, it is difficult to handle multiple failures. Furthermore this protection technique can be quite expensive due to the need of duplicate network equipment.

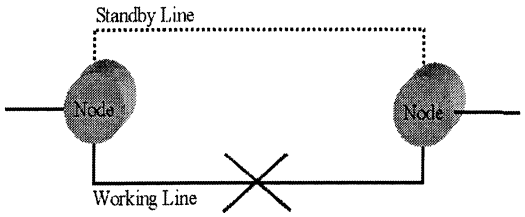
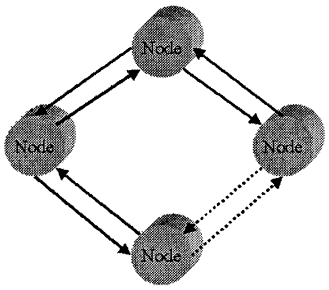
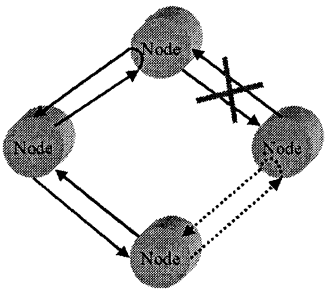


Figure 1: *Route diversity*

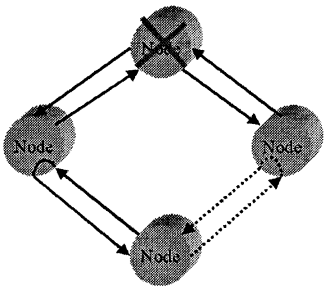
Self-healing ring (SHR), is considered one of the most efficient restoration techniques for fiber optic networks (e.g. SONET [3] and FDDI [4]). Literally, ring protocols are applied to a network with dual-ring topology, as shown in figure 2. It is considered a static restoration since restoration routes are pre-determined. In case of a failure the affected traffic can be rerouted in the opposite direction. Thus, the spare protection capacity of the ring is shared by all of the working traffic on the ring. The ring architecture provides fast restoration within 50ms [3], thus fast enough to avoid dropped calls. However, it cannot always handle multiple simultaneous failures. Note that, theoretically, the point-to-point topology is a two-node ring.



a) Dual-ring: normal operation



b) Ring recovery after span hit



c) Ring recovery after node failure

Figure 2: *Ring restoration.*

Mesh is defined as a network where the degree of connectivity exceeds two (i.e., the number of links coincident on a node) for at least one node. In case of a failure, the affected traffic can be rerouted over a variety of diverse paths, in contrast to a single alternate path for the point-to-point and ring architectures. The restoration routes are dynamically and flexibly formed around the failed facility upon detection of the failure. This scheme can provide relatively fast and flexible restoration and has the ability to handle multiple failures. This restoration can be done either by diverse link or by diverse path (see figure 3).

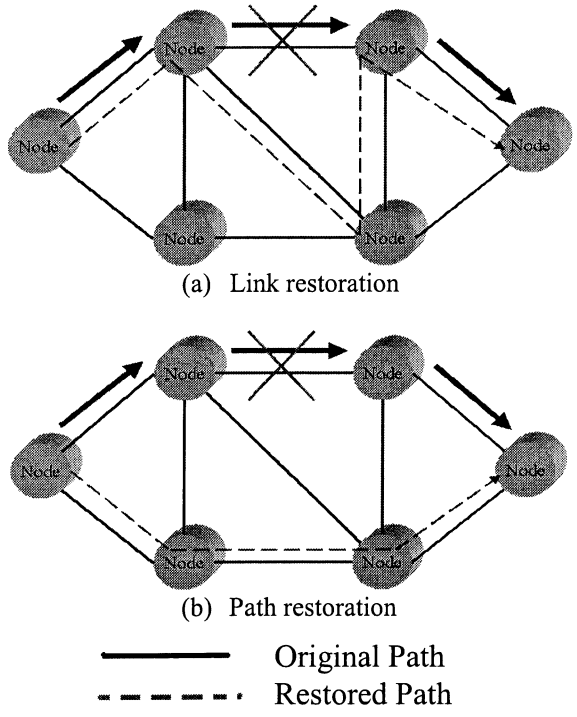


Figure 3: *Link and path restoration*

Table 1: *Reliability data*

Network element or device	Failure rate (FIT*)	U***	MTTR [hours]
Optical fiber amplifier	2,485 [6]	$U_{OFA} = 1.49\ 10^{-5}$	6
Fiber cable**	114 /km to 570/km	U_{FC}	variable
1:2 coupler	50 [8]	$U_C = 10^{-7}$	2
Transmitter	180 [8]	$U_T = 3.6\ 10^{-7}$	2
Receiver	70 [8]	$U_R = 1.4\ 10^{-7}$	2
Network node	5,000 [7]	$U_{node} = 10^{-5}$	2

*1 FIT = 1 failure/109h,
** This corresponds to 1 to 5 cable breaks per year in a 1,000km network,
*** U = FIT x MTTR/109h

4 Network elements and reliability data

Optical transport networks consist mainly of fiber cables, optical fiber amplifiers and network nodes. Nodes can be either all-optical, e.g. optical cross-connects (OXC) with optical switching core and optical add-drop multiplexers (OADM), or those that require opto-electrical conversion.

Typical values for reliability data are shown in Table 1. The results in Section 8 have been calculated with these figures. We would like to point out that these values are used for exemplary purposes. Our availability models presented in Section 6 are general and can be applied to any network with any reliability data.

A node in our models represents either an add-drop multiplexer (ADM) or an optical cross-connect (OXC) with add-drop capability. Reliability data for these network elements are average values and do not correspond to a specific device or system.

In our analysis we vary the cable failure rate, cable MTTR and cable length keeping the overall parameters constant. It is because we concentrate on limitation that cable MTTR and cable failure rate can pose on network geographical size.

The links employ optical amplifiers spaced 50km apart. The fiber cables are assumed to endure 1-5 cable breaks a year for every 1,000km (cable kilometers, not fiber kilometers).

5 Assumptions

In our analysis we assume that:

- The network elements are defect-free from the beginning
- Failure and repair rates of the network elements are time independent
- Failures are caused by non self-healing outages, i.e., every failure has to be repaired
- There must be a method for detecting a failure, and a technique for recovery from faults
- Failure duration is short relative to the time between failures
- Times between failures and the duration of the component failures are independently distributed
- Mean time to repair (MTTR) is 2 hours for in-office equipment [5].
- ·Amplifier spacing is 50km [9]. For purposes of simplicity, this is kept at 50km for all the cases considered. In real networks, this would depend on the length of the links, the number of optical channels employed and the data rate of each channel.

6 Reliability models

We consider connection availability in a mesh network and network availability for a dual ring. We present general availability models for the following survivability schemes:

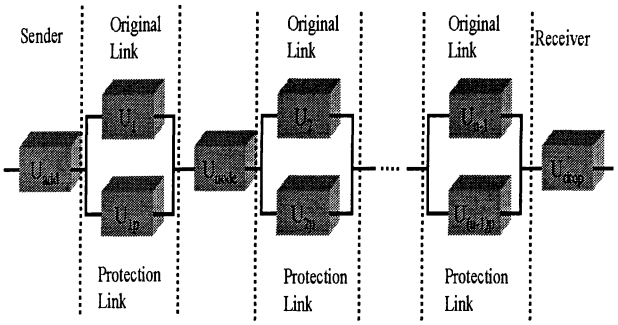
- 1+1 link diverse protection (Fig.4)
- 1+1 path diverse protection (Fig. 5)
- self-healing ring SHR (Fig.6)

Figures 4 through 6 present reliability block diagrams derived for mesh networks with link and path protection and for SHR. Each block in the figures represents either a piece of equipment or a path. In the diagrams the asymptotic unavailability for each block is indicated. For a description of the symbols and their values relate to Table 1 and the figure captions.

In figure 4 the availability model derived for the link protection scheme is shown. Blocks representing links and protecting paths are depicted in this figure. Notice that nodes are not protected in this scheme. Link protection provides redundancy for the links between the nodes whilst the nodes themselves remain unprotected.

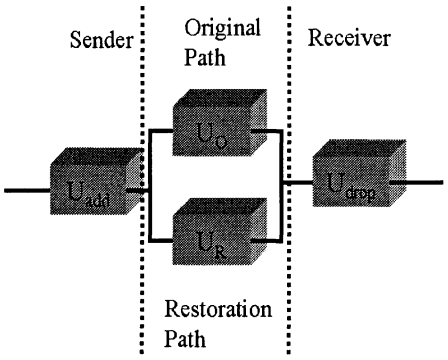
Figure 5 shows the availability model derived for the path protection scheme. The working and protecting paths are indicated in this figure. In this scheme the protecting path provides the redundancy for all equipment located on the working path.

In figure 6 the availability model derived for the SHR is presented. Notice, that for SHR there are two ways reconfiguring the ring, i.e., one in the case of span hit (SH), after the ring has reconfigured in order to isolate the broken fiber, and second in the case of node failure after isolation of the failed node (see figure 2)



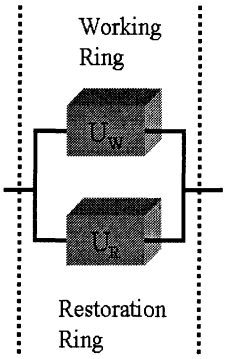
$U_{Add} = U_C + U_T$, $U_{Drop} = U_C + U_R$,
$U_1 = U_2 = \ldots = U_{n-1} = LU_{FC} + (L/50 - 1)U_{OFA}$,
$U_{1p} = U_{2p} = \ldots = U_{(n-1)p} = L_pU_{FC} + (L_p/50 - 1)U_{OFA}$,
(n-1)L = distance between the sender and receiver
where:
n is the number of nodes inclusive the sender and receiver
L is the length of fiber link between nodes and Lp is the length of protection fiber

Figure 4. *Reliability block diagram for connection with link restoration*



$U_{Add} = U_C + U_T$, $U_{drop} = U_C + U_R$
$U_O = (n_O-1)LU_{FC} + (n_O-2)U_{node} + (L/50 - 1)(n_O-1)U_{OFA}$,
$U_R = (n_R-1)LU_{FC} + (n_R-2)U_{node} + (L/50- 1)(n_R-1)$
where:
n_O and n_R are the number of nodes in the original respective restoration path inclusive the sender and receiver
L is the length of fiber between nodes

Figure 5: *Reliability block diagram for connection with path restoration*



$U_W = nLU_{FC} + nU_{node} + n(L/50 - 1)U_{OFA}$
$U_R = U_{R(SH)} = (n-1)U_{FC} + (n-1)(L/50 - 1)U_{OFA} + nU_{node}$
in the case of span hit (SH)
or
$U_R = U_{R(NF)} = (n-2)U_{FC} + (n-2)(L/50 - 1)U_{OFA} + (n-1)U_{node}$
in the case of node failure (NF)
where:
n is the number of nodes on the ring
L is the length of fiber between the nodes

Figure 6. *Reliability block diagram for SHR*

7 Reliability requirements

Today's requirements for network reliability depend on the market and the applications we are considering. For voice circuits the requirements are not as stringent as for data traffic. With new services being offered where voice and data applications merge (MPLS, VoIP) the distinction between these two types of traffic becomes

less clear. No general standards exit for network reliability and it is up to the network operators themselves to define to which standards their networks are to be designed.
However, some de facto standards do exist though they are from a market point of view. Operators talk about 5 nine service, which is service where the availability is guaranteed to be greater than 99.999%, or there is Verizon's IntelliLight offering 99.9994% availability. 99.999% availability corresponds to a connection down-time of no more than 6 min/year! Most operators will offer a suite of services, with different levels of QoS or availability. This paper considers three levels of availability being 99.999%, 99.9994% and 99.99%

8 Results and discussion

There are a variety of measures and functions of reliability that can be derived using our general models. In this paper we consider asymptotic (steady-state) connection/network availability (A), defined as the probability that, in the steady state, the connection/network is available. From reliability point of view there are a number of limiting factors for expanding a network size. In this work we concentrate on efficiency of a reparation process and more specifically on fiber cable MTTR that is probably one of the most critical parameter for the network scalability.

We keep the number of nodes constant and vary MTTR for optical cables in order to investigate:

1. The maximum network size satisfying the availability requirements presented in Section 7.
2. The network/connection availability as a function of MTTR for optical cables.
3. The probability of a second failure during the reparation time.

Our calculations are based on the reliability data given in Table 1 (see Section 4).

8.1 Limitations on the network size

In this Section we study the limitation on the network size due to cable MTTR. We show the results derived for three values of cable failure rate: 1, 3, and 5 cable breaks per year in a 1,000km network that corresponds to 114 FITs /km, 342 FITs/km, and 570FITs/km.

8.1.1 Mesh network with a link protection

The diagrams in Fig. 7 show the dependence of maximum distance - between nodes in mesh networks with link protection - on allowable cable MTTR for three connection availability requirements: 99.9994%, 99.999%, and 99.99%. Notice that for a connection availability of 99.9994% and 99.999% the largest number of nodes in one connection is 2, i.e., n = 2, giving point-to-point topology. This is because we assumed the node unavailability equal to 10-5 (see Table 1), which exceeds the allowed connection unavailability and prohibits having any nodes between sender and receiver.

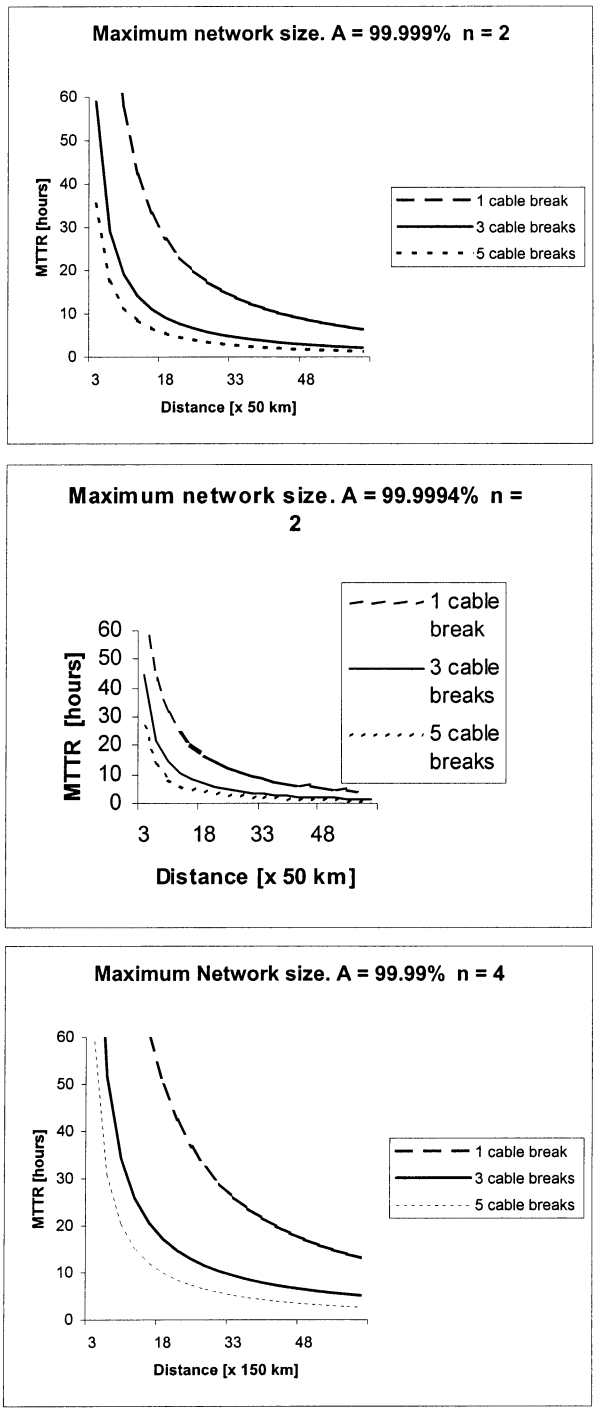


Figure 7 *Maximum network size for mesh network with link protection as a function of MTTR for different cable failure rates.*

Figure 7 shows that for the link protection scheme allowable network size increases by a factor of about 6

when going from an availability of 99.999% to 99.99%. Availability requirement of 99.99% allows for having nodes between the sender and the receiver (i.e. increasing the number of nodes in a connection). The dependence of the allowable network size on cable MTTR is strong. Our results show that for small values of cable MTTR (in this case for MTTR below 5 hours) even small variations of MTTR can give a huge difference in the maximal network size. In our example, for cable MTTR below 2 hours the network size becomes practically unlimited. Conversely, for large MTTR the network size has to be quite small in order to satisfy availability requirements. Any means of reducing cable MTTR is important in ensuring a network above a certain size with a guaranteed availability.

8.1.2 Mesh network with a path protection

Figure 8 shows that compared with the results for networks with link protection the network size can be larger. This difference lies in the protection of nodes, i.e. link restoration provides redundancy for the links between the nodes while nodes themselves remain unprotected. For path protection the protecting path gives redundancy for all equipment located on the working path.

The same dependence of allowable network size on cable MTTR is seen, with the clear conclusion that even with path protection cable MTTR is a very important parameter in network planning.

8.1.3 Self-healing ring (SHR)

Figure 9 shows that the allowable network size for this topology depends on the cable MTTR in a similar way as for the mesh networks with link and/or path protection. As for the self-healing mesh networks, cable MTTR is an important factor in particular for networks located in areas with high cable failure rates.

8.2 Network/connection availability

Figure 10 shows a comparison of network/connection availability between the three network types with built-in redundancy described in Section 3, i.e. SHR and mesh with link and path protection, as a function of cable MTTR for a fixed network size and for cable failure rate of 570 FIT/km.

For cable MTTR higher than 5 hours, networks with link protection performs much better than the others. Protection for each link makes that this type of network can survive multiple simultaneous cable breaks as long as there is not a simultaneous failure of working and protecting link. Thus, the probability of a simultaneous failure of both working and protection links is lower than the simultaneous failure of working and protection path. However, for very small values of MTTR, the link protection scheme is characterized by a lower availability than for SHR and path protection. With these very small values of MTTR – when cable breaks are repaired very fast - the reliability of the unprotected nodes dominate, and link protection becomes less secure than path protection and SHR.

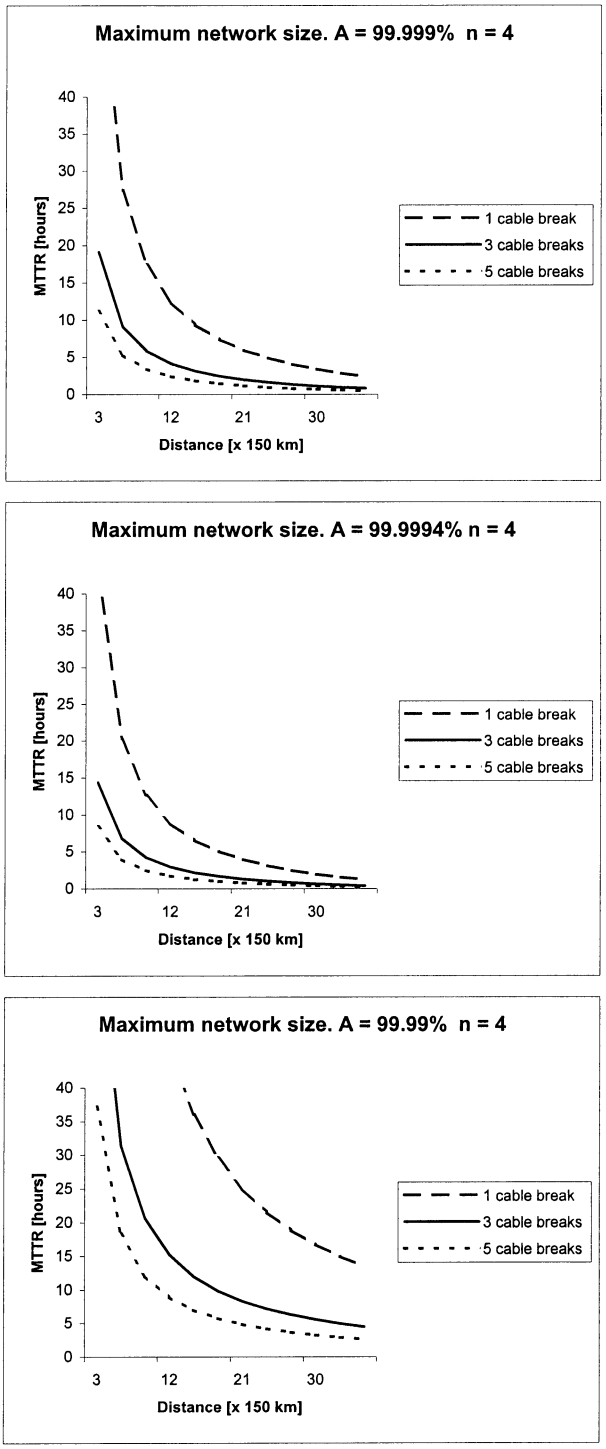


Figure 8: *Maximum network size for mesh network with path protection as a function of MTTR for different cable failure rates.*

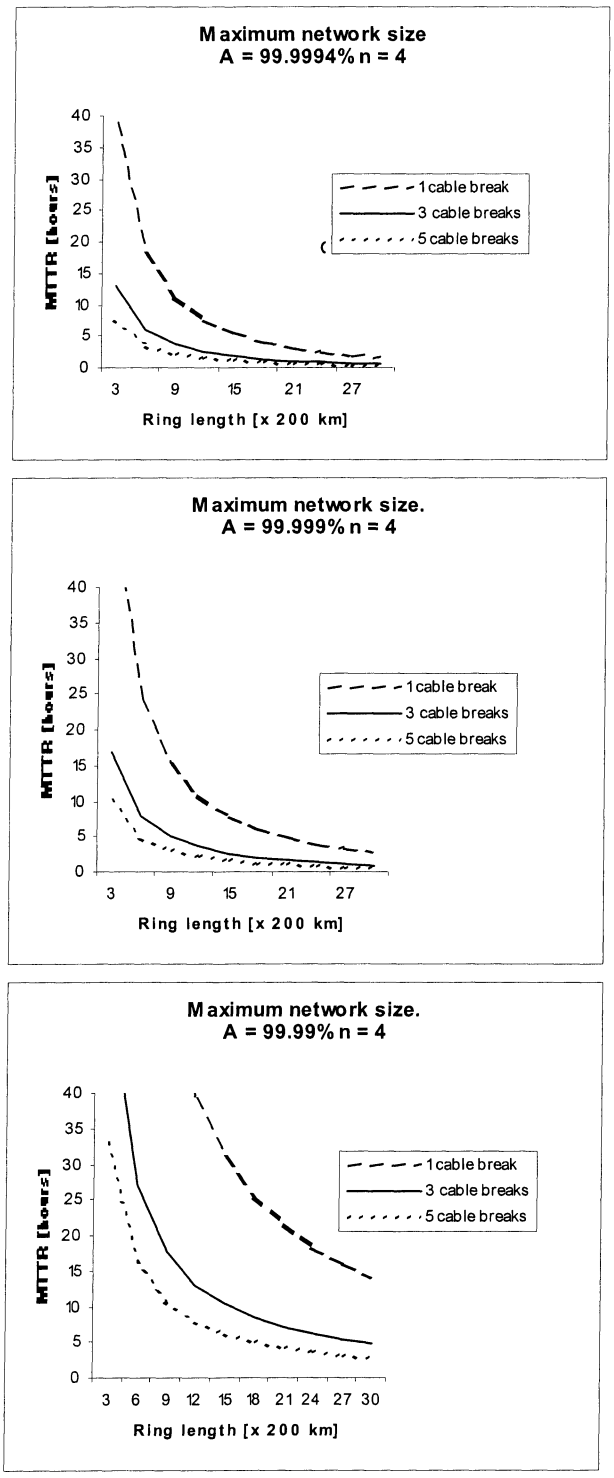


Figure 9: *Maximum network size for self-healing ring network as a function of MTTR for different cable failure rates.*

The availability of network/connection drops rapidly with cable MTTR for all three types of networks. If we assume that the cable MTTR is reduced from 10 hours to 4 hours, the availability of a self-healing ring network improves from 99.989% to 99.998%. This may allow the network operator to offer at least 99.99% availability.

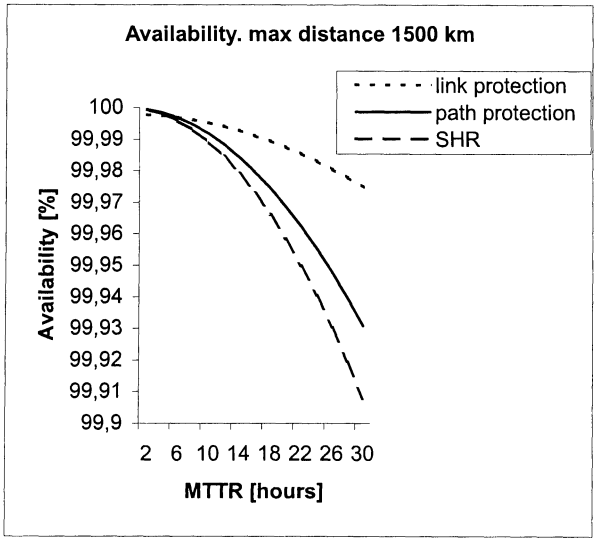
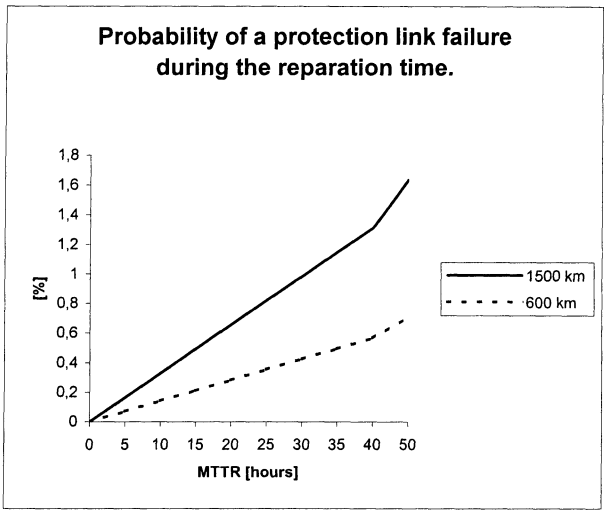


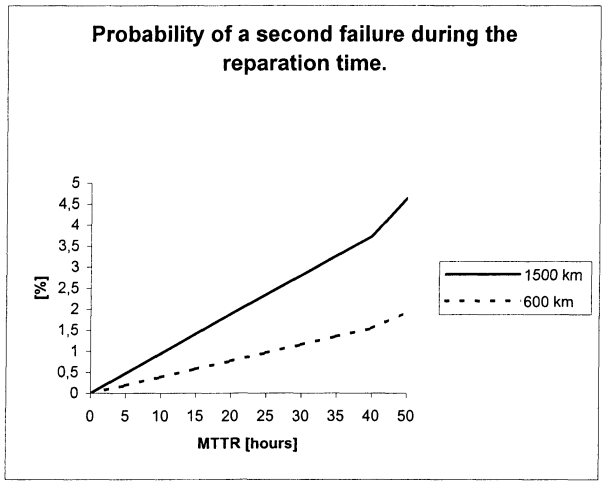
Figure 10: *Network availability as a function of MTTR for the three different protection schemes for a maximum allowable network size of 1,500km*

8.3 Failure probability of a protection faciliy

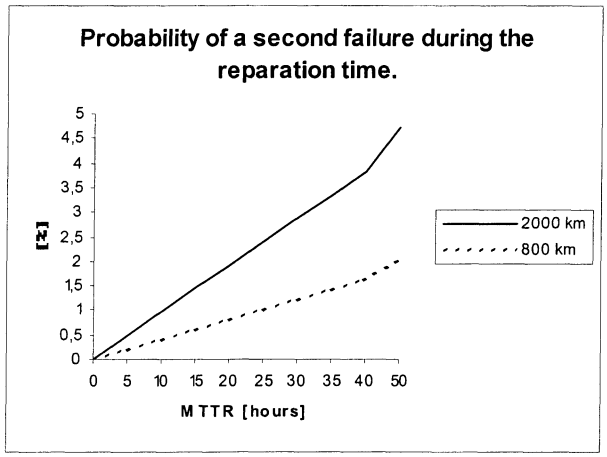
The diagrams in figure 11 (a) shows the probability of a protection link failure during the time the original link is repaired as a function of cable MTTR. A break of the protection link (not necessarily the second failure) will cause the failure of the connection. Figure 11(b) and (c) presents the probability of a second failure during the reparation time as a function of cable MTTR. In this case, the second failure will render the network inoperable.



(a) Link protection



(b) Path protection



(c) SHR

Figure 11: *Probability of protection facility failure during the reparation of the original facility of first failure for the different protection schemes.*

9 Conclusions

With the vast increase in possible link capacity through recent technological achievements, and the simultaneous increased focus on Quality of Service as a differentiating parameter within the telecommunications industry, network operators have begun to offer network availabilities in the order of 99.999%. Networks architectures are deployed that have built-in redundancy at least to a first degree, and these structures offer very robust service. The availability requirements are in most cases at least theoretically met, when all factors except cable MTTR are considered. Including cable MTTR as is done in this paper shows that most networks offering certain levels of availability may not meet these standards.

The availability models presented in Section 6 are general and can be applied to any network and any reliability data. Applying these models to different types of networks with built-in redundancy has shown in all cases, that cable MTTR is a very crucial parameter in network planning. We have shown that cable MTTR below 2 hours allows for practically unlimited distance between the users while cable MTTR above 20 hours gives hard restrictions on the network size.

The importance of MTTR in network planning is in many cases not taken into account, since the protection is implemented to alleviate the effects of cable failures and other failures in the first place. However, as shown in this paper, with today's availability requirements (such as 99.999% uptime), the cable MTTR has to be considered.

References

[1] L. Wosinska and T. Svensson, *Analysis of Connection Availability in All-Optical Networks*, in Proc. NFOEC'2000, September 2000

[2] European project COST 239: "Ultra High-Capacity Optical Transmission Networks"

[3] Tsong-Ho Wu, *Fiber Network Service Survivability,* Artech House, Inc. 1992.

[4] F.E. Ross, *FDDI-A tutorial*" IEEE Commun. Mag., vol.24, pp 10-17, May 1986.

[5] Bellcore GR-418-CORE, Issue1, *Generic Reliability Assurance requirements for Fiber Optics Transport Systems*, Dec.1997

[6] L. Wosinska and R. Holmstrom, *Reliability performances of large capacity multiwavelength optical cross-connects. Comparison between electronic and optical switch core*, in Proc. Photonics in Switching, PS'99, Santa Barbara, CA, July 21-23, 1999.

[7] M.J. Van Tuyl, *Building a DWDM Based Long Distance Network with RBOC Parts,* in Proc. NFOEC'99, October 1999

[8] Ericsson, ERION NETWORKER

[9] T. Svensson and B. Hermansson, *Reliability of an unprotected optical ring network*, in Proc. SPIE, vol 3861, pp 80-91, Sept 20, 1999

Wide Area Measurements of Voice Over IP Quality Revisited

Ian Marsh[1] and Fengyi Li[2]

[1]Swedish Institute of Computer Science, Kista S-164 29, Sweden. E-mail: ianm@sics.se.

[2]KTH IMIT, Kista S-164 40, Sweden. E-mail: d97-fli@nada.kth.se.

Abstract

Time, day, location and instantaneous network conditions largely dictate the quality of Voice over IP calls. In this paper we present results of VoIP quality mainly in terms of packet jitter and loss. This work revisits our objective of measuring VoIP quality in 1999 enabling us to make comparisons with the results we obtained three years ago. It also allows us to improve the measurement infra-structure we used previously through adding new functionality plus using more test sites. Furthermore we present findings on the influence of parameters such as packet size, asymmetry, silence suppression and location as well as much more comprehensive loss and jitter results. Our results conclude that the quality of Voice over IP sessions have improved over the last three years. In spite of this fact however, achieving acceptable *global* Voice over IP quality still remains a challenge.

1 Introduction

No standardised scheme exists for monitoring VoIP quality. It is well known that VoIP users of the Internet are susceptible and sensitive to quality changes, causing them to abandon calls when the quality deteriorates below an acceptable level [1]. The quality of VoIP sessions can be often quantified by packet delay, loss and jitter.

A wide area monitoring scheme, managed over multiple domains, would be beneficial to allow "administrators" to adjust resources for users of an Internet telephony service. Therefore the objective of this work is to describe a VoIP quality monitoring scheme, summarise results we have obtained, plus motivate how to chose certain sites, packet sizes, sampling frequencies and the use of silence suppression. We performed the measurements continuously over a number of weeks, providing us with more than 10000 simulated phone calls. It should be stressed that our work concentrates on the data transport of VoIP sessions, it does not include any call signalling in the measurements.

Using off-line techniques to probe and monitor the Internet for quality changes is non-trivial. By off-line we mean periodically sending pre-recorded conversations and measuring their received quality. It is non-trivial as exact call durations are not well known when there is no payment structure. Also no restrictions are made on the voice coding thus resulting in non-standard packet sizes being used, particularly by commercial vendors. Our contribution in this area is to motivate where choices of measurement parameters such as packet size, probe duration, network asymmetry can affect the results, in addition to collating a large repository traces from widespread sites. We also state the quality of connections in 2002 and compare them with those we gathered in late 1999.

This paper starts with a brief description of the current measurement infra-structure in Section 2. We show the locations of the hosts used now and those used three years ago. Also included are details of the sample file we used to probe connections between sites and a summary of data we gathered as a result of this probing. Our measurement efforts form the main body of this paper and are presented in Section 3, with particular attention to loss and jitter measurements. The related work section compares results obtained by other researchers in similar efforts and ours. Finally, we round off the paper with conclusions and a short discussion in Section 5.

2 Simulating IP Telephony Calls

Our method to measure call quality is to send pre-recorded calls between participating sites. Through modification of our own VoIP tool (called Sicsophone) we probe the intervening connections for variations in VoIP quality. We have carefully chosen nine sites, with large variations in both Internet hops and geographic distance to obtain a wide cross-section of distributed points. The location of the sites is shown in Figure 1.

One of the major changes we made since the last measurements [2] was to fully inter-connect the hosts we had available to us. Nine hosts connected to each other allows us eight connections to the other hosts giving a possible 72 interconnections. This partly accounts for the much larger amount of traces we were able to obtain. In almost all the connections tested we initiated a bidirectional call, simultaneously joining two parties. This is also an improvement over measurements made in 1999 where most sessions were only unidirectional. This scheme is also responsible for the large number of sessions we have.

We found that ten (of the 72) routes gave an unequal number of hops between the same hosts in opposite direc-

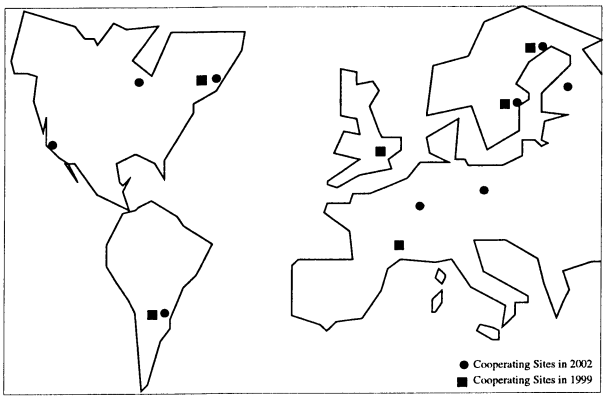


Figure 1: Locations of Test Sites

tions. Route asymmetry can be troublesome for IP telephony as one party might be experiencing different quality than the other, possibly causing one party to terminate the connection. We could not test the full mesh for asymmetry due to restrictions on the use of traceroute and firewalls. The frequency of calls was chosen to be one per hour between each cooperating sites, due to load considerations on remote machines. For most of our measurements we found this sufficient, especially when repeated over several weeks. In some cases, we performed simultaneous tests to assess whether schemes such as silence suppression or packet size had an impact. In these cases four or more connections were conducted in parallel. Table 1 summarises the call characteristics in the top half and the data we obtained from it in the bottom half[1]. We purposely did not change the trace file used in 1999 to keep comparisons simple and point out that other researchers have investigated the affect of call duration on quality (Section 4).

Table 1: Used and Obtained Data

Used Data	Details
Call duration	70 secs
Silence Sup.	2043 packets
No Silence Sup.	3653 packets
Coding	8 bit PCM
File Size	584480 bytes
Obtained Data	
Total Trace Files	10288
Bidirectional Tests	3742
Differing Packet Size	2491
Suppression Tests	1251
Measurement Duration	8 weeks

[1] Available from http://www.sics.se/~ianm/COST263/cost263.html.

3 Results

3.1 1999 and 2002 Differences

This study is a follow up of work we started in 1999 [2]. We concluded then most calls were completed with a loss percentage less than 5% and a network delay below 150 ms. These results still hold in Spring 2002. For the hosts outside South America, all but one exhibited a delay less than 100ms. The average one-way delay between Argentina and the other sites is 133ms. Broadly speaking we have seen an improvement in the quality of VoIP sessions over the *same* connections. Table 2 shows the delay, jitter and loss for all the traces we obtained and their standard deviations in brackets.

Table 2: 1999 and 2002 VoIP Quality Differences

Quality	1999	2002	% Diff. (+/-)
Jitter	45.1ms	22.6ms ($\pm$13.7)	-50.0%
Loss	1.2%	0.5% ($\pm$2.6)	-58.3%
Delay	115 ms	84.95ms ($\pm$44.85)	-26.1%

The remainder of the results presented refer to our measurements taken in 2002.

3.2 Loss

Generally the measurements produced very good quality. 75% of the conversations showed zero percent loss and over 96% with a loss rate less than 10%. We quote 10% as speech coding schemes with tolerance to lost data, forward error correction mechanisms and human tolerance. Figure 2 shows a typical loss pattern over one day for two different connections, the x-axis starts from 0:00 up to 24:00. Note we use our local time (Swedish) to be able to plot the two on the same figure. In relative terms the loss differences is large but both are below 0.5% packet loss. Connections to and from Europe/USA and South America were slightly better than we reported previously with an average loss rate of just over 3%.

Figure 3 shows the consecutive losses for all of the traces. The y-axis is a $\log_{10}$ scale. The majority of losses are one to four consecutive losses. We also observed some very long loss sequences almost inevitably caused by temporary outages at either the sender or the receiver. From examination of the trace files they were nearly always at the end of the session resulting in no more packets being sent or received. We therefore cannot attribute these losses to dropped packets in inter-mediatory routers. The number of files exhibiting these large losses is small compared to the 10000 files we collected, and 40-50 packets is still only 1.5% of the size of the sample call.

3.3 Jitter

Most traces exhibited low jitter values. The average across all files sites was 22.6 ms ($\pm 13.7ms$). Figure 4 shows a

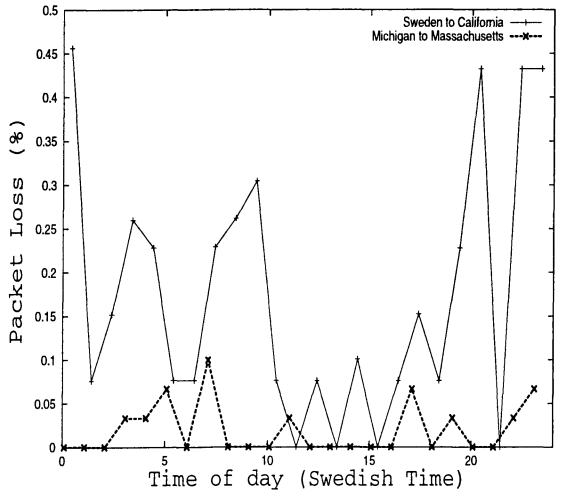


Figure 2: 24 hour Loss in 2002

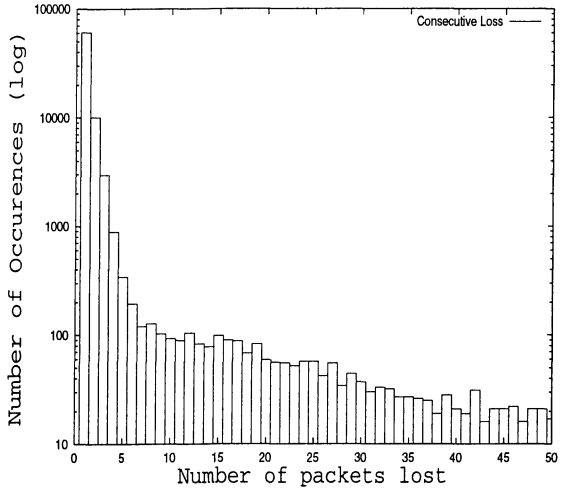


Figure 3: Consecutive Loss

jitter plot for four of the connections. The plot shows three columns, the minimum, average and maximum jitter for a connection. The first labelled site is the sender and the second the receiver (i.e. from Finland to Sweden). The jitter values are an average of all the sessions we recorded. None of the connections were tested less than 400 times, approximately 700 was typical. In the rightmost two plots the minimum jitter was so small that it is barely noticeable. The difference *to* the South American site and the others however is clearly evident. This message is repeated in the next experiment we conducted.

Figure 5 shows the sites grouped by location. In this plot we show the maximum jitter experienced by each chosen connection. The key is ordered about the plot, Massachusetts to Michigan is the first column and so on for all 16 data sets. Most columns contain between 100 and 300 data points. The dispersion of the data is the facet we want

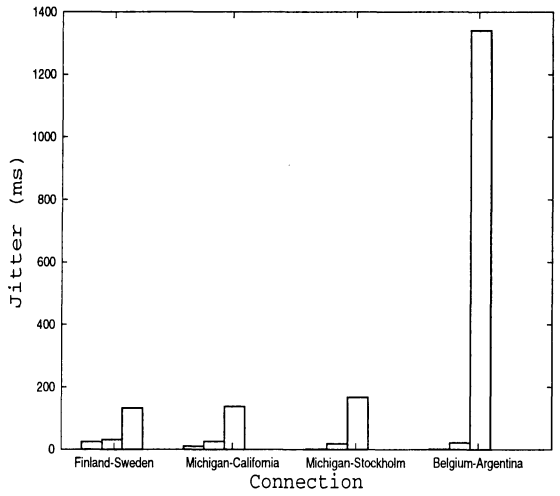


Figure 4: Jitter for Selected Sites

to highlight. Quite clearly the message of poor quality *to* South America is repeated in this plot, interesting to note is the maximum jitter *from* Buenos Aires is clearly much less than to it. Unfortunately the host in Argentina was one which we could not perform a traceroute from it.

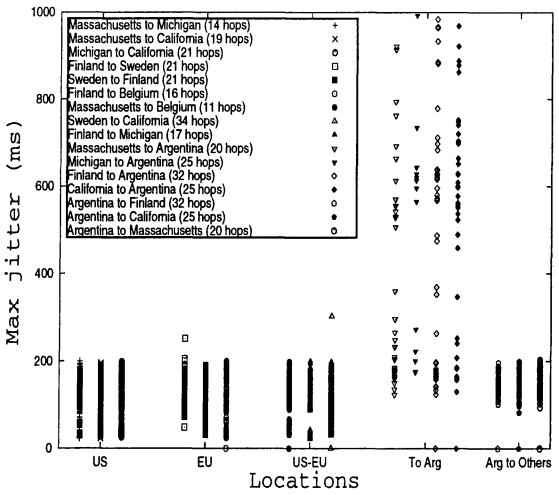


Figure 5: Maximum Jitter per Location

3.4 Network Asymmetry

By network asymmetry we mean the connections from one site to another is not the same in the opposite direction. This can result in differing delays, loss rates and jitter for data on the Internet [3]. In our previous work we concluded that network asymmetry could play a significant role in the perceived quality of one of the callers. We found that on average there was an asymmetry. However it is not so serious in terms of VoIP quality as reported for

the traces we compared. The loss values were up to 0.5-1.25% different for calls taken at the same time in opposite directions. An example is shown in Figure 6.

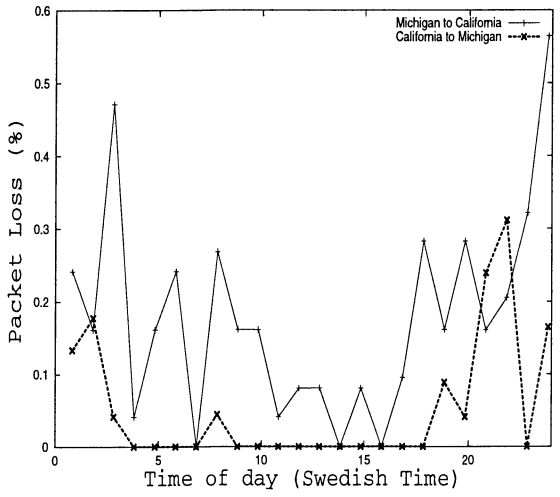


Figure 6: Route Asymmetry

3.5 Silence Suppression

Measuring loss and jitter were not the only goals of this work. In our previous measurements we did not address the effect of silence suppression on measuring call quality. Not using silence suppression (or a sample file with no silence periods) enables more sampling of the network quality, as packets are sent even when the sender is silent. The cost is however more load on the network. Table 1 shows that only 56% of the packets need to be sent using silence suppression, a significant saving on a congested network. Figures 7 and 8 show the jitter of two simultaneous calls at the receiver, the top plot not using silence suppression and the bottom one utilising it. The differences are small, slightly more jitter is experienced by the call using silence suppression, indicated by the spikes, we can attribute this to a slight timing mismatch of the two calls, as the packets are not perfectly interleaved, possibly allowing data packets to be inserted or the streams to interfere with each other. Our recommendation is not to use silence suppression where no loss is being experienced, and use silence suppression if losses are being encountered. Loss in itself was not affected when using silence suppression over not using it (packet size 160 bytes).

3.6 Packet Size

Another factor is the choice of packet size. Finding the optimal size for measurement partly depends on the applications that will use the network in an operational setting. Different codings, applications and standards can dictate the final choice of packet size. Some applications are even

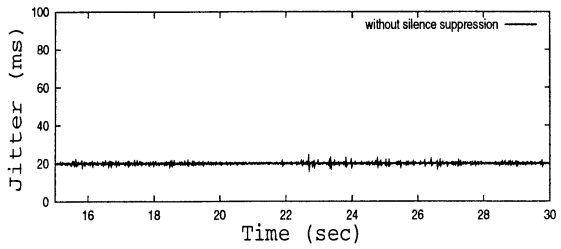


Figure 7: Without Silence Suppression

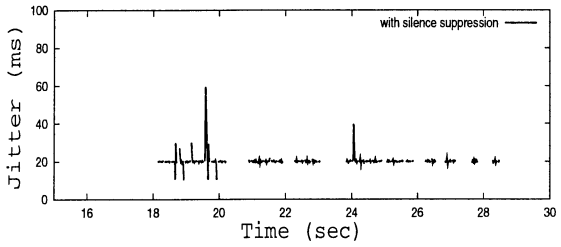


Figure 8: With Silence Suppression

capable of selecting different coding styles resulting in different packet sizes. In order to assess if a correlation exists between packet size and the loss rate we ran several parallel experiments with packet sizes ranging from 160 to 1280 bytes. Our conclusion is that no noticeable difference in loss was experienced. Figure 9 shows the loss rates between using 160 and 1280 bytes over a 24 hour period. One drawback with using large packets is less data is produced for sampling due to larger chunks, but the overall load is lower due to less space used for each header (20 IP + 8 UDP and 12 RTP = 40 bytes).

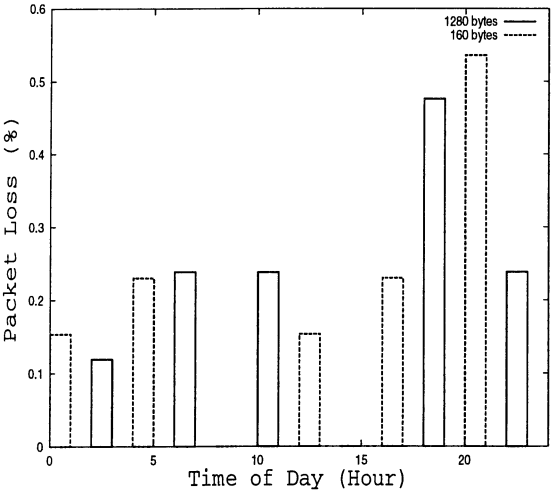


Figure 9: Effect of Packet Size

4 Related Work

Bolot and Hardman looked at consecutive loss, for a FEC scheme [4] [5]. They concluded that the number of consecutive losses is quite low, Bolot stated that between one and five at 8am and between one and ten at 4pm. Our results (shown in Figure 3) are in agreement with these findings. We also have seen traces with more losses, but we had many more traces and can attribute some losses due to reasons other than buffer overflow in routers. Maxemchuk measured both loss and delay variation for intra-state connections, within the USA and international links [6]. Their conclusion was the quality depends on the length of the connection and the time of day. We did not try different connection lengths but found variations during a 24 hour cycle (shown in Figures 2 and 6). It is worthy to point out our loss rates were considerably less than theirs (3-4%). Dong Lin had similar conclusions in [7], stating that in fact even calls within the USA could suffer large jitter delays. Her results on packet loss also agree with those in [4], which is interesting, as the measurements were taken four years later. Her thesis also contains results on trying different packet sizes and concludes that loss patterns were almost indistinguishable for packet sizes from 5 to 512 bytes. We also did not observe very differing losses for packet sizes of 160, 640 and 1280 bytes, as shown in Figure 9.

5 Conclusions and Discussion

Our second measurement attempt took in more hosts, nine rather than six and more importantly connected them in a full-mesh configuration. We now have over 15000 simulated calls (10000 were used in the preparation of this paper) compared to several hundred in 1999. Performing measurements on this scale is not an easy task. Different access mechanisms, firewalls, NATs even non super-user access complicates the work to obtaining measurements.

Voice over IP is still feasible in 2002, despite the dramatic rise of WWW traffic. The majority of calls we tested exhibited excellent quality; the vast number of calls completed with less than 5% packet loss and under 50ms jitter. Compared to 1999 we have seen an improvement in quality of telephony communication using the Internet. The delay has improved by 26% and the jitter and loss by over 50%. We can only attribute this Internet Service Providers upgrading and improving their networks. We have seen higher network asymmetry, measured in hops, between the same sites we used in 1999 and 2002, but with the exception of one host has not significantly affected our results. We still see a disparity between hosts in the EU countries and the USA and those outside these areas. Despite our careful choice of sites in 1999 and 2002 we still see a stark separation into two categories. We therefore plan to investigate this further and have obtained access to sites in Turkey and New Zealand. We have also observed factors such as packet size and the use of silence suppression do not have a significant impact of quality of Internet connections.

References

[1] D. De Vleeschauwer, J. Janssen, G. H. Petit, and F. Poppe. Quality bounds for packetized voice transport. *Alcatel Telecommunications Review*, pages 19–24, 1st Quarter 2000.

[2] Olof Hagsand, Kjell Hansson, and Ian Marsh. Measuring internet telephone quality: Where are we today ? In *Proceedings of the IEEE Conference on Global Communications (GLOBECOM)*, Rio, Brazil, November 1999. IEEE.

[3] Vern Paxson. End-to-end routing behavior in the internet. In *SIGCOMM Symposium on Communications Architectures and Protocols*, Stanford, California, August 1996.

[4] J. Bolot, H. Crepin, and A. Garcia. Analysis of audio packet loss in the internet. In *Proc. International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, Lecture Notes in Computer Science, pages 163–174, Durham, New Hampshire, April 1995. Springer.

[5] Vicky Hardman, Angela Sasse, Mark Handley, and Anna Watson. Reliable audio for use over the internet. In *Proc. of INET'95*, Honolulu, Hawaii, June 1995.

[6] N. F. Maxemchuk and S. Lo. Measurement and interpretation of voice traffic on the internet. In *Conference Record of the International Conference on Communications (ICC)*, Montreal, Canada, June 1997.

[7] Dong Lin. Real-time voice transmissions over the internet. Master's thesis, Univ. of Illinois at Urbana-Champaign, 1999.

GEMS and SMART — Swedish pioneers in state-of-the-art electromagnetics

J. Gustafsson[1], U. Thibblin[2]

[1] Saab Avionics AB. E-mail: jonas.gustafsson@avionics.saab.se.
[2] Saab Avionics AB. E-mail: ulf.thibblin@avionics.saab.se.

Abstract

The GEMS (General ElectroMagnetic Solvers) project, the first large software development project in Sweden in the field of computational electromagnetics (CEM) is presented. Examples of industrial applications of the resulting software suite are given. The continuation of the project with the SMART programme is discussed together with recent results.

1 Introduction

During the mid 90's it became clear that Sweden no longer had access to state-of-the-art software for electromagnetic simulations. At the same time CEM was becoming a more integrated part of the development process and an enabling technology for developing new platforms with for example low signature. Moreover, other areas of the industry started to have a larger need for advanced CEM tools. Therefore the decision was taken to carry out a study of the future needs of Saab and Swedish industry in general.

It was evident that the area was not, and still is not, mature and that the frontier was moving very rapidly and would continue to do so for many years. Moreover, it was difficult to find state-of-the-art software on the market, in particular source code. Therefore it was decided to develop a new software suite. The final group that started the GEMS project was Saab Avionics, Saab Aerospace, Saab Ericsson Space, Ericsson Microwave Systems, the Swedish Defence Research Establishment (FOI), the Institute for Applied Mathematics (ITM), and the Parallel Scientific Computing Institute (PSCI). The objective was to develop a state-of-the-art electromagnetic software suite mainly addressing the applications radar cross section (RCS) management, electromagnetic compatibility, interoperability, antenna and microwave component design and antenna integration. The final result should be of industrial standard and ready to use.

The methods, algorithms and solvers were mainly developed by PSCI, ITM and FOI while the industrial partners worked with project management, development of pre processing and post processing tools and validation. The GEMS project run between 1998 and 2000. It was funded by the participating partners and by NUTEK and NFFP.

This paper presents the GEMS project, applications of the codes and further developments after the project was finished.

2 The GEMS project

2.1 Requirements

The GEMS project encompassed antenna analysis, scattering analysis, and electromagnetic effects (EME) simulations. The objective was to handle every different type of antenna including interactions with the surroundings. EME involved electromagnetic effects in general such as interference, lightning, emission, and High Intensity Radiated Fields (HIRF).

A fundamental requirement was that an experienced engineer should be able to use the software without too much introduction. Hence, a lot of effort was put into a well structured user interface and informative error messages. On the other hand, in order to keep the flexibility of the program, the choice of methods, the accuracy of the geometrical models, the meshing, etc. was left to the expertise of the engineer using the software.

In short, the goal was to make the GEMS software accurate, versatile, and easy to use. The strategy employed to obtain this was:

- Use specialists in electromagnetics, numerical analysis, and programming.
- Have a close co-operation between the users and the developers in all phases.
- Reuse existing software where possible, but rewrite all codes to get a solid structure and use one program language, namely FORTRAN 90.
- Develop interfaces to CAD programs to allow arbitrary geometries.
- In order to tackle a wide range of problems, use several different solvers.

- Define realistic test cases, covering the industrial needs, early on in the project.

Although the requirements were set out very high, a majority of them were eventually achieved. The basic concept with GEMS and the strategies described above proved to be very successful. Even so, further development is required after the project was completed, as will be described in Section 3.

2.2 Codes

The idea behind the codes was that no one single algorithm can efficiently be applied to all of the required applications. Instead, two hybrid codes were developed; one in the time domain (TD), with FDTD as the core, and one in the frequency domain (FD), with a boundary element method as the core. Both hybrid codes were parallelized using message passing (MPI) and run on simple PC clusters as well as massively parallel super computer systems.

The multiblock TD code was based on finite differences (FDTD) on structured grids, hybridised with explicit finite volumes (FVTD) and implicit finite elements (FETD) on unstructured grids. The original approach was to use FVTD to get a body conforming mesh in regions where the cell size was of the same order as in the structured region. Since the method is unconditionally stable, FETD puts no restriction on the time step, and was hence envisaged for regions where fine details had to be resolved. A general grid generation tool was developed in co-operation with the UK based company FEGS Limited, to enable industrial use of the hybrid TD solvers for arbitrary geometries.

The FVTD solver was eventually retired. The main reason for this was stability problems. Artificial dispersion had to be imposed on the solution in order to keep the instabilities in place, but even so stability could not be guaranteed at all times. The explicit nature of FVTD also set quite high demands for the quality of the unstructured grids. A few bad cells can force the user to use a small time step, and thus cause very large execution times. Generally, it is impossible to generate an unstructured grid around a realistic object (e.g. an aircraft) without getting some poor quality elements. Another reason for the failure of FVTD to meet the requirements was the success of FETD, which proved to be more efficient than anticipated by the use of iterative techniques.

In development and research projects such as GEMS there is always an element of uncertainty. This is something that is inherent in the very nature of research and is certainly inevitable. The important lesson to draw from the FVTD experience, is that it is vital to detect the dead end tracks as early as possible. And when evidence become available that perhaps FETD is the way to go instead of FVTD, there must be a flexibility in the project organisation to allow for a rapid transition to the new path.

The FD hybrid code developed in GEMS was based on the Method of Moments (MM), Physical Optics (PO), and the General Theory of Diffraction (GTD). Because of its scaling properties, MM is very expensive for medium and high frequency problems. Therefore, the strategy was a hybrid between MM and the asymptotic methods.

In order to allow interfaces to arbitrary ports, a general numerical wave guide mode solver was developed. The mode solver enabled interfaces to wave guides with arbitrary cross section, coaxial lines and micro strip lines, and was used as to define excitations in both the TD code and the FD code.

2.3 Applications

As already mentioned, the GEMS software suite was intended to be widely applicable, but with focus on antennas, scattering and EME. All types of antennas should be handled, as well as the interaction with complex structures, e.g. aircraft, ships, masts, and buildings. For scattering the focus was on radar cross section computations for complex systems such as aircraft. The EME applications concerned aircraft, cars, ships, antenna systems, and electronic equipment.

Below, two different examples of applications where the GEMS software is used are given. In Figure 1 the results from an antenna computation for a Vivaldi element is shown, and in Figure 2 the currents induced by lightning in the Saab 2000 commercial aircraft is presented.

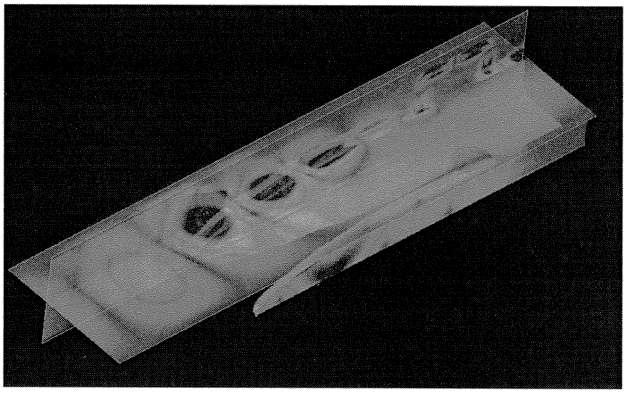


Figure 1. Antenna calculation performed using the GEMS tools.

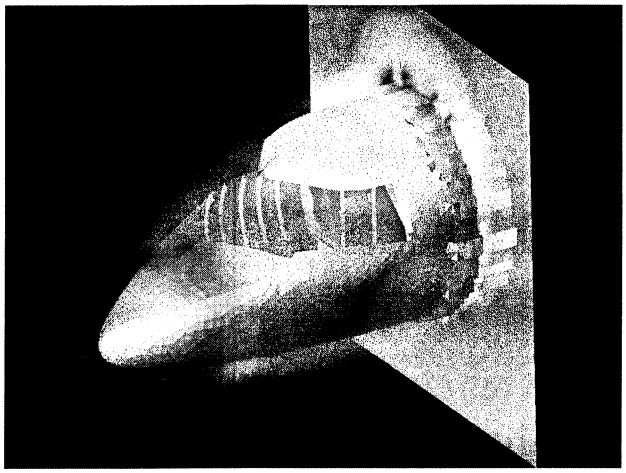


Figure 2. EME computation performed using the GEMS tools.

3 Post GEMS — SMART

The software tools developed in GEMS were aiming at a wide variety of CEM applications. As implied by the abbreviation, the solvers are *general* in the sense that they are able to treat many different type of problems. Upon completion of the GEMS project the need to take the tools one step further became evident. In particular, certain applications are driven by a handful of key requirements that outweigh everything else.

The SMART (Signature Modelling And Reduction Tools) project picks up where the GEMS project left off, but the focus is entirely on scattering (i.e. radar cross section). The organisation of SMART is very similar to the successful concept initiated with GEMS; numerical expertise within the Swedish academia develops the solvers, while the pre processing, post processing, and validation is handled by the industry. The partners in SMART are Uppsala University, Chalmers, and KTH — all participating through PSCI — on the academic side, and the industrial partners are Saab Avionics and Saab Aerospace. The project started in July 2001 and it is a three year programme that is funded by the participating partners and by NFFP and FMV.

2.1 Requirements

The driver behind the SMART project is the increasing interest for scattering computations in the military industry; there is a substantial demand for a low signature on the next generation of aircraft, manned or unmanned, as well as other platforms. The contributions from various minor scattering sources such as antennas, sensors, joints, small inlets, etc. will increase in importance when the RCS of the major contributors (e.g. the inlets and leading wing edges) is reduced. New computational models and methods must be developed to deal with these sources. There is also an increasing use of different kinds of radar absorbing materials and structures (RAM/RAS), both dielectric and magnetic ones. These new materials and structures have to be modelled by the computer codes.

Almost all numerical modelling involves performance issues. In computational radar cross section three user requirements are especially tightly linked with the performance of the available hardware and software:

- The frequency of the incident wave and the number of frequencies.
- The number of directions of the incident wave.
- The accuracy of the solution.

In order for RCS simulations to be a feasible tool for industrial problems the run times should not be longer than 24 – 48 hours, and the demands for primary and secondary memory must be within the limits of the hardware at hand. This imposes some restrictions on what is possible to achieve with respect to the three requirements above.

The frequencies of interest in RCS computations are mainly between 1 and 20 GHz. In computational RCS one is, today, almost solely interested in monostatic properties (i.e. the case where the radar source and receiver are collocated). The directions of the incident wave that has to be calculated can be anything from the entire sphere to only a small sector, depending on the threat defined for the object in question. The RCS for one frequency and a single angle is of no practical interest. The number of angles that has to be computed is proportional to the size of the object and to the frequency. The number of frequencies is proportional to the size of the object.

It is also important to remember that accuracy is a performance related issue. For the highest frequencies of interest, say 10 – 20 GHz, asymptotical methods might have to be employed. In that case larger errors due to the physical approximations inherent in the method can be acceptable, if no other alternative solution is possible.

2.2 Codes

The objective of the SMART project is to develop a state-of-the-art software suite for RCS applications, with respect to the requirements given above in section 2.1, in order to compute and optimise the RCS performance of future low observable vehicles.

In order to facilitate this objective, new and improved solvers are required. The frequency domain solvers from GEMS will therefore be further developed. In summary, the solvers involved are:

- The hybrid between Method of Moments (MM) and Physical Optics (PO). Improvements will be made of the accuracy in the PO region and of the overall performance of the hybrid. New materials will be investigated.
- Geometrical Theory of Diffraction (GTD). Extensions of the method for monostatic RCS computations will be made. An MM-GTD hybrid will also be developed.

- The Fast Multipole Method (FMM) will be employed to bridge the gap between the asymptotic methods and the exact ones.

A completely new aspect of SMART, that was not treated in GEMS, is that of optimisation. The objective here is to provide an automatic tool for the reduction of an objects RCS. The different optimisation approaches that can be tested include:

- Gradient methods.
- Genetic algorithms.
- Particle swarm methods.

Initially, gradient methods for shape optimisation will be tested in two dimensions. As a second step optimisation of material parameters will be investigated.

2.3 Applications

The application in SMART is primarily RCS computations of airborne vehicles. The emphasis is mainly on design testing and optimisation, rather than verification of the RCS of a product. Below, examples of RCS applications are given:

- Predict the cross section for parts of or for complete vehicles.
- Verify that the products fulfil the specifications of the customers.
- Optimise new designs.
- Support testing by advise for test set-ups and support for interpretation of results.
- Visualise the physics both in real and in generic situations which leads to a better understanding of the scattering mechanisms, new and better design rules, and innovations.

To get a representative set of applications three different objects are used for the final validation and testing of the tools developed in SMART. The different sizes of the objects ensure that all the aspects of the requirements are taken into account, and various configurations of the objects provide a representative way to test the features of the codes. In Figure 3 and Figure 4 two concept Unmanned Aerial Vehicles (UAV) are shown. The final object, shown in Figure 5, is a conventional full size aircraft.

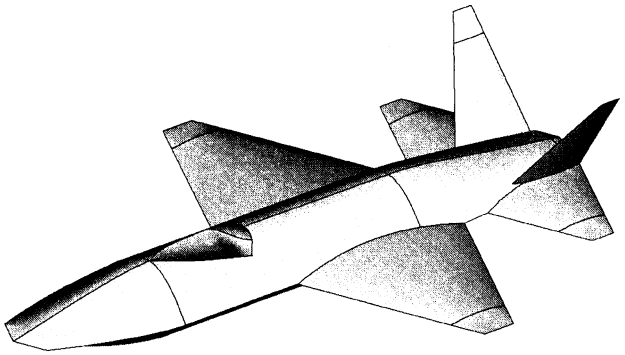


Figure 3. UAV 2.6 x 1.5 x 0.5 m.

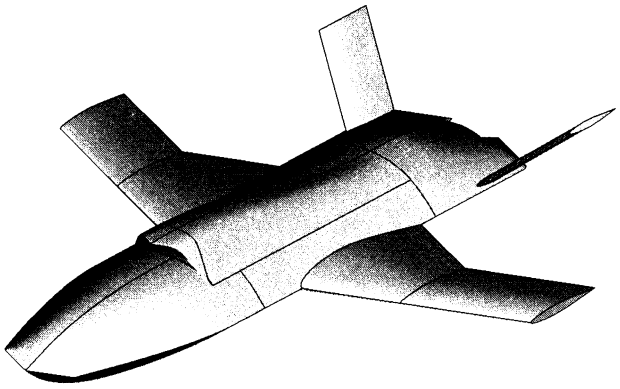


Figure 4. UAV 7.3 x 5.0 x 1.1 m.

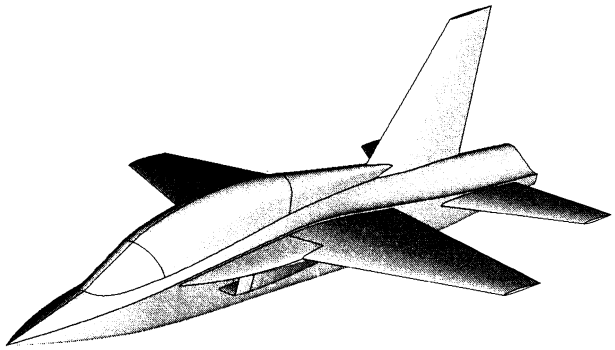


Figure 5. Aircraft 11.0 x 8.0 x 3.4 m.

Finally, in Figure 6 the surface currents, together with the resulting radiation pattern, induced by a wave with a frequency of 3 GHz computed using the SMART multipole solver is presented. This was the first computation with more than one million unknown variables, which is something of milestone for computational RCS in Sweden and for the SMART project.

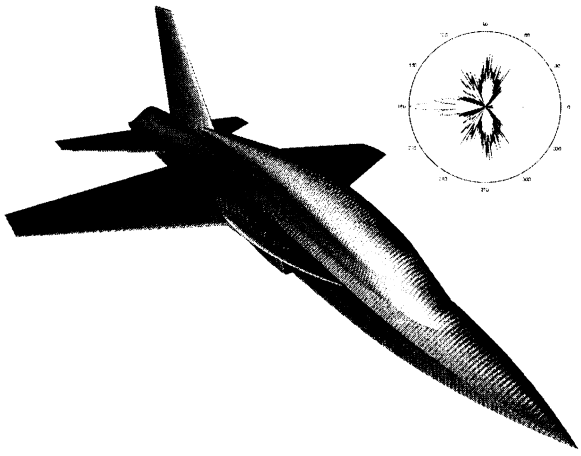


Figure 6. RCS computation performed in SMART.

A hybrid FDFD-BIE approach to two-dimensional scattering from an inhomogeneous biisotropic cylinder

Martin Norgren,

Div. of Electromagnetic Theory, Alfvén Laboratory, Kgl Tekniska Högskolan. E-mail: martin@tet.kth.se.

Abstract

The scattering problem for an inhomogeneous two-dimensional biisotropic cylinder is solved in the frequency domain by means of a hybrid method, in which finite difference equations in the interior region are combined with a mesh truncation in terms of a boundary integral equation. Numerical results for the bistatic echo widths are presented and compared with a reference solutions in the circular cylinder cases and it is found that the method yields more accurate results than what can be achieved with a local absorbing boundary condition.

1 Introduction

A scattering problem involving inhomogeneous media can be solved using a volume integral equation (VIE) that is discretized into a matrix equation by using the method of moments (MoM). However, a drawback with VIE is that the integral operator discretizes into a filled matrix, that becomes very large and exceedingly time consuming to invert if the geometry is several wavelengths in size. An alternative is to use a partial differential equation (PDE) method, in which the differential operator discretizes into a sparse matrix by using finite differences (FD) or by using the finite element method (FEM). However, when using FD or FEM in an open region a radiation condition must be supplemented by truncating the computational domain with an absorbing boundary condition (ABC), which can be either an approximate (local) ABC or an exact (global) ABC. One example of a local ABC is the perfectly matched layer (PML), which has been extended recently to include bianisotropic media [1]. A global ABC is usually formulated in terms of an integral operator on the boundary of the computational domain, which yields that the FD or FEM equations then must be solved in conjunction with a boundary integral equation (BIE). Examples of such combined methods, referred to as hybrid methods, are e.g. FEM-BIE [2] and FDTD-BIE [3]. In this paper, we consider a hybrid FD FrequencyDomain-BIE approach to a two-dimensional scattering problem for an inhomogeneous cylinder of a biisotropic material. Interior to the region containing the biisotropic cylinder, we use finite differences and on the boundary of the region we use a BIE, which reduces to a contour integral equation.

2 The scattering problem

Consider an in free space located biisotropic scatterer (see figure 1) described by the following constitutive relations in the frequency-domain:

$$\boldsymbol{D}(\boldsymbol{r}) = \varepsilon_0 \varepsilon(\boldsymbol{r})\,\boldsymbol{E}(\boldsymbol{r}) + \sqrt{\varepsilon_0\mu_0}\,\xi(\boldsymbol{r})\,\boldsymbol{H}(\boldsymbol{r}), \quad (1)$$

$$\boldsymbol{B}(\boldsymbol{r}) = \mu_0 \mu(\boldsymbol{r})\,\boldsymbol{H}(\boldsymbol{r}) + \sqrt{\varepsilon_0\mu_0}\,\zeta(\boldsymbol{r})\,\boldsymbol{E}(\boldsymbol{r}), \quad (2)$$

where the parameters ε, μ, ξ and ζ are arbitrary functions of the position variable $\boldsymbol{r}$.

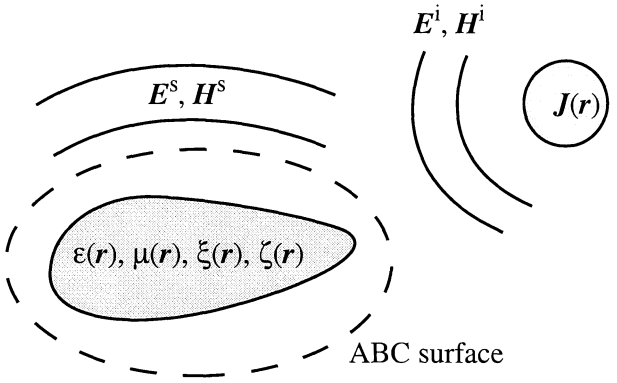


Figure 1: The scattering configuration.

The scattering problem is to determine the scattered fields $\boldsymbol{E}^{\mathrm{s}}, \boldsymbol{H}^{\mathrm{s}}$ when the scatterer is excited by known incident fields $\boldsymbol{E}^{\mathrm{i}}, \boldsymbol{H}^{\mathrm{i}}$, created by an extraneous time-harmonic ($\exp(\mathrm{j}\omega t)$) current density $\boldsymbol{J}$; see figure 1. Substituting the constitutive relations (1) and (2) into the Maxwell's equations, we arrive at the following equations for the total fields:

$$\nabla \times \boldsymbol{E} + \mathrm{j}k_0\left(\mu\eta_0\boldsymbol{H} + \zeta\boldsymbol{E}\right) = \boldsymbol{0}, \quad (3)$$

$$\nabla \times \eta_0\boldsymbol{H} - \mathrm{j}k_0\left(\varepsilon\boldsymbol{E} + \xi\eta_0\boldsymbol{H}\right) = \boldsymbol{0}, \quad (4)$$

where $k_0 = \omega\sqrt{\varepsilon_0\mu_0}$ and $\eta_0 = \sqrt{\mu_0/\varepsilon_0}$.

3 The FDFD equations in the two-dimensional case

The 2D FDFD equations are derived in a similar way as in [4]. The region containing the scatterer is divided into square cells (see figure 2), in which the constitutive parameters are approximated as homogeneous and equal to their true values at the mid points. In the outermost layer of cells, the parameters assume their vacuum values: $\varepsilon = \mu = 1, \xi = \zeta = 0$. The FDFD approximation results in a system of equations for the everywhere continuous z-components of the fields [5]:

$$\mathbf{M}'_{\mathrm{EE}}\bar{E}_{z,\mathrm{inside}} + \mathbf{B}\bar{E}_{z,\mathrm{outside}} + \mathbf{M}_{\mathrm{EH}}\bar{H}_{z,\mathrm{inside}} = \bar{0}, \quad (5)$$

$$\mathbf{M}'_{\mathrm{HH}}\bar{H}_{z,\mathrm{inside}} + \mathbf{B}\bar{H}_{z,\mathrm{outside}} + \mathbf{M}_{\mathrm{HE}}\bar{E}_{z,\mathrm{inside}} = \bar{0}, \quad (6)$$

where the components of the vectors $\bar{E}_{z,\mathrm{inside}}, \bar{H}_{z,\mathrm{inside}}$ are evaluated at the interior FDFD points, indicated with filled circles in figure 2, and $\bar{E}_{z,\mathrm{outside}}, \bar{H}_{z,\mathrm{outside}}$ are evaluated at the extruding FDFD points, indicated with blank circles in figure 2. $\mathbf{M}'_{\mathrm{EE}}, \mathbf{M}_{\mathrm{EH}}, \mathbf{M}_{\mathrm{HE}}$ and $\mathbf{M}'_{\mathrm{HH}}$ are sparsely filled square matrices, describing the FDFD operators at the interior points and the matrix $\mathbf{B}$ describes the FDFD operators on the extruding points.

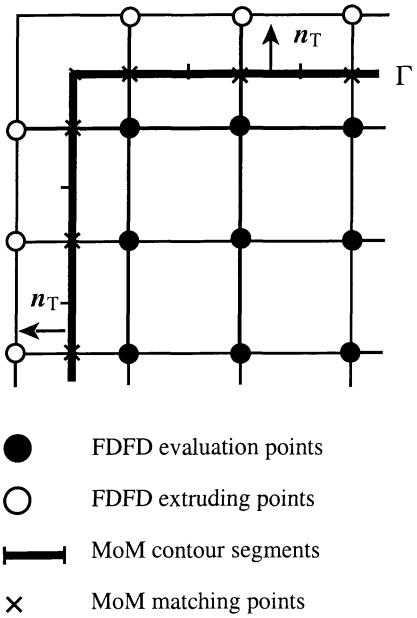


Figure 2: One corner region of the computational domain. Γ is the contour at which the global absorbing boundary condition is calculated.

4 Absorbing boundary condition in terms of a BIE

Inside the contour Γ, depicted with a thick solid line in figure 2, the z-component of the fields satisfy the integral equation

$$\frac{1}{2}F_z(\boldsymbol{r}) + \oint_\Gamma \left[G(\boldsymbol{r},\boldsymbol{r}') \frac{\partial F_z}{\partial n'}(\boldsymbol{r}') - F_z(\boldsymbol{r}') \frac{\partial G}{\partial n'}(\boldsymbol{r},\boldsymbol{r}') \right] \mathrm{d}\Gamma' = F_z^{\mathrm{i}}, \quad \boldsymbol{r} \text{ on } \Gamma, \quad (7)$$

where F_z denotes E_z or H_z, F_z^{i} their incident parts and $G(\boldsymbol{r},\boldsymbol{r}') = -\frac{j}{4}\mathrm{H}_0^{(2)}(k_0|\boldsymbol{r}-\boldsymbol{r}'|)$ is the 2D free space Green's function. After a simple MoM approximation using pulse functions and point-matching at the points indicated with crosses in figure 2, we obtain the following relation (the numerical ABC) between the scattered field and its normal derivative:

$$\mathbf{P}\bar{F}_z + \mathbf{Q}\partial_n\bar{F}_z = \bar{F}_z^{\mathrm{i}}, \quad (8)$$

where $\bar{F}_z, \partial_n\bar{F}_z$ and $\bar{F}_z^{\mathrm{i}}$ are vectors containing the total field, its normal derivative and the incident field, respectively, and where $\mathbf{P}$ and $\mathbf{Q}$ are matrices. To connect the ABC to the FDFD scheme, the elements in $\bar{F}_z$ and $\partial_n\bar{F}_z$ are approximated from the two nearest FDFD points on each side of Γ by linear interpolations and central differences, respectively, i.e.

$$\bar{F}_z \approx \frac{F_{z,\mathrm{outside}} + F_{z,\mathrm{inside}}}{2}, \quad (9)$$

$$\partial_n\bar{F}_z \approx \frac{F_{z,\mathrm{outside}} - F_{z,\mathrm{inside}}}{h}, \quad (10)$$

where h (the mesh parameter) is the side-length of a cell. With (10) in (8), we obtain the following matrix relation between the field values on and adjacent to Γ:

$$\mathbf{D}\bar{E}_{z,\mathrm{outside}} + \mathbf{C}\bar{E}_{z,\mathrm{inside}} = \bar{E}_z^{\mathrm{i}}, \quad (11)$$

$$\mathbf{D}\bar{H}_{z,\mathrm{outside}} + \mathbf{C}\bar{H}_{z,\mathrm{inside}} = \bar{H}_z^{\mathrm{i}}, \quad (12)$$

where the square matrix $\mathbf{D}$ appears to be well-conditioned and hence invertible. Using (12) in (5) and (6), it thus follows that the fields at the interior points are obtained from the following matrix equation:

$$\begin{bmatrix} \mathbf{M}_{\mathrm{EE}} & \mathbf{M}_{\mathrm{EH}} \\ \mathbf{M}_{\mathrm{HE}} & \mathbf{M}_{\mathrm{HH}} \end{bmatrix} \cdot \begin{bmatrix} \bar{E}_{z,\mathrm{inside}} \\ \bar{H}_{z,\mathrm{inside}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}\bar{E}_z^{\mathrm{i}} \\ \mathbf{X}\bar{H}_z^{\mathrm{i}} \end{bmatrix}, \quad (13)$$

where

$$\mathbf{M}_{\mathrm{EE}} = \mathbf{M}'_{\mathrm{EE}} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}, \quad (14)$$

$$\mathbf{M}_{\mathrm{HH}} = \mathbf{M}'_{\mathrm{HH}} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}, \quad (15)$$

$$\mathbf{X} = -\mathbf{B}\mathbf{D}^{-1}. \quad (16)$$

Note that since neither $\mathbf{B}$ nor $\mathbf{C}$ nor $\mathbf{D}$ depend on the properties of the scatterer, the operation $\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ in (16) (and the calculation of $\mathbf{X}$) has to be carried out only once. Thus, for a given mesh and frequency the ABC can be precalculated and stored for further usage when considering different properties of the scatterer, as long as the scatterer is confined within the boundary Γ.

5 Numerical example

In this one example, consider a circular chiral cylinder with the radius $= \lambda$, where λ is the wavelength in free space. The constitutive parameters are chosen to be $\varepsilon = 2.6 - 0.2\mathrm{j}, \mu = 1.5 - 0.1\mathrm{j}, \xi = -\zeta = -\mathrm{j}\kappa = -\mathrm{j}\,(0.6 - 0.1\mathrm{j})$. For the computational region we consider two different mesh sizes, 41×41 FDFD points with spacing $h = 0.05\lambda$ and 81×81 FDFD points with spacing $h = 0.025\lambda$. The numerical results are compared with a reference solution obtained by means of an expansion in cylindrical eigenwaves. The cross-polarized bistatic echo width $\sigma^{\mathrm{E}}_{\mathrm{cross}} = \sigma^{\mathrm{H}}_{\mathrm{cross}}$ for waves impinging in the direction $\phi = 0$ is depicted in figure 3, with the solid lines for the reference solution, the dotted lines for $h = 0.05\lambda$ and the dashed lines for $h = 0.025\lambda$. The numerical results indicate convergence and are in good agreements with the reference solution.

6 Conclusion

The scattering problem for inhomogeneous biisotropic cylinders has been considered. Finite difference equations for the interior region as well as a contour integral equation realizing a global absorbing boundary condition have been derived and implemented numerically. Computationally, the present FDFD-BIE method is much faster than a moment method using a filled matrix of the same size. Hence, the present method can be used for solving 2D scattering problems involving scatterers of intermediate sizes (a couple of wavelengths) in both directions whereas MoM under, the same computation time, only can be used for considerably smaller or thinner structures.

Since our global ABC is unlikely to generate artifacts from the boundary in the solution, the usefulness of the hybrid method as a fast solver in an optimization approach for solving the corresponding inverse problem is of interest for forthcoming research.

References

[1] M. Kuzuoglu and R. Mittra. A systematic approach to the derivation of constitutive parameters of a perfectly matched absorber. *IEEE Microw. Guided Wave Lett.*, 8(9):313-315, June 1998.

[2] X.-Q. Sheng *et al.*. On the formulation of hybrid finite-element and boundary-integral methods for 3-D scattering. *IEEE Trans. Ant. Prop.*, 46(3):303-311, March 1998.

[3] H. Rogier *et al.*. A new hybrid FDTD-BIE approach to model electromagnetic scattering problems. *IEEE Microw. Guided Wave Lett.*, 8(3):138-140, March 1998.

[4] W. Hong *et al.*. Application of the measured equation of invariance to solve scattering problems involving a penetrable medium. *Radio Science*, 29(4):897-906, July-August 1994.

[5] M. Norgren. A hybrid FDFD-BIE approach to two-dimensional scattering from an inhomogeneous biisotropic cylinder. Technical Report TRITA-TET 00-1, Royal Institute of Technology, SE-100 44, Stockholm.

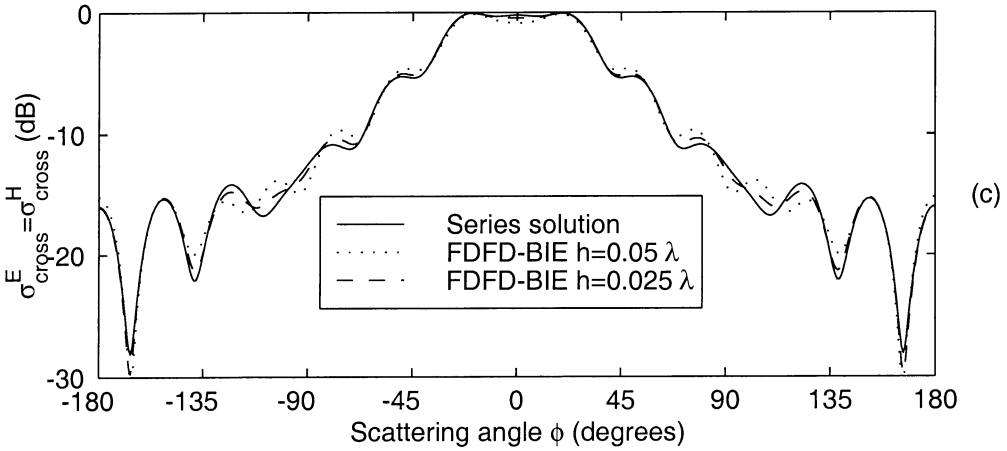


Figure 3: The cross-polarized bistatic echo width for a chiral circular cylinder with the radius $= \lambda$ and the parameters $\varepsilon = 2.6 - 0.2\mathrm{j}, \mu = 1.5 - 0.1\mathrm{j}$ and $\kappa = 0.6 - 0.1\mathrm{j}$. The cylinder is illuminated in the direction $\phi = 0°$.

Dissipation in Frequency Selective Structures

Björn Widenberg

Dept. of Electroscience, Lund University, 221 00 Lund, Sweden. E-mail: bjorn.widenberg@es.lth.se

Abstract

Dissipation in Frequency Selective Structures (FSS) of aperture/slot type is studied. The dissipation in an FSS is due to losses in the dielectric material, and losses due to finite conductivity in the metallic plate. The dissipation in the dielectric medium is taken into account by the complex permittivity. The dissipation on the metallic structure arises both on the plane metallic surface and on the walls of the apertures. The method used to analyze the FSS is based on the mode matching technique and the Finite Element Method (FEM). The apertures are treated as waveguides, and can have arbitrary cross-section. The attenuation and the power losses are calculated for a number of different FSS, and based on these results the performance of an FSS with losses is discussed.

1 Introduction

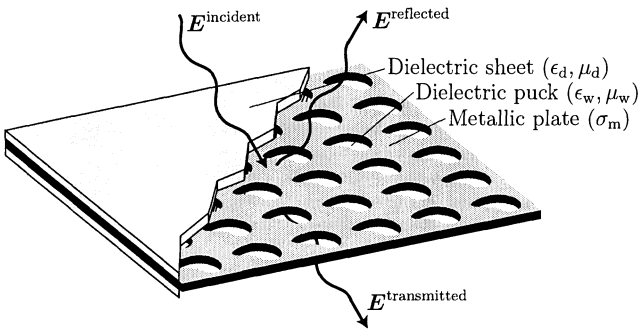


Figure 1: A Frequency Selective Structure (FSS).

In this paper the dissipation of power in Frequency Selective Structures (FSS) of aperture/slot type is studied by a mode matching technique. The dissipation in an FSS is due to losses in the dielectric material, and losses due to the finite conductivity in the metallic plate. The dissipation in the dielectric parts is taken into account by complex permittivities. The dissipation on the metallic plate arises both on the plane metallic surface and on the walls of the apertures, and is modelled by the introduction of a surface resistance.

2 Geometry and Method

The geometry of a simple FSS, that consists of a perforated conducting plate sandwiched between two dielectric slabs, is depicted in Figure 1. The screen can have an arbitrary number of aperture layers and dielectric layers. An aperture layer consists of an electrically conducting plate perforated with a periodic array of apertures. The apertures are treated as waveguides, can have arbitrary cross-section, and can be filled with a dielectric material. The structure is infinite, and the z-axis is assumed to be orthogonal to the surface of the conducting screen.

The method for analyzing the FSS is based on a general mode-matching technique. The FSS is divided into a number of boundaries and uniform layers. The electric and magnetic fields outside the screen and inside the dielectric layers are expanded in tangential plane waves, Floquet modes. The tangential fields inside the aperture layers, are expanded in waveguide modes obtained by the Finite Element Method (FEM). The tangential electric and magnetic fields are matched by the boundary conditions at every boundary surface between uniform sections of the structure, in order to obtain a scattering matrix. For every layer a propagation matrix is calculated. These scattering matrices are cascade coupled to form an overall scattering matrix for the complete FSS. The mode coefficients and fields can be found at every boundary and inside every layer. Extensive discussions on the geometry and the method can be found in [1, 2].

In all examples in the paper, the incident field is a TE-polarized plane wave at normal incidence, and it impinges from the left side. The losses are calculated at the frequency with maximum transmission, and the metallic screen is made of copper.

3 Power flow

The time-average power flow across a surface Ω in the z-direction is given by the integral of the complex Poyntings vector over that surface,

$$P = \int_\Omega \hat{\boldsymbol{z}} \cdot < \boldsymbol{S} > \mathrm{d}a = \frac{1}{2}\,\mathrm{Re}\int_\Omega \hat{\boldsymbol{z}} \cdot \{\boldsymbol{E} \times \boldsymbol{H}^*\}\,\mathrm{d}a$$
$$= \frac{1}{2}\,\mathrm{Re}\int_\Omega \hat{\boldsymbol{z}} \cdot \{\boldsymbol{E}_\mathrm{T} \times \boldsymbol{H}_\mathrm{T}^*\}\,\mathrm{d}a, \tag{1}$$

where $*$ denotes complex conjugate, and $\boldsymbol{E}_{\mathrm{T}}, \boldsymbol{H}_{\mathrm{T}}$ are the transverse fields. The time-average power flow across a surface Ω in the z-direction can be written as

$$P = \mathrm{Re} \sum_n \frac{Y_n^*}{2\eta_0} \{ |a_n^+|^2 \mathrm{e}^{-2\,\mathrm{Im}(k_{zn})z} - |a_n^-|^2 \mathrm{e}^{2\,\mathrm{Im}(k_{zn})z} - \mathrm{i}2\,\mathrm{Im}(a_n^+ a_n^{-*} \mathrm{e}^{\mathrm{i}2\,\mathrm{Re}(k_{zn})z}) \}, \tag{2}$$

where $a_n^{\pm}$ are the coefficients of the expansion and Y_n is the wave impedance. In the apertures the summation is done over all of the waveguide modes, and in the dielectric layers the summation is done over all of the Floquet modes.

3.1 Conservation of power

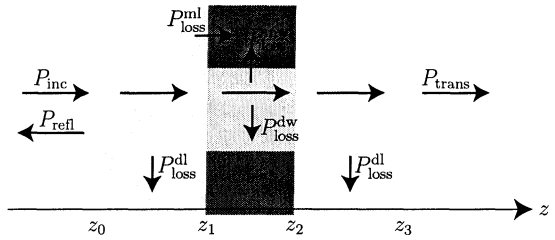


Figure 2: Power flow through a FSS.

For the FSS it yields that, see Figure 2,

$$P_{\mathrm{inc}} = P_{\mathrm{refl}} + P_{\mathrm{trans}} + P_{\mathrm{diss}}, \tag{3}$$

where the dissipated power is

$$P_{\mathrm{diss}} = P_{\mathrm{loss}}^{\mathrm{dl}} + P_{\mathrm{loss}}^{\mathrm{dw}} + P_{\mathrm{loss}}^{\mathrm{ml}} + P_{\mathrm{loss}}^{\mathrm{mw}}, \tag{4}$$

and $P_{\mathrm{loss}}^{\mathrm{dl}}$ is due to losses in the dielectric layer, $P_{\mathrm{loss}}^{\mathrm{dw}}$ is due to losses in the dielectric waveguide puck, $P_{\mathrm{loss}}^{\mathrm{ml}}$ is due to losses in the plane metallic surfaces, and $P_{\mathrm{loss}}^{\mathrm{mw}}$ is due to losses in the walls of the waveguide.

3.2 Numerical plots of power

The power flow through an FSS is illustrated in Figure 3. The structure in the example consists of a hexagonal FSS imbedded in two dielectric layers, as in Figure 1. The power plots are symmetric, *i.e.*, they are the same at corresponding boundary on the other half of the FSS, this follows from the fact that the FSS is symmetric, the reflected power is zero, and time reversal.

4 Dielectric losses

Electric-type dissipation in the dielectric medium is modelled by the complex relative permittivity

$$\epsilon = \epsilon_{\mathrm{r}} + \mathrm{i}\epsilon_{\mathrm{i}} = \epsilon_{\mathrm{r}}(1 + \mathrm{i}\tan\delta), \tag{5}$$

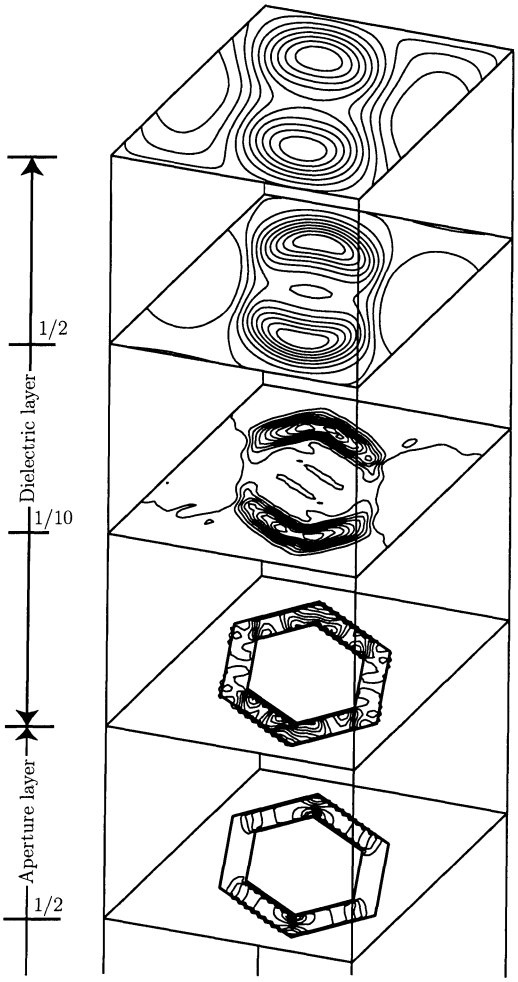


Figure 3: Contour plot of the power flow for a hexagonal FSS.

where ϵ_r is the real relative dielectric constant and $\tan\delta = \frac{\epsilon_{\mathrm{i}}}{\epsilon_r}$ is the loss tangent for the material. For a dissipative medium it is possible to introduce a conductivity σ that is related to the imaginary part of the permittivity ϵ_{i} as

$$\epsilon = \epsilon_{\mathrm{r}} + \mathrm{i}\frac{\sigma}{\omega\epsilon_0} = \epsilon_{\mathrm{r}} + \mathrm{i}\frac{\sigma}{2\pi f \epsilon_0}. \tag{6}$$

The propagation of a mode is described by

$$\mathrm{e}^{\pm \mathrm{i}k_{zn}z} = \mathrm{e}^{\pm \mathrm{i}\beta_n z} \mathrm{e}^{\mp \alpha_{\mathrm{d}n} z}, \tag{7}$$

where k_{zn} is the complex propagation constant, β_n is the phase constant, and $\alpha_{\mathrm{d}n}$ is the attenuation constant due to dissipation in the dielectric medium.

The total transmitted power through the FSS is

$$\begin{aligned} P_{\mathrm{trans}} &= P_{\mathrm{inc}} - P_{\mathrm{loss}}^{\mathrm{d}} = P_{\mathrm{inc}} - (P_{\mathrm{loss}}^{\mathrm{dl}} + P_{\mathrm{loss}}^{\mathrm{dw}}) \\ &\simeq P_{\mathrm{inc}}(1 - 2\alpha_{\mathrm{dl}}\Delta z_{\mathrm{dl}})(1 - 2\alpha_{\mathrm{dw}}\Delta z_{\mathrm{dw}}), \end{aligned} \tag{8}$$

where P_{inc} is the incident power on the FSS, $P_{\mathrm{loss}}^{\mathrm{dl}} \simeq P_{\mathrm{inc}} 2\alpha_{\mathrm{dl}}\Delta z_{\mathrm{dl}}$ is the total power loss in the dielectric

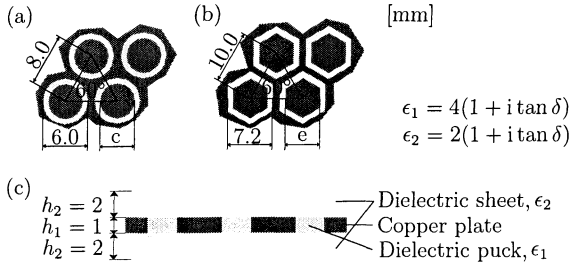


Figure 4: Two examples of FSS patterns; (a) annular FSS, (b) hexagonal FSS. In (c) the vertical cross-section of an FSS is depicted. Cross-sections of type 1 has $c = 5.0\,\text{mm}$ and $e = 5.2\,\text{mm}$, while cross-sections of type 2 has $c = 4.0\,\text{mm}$, and $e = 3.2\,\text{mm}$.

$\tan\delta$		$P^{\text{d}}_{\text{loss}}$ [dB]		
Dielectric layer	Aperture layer	Annular 1	Annular 2	Hexagonal 1
0	0	0	0	0
0	0.001	0.046	0.026	0.035
0	0.01	0.40	0.24	0.35
0	0.1	3.27	2.08	2.94
0.001	0	0.025	0.019	0.023
0.001	0.001	0.065	0.043	0.058
0.001	0.01	0.41	0.25	0.37
0.001	0.1	3.29	2.09	2.96
0.01	0	0.20	0.17	0.23
0.01	0.001	0.23	0.20	0.26
0.01	0.01	0.58	0.40	0.56
0.01	0.1	3.41	2.21	3.10
0.1	0	1.75	1.57	2.00
0.1	0.001	1.79	1.59	2.03
0.1	0.01	2.08	1.76	2.28
0.1	0.1	4.54	3.37	4.44

Table 1: Losses in the dielectric medium.

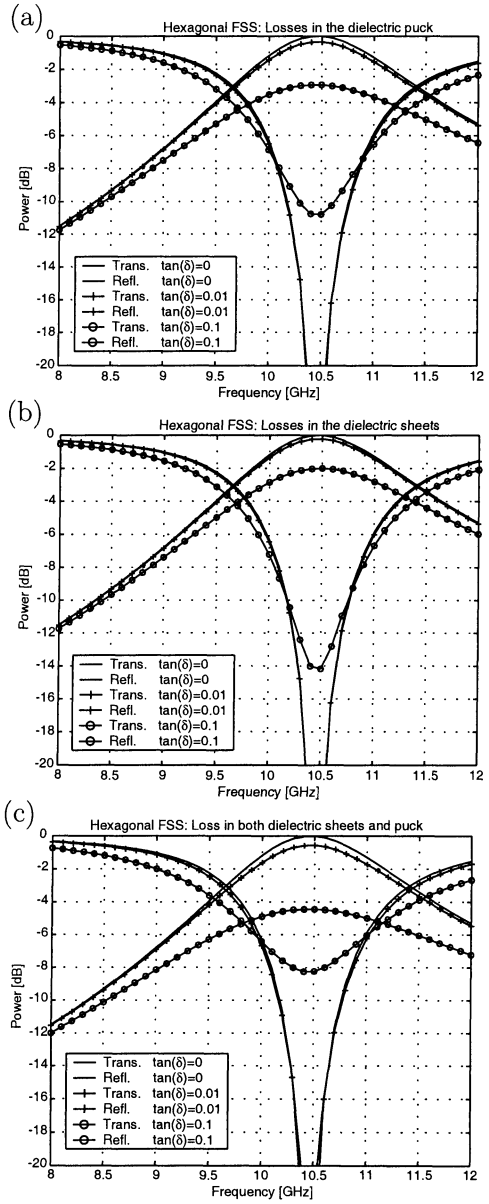


Figure 5: Dissipation in a hexagonal FSS (type 1).

sheets, and $P^{\text{dw}}_{\text{loss}} \simeq P_{\text{inc}} 2\alpha_{\text{dw}} \Delta z_{\text{dw}}$ is the total power loss in the dielectric waveguide puck. The total attenuation constant due to dissipation in the dielectric sheets can be written as $\alpha_{\text{dl}} = \sum_{n=1}^{\infty} \frac{P^{\text{dl}}_n}{P_{\text{inc}}} \alpha_{\text{dl}n}$, and the total attenuation constant due to dissipation in the dielectric waveguide puck $\alpha_{\text{dw}} = \sum_{n=1}^{\infty} \frac{P^{\text{dw}}_n}{P_{\text{inc}}} \alpha_{\text{dw}n}$.

4.1 Numerical examples of dielectric losses

The reflected and transmitted power for a lossy structure are plotted in Figure 5. The structure consists of a lossy hexagonal-puck aperture imbedded in two lossy dielectric layers, as in Figure 4. For obvious reasons the sharp null in the reflection power of the lossless structure does not exist when the losses are added. In Table 1 numerical values of losses in the dielectric medium are presented. Numerical values of the attenuation factors for dissipation in the dielectric are presented in Table 2.

Attenuation constant [mm^{-1}]				
$\tan\delta$	0.001		0.01	
	α_{dl}	α_{dw}	α_{dl}	α_{dw}
Annular 1	0.00073	0.0053	0.0058	0.046
Annular 2	0.00056	0.0030	0.0049	0.028
Hexagonal 1	0.00066	0.0041	0.0066	0.040

Table 2: The attenuation constant due to dissipation in the dielectric medium.

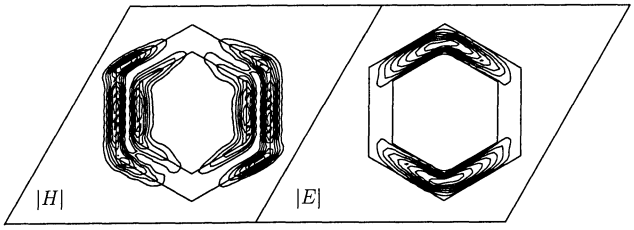


Figure 6: Contour plot of the electric and the magnetic fields at the metallic surface.

5 Losses in plane metallic screen

The metal is supposed to have a finite conductivity, σ_{m}, and to fulfill the condition for a good conductor, *i.e.*, $\sigma_{\mathrm{m}} \gg \omega\epsilon_0\epsilon_{\mathrm{m}}$. The power loss on the plane metallic part of the cell can be written as

$$P_{\mathrm{loss}}^{\mathrm{ml}} = R_{\mathrm{S}} \int_{D\setminus\Omega} |\boldsymbol{H}_{\parallel}|^2 \,\mathrm{d}a, \tag{9}$$

where $R_{\mathrm{S}} = \frac{1}{2\sigma_{\mathrm{m}}\delta}$ is a surface resistance, δ is the skin depth, and $\boldsymbol{H}_{\parallel}$ is the tangential component of the magnetic field at the surface of the metallic screen. A very good approximation is that the value of $\boldsymbol{H}_{\parallel}$ is the same as for a perfectly conducting screen. The loss factor L_{f} due to the finite conductivity in the plane metallic screen is defined as

$$L_{\mathrm{f}} = \frac{P_{\mathrm{loss}}}{P_{\mathrm{inc}}}, \tag{10}$$

where P_{inc} incident power on the aperture layer.

5.1 Surface field plots and numerical examples of losses

In Figure 6 contour plots for the electric and the magnetic fields at the metallic surface for a hexagonal FSS are depicted. For obvious reasons the electric field exists only in the aperture and is zero at the metallic surface, but the magnetic field can exist both in the aperture and at the metallic surface.

Numerical values of the incident power at the metallic surfaces, and the loss factor for the metallic surface are presented in Table 3. In the table dissipation in the dielectric material is also included. The loss tangent is the same for both the puck and the layers.

6 Losses in waveguide walls

Assume that the walls of the waveguide are good, but not perfect, conductors, *i.e.*, the conductivity, σ, is large, but finite. The power loss per unit length can be written as

$$\frac{\mathrm{d}P}{\mathrm{d}z} = -\oint_{\Gamma} \frac{\mathrm{d}P}{\mathrm{d}a}\,\mathrm{d}l = -\frac{1}{2\sigma\delta}\oint_{\Gamma} |\boldsymbol{H}^{\pm}|^2\,\mathrm{d}l, \tag{11}$$

Incident power, P^{d} and Loss factor, L_{f}				
	Left side		Right side	
	$P_{\mathrm{l}}^{\mathrm{d}}$	L_{f}	$P_{\mathrm{r}}^{\mathrm{d}}$	L_{f}
$\tan\delta = 0$:				
Annular 1	1.00	0.00072	1.00	0.00072
Annular 2	1.00	0.00034	1.00	0.00034
Hexagonal 1	1.00	0.00048	1.00	0.00048
$\tan\delta = 0.001$:				
Annular 1	1.00	0.00071	0.99	0.00072
Annular 2	1.00	0.00034	0.99	0.00034
Hexagonal 1	0.99	0.00048	0.98	0.00049
$\tan\delta = 0.01$:				
Annular 1	0.98	0.00065	0.90	0.00071
Annular 2	0.98	0.00032	0.93	0.00033
Hexagonal 1	0.97	0.00045	0.90	0.00048

Table 3: The incident power and the loss factor at the plane metallic surface. Dissipation in the dielectric material is included with $\tan\delta$.

where $|\boldsymbol{H}^{\pm}|^2 = \sum_{vp} |\boldsymbol{H}_{vp}^{\pm}|^2$ is the total magnetic field at the metallic waveguide walls and

$$|\boldsymbol{H}_{vp}^{\pm}|^2 = |a_{vp}^{\pm}|^2\{|\boldsymbol{H}_{\mathrm{T}vp}|^2 + \frac{1}{\eta_0^2}|H_{zvp}|^2\}, \tag{12}$$

where $a_{vp}^{\pm}$ are expansion coefficients. The total power loss in a section of length Δz of the waveguide is given by

$$P_{\mathrm{loss}}^{\mathrm{mw}} = -\frac{\mathrm{d}P}{\mathrm{d}z}\Delta z = P_{\mathrm{inc}} 2\alpha_c \Delta z, \tag{13}$$

where P_{inc} is the incident power through the aperture, and α_{c} is the corresponding attenuation constant of the field due to finite conductivity in the walls of the waveguide.

6.1 Numerical examples of losses in aperture walls

Numerical values of the power flow through the aperture, and the power losses in the walls of the waveguide are presented in Table 4. The cross-sections are given by Figure 4. In the table dissipation in the dielectric material is also included. The loss tangent is the same for both the puck and the layers. If the dielectric materials are lossy, *i.e.*, $\tan\delta > 0$, the power flow is attenuated in the dielectric layer. As a result of the dissipation in the waveguide puck, the power flow through aperture on the right side is smaller than the power flow through aperture on the left side.

7 Conclusions

The power transmission through a hexagonal FSS is depicted in Figure 7. The power decreases as it propagates through the FSS. First the power is attenuated

	$P_{\mathrm{l}}^{\mathrm{w}}$ [W]	$P_{\mathrm{r}}^{\mathrm{w}}$ [W]	P_{loss} [dB]	$2\alpha_{\mathrm{c}}$ [m^{-1}]
$\tan\delta = 0$:				
Annular 1	1.00	1.00	0.0205	1.024
Annular 2	1.00	1.00	0.0064	0.322
Hexagonal 1	1.00	1.00	0.0121	0.605
$\tan\delta = 0.001$:				
Annular 1	0.996	0.987	0.0203	1.018
Annular 2	0.997	0.991	0.0064	0.322
Hexagonal 1	0.995	0.987	0.0120	0.604
$\tan\delta = 0.01$:				
Annular 1	0.975	0.895	0.0181	0.927
Annular 2	0.979	0.930	0.0060	0.306
Hexagonal 1	0.972	0.901	0.0108	0.553

Table 4: The power losses in the walls of the waveguide due to finite conductivity. Dissipation in the dielectric material is included with $\tan\delta$. $P_{\mathrm{l}}^{\mathrm{w}}$ and $P_{\mathrm{r}}^{\mathrm{w}}$ are the power flow through the apertures of the waveguide on left and right side, respectively. P_{loss} is the total power loss in the walls of the waveguide, and α_c is the attenuation constant.

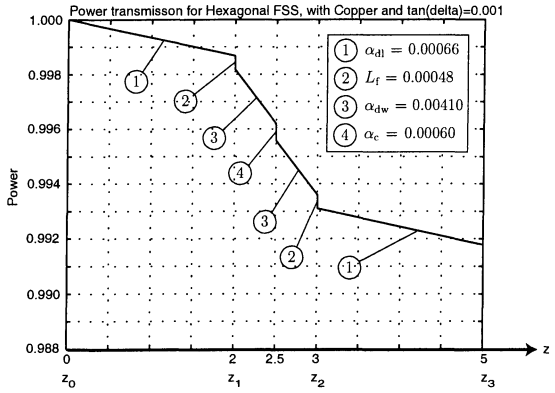


Figure 7: Power transmission through a FSS.

in the dielectric sheet (1), then there is some losses in the plane metallic surface (2) , and after that the power is attenuated in the dielectric puck (3). The losses in the metallic waveguide walls (4) are plotted all at ones in the middle of the waveguide, but in reality they are spread all over the walls of the waveguide.

The presence of dissipation in the dielectric medium and the metallic surfaces in an FSS slightly modifies the usual description of scattering from an FSS. This modification takes the form of a complex rather than an real propagation constant, and by introduction of a surface resistance. The transmitted power can approximately be written as

$$P_{\mathrm{trans}} = P_{\mathrm{inc}} t_{\mathrm{dl}} t_{\mathrm{dw}} t_{\mathrm{c}} L_{\mathrm{f}}, \qquad (14)$$

where $t_{\mathrm{dl}}(\Delta z_{\mathrm{dl}})$ is the transmissivity due to dissipation in the dielectric layers, $t_{\mathrm{dw}}(\Delta z_{\mathrm{dw}})$ is the transmissivity due to dissipation in the dielectric waveguide puck , $t_{\mathrm{c}}(\Delta z_{\mathrm{c}})$ is the transmissivity due to finite conductivity in the walls of the waveguide, and L_{f} is the loss factor due to finite conductivity in the plane metallic surface. The transmissivities and the loss factor can with good accuracy be determined separately. The transmissivities due to dissipation in the dielectric medium, t_{dl} and t_{dw}, are determined by assuming infinite conductivity for the metallic surfaces, and the transmissivity due to finite conductivity in the walls of the waveguide t_{c} and the loss factor due to finite conductivity in metallic medium L_{f} is determined by assuming lossless dielectric medium.

The losses in the dielectric material are rather small for common "good dielectrics", but can be quite large for "lossy dielectrics", if the thickness of the layers are large. The attenuation is proportionately larger in the dielectric waveguide puck than in the dielectric layer, and the attenuation become smaller if the size of the aperture become larger.

The losses due to finite conductivity in the metallic plate are also rather small, if the metallic plate is made of a good conductor, such as copper or aluminium. The losses in the waveguide walls are proportionately larger than the losses in the plane metallic surface. It has also been seen that both losses on the plane metallic screen and losses on the waveguide walls are smaller, if the aperture are larger. In general, the losses in the dielectric material are dominating over the losses due to finite conductivity in the metallic plate.

References

[1] B. Widenberg, S. Poulsen, and A. Karlsson. Scattering from Thick Frequency Selectiv Screens. *J. Electro. Waves Applic.*, **14**, 1301–1327, 2000.

[2] B. Widenberg. A General Mode Matching Technique applied to Bandpass Radomes. Technical Report LUTEDX/(TEAT-7098)/1–33/(2001), Dept. of Electroscience, Lund Institute of Technology, P.O. Box 118, SE-211 00 Lund, Sweden. 2001.

Identification of the current distribution at a radome from near-field measurements

Kristin Persson[1], Mats Gustafsson[2], and Michael Andersson[3]

[1]Electroscience, LTH, Box 118, SE-221 00 Lund, Sweden. E-mail: f96kp@es.lth.se.

[2]Electroscience, LTH, Box 118, SE-221 00 Lund, Sweden. E-mail: mats@es.lth.se.

[3]Applied Composites AB, Box 163, 341 23 Ljungby, Sweden. E-mail: michael.andersson@acab.se.

Abstract

Information about the equivalent current distribution at a radome can be used to improve radome design, detect manufacturing errors, and to verify numerical simulations. In this paper, the transformation from near-field data to the equivalent current distribution is analyzed. The transformation is based on inversion of the surface integral representation that relates the equivalent current to the near-field data. The attempt is to develop a mathematical model that easily can be used for arbitrary geometric structures. The symmetries of a specific problem is used to reduce the computational complexity. Both synthetic data and measured data are used to verify the algorithm.

1 Introduction

There are several applications of a near field to equivalent current transformation. For example, if the radiation pattern from a radiating body (radome, antenna arrays) does not exhibit the expected form. One can determine the equivalent current at the radiating body and directly find the malfunctioning components instead of dismounting the radiating body and looking for defect components. Another application is in the radome industry where computer simulations are used to analyze radome structures. These simulations need input data depicting the field from the antenna placed inside the radome in an accurate way. A person not initiated in this area might wonder why one cannot simply measure the field from the antenna at the radius where the radome should be placed. The reason for this is that the radome often is placed very close to the antenna and at this distance, there is a substantial interaction between the antenna and the measuring probe, see Ref. [3].

In this paper, the transform of near-field data to its equivalent current distribution on a surface close to the radiating body is analyzed. This transformation is a linear inverse source problem. As most inverse problems it is ill posed, *i.e.*, small errors in the near-field data can produce large errors in the equivalent current. The method used to reverse the electromagnetic field, is the surface integral representation (SIR),see Refs [4, 5]. This representation is a way of describing an electromagnetic field. If the field is known at a specific surface, defining a volume, the field in the volume can be determined. Within the representation, all fields outside the volume cancel out. The field at the surface is identified as its equivalent current, see Ref. [2]. In this paper, the field is presumed to be scalar. This assumption is acceptable since the measured near-field data clearly have one dominating component.

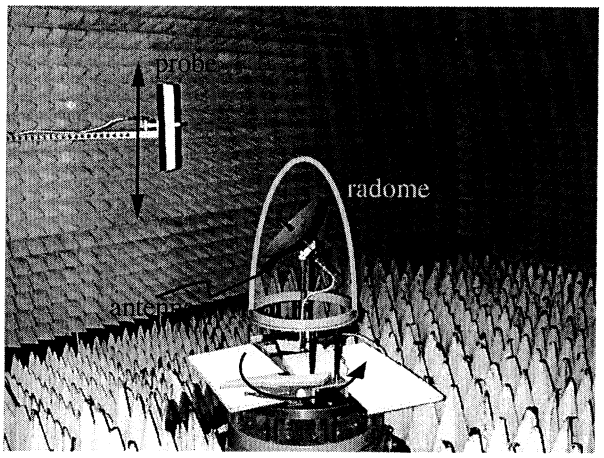


Figure 1: Photo of the cylindrical near-field range at SAAB Bofors Dynamics and Applied Composites AB (ACAB), Sweden. (Courtesy of SAAB Bofors Dynamics and Applied Composites AB, Sweden.)

To our disposal, we have near-field data supplied by SAAB Bofors Dynamics and Applied Composites AB (ACAB), Sweden. The near-field data is given for three different antenna and radome configurations. These three configurations are antenna, antenna together with radome, and antenna together with a defect radome.

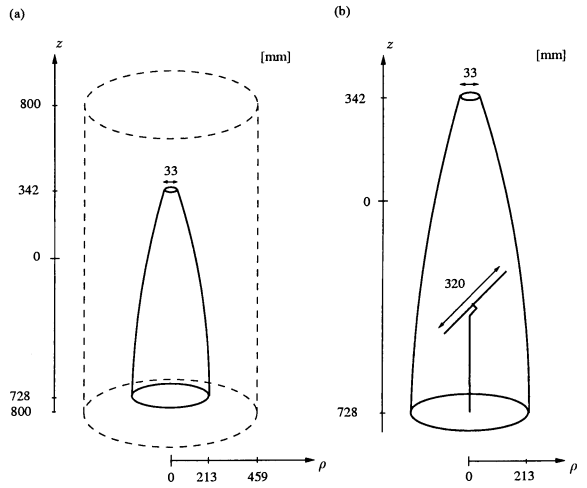


Figure 2: (a) The dimensions of the radome and the cylindrical surface where the near field is measured. (b) The dimensions of the radome and the reflector antenna.

2 Radome example

The used near-field data is supplied by SAAB Bofors Dynamics and Applied Composites AB (ACAB), Sweden, see Figure 1. The measured electric field has a dominating z-component. Three different measurements are done, data sampled without the radome, data sampled with the radome, and data sampled with the defect radome. The defect radome is constructed by attaching two copper plates on the radome surface. The setup with relevant dimensions indicated are depicted in Figure 2. The antenna is a reflector antenna with a focal distance of 0.1 m. It has vertical polarization perpendicular to the angular direction. The Standing Wave Ratio (SWR) is 1.4 in the frequency range $8.2 - 9.5$ GHz. The antenna is poorly adapted for other frequencies. A 10 dB reflection attenuator is connected to the antenna.

The radome surface is described by $\rho(z) = z_0$ for $z_1 \leq z \leq z_2$ and

$$\rho(z) = -(bz' + d) + \sqrt{(bz' + d)^2 - a(z')^2 - 2cz' - e}$$

for $z_2 < z \leq z_3$ where $z' = z - z_1$ and the constants are tabulated in (2.1).

$$\begin{aligned} a &= 1.223 \cdot 10^{-1} & z_0 &= 0.213\,\mathrm{m} \\ b &= 4.830 \cdot 10^{-2} & z_1 &= -0.728\,\mathrm{m} \\ c &= -1.823 \cdot 10^{-2}\,\mathrm{m} & z_2 &= -0.663\,\mathrm{m} \\ d &= 1.484 \cdot 10^{-1}\,\mathrm{m} & z_3 &= 0.342\,\mathrm{m} \\ e &= -1.079 \cdot 10^{-1}\,\mathrm{m}^2 \end{aligned} \tag{2.1}$$

The near-field probe consists of a waveguide for which no compensation is made in the final data [3]. The cylinder surface, where the field is sampled, is placed in both the reactive near-field zone and the radiating near-field zone [1]. In this way the coupling between the measuring near-field probe and the object is reduced.

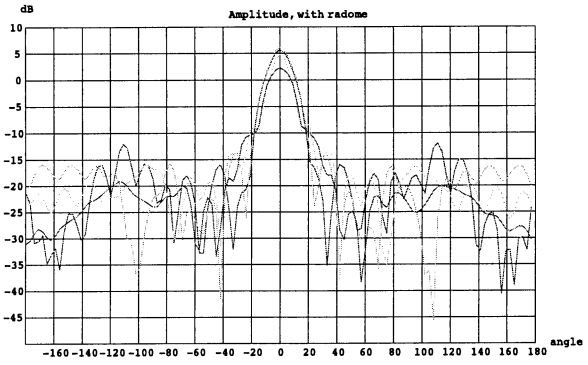


Figure 3: Example of the measured field in the angular direction when the radome is used.

The amplitude and phase of the electromagnetic field are measured in the frequency range $8 - 12$ GHz. The measurements are done on a cylindrical surface by moving the probe in the z-direction and rotating the antenna under test, see Figure 1. Angle wise, 120 points is measured going from $-180°$ to $180°$ in steps of $3°$. The z-dimension is divided into 129 points, each separated by 12.5 mm. This means that at 8 GHz the field is sampled three times per wavelength and at 12 GHz twice per wavelength. Together, a total of $120 \cdot 129 = 15480$ measuring points are sampled. Example of the field amplitude a function of the angle, when the radome is used, is shown in Figure 3.

3 The surface integral representation

The scalar surface integral representation

$$\iint_{\text{radome}} \frac{\partial g(k, \boldsymbol{r}, \boldsymbol{r}')}{\partial n} E_z(\boldsymbol{r}) - g(k, \boldsymbol{r}, \boldsymbol{r}') \frac{\partial E_z(\boldsymbol{r})}{\partial n}\, \mathrm{d}S = \begin{cases} -E_z(\boldsymbol{r}') & \boldsymbol{r}' \in V \\ 0 & \boldsymbol{r}' \notin V \end{cases} \tag{3.1}$$

is used to relate the field at the radome surface to the measured field. Observe that the true field does not have to be zero outside the volume. The SIR (3.1) only says that the lefthandside of the equation turns out to be zero if the vector $\boldsymbol{r}'$ points outside the volume V.

The equivalent surface currents are introduced as $M(\boldsymbol{r}) = E_z(\boldsymbol{r})$ and $J(\boldsymbol{r}) = \frac{\partial E_z(\boldsymbol{r})}{\partial n}$. These designations are not in total accordance with the actual def-

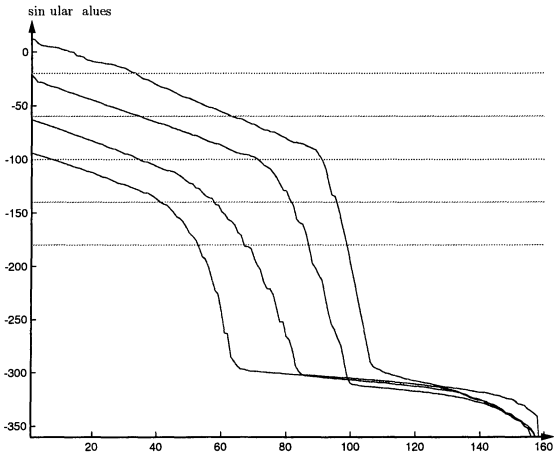


Figure 4: The singular values of $\hat{G}$ in Db scale for four of the Fourier transformed angle components. The horizontal lines show the used cut-off values.

initions of the electric and magnetic equivalent currents which are $\boldsymbol{M}(\boldsymbol{r}) = -\hat{\boldsymbol{n}}\times\boldsymbol{E}(\boldsymbol{r})$ and $\boldsymbol{J}(\boldsymbol{r}) = \hat{\boldsymbol{n}}\times\boldsymbol{H}(\boldsymbol{r})$, see also Ref. [2]. Even so, the similarities with the quantities in (3.1) make them plausible designations for the purposes of this paper. Bear also in mind that these denominations are not true currents. The equivalent currents M and J are only a way of describing the electromagnetic field so that the field in the volume enclosed by the surface at which the current exists can be set to zero.

The transformation (Green's function) is rotational symmetric sine the measurement setup is rotational symmetric. This reduces (3.1) to an angle wise convolution which makes it possible to evaluate the SIR by an angle wise Fourier transform

$$\sum_m \hat{G}_{i,m,k_j} \begin{pmatrix} \hat{M}_{m,k_j} \\ \hat{J}_{m,k_j} \end{pmatrix} = \frac{1}{N} \begin{pmatrix} -\hat{E}^{\text{cyl}}_{i,k_j} \\ 0 \end{pmatrix} \qquad (3.2)$$

for all i, j.

Since the matrix $\hat{G}$ in (3.2) is not quadratic a regular inversion can not be done. A fast and easy way to solve this is to use the Singular Value Decomposition (SVD), see Ref. [6]. The matrix $\hat{G}$ can by this method be divided into three matrices so that $\hat{G} = \hat{U}\hat{S}\hat{V}'$ where $\hat{U}$ and $\hat{V}'$ are orthogonal matrices, *i.e.*, $\hat{U}^{-1} = \hat{U}'$ and $\hat{V}^{-1} = \hat{V}'$ where $\hat{V}'$ denotes the Hermitian conjugate of $\hat{V}$. A cut-off value δ is chosen and all singular values below this value are set to zero. If this is not done the small singular values will create tremendously growth of non-radiation currents.

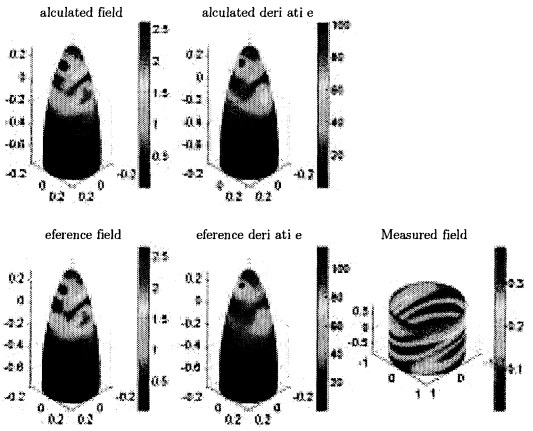


Figure 5: Synthetic data originating from five sources inside the radome. The data is reversed to its equivalent current on the radome surface. The frequency is 1.7 GHz and the cut-off value $\delta = 10^{-8}$. The number of sample points per wavelength are 6.7 in the angular direction and 7.7 in the z-direction.

4 Inversion results with synthetic data

The inversion algorithm is implemented in MATLAB. One of the major computational problems is due to the storage of the matrix G, *i.e.*, G has approximately 10^8 elements. To verify the computer code a rather low frequency is used to reduce the amount of necessary sample points and hence the memory requirement. For a frequency of 1.7 GHz a sample density of 15.7 points per wavelength in the angular direction and 15.9 in the z-direction can easily be attained. For easier calculations the radome surface is reshaped in a continuous way, where a smooth top and bottom surface have been added to the radome. The synthetic data is calculated from five point sources placed inside the radome. The inversion result is shown in Figure 5. The corresponding far field pattern is depicted in Figure 6.

5 Inversion results with experimental data

Inversion with the measured near-field data, described in Section 2, is shown in Figure 7. The results are preliminary and more analysis is needed before the results become reliable. To be able to rule out errors originating from low sample density our aim is to compare the results with synthetic data and to use a more powerful computer (the present computer has 4 GB of

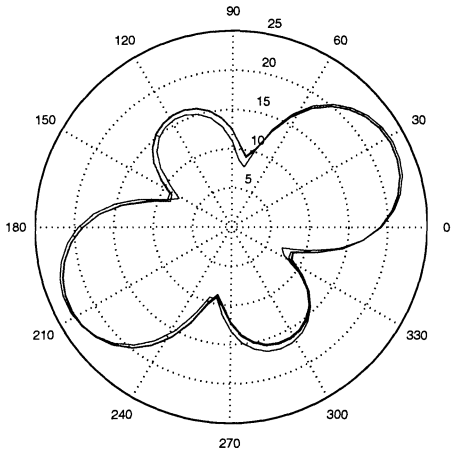


Figure 6: Far-field pattern from the equivalent current distribution depicted in Figure 5 for the different cut-off values in Figure 4.

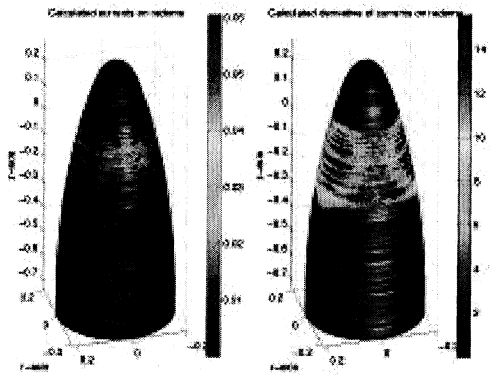


Figure 7: Measured data reversed to its equivalent current at the radome surface. The frequency is 8 GHz and the cut-off value is 1. The number of sample points per wavelength are 6.7 in the angular direction and 10.2 in the z-direction.

RAM). Another source of error are reflections in the measured data from the walls and floors surrounding the radome.

6 Conclusions

The used surface integral representation gives a linear map between the equivalent current and the near-field data for general structures. It is shown that this map can be inverted for rotational symmetric geometries. The inversion results for synthetic data are very good. The preliminary inversion results for the measured data are promising but more analysis is needed in this case. The good results for this scalar approximation is also very promising for a full vector implementation of the near field to equivalent current transformation.

7 Acknowledgments

SAAB Bofors Dynamics and Applied Composites AB (ACAB), Sweden are gratefully acknowledged for supplying the experimental data and the photo depicted in Figure 1.

References

[1] C. A. Balanis. *Antenna Theory.* John Wiley & Sons, New York, second edition, 1997.

[2] C. A. Balanis. *Advanced Engineering Electromagnetics.* John Wiley & Sons, New York, 1989.

[3] J. E. Hansen, editor. *Spherical Near-Field Antenna Measurements.* Number 26 in IEE electromagnetic waves series. Peter Peregrinus Ltd., Stevenage, UK, 1988. ISBN: 0-86341-110-X.

[4] G. Kristensson. *Spridningsteori med antennttillämpningar.* Studentlitteratur, Lund, 1999. (In Swedish).

[5] T. K. Sarkar and A. Taaghol. Near-field to near/far-field transformation for arbitrary near-field geometry utilizing an equivalent electric current and MoM. *IEEE Trans. Antennas Propagat.*, **47**(3), 566–573, March 1999.

[6] G. Strang. *Introduction to applied mathematics.* Wellesley-Cambridge Press, Box 157, Wellesley MA 02181, 1986.

Microwaves in a circularly bend waveguide of rectangular cross-section

Thomas Biro[1] and Börje Nilsson[2]
[1]School of Mathematics and Systems Engineering, Växjö University, SE-351 95 VÄXJÖ.
E-mail: thomas.biro@msi.vxu.se
[2]School of Mathematics and Systems Engineering, Växjö University, SE-351 95 VÄXJÖ.
E-mail: borje.nilsson@msi.vxu.se

Abstract

A new full-wave method is used to solve the Maxwell equations in the microwave region of a circular waveguide bend with a rectangular cross section. The description in both the bend and the straight waveguides are expressed in the same type of operator formalism and basis functions. This feature is important for generalisations to complicated geometries. In addition, a compact set of continuity equations of the transverse fields at the bend interfaces with the straight waveguides is derived. The scattering matrices for the key building blocks, the H-plane and the E-plane bends, are calculated. Also, a multiple bend structure is analysed by cascade matrix techniques. Although the numerical calculations of the scattering matrices is made for the lowest propagating mode, the fundamental equations are general and suitable for an analysis of more complex structures, such as smoothly bent E-plane filters.

1 Introduction

In order to construct a compact product, waveguide bends are important in the design of passive microwave components such as filters. Some papers, concerning this topic, have been written during the last decade. The first to present an accurate and numerically efficient solution were Weisshaar *et al*, [1]. Their method is based on mode-matching technique where a Fourier expansion of the eigenfunctions of the Helmholtz equation is employed. Cornet *et al*, [2], used a tensorial covariant formalism of the Maxwell equations to determine the scattering characteristics of the bend. Their method is more efficient compared to the method of [1] for simulations of bends with zero inner radius. Recently, MacPhie *et al*, [3], used Bessel-Fourier modal functions to derive solutions for bends having zero inner radius. However, the methods presented by [1], [2] and [3] are not easily generalized to include multiple scattering of a discontinuity in the bend.

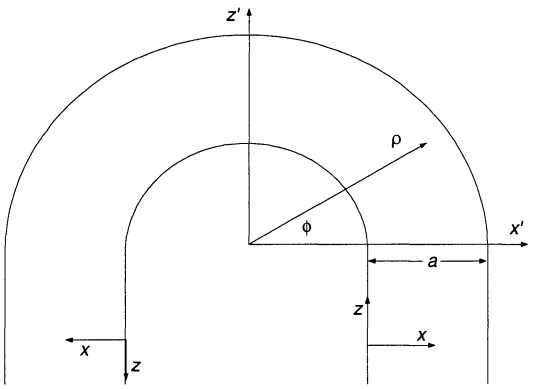


Figure 1: Geometry of a circularly bend waveguide of rectangular cross-section.

In this paper we introduce a new method, that is easier to generalize, to solve the Maxwell equations in the microwave region of a circular bend in a waveguide having a rectangular cross section, see Figure 1. Thereby, we determine the reflection and transmission properties of the bend for incoming waves. Although the derivation of the scattering matrices is made without discontinuities in the bend, the fundamental equations are general and form a suitable starting point for an analysis of a more complex structure, e.g. E-plane filter in a circularly bend geometry.

2 The fundamental equations

We solve the problem of uniform and rectangular H- and E-plane bends. The concept of H-plane bend originates from the fact that, for the lowest propagating mode, the **H** field is in the plane defined by the bend, if $a > b$, where a and b are the transverse rectangular measures of the waveguide. For the E-plane bend we have the condition that $a < b$, and that the **E** field is in the plane defined by the bend. In our geometry, the lowest propagating mode for the H-plane bend is the TE_{10} mode. For the E-plane bend the TE_{01} mode

is the lowest propagating mode.

With a harmonic time dependence, $\exp(-i\omega t)$, and an isotropic homogeneous medium the $\mathbf{E}$ and $\mathbf{H}$ fields satisfy the vector Helmholtz equation

$$(\nabla^2 + k^2)\begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix} = 0, \tag{1}$$

where $k = \omega/c$. By taking the y component of (1) we get the scalar Helmholtz equation

$$(\nabla^2 + k^2)\begin{pmatrix}E_y\\ H_y\end{pmatrix} = 0. \tag{2}$$

If this equation is solved, then E_y and H_y can be used to generate the electric and magnetic field components, $\mathbf{E}_\perp$ and $\mathbf{H}_\perp$, in the $x-z$ plane. To see this, we separate the fields into components parallel and perpendicular to the y axis, and rewrite the Maxwell equations in terms of these components to get

$$\begin{aligned}\left(\tfrac{\partial^2}{\partial y^2} + k^2\right)\mathbf{E}_\perp &= \boldsymbol{\nabla}_\perp \tfrac{\partial}{\partial y} E_y - i\omega\mu\hat{\mathbf{y}} \times \boldsymbol{\nabla}_\perp H_y,\\ \left(\tfrac{\partial^2}{\partial y^2} + k^2\right)\mathbf{H}_\perp &= \boldsymbol{\nabla}_\perp \tfrac{\partial}{\partial y} H_y + i\omega\epsilon\hat{\mathbf{y}} \times \boldsymbol{\nabla}_\perp E_y,\end{aligned} \tag{3}$$

where ϵ and μ are the permittivity and the permeability of the medium, respectively. The operator $\boldsymbol{\nabla}_\perp$ is in cylindrical coordinates given by

$$\boldsymbol{\nabla}_\perp \begin{pmatrix}E_y\\ H_y\end{pmatrix} = \hat{\boldsymbol{\rho}}\frac{1}{\rho}\frac{\partial}{\partial\rho}\begin{pmatrix}\rho E_y\\ \rho H_y\end{pmatrix} + \hat{\boldsymbol{\phi}}\frac{1}{\rho}\frac{\partial}{\partial\phi}\begin{pmatrix}E_y\\ H_y\end{pmatrix}. \tag{4}$$

From (3) it is possible to determine the electric field $\mathbf{E}_\perp$ and the magnetic field $\mathbf{H}_\perp$ in the $x-z$ plane.

Let us denote the operator $B^2 = \rho^2(\nabla^2 + k^2) - \partial^2/\partial\phi^2$. Then (2) for H_y can be written as

$$\left(\frac{\partial^2}{\partial\phi^2} + B^2\right) H_y = 0. \tag{5}$$

Since B^2 is independent of ϕ we can solve (5) for H_y and get

$$H_y(\rho,\phi,y) = e^{iB\phi}H^+(\rho,y) + e^{-iB\phi}H^-(\rho,y), \tag{6}$$

where H^+ and H^- are the wave amplitude functions, associated with propagation in the positive and negative directions, respectively. The wave splitting into right- and left-going waves guarantees a stable formulation of the problem.

For the explicit calculation of H_y we use trigonometric basis functions, in the transverse variable ρ. Since there is no discontinuity in the y direction in the bend, we may assume that the y dependence is of same form in the bend as in the straight waveguide feeding it, i.e. $\sin(\pi y/b)$. The basis functions satisfy the boundary conditions of an electric wall, i.e. $H^\pm(\rho,y) = 0$, if $y = 0$ or $y = b$, and $(\partial/\partial\rho)H^\pm(\rho,y) = 0$, if $\rho = R - a/2$ or $\rho = R + a/2$, where R is the radius of the circle centered at the origin and equidistant from the waveguide walls in the plane $y = 0$. The operator B is then represented by a matrix $\mathbf{B}$ and the functions H^+ and H^- transfers to amplitude coefficients in vector form $\mathbf{H}^+$ and $\mathbf{H}^-$. Equation (6) then becomes

$$H_y(\rho,\phi,y) = \mathbf{x}^T\cdot(e^{i\mathbf{B}\phi}\cdot\mathbf{H}^+ + e^{-i\mathbf{B}\phi}\cdot\mathbf{H}^-), \tag{7}$$

where the $\mathbf{x}$ is basis functions in vector form, whose components are given by $x_m = \cos((m-1)\pi(\rho-(R-a/2))/a)\sin(\pi y/b)$. We do not use expansion of the fields in terms of the eigenfunctions of B^2 in order to avoid numerical problems associated with Bessel functions when the radius of curvature for the bend is large. Also, more importantly, matching procedures are easier to handle with trigonometric basis functions. This applies to matching of the bend with straight waveguides as well as a coming matching of discontinuities in the bend.

The solution for E_y for the H-plane bend is obtained by the same procedure and becomes

$$E_y(\rho,\phi,y) = \mathbf{x}^T\cdot(e^{i\mathbf{B}\phi}\cdot\mathbf{E}^+ + e^{-i\mathbf{B}\phi}\cdot\mathbf{E}^-), \tag{8}$$

where the components of $\mathbf{x}$ in this case are given by $x_m = \sin(m\pi(\rho-(R-a/2))/a)$. The matrix $\mathbf{B}$ is determined by expansion of the operator B in these basis functions.

3 Matching procedures

The solution to (2) in the straight waveguide connecting to the right in Figure 1 for the E-plane case is given by

$$H_y(x,y,z) = \mathbf{x}^T\cdot(e^{i\boldsymbol{\beta}z}\cdot\mathbf{A}_1^+ + e^{-i\boldsymbol{\beta}z}\cdot\mathbf{A}_1^-), \tag{9}$$

where $\boldsymbol{\beta}$ is a diagonal matrix, and $\mathbf{A}_1^\pm$ are the amplitude coefficients associated with propagation in the positive and negative z direction, respectively. The matrix $\boldsymbol{\beta}$ is diagonal due to the fact that the basis functions $x_m = \cos((m-1)\pi x/a)\sin(\pi y/b)$ are eigenfunctions of the operator $(\nabla^2 + k^2) - \partial^2/\partial z^2$ with eigenvalues $k^2 - ((m-1)\pi/a)^2 - (\pi/b)^2$. In the other straight waveguide connecting to the bend we have the same representation for the magnetic field component, but the amplitude coefficients are of course different, $\mathbf{A}_1^\pm \rightarrow \mathbf{A}_2^\pm$. The amplitudes $\mathbf{A}_1^+$ and $\mathbf{A}_2^-$ are associated with waves coming into the bend from the two straight waveguides, while $\mathbf{A}_2^+$ and $\mathbf{A}_1^-$ are the unknown scattered wave amplitudes.

To have a complete set of equations for the scattered wave amplitudes we have to use the continuity condition for the transverse electric field components across the bend interfaces with the straight waveguides. The electric field components $E_\rho(\rho,\phi,y)$ in the bend and $E_x(x,y,z)$ in the straight waveguide,

are determined from (7) and (9) by using (3), where the operator $\nabla_\perp = \hat{\mathbf{x}}\,\partial/\partial x + \hat{\mathbf{z}}\,\partial/\partial z$ in the straight waveguide. The four continuity equations for the transverse magnetic and electric field components at the bend interfaces $\phi = 0$ and $\phi = \phi_0$, with $z = 0$ at both ends of the bend, then become

$$\begin{aligned} &\mathbf{A}_1^+ + \mathbf{A}_1^- = \mathbf{H}^+ + \mathbf{H}^- \\ &\mathbf{A}_2^+ + \mathbf{A}_2^- = e^{i\mathbf{B}\phi_0} \cdot \mathbf{H}^+ + e^{-i\mathbf{B}\phi_0} \cdot \mathbf{H}^- \\ &\mathbf{S}\cdot(\mathbf{A}_1^+ - \mathbf{A}_1^-) = \mathbf{H}^+ - \mathbf{H}^- \\ &\mathbf{S}\cdot(\mathbf{A}_2^+ - \mathbf{A}_2^-) = e^{i\mathbf{B}\phi_0} \cdot \mathbf{H}^+ - e^{-i\mathbf{B}\phi_0} \cdot \mathbf{H}^-. \end{aligned} \tag{10}$$

S is determined from the composition of the operators $(\rho^{-1}\partial/\partial\phi)^{-1} \circ (\partial/\partial z)$, where $(\partial/\partial z)$ is acting on H_y, given by (9), in the straight waveguide and $\rho^{-1}\partial/\partial\phi$ is acting on H_y, given by (7), in the bend. The condition that $z = 0$ at both ends of the bend is equivalent to regard the bend as being a discontinuity in the straight waveguide located at the plane $z = 0$. The solution of this set of equations can be found by straightforward matrix algebra and is represented as

$$\begin{pmatrix} \mathbf{A}_1^- \\ \mathbf{A}_2^+ \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{T} & \mathbf{R} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{A}_1^+ \\ \mathbf{A}_2^- \end{pmatrix}. \tag{11}$$

The reflection and transmission matrices $\mathbf{R}$ and $\mathbf{T}$ depend on the matrices $\mathbf{B}$ and $\mathbf{S}$ and the angle ϕ_0.

The scattering properties of a multiple bend structure can be deduced from (11) by cascade techniques. There are two main blocks from where any multiple bend structure can be constructed. They are the cascade U-block and the cascade S-block. In the U-block the two centres of curvatures are on the same side of the waveguide, while for the S-block the two centres are on opposite sides of the waveguide. If there is no straight guide between the two bends then the cascade reflection matrix for the U-block is given by

$$\mathbf{R}_1 + \mathbf{T}_1\mathbf{R}_2(\mathbf{I} - \mathbf{R}_1\mathbf{R}_2)^{-1}\mathbf{T}_1, \tag{12}$$

where $\mathbf{I}$ is the identity matrix. The indices 1 and 2 refer to the first and second bend encountered by a wave propagating in the positive z direction. The cascade reflection matrix for S-block, when no straight guide is between the bends, is given by

$$\mathbf{R}_1 + \mathbf{T}_1\mathbf{R}_2^s(\mathbf{I} - \mathbf{R}_1\mathbf{R}_2^s)^{-1}\mathbf{T}_1, \tag{13}$$

where $\mathbf{R}_2^s = \mathbf{M}^{-1}\mathbf{R}_2\mathbf{M}$. The diagonal matrix $\mathbf{M}$ is a transformation matrix of the basis vector (basis functions) from the coordinate system of the second bend to the coordinate system of the first bend. The explicit expression for the transformed basis functions is

$$x_m^s = (-1)^{(m-1)}\, x_m, \tag{14}$$

hence the diagonal elements of the matrix $\mathbf{M}$ are given by $(-1)^{(m-1)}$.

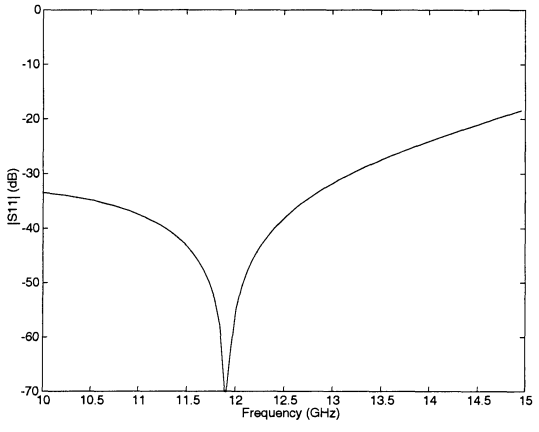


Figure 2: Reflection coefficient of an S-shaped cascaded 45^0 E-plane bend in a WR-75 rectangular waveguide, with $R = 8$ mm, $a = 9,525$ mm and $b = 19,05$ mm..

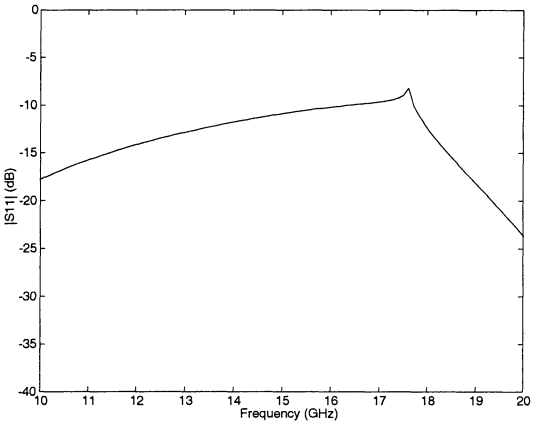


Figure 3: Reflection coefficient of an E-plane bend with right angle in a WR-75 rectangular waveguide with $R = a/2$, $a = 9,525$ mm and $b = 19,05$ mm.

4 Numerical results

To verify the method presented in this paper, we compare our results with [1] for the reflection coefficient of an S-shaped E-plane bend (Figure 2), and with [2] for a single E- and H-plane bend with inner radii of curvature equal to zero (Figure 3 and Figure 4). We find that our results are in very good agreement with the results of [1] and [2].

The numerical efficiency of the method presented here is very good since the relative error of the reflection coefficient is less than 0,1% if five expansion terms are used, and less than 0,01% if nine expansion terms are used. The analysis for the S-shaped E-plane bend in Figure 2 with nine expansion terms and 100 data points takes approximately 10 seconds on a PC with a Pentium *III* processor.

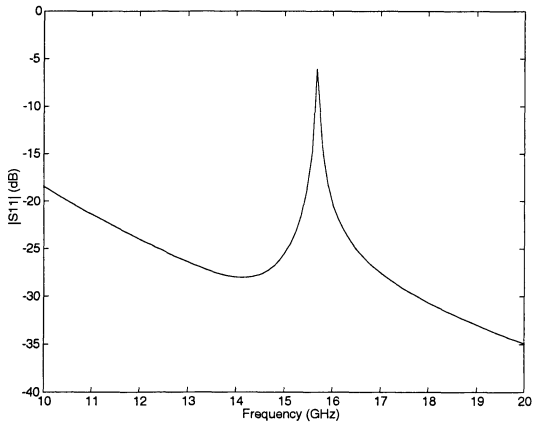


Figure 4: Reflection coefficient of an H-plane bend with right angle in a WR-75 rectangular waveguide, with $R = a/2$, $a = 19,05$ mm and $b = 9,525$ mm.

References

[1] A. Weisshaar, S. Goodnick and V. K. Tripathi, "A Rigorous and Efficient Method of Moments Solution for Curved Waveguide Bends", *IEEE Trans. Microwave Theory Tech.*, vol. 40, pp. 2200-2206, December1992.

[2] P. Cornet, R. Dusséaux and J. Chandezon, "Wave Propagation in Curved Waveguides of Rectangular Cross Section", *IEEE Trans. Microwave Theory Tech.*, vol. 47, pp. 965-972, July 1999.

[3] R. H. MacPhie and K. L. Wu, "A Full-Wave Modal Analysis of Inhomogeneous Waveguide Discontinuities with Both Planar and Circular Cylindrical Boundaries", *IEEE Trans. Microwave Theory Tech.*, vol. 49, pp. 1132-1136, June 2001.

Microwaves in waveguides with curvature and varying rectangular cross-section

Börje Nilsson[1] and Thomas Biro[2]
[1]School of Mathematics and Systems Engineering, Växjö University, SE-351 95 VÄXJÖ.
E-mail: borje.nilsson@msi.vxu.se
[2]School of Mathematics and Systems Engineering, Växjö University, SE-351 95 VÄXJÖ.
E-mail: thomas.biro@msi.vxu.se

Abstract

A theory is presented for harmonic waves in waveguides with rectangular cross-sections where curvature and cross-section vary in a plane that may change along the waveguide; both smooth variations and corners are included. Therefore, undulating threedimensional waveguides with straight ends are accessible. A cascade technique reduces the problem to flat waveguides with all variations in one plane. Since the electric and magnetic fields can be calculated from their components normal to the waveguide plane, a Fourier expansion reduces the transmission problem to a two-dimensional set of scalar problems. The waveguide is transformed to a straight one using a conformal mapping, which gives a safe treatment of complicated waveguides. To handle the numerical instability in variable waveguides, a splitting technique gives stable ordinary differential equations for reflection and transmission matrices; standard MATLAB-routines are used for their computation. The size of the required matrices is particularly small if the waveguide is smooth.

1 Introduction

A microwave filter is often based on a system of waveguides. With a demand for higher performance within a more confined space there is an increasing interest for exploring more complicated geometries. For this purpose reliable design tools are required.

Waveguides with sudden area changes and sharp edges are conveniently treated, [1], with mode matching, Wiener-Hopf or modified Wiener-Hopf methods in combination with cascade techniques. FEM is popular mainly due to its great potential for attacking complicated geometries. Since FEM calculations are time consuming and the method requires regularization of the inherent illposedness for stationary problems, alternative methods are of interest.

Recently, a few papers are published in which the waveguide is divided into a number of parts and the boundary is approximated by piecewise straight lines or circle segments. The fields are then expanded in a suitable function system, often of Fourier-Bessel type, followed by a mode matching procedure. This was first done by [2] and has then been further developed by [3] and [4].

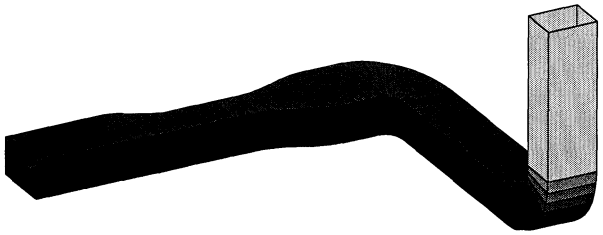


Figure 1: Rectangular waveguide with curvature and varying cross-section. In the left part the variations take place in the horisontal plane while the last bend is in a vertical plane.

For waveguides with a slowly varying cross-section, asymptotic and related analytical methods have been used [5, 6] for scalar waves. Using a theory based on first order differential equations for acoustic waves, [7] treats straight waveguides with a continuously variation of cross-section that is not slow. However, inconsistent results are reported by [7] for applications on second order differential equations. Inspired by [7] and [8], [9] gives a theory for the acoustic transmission in curved waveguides with a varying cross-section, also proving asymptotic stability for the method. The present work is a generalization of these results to electromagnetic waves but restricted to waveguides with rectangular cross-sections.

To this end we develop a stable method for calculating the transmission of harmonic electromagnetic waves in the waveguide, with perfectly electric or magnetic walls, depicted in figure 1. The cross-section is rectangular in form but may vary in size. In addition, the waveguide may be curved. Variations in cross-



Figure 2: Part waveguide parallel to the $x-y$ plane and with constant heigth.

section and curvature can only take place in one plane but that plane may vary along the waveguide. In this way only singly curved boundaries of the waveguide are allowed. Both ends of the waveguide are asymptotically straight with a constant cross-section.

2 Formulation in terms of two-dimensional scalar problems

It is possible to greatly reduce the complexity of the transmission or scattering problem by using a cascade technique. Divide the waveguide into a number of part waveguides so that each part is located in one plane only. Now, the scattering properties of the full waveguide can be determined from the corresponding properties of its parts. Therefore, it is sufficient to be able to handle waveguides of the type shown in figure 2, where the waveguide is parallel to the $x-y$ plane.

Let us divide the electric and magnetic fields into components parallel and perpendicular to the $x-y$ plane such that $\mathbf{E}=\mathbf{E}_\perp+\widehat{\mathbf{z}}E_z$ and $\mathbf{H}=\mathbf{H}_\perp+\widehat{\mathbf{z}}H_z$. Since the boundary conditions for E_z and H_z cannot generally be satisfied simultaneously, the fields divide themselves into two distinct categories:

H-plane wave
$H_z=0$ everywhere
$\mathbf{E}_\perp$ and $\mathbf{H}_\perp$ are expressed in terms of E_z

E-plane wave
$E_z=0$ everywhere
$\mathbf{E}_\perp$ and $\mathbf{H}_\perp$ are expressed in terms of H_z

Note that this division of the fields into parallel and perpendicular components differs from the standard TE/TM/TEM theory for straight waveguides where the reference direction coincide with the waveguide axis.

Denoting the perpendicular component $-\widehat{\mathbf{z}}\times(\widehat{\mathbf{z}}\times\boldsymbol{\nabla})$ of the nabla operator as $\nabla_\perp$ we get from Maxwell's equations that

$$\begin{pmatrix}\frac{\partial^2}{\partial z^2}+k^2\end{pmatrix}\mathbf{E}_\perp=\boldsymbol{\nabla}_\perp\tfrac{\partial}{\partial z}E_z-\mathrm{i}\omega\mu\widehat{\mathbf{z}}\times\boldsymbol{\nabla}_\perp H_z,$$
$$\begin{pmatrix}\frac{\partial^2}{\partial z^2}+k^2\end{pmatrix}\mathbf{H}_\perp=\boldsymbol{\nabla}_\perp\tfrac{\partial}{\partial z}H_z+\mathrm{i}\omega\epsilon\widehat{\mathbf{z}}\times\boldsymbol{\nabla}_\perp E_z, \quad (1)$$

It is assumed a time harmonic dependence $\exp(-\mathrm{i}\omega t)$ and that the waveguide is filled with a uniform nondissipative medium having dielectric constant ε and permeability μ. As usual $k=\omega/c$ is the free field wavenumber.

It is a great simplification to treat each waveguide part like the one in figure 2 separately, having no variation in the z-direction. To this end we expand the electric and magnetic fields

$$\begin{cases}\mathbf{E}=\sum_l\widetilde{\mathbf{E}}^l\varphi(l\pi z/d)\\ \mathbf{H}=\sum_l\widetilde{\mathbf{H}}^l\xi(l\pi z/d)\end{cases}, \quad (2)$$

with d being the height of the waveguide. If the plane horizontal walls are electric then $\varphi=\cos$ and $\xi=\sin$, while $\varphi=\sin$ and $\xi=\cos$ for magnetic walls. There is no coupling between different l-components of $\widetilde{\mathbf{E}}^l$ and $\widetilde{\mathbf{H}}^l$ inside each waveguide part. Such couplings do occur between waveguide parts lying in different planes and are handled with the cascade technique. Consequently, there is usually no risk for confusion to suppress the super index l. With $\kappa_l^2=k^2-l^2\pi^2/d^2$ and $\nabla_\perp^2=\nabla^2-\partial^2/\partial z^2$ we get

$$(\nabla_\perp^2+\kappa_l^2)\begin{pmatrix}\widetilde{E}_z^l\\ \widetilde{H}_z^l\end{pmatrix}=0, \quad (3)$$

which together with boundary and radiation conditions can be used to determine $\widetilde{E}_z^l$ and $\widetilde{H}_z^l$, and from (2) and (1) the entire electric and magnetic fields $\mathbf{E}$ and $\mathbf{H}$, respectively. We restrict ourselves in this paper to the E-plane case with electrical vertical walls on which $\partial\widetilde{H}_z^l/\partial n=0$.

In order to get a stable formulation we decompose $\widetilde{H}_z^l=\widetilde{H}_z^{l+}+\widetilde{H}_z^{l-}$ in right- and leftgoing waves in the straight terminating waveguides. Assume now that the incoming waves $\widetilde{H}_{z,\mathrm{in}}^{l+}$ and $\widetilde{H}_{z,\mathrm{in}}^{l-}$, from left and right, respectively, are known. Then we write

$$\begin{cases}\widetilde{H}_z^l=\widetilde{H}_{z,\mathrm{in}}^{l+}+R^+\widetilde{H}_{z,\mathrm{in}}^{l+}+T^-\widetilde{H}_{z,\mathrm{in}}^{l-}, & \text{in the left part}\\ \widetilde{H}_z^l=\widetilde{H}_{z,\mathrm{in}}^{l-}+R^-\widetilde{H}_{z,\mathrm{in}}^{l-}+T^+\widetilde{H}_{z,\mathrm{in}}^{l+}, & \text{in the right part}\end{cases} \quad (4)$$

showing that the scattering problem can be expressed in terms of four scattering operators R^+,T^-,R^- and T^+. To find these scattering operators we shall solve

$$\begin{cases}\left(\nabla_\perp^2+\kappa_l^2\right)\widetilde{H}_z^l=0 \text{ in the waveguide}\\ \partial\widetilde{H}_z^l/\partial n=0 \text{ on the boundary}\\ \widetilde{H}_z^{l+}=\widetilde{H}_{z,\mathrm{in}}^{l+} \text{ in the left part}\\ \widetilde{H}_z^{l-}=\widetilde{H}_{z,\mathrm{in}}^{l-} \text{ in the rigth part}\end{cases} \quad (5)$$

There is always a solution to (5), [10]. Except for a discrete number of eigenfrequencies $\omega=\omega_i$, $i=1,2,3,...$, the solution is also unique. While it is possible, [9], to find ω_i and the related trapped mode we will assume $\omega\neq\omega_i$ and concentrate on the scattering problem.

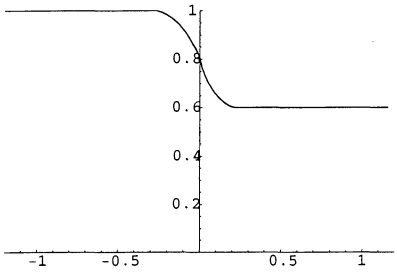


Figure 3: View in the $\zeta = x+iy$ plane of the waveguide of figure 2. $b/a = 0.6$

3 Equivalent straight waveguide with constant cross-section

There are two difficulties associated with the problem formulated in (5). First, as demonstrated by [7], the application of the Neumann boundary condition on a curved boundary may pose a mathematical difficulty. Secondly, the instability originating from a numerical propagation of a damped mode in the backward direction can give large errors. To solve the first difficulty the curved boundary is made straight by a coordinate transformation. A splitting technique, presented in section 4, avoids the instability.

To this end we use a conformal mapping $w(\zeta)$ transforming the interior of the two-dimensional waveguide with a varying cross-section in the $\zeta = x+iy$ plane. Introducing $\phi(u,v) = \widetilde{H}_z^l(x,y)$ we get

$$\begin{cases} \Delta_{u,v}\phi + \kappa_l^2\mu(u,v)\phi = 0 \text{ in the waveguide} \\ \partial_v\phi(u,0) = \partial_v\phi(u,a) = 0, u \in \mathbb{R}, \end{cases} \tag{6}$$

with $\mu(u,v) = |\mathrm{d}\zeta/\mathrm{d}w|^2$ and $\Delta_{u,v}$ denoting the Laplacian in u,v coordinates. Physically, $\mu(u,v)^{1/2}$ is the refractive index for the new waveguide. Note that the coordinate transformation $w(\zeta)$ is determined once for each geometry while the scattering problem is to be solved for all κ_l.

There are numerical packages available to produce conformal mappings. However, it is rather easy to produce such mappings for many technically interesting waveguides. We will illustrate this for the waveguide (two-dimensional view parallel to the $x-y$ plane) in figure 2. A view of of the waveguide in the $\zeta = x+iy$plane is shown in figure 3.

First we replace the smooth two-dimensional waveguide with a waveguide with two corners, having a sudden change in width at $x = 0$. The Schwarz-Christoffel mapping $\zeta = \zeta(w_1)$ defined via

$$\frac{\mathrm{d}\zeta}{\mathrm{d}w_1} = A\frac{\sqrt{w_1+g}}{\sqrt{w_1+f}}\frac{1}{w_1}, \tag{7}$$

with suitably chosen constants A, f and g, maps the upper half w_1-plane to the interior of the waveguide with two corners. By combining this with the mapping $w = (a/\pi)\log w_1$ the mapping $\zeta = \zeta(w)$ for the waveguide with two corners is found.

By replacing $\sqrt{w_1+g}$ with $\sqrt{w_1+g}+\gamma\sqrt{w_1+g'}$ and $1/\sqrt{w_1+f}$ with $\sqrt{w_1+f}-\sqrt{w_1+f'}$ in (7), the corners get rounded affecting only the boundary near the corners, [11]. The integration to find $\zeta = \zeta(w_1)$ can be done analytically, in our case using MATHEMATICA, and the shape of the rounding can be affected by changing the parameters in the mapping.

If a higher accuracy is requested iterative schemes like those in [12] for transforming circle-like figures to circles should be possible to adapt using the modified Schwarz-Christoffel mapping as a starting point.

An alternative to a conformal mapping is to use a local orthogonal transformation, [13]. It is a simpler procedure to find the transformation but numerical difficulties are expected when applying the transformation if the variation of the waveguide is large. In addition, the formulation (5) is altered which, however, is possible to handle. We use the conformal approach in the following.

4 Reformulation of the problem

In this section we will transform the partial differential equation problem (6) to a set of ordinary differential equations for the scattering operators closely related to those in (4). To this end we rewrite (6) as

$$\begin{cases} \partial_u^2\phi - B^2(u)\phi = 0 \\ \partial_v\phi(u,0) = \partial_v\phi(u,a) = 0 \end{cases}, u \in \mathbb{R}. \tag{8}$$

written in operator form, where we have introduced

$$\begin{cases} -B^2(u) = \partial_v^2 + \kappa_l^2\mu \\ \mu = \left|\dfrac{\mathrm{d}z}{\mathrm{d}w}\right|^2 \end{cases}. \tag{9}$$

The tube is asymptotically straight and we have that μ tends to constants $\mu_\pm$ when u tends to $\pm\infty$. Therefore, we denote $-B_\pm^2 = \partial_v^2 + \kappa_l^2\mu_\pm$. Introduce the operator

$$C = B_- + f(u)(B_+ - B_-), \tag{10}$$

where f is increasing with $\lim_{u\to-\infty} f(u) = 0$ and $\lim_{u\to\infty} f(u) = 1$. Consequently, the operator C varies continuously and independently of u from B_- in the left part to B_+ in the right part. To arrive at asymptotic stability a wave splitting in right $\phi^+$and left waves ϕ^- is introduced by

$$\begin{pmatrix} \phi \\ \partial_u\phi \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -C & C \end{pmatrix}\begin{pmatrix} \phi^+ \\ \phi^- \end{pmatrix}. \tag{11}$$

Note that the splitting into right and left waves is unique only in the straight parts of the waveguide.

From the dynamic equation (8) for ϕ follows the dynamic equations for ϕ^+ and ϕ^-:

$$\partial_u \begin{pmatrix} \phi^+ \\ \phi^- \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} \phi^+ \\ \phi^- \end{pmatrix}, \qquad (12)$$

where the u-dependent operators are given by:

$$\begin{cases} \alpha = \frac{1}{2}[(\partial_u C^{-1})C + iC^{-1}B^2 + iC] \\ \beta = \frac{1}{2}[-(\partial_u C^{-1})C + iC^{-1}B^2 - iC] \\ \gamma = \frac{1}{2}[-(\partial_u C^{-1})C - iC^{-1}B^2 + iC] \\ \delta = \frac{1}{2}[(\partial_u C^{-1})C - iC^{-1}B^2 - iC] \end{cases} . \qquad (13)$$

Using the dynamic equation (12) for ϕ^+ and ϕ^-, we will now derive ordinary differential equations for transmission and reflection operators. Let the 2 by 2 scattering operator $S(u_2, u_1)$,

$$S(u_2, u_1) = \begin{pmatrix} T^+(u_2, u_1) & R^-(u_1, u_2) \\ R^+(u_2, u_1) & T^-(u_1, u_2) \end{pmatrix}, \qquad (14)$$

be composed of the transmission and reflection operators T^+, T^-, R^+ and R^-, assuming $u_1 \leq u_2$. Then it is convenient to define the scattering operator $S(u_2, u_1)$ via

$$\begin{pmatrix} \phi^+(u_2) \\ \phi^-(u_1) \end{pmatrix} = S(u_2, u_1) \begin{pmatrix} \phi^+(u_1) \\ \phi^-(u_2) \end{pmatrix}. \qquad (15)$$

The waves on the right hand side of (15) are incoming waves while those on the left are scattered ones.

We start the derivation of the differential equations by requiring that the incoming wave from right $\phi^-(u_2)$ vanishes, which means that $R^+(u_2, u_2) = 0$ and $T^+(u_2, u_2) = I$. Ordinary differential equations of Riccati type can then be derived for $R^+(u_2, u)$ and $T^+(u_2, u)$. The full result is

$$\begin{cases} \partial_u R^+(u_2, u) = \gamma + \delta R^+(u_2, u) - R^+(u_2, u)\alpha \\ \qquad\qquad - R^+(u_2, u)\beta R^+(u_2, u) \\ \partial_u T^+(u_2, u) = -T^+(u_2, u)\alpha - T^+(u_2, u)\beta R^+(u_2, u) \\ \partial_u R^-(u_1, u) = \beta + \alpha R^-(u_1, u) - R^-(u_1, u)\delta \\ \qquad\qquad - R^-(u_1, u)\gamma R^-(u_1, u) \\ \partial_u T^-(u_1, u) = -T^-(u_1, u)\delta - T^-(u_1, u)\gamma R^-(u_1, u) \end{cases} \qquad (16)$$

with the starting conditions

$$\begin{cases} R^+(u_2, u_2) = 0 \\ T^+(u_2, u_2) = I \\ R^-(u_1, u_1) = 0 \\ T^-(u_1, u_1) = I \end{cases} . \qquad (17)$$

This ends the derivation of the operator equations.

5 Numerical results

In order to use numerical methods we expand ϕ and μ in Fourier cosine series

$$\begin{cases} \phi(u, v) = \sum_{n=0}^{\infty} \phi_n(u) \cos(n\pi v/a), \\ \mu(u, v) = \sum_{n=0}^{\infty} \mu_n(u) \cos(n\pi v/a). \end{cases} \qquad (18)$$

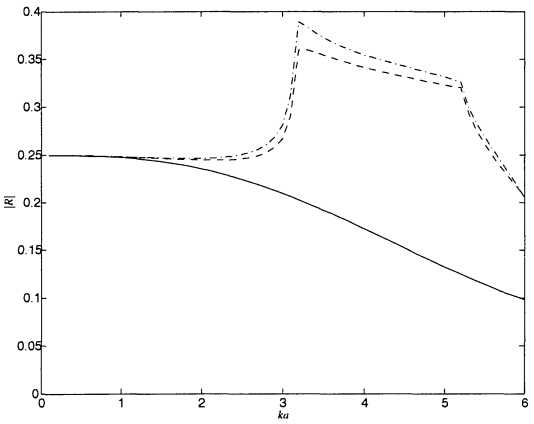


Figure 4: Magnitude of reflection coeefficient for the lowest mode in the waveguide in figure 2 as function of frequency ka with truncation N as parameter. Vertical walls are electric while horisontal are magnetic. — $N = 1$; − − − $N = 2$; · · · · $N = 3$; − · −· $N = 5$.

By introducing the vector $\boldsymbol{\phi} = (\phi_0, \phi_1, ...)^T$ we find that

$$\frac{d^2\boldsymbol{\phi}(u)}{du^2} - \mathbf{B}^2(u)\boldsymbol{\phi}(u) = 0, \qquad (19)$$

where the matrix elements of $\mathbf{B}^2(u)$ are, assuming $\mu_p(u) = 0$ for negative p,

$$\mathbf{B}^2(u)_{nm} = -\frac{\kappa_l^2}{2}[\mu_{m+n}(u) + \mu_{m-n}(u)(1 - \delta_{n0}) + \mu_{n-m}(u)] + \frac{n^2\pi^2}{a^2}\delta_{nm}, \; n, m = 0, 1, 2, ... \qquad (20)$$

In this way the operator differential equations (16) are transformed to matrix differential equations, which are truncated to size N and results are presented for size N_0, with $N \gg N_0$. [9] shows that results for size N_0, N_0 fixed, converges when $N \to \infty$. Convergence is to be understood in the sense of properly defined Sobolev spaces.

To illustrate the method, reflection data are presented in figure 3 for the E-bend of figure 2. While the horizontal walls are magnetic, the vertical walls are electric. As expected a scalar equation, i.e. truncation $N = 1$, is sufficient for low frequencies. More surprising is perhaps that already $N = 2$ gives a qualitatively correct behaviour even above the first cut-on at $ka = \pi$. For $N \geq 5$ there is no visible change in the magnitude of the reflection due to difference in truncation.

[9] has used the method for the transmission of acoustic waves in an L-bend and has demonstrated how to solve the transmission of acoustic waves in waveguides with non-rectangular cross-sections. The first case transfers easily to electromagnetic waves while the latter is more difficult in this respect due to complicated polarization in such waveguides.

References

[1] R. Mittra & S. W. Lee 1971 *Analytical Techniques in the Theory of Guided Waves.* New York: Macmillan.

[2] J. M. Reiter 1995 Rigorous Analysis of Arbitrarily Shaped H- and E-plane Discontinuities. in Rectangular Waveguides by a Full-Wave Boundary Contour Mode-Matching method. *IEEE Trans. Microwave Theory Tech.* , vol. 43, pp 796-801.

[3] R. H. MacPhie & K.-L. Wu 1999 A Full-Wave Modal Analysis of Arbitrarily Shaped Waveguide Discontinuities. Using the Finite Plane-Wave Series Expansion. *IEEE Trans. Microwave Theory Tech.* , vol. 47, pp 232-237

[4] R. H. MacPhie & K.-L. Wu 2001 A Full-Wave Modal Analysis of Inhomogeneous Waveguide Discontinuities. with Both Planar and Circular Cylindrical Boundaries. *IEEE Trans. Microwave Theory Tech.* , vol. 49, pp 1132-1136.

[5] C. K. W. Tam 1971 Transmission of spinning acoustic modes in a slightly non-uniform duct. *J. Sound Vib.* **18**, 339-351.

[6] A. H. Nayfeh & D. P. Telionis 1973 Acoustic propagation in ducts with varying cross-sections. *J. Acoust. Soc. Am.* **54**, 1654-1661.

[7] V. Pagneux, N. Amir & J. Kergomard 1996 A study of wave propagation in varying cross-section waveguides by modal decomposition .1. Theory and validation. *JASA* **100** (4), 2034-2048.

[8] L. Fishman 1993 One-way propagation methods in direct and inverse scalar wave propagation modeling. *Radio Science* **28(5)**, 865-876.

[9] B. Nilsson 2002. Acoustic transmission in curved ducts with varying cross-sections. Accepted for publication in the proceedings of Royal Society A.

[10] M. Cessenat 1996 *Mathematical methods in electromagnetism.* Singapore: World Scientific Publishing Co.

[11] P. Henrici 1988 *Applied and computational complex analysis. Volume* **I**. New York: John Wiley & Sons.

[12] R. Schinzinger & P. Laura 1991 *Conformal mapping: Methods and applications.* Amsterdam: Elsevier.

[13] Y. Y. Lu, J. Huang & J. R. McLaughlin 2001 Local Orthogonal Transformation and One-way Methods for Acoustics Waveguides, *Wave Motion* , Vol 34/2, pp. 193-207.

Numerical techniques for mixed high-low frequency EM-problems*

Stefan Hagdahl[1,2], Bjorn Engquist[2]

[1]Saab Avionics AB and Electromagnetic Technology, Linkoping. stefanh@nada.kth.se.

[2]Dep. of comp. science and num. analysis at Royal institute of tech., Stockholm. engquist@nada.kth.se.

Abstract

Our poster present numerical methods to be used for wave propagation problems where both relatively short and long wavelengths, λ, are present. Short and long wavelengths is referred to the characteristic sizes of the scales in the boundary conditions (BC) and material properties (MP). We propose that, depending on properties of the BC, one of three different Boundary Element Methods (BEM) are hybridised with a wave front, as opposed to a ray, implementation of Geometrical Theory of Diffraction (GTD). We will only consider the scattering problem in the presentation.

Sometimes the electromagnetic engineer needs to know how a small size feature, e.g. a small perturbation of the BC, affects scattering parameters. We indicate the original Scatterer by an upper case S and the small scatterer, i.e. the perturbation, with lower case s. If the perturbation together with the Scatterer is small, BEM will serve as an excellent tool for numerical computations. However, when the complexity of the problem overcome the computer resources, other methods must be used. We propose that either of the three following boundary element methods are hybridised with GTD.

If the perturbation only consist in an addition of a geometrical detail, i.e. the Scatterer's BC and MP is unaffected by the perturbation, the perturbation can be discretized with local boundary elements and the impact on scattering parameters can be taken into account with GTD. If a boundary element is close to the scatterer, hence does not illuminate the scatterer with a ray field, we propose to use a linear combination of the image solution and the GTD to model the illumination. From an implementation point of view the interaction between the perturbation and the scattering is preferably done out-of-core and iteratively thereby leaving the GTD-solver and the BEM-solver separated.

If the perturbation actually consist of a modifications of the Scatterer we propose that, at least in the two dimensional case, edge diffraction is applied to the artificial edges that indicate the points separating the perturbed and unperturbed parts of the Scatterer. No iteration needs to be done in this case since we assume that all scattered waves are outgoing.

The third BEM is to be used when the perturbation actually changes the scatterer and the shape of the Scatterer close to the perturbation is crucial and needs to be rigorously taken into account. We then need to have mixed local and global boundary elements in the BEM-domain to model the interaction properly. We discuss how to formulate the boundary integral problem for a small change in a infinite electric perfect conducting ground plane.

To take into account features that are far away from the perturbation can, as mentioned above, GTD be used. GTD has its origin in the method of characteristics. The characteristics are typically found with two types of methods in CEM-software.

- By using analytical or numerical methods applied to a variational problem with constraints. The method will fail in caustics since we then have infinite many solutions.
- By using the so-called shooting method. The shooting method succeed in caustics but may produce bad resolution for certain problems.

We propose that our implementation of GTD, which we call the Wave Front method (WF-method), is used instead of the two methods presented above since we then will avoid the problems with efficiency and accuracy. It can be used in a CAD-setting and also be applied to smooth surface diffraction and inhomogeneous-an isotropic materials.

We will present numerical results from the presented methods in future publications.

1 Introduction

Every electrical engineer knows that it is hard to say, ad hoc, that geometrically small sized features necessarily make a relatively small impact on electromagnetic parameters. In extreme cases such as if small

*This work was supported by the CEM program at the Parallel Scientific Computing Institute at the Royal Institute of Technology. Project name SMART.

sized features in the order of wavelength size are installed on stealth fighter aircraft they can actually affect the electromagnetic signature tremendously in specific sectors. The engineer thereby have to include all small features of order $\sim \lambda$ when the signature analysis is done. This means of course great demands on computer resources and engineering time. To complement rigorous methods we therefore present three methods to be used on for example new editions or updates of fighter aircrafts.

As an illustration of what type of problems that can arise in signature analysis we can think of two types of antennas. Exterior antennas, objects that actually are attached to the geometry and therefore leave the initial geometry unaffected, and structure integrated antennas, objects that we integrate in the skin. We think that the signature analysis can be done with different types of numerical methods for each type of antenna and will therefore split up the presentation in two parts, paragraph 2 and 3.

Although we will focus on the scattering case in this presentation we emphasis that the methodology we use is also applicable to antenna analysis such as antenna performance- and antenna-to-antenna isolation computations. Other disciplines where large geometries are present together with small ones are the photographic cards used by the integrated circuit manufacturers. Wavelength size notches are present in a ground plane of hundreds or thousands of wavelengths.

In contrast to the rather intricate and complicated geometries that are present in the applications mentioned above we will focus on very simple geometries in the discussion below. We do not think though that there are any difference of principle with the applications in terms of the numerical method to be used.

2 Unperturbed Scatterer

For the case when the Scatterer is unperturbed we suggest that a Boundary Element Method is combined with GTD. Referring to figure 2 our representative problem will be a metal strip above a infinite ground plane. We want to compute how much the scattered field change when the boundary conditions associated to the metal strip is added. Mathematically we search

$$\lim_{r\to\infty} 4\pi r^2 E^{scatt}_{partial} \sim \lim_{r\to\infty} 4\pi r^2 (E^{scatt}_{entire} - E^{scatt}_{large}) \quad (1)$$

2.1 Theory

We will not mathematically present the boundary integral equation but just say that when it is discretized it will give us a matrix equation $MJ = E^{inc}$ where J represent the unknown currents, M the mass matrix and E^{inc} the exciting field. If we would solve the problem with BEM solely the exciting field would be a plane wave for the signature problem. To avoid discretizing the scatterer we represent it by two extra term in the right hand side in the matrix equation. One is developed by computing all significant terms in the GTD expansion for the true source (the plane wave) and the receivers (the boundary elements) on the scatterer (with the Scatterer present). Differently stated this part of the excitation asymptotically fulfil the BC on the Scatterer if the scatterer where not present. The third term represent the interaction between the Scatterer and the scatterer where the boundary elements represent sources to itself, i.e. self-illumination. In principle we could modify the mass matrix $M \to M + M^{GTD}$ to incorporate the interaction with the Scatterer. This is in practice done by computing all relevant GTD-terms for all unit base functions represented on the scatterer. This has been investigated, especially for one dimensional boundary elements, in some papers in the 70s.

Instead we choose for practical reasons to incorporate the self interaction iteratively, i.e.

$$MJ_{j+1} = E^{Plw} + E^{GTD} + E^{Slf}_j \quad (2)$$

$$E^{Slf}_{j+1} = F(J_{j+1}) \quad (3)$$

where $E^{Slf}_0 = 0$. If the scatterer where placed far away from the Scatterer we then could use GTD straight forward to compute the self term. However we know that that the field from the currents on a BEM-triangle will not produce a Ray-field close to the triangle. Therefore we use image theory to compute E^{Slf}_j when a BEM-triangle is close to the Scatterer.

Since image theory only holds for infinite ground planes we cannot expect that the method holds exactly for the more realistic cases but if the BEM-triangle is close to the ground plane we think that the errors are negligible.

In the end we want to achieve a linear combination between the GTD-field (which holds far away from a curved Scatterer) and the Image-field (which holds close to a flat Scatterer).

Mathematically we search

$$E^{Slf}_j = \alpha(\rho) E^{Image}_j + (1 - \alpha(\rho)) E^{GTD}_j, \quad (4)$$

where we defined $\rho = |\hat{\mathbf{n}} \cdot (\mathbf{r}_{grpl} - \mathbf{r}_{tri})|/\lambda$ and $\hat{\mathbf{n}}$ as normal to ground plane. Probably there are optimal functions α, to be used for each specific problem. However we propose that the most simple is as good as anyone for a general solver. We therefore propose that $\alpha(\rho) = \rho$ for $\rho = [0 \dots c]$ ($\rho = 0$ otherwise). c can be decided and optimised numerically for a infinite ground plane.

Although the theory are developped for PEC boundary conditions we see no problem with applying the method to dielectric material.

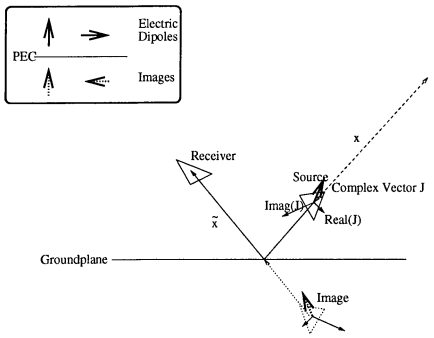


Figure 1: Approximation of image theory for a small source of electric current.

We want to stress that the method presented here are not suited for problems where the scatterer is buried in the Scatterer since then we have a type of cavity and resonant phenomenon appear which neither GTD or Image-theory model well.

2.1.1 Image field

We know from image theory that a boundary value problem represented by an infinitesimal dipole above an infinite ground plane can be solved by removing the ground plane and add a rotated dipole in the image point. As we see in figure 1, we do not have a dipole but a triangle with a prescribed complex valued surface current, $\mathbf{J}(\mathbf{x})$, with $\mathbf{x} \in$ triangle. Approximately though, we can model the current with two dipoles (real and imaginary part) placed in the triangle's Gauss node. Our numerical tests in the future will try to answer the question whether the image theory work in this case also. By iteration we can search a solution to the boundary value problem that approximately fulfil the boundary condition along the ground plane, i.e. $\mathbf{E}_{tan} = 0$. Notice that if we actually would have used electrical dipoles to discretize the scatterer in the BEM-solver then we did not have to iterate between the scatterer and its image since image theory provide us with the solution.

Let us call the surface current density on triangle T_i for $\mathbf{J}_i(x)$ where $\mathbf{x} \in T_i$. If the triangle is small and oblong we can expect that it is a good approximation to replace the surface current on triangle T_i with two infinitesimal dipoles $\mathbf{D}_i^{(1)}$ and $\mathbf{D}_i^{(2)}$ in the first Gauss node. We compute them by doing

$$\mathbf{D}_i^{(1)} = Real(\bar{\mathbf{J}}_i) = Real(\int_{Tri_j} \mathbf{J}(\mathbf{x})d\mathbf{x}) \qquad (5)$$

$$\mathbf{D}_i^{(2)} = Imag(\bar{\mathbf{J}}_i) = Imag(\int_{Tri_j} \mathbf{J}(\mathbf{x})d\mathbf{x}). \qquad (6)$$

What we need in the algorithm is a mean to compute how the Image triangle radiate, i.e. we need a function to be called. Either we could create such function by creating the image $\tilde{T}_i^{(j)}$ as a data structure of its own and associate a function to it, $(\mathbf{E}, \mathbf{H}) = \tilde{T}_\text{field}_j(\mathbf{r})$, which return the field in a point $\mathbf{r}$. However, from an implementation point of view it is easier to define a transformation that transform $\mathbf{r}$ to a new point $\mathbf{r}'$ which fulfil

$$\tilde{T}_\text{field}_j(\mathbf{r}) = T_\text{field}_j(\mathbf{r}') \qquad (7)$$

for all points $\mathbf{r}$. We can associate a coordinate system, $(\hat{e}_1^{(j)}, \hat{e}_2^{(j)}, \hat{e}_3^{(j)})_i$, to each of the triangles. Let both of the $\mathbf{D}_i$s point in the $(\hat{e}_3^{(j)})_i$-direction. By associating standard polar coordinates to each coordinate system we realize that we cannot expect that $T_\text{field}_j(\theta, \phi) = T_\text{field}_j(\theta)$ holds for a triangle current. However we can expect that the approximation $T_\text{field}_j(\theta, \phi) = T_\text{field}_j(\pi - \theta, \phi)$ may hold.

By inspection we realize that transformation $\mathbf{r} \rightarrow \mathbf{r}'$ can be computed.

According to image theory, see figure 1, we know that the image dipole of $D_i^{(j)}$ is achieved by multiplying $\mathbf{D}_i^{(j)}$ with

$$T = \begin{pmatrix} 2\hat{n}_1\hat{\mathbf{n}} - 1 \\ 2\hat{n}_2\hat{\mathbf{n}} - 1 \\ 2\hat{n}_3\hat{\mathbf{n}} - 1 \end{pmatrix} \qquad (8)$$

where we have defined $\hat{n}$ as the normal to the PEC surface in the reflection point. The image source is put in the image point of the ground plane. However T is not the only matrix that do the job. If we interpret T as a rotation we realize, since $\hat{\mathbf{n}}$ is an eigenvector to T, it prescribes a rotation around $\hat{\mathbf{n}}$, π radians. To comply with the assumed symmetry mentioned above we should perform the rotation around $\hat{\mathbf{p}}_i^{(j)} = \hat{D}_i^{(j)} \times \hat{n}_i$. The rotation matrix, R, is achieved by using Rodrigue's formula [2]. To get the field at the receiver we then compute the position vector of the receiver relative to the image system. Let us call it $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$. By doing $\mathbf{x} = R^T\tilde{\mathbf{x}}$ we can finally get the field point that we search if we add the source origin $\mathbf{r}' = \mathbf{r} + \mathbf{x}$, where $\mathbf{r}$ is the position vector of the source. When we have computed the field in $\mathbf{r}'$ we have to apply R to $(\mathbf{E}, \mathbf{H})$ to get the correct polarisation of it. When convergence for $\mathbf{J}$ has been achieved the scattered far field is computed with a similar method, i.e. the image source is used for triangles close to the Scatterer.

The above method is applicable to gemetries that do not have electrical contact with the Scatterer. If the scatterer is actually attached we have to modify the base functions on the triangle(s) that are attached to the Scatterer. Practically it is done by adding a degree of freedom to each edge that have electrical contact with the Scatterer.

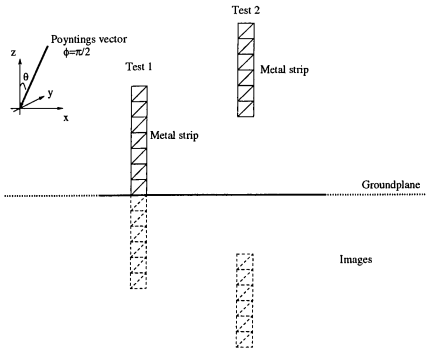


Figure 2: Two test cases.

2.1.2 GTD field

From Keller's GTD we know that any first order Ray field is described by

$$\mathbf{E}(\sigma) = \mathbf{E}(\sigma_0)\left[\frac{(\rho_1+\sigma_0)(\rho_2+\sigma_0)}{(\rho_1+\sigma)(\rho_1+\sigma)}\right]^{\frac{1}{2}} e^{-ik\sigma}. \quad (9)$$

In our method we assume that the field is spherical at the first reflection/diffraction point ($\rho_1 = \rho_2$). In this point we also put phase center and compute, with the BEM-solver, $\mathbf{E}(\sigma_0)$.

3 Perturbed Scatterer

Two methods will be presented for the perturbed Scatterer case. The first one stems from [1] and it combines non-local base functions with local ones. The second method is a variation of the method discussed in paragraph 2. Both of the methods will only be presented for the 2D-case.

3.1 Global and local base functions

Let us call the unknown $\mathbf{E}(\mathbf{r})$-field for $u(\mathbf{r})$. For the BC pictured in figure 3 we may split up u in three terms by defining the incoming field and the reflected field for the unperturbed boundary value problem. We have

$$u = u^i + u^r + u^s \quad (10)$$

where u^s is the scattered field which the perturbation give rise to. For the unknown field we have by Green's theorem

$$\begin{aligned} u(r) &= \int_\Gamma (u(r')\partial_n G(k|r-r'|) \\ &- \int_\Gamma G(k|r-r'|)\partial_n u(r'))\, dr'. \end{aligned} \quad (11)$$

Since we search the solution u to Maxwell's equation and we know that for a plane wave illumination both u^i and u^r are solutions to the equations we may replace u by u^s in (11). By using the boundary conditions in the figure we have

$$u^s = \int_{-c}^{c} -(u^i + u^r)\partial_n G dr' - \int_{-\infty}^{\infty} G\partial_n u^s dr' \quad (12)$$

The second integral has an infinite extent which cause problem when we want solve the problem numerically. We therefore hope for that $\partial_n u^s$ may be approximated by zero for $|r| > r_{max}$. If this is the case reference [1] suggests that we write u^s in polar coordinates as

$$u^s(\theta,\theta) = \sum_{n=1}^{\infty} c_n \sin(n\theta) H_n^{(2)}(k\rho) \quad |\rho| > c. \quad (13)$$

The sin-function has been chosen because the unperturbed problem require that $u^s = 0$ at $\theta = 0$ and $\theta = \pi$.

If we plug in expression (13) into equation (12) and apply the collocation method in N points for the Rooftop and the Hankel basis functions we will get a matrix problem with three different type of matrix elements. One type is the usual one showing up in BEM, the second one is an integral with one local and global one base function. The last one is an integral with two global functions. If the method is to be implemented you could start with a standard BEM-solver and then add the two new types of elements.

3.2 Local base functions

We see that in the last method the normal derivatives of the solution in the unperturbed domain is coupled to the solution in the perturbed part. That is, if the unperturbed part where curved we would, by substituting the sin-function with another appropriate one, actually take it into account when we compute the currents in the perturbed part. Differently stated, the two domains are rigorously coupled. This may be important in some cases. In other cases we may apply edge diffraction to the points which indicate where global base functions have support. The following discussion make this evident.

Let us define

$$u = u^d + u^i + u^r + u^s \quad (14)$$

where u^d solves the slot problem defined by the figure when all BC's between $-c$ and c have been removed. For $\mathbf{x}$ that fullfill $|\mathbf{x}| > c$ we know that $u^s = 0$. We then know that $u = u^i + u^r = 0$ for points outside

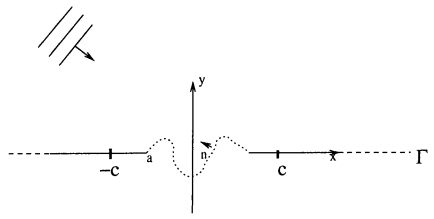


Figure 3: General case

$|\mathbf{x}| > c$. By inspecting the GTD edge diffraction coefficients we see that $\partial_n u^d = 0$ (this is true for reason of symmetry) then we also know that $\partial_n u^s = 0$ since

$$\begin{aligned} \partial_n u = \partial_n(u^i + u^r) &= \partial_n(u^i + u^r) + 0 + \partial_n u^s \\ &\Rightarrow \\ \partial_n u^s &= 0. \end{aligned} \tag{15}$$

It means that we do not have any infinite integral in (12) and we may use standard BEM-solvers to solve the integral equation. To apply this method to the three dimensional Maxwell problems we foresee some more analysis.

Notice that by making use of the method to incorporate u^d in the excitation u^s will actually represent the increase of scattered field compared with the slot problem (though the slot can be several wave length wide). However if we search the increase of scattered field compared with the ground plane we just do a computation on the infinite ground plane problem first and then take the difference.

4 Wave Fronts

To find the characteristics in the GTD we suggest that the wave front method presented in [3] is used. The WF-method is based on the method of characteristics. However, taking the 2D problem as an example, it does not represent the solution at each point s along a ray $X(\alpha_0, s)$ but on a curve orthogonal to the characteristics, i.e. the equi-phase front $X(u, s)$. With u we mean curve parameter for the equi-phase front and α_0 is the initial condition for the ray. A wave front can be represented in different spaces. If the wave front is not allowed to cross itself it must be represented by at least the three numbers (x_1, x_2, α). Here $\mathbf{x} = (x_1, x_2)$ represent the position vector for a marker and α the angle its Poynting vector make with $\hat{x}_1$.

GTD is applicable to high frequency wave propagation in inhomogenous - an isotropic material and on smooth surfaces. With inhomogeneous-an isotropic we mean that $c = c(\lambda, P, x)$ where P is the Poynting vector and c the wave speed of the ray field. Having in mind that a point on a surface can be specified by two number (x_1, x_2) in the parametric space we can show by using standard variational methods that for a marker moving with speed one, $\frac{\partial \alpha}{\partial s}$ is equal to, [3],

$$\begin{aligned} - \quad & (\Gamma^1_{11}\cos^2\alpha + \Gamma^1_{12}\sin 2\alpha + \Gamma^1_{22}\ \sin^2\alpha)\sin\alpha \\ + \quad & (\Gamma^2_{11}\cos^2\alpha + \Gamma^2_{12}\sin 2\alpha + \Gamma^2_{22}\sin^2\alpha)\cos\alpha. \end{aligned}$$

for smooth surface rays and [4]

$$\frac{n_{x_2} - n_{x_1\alpha}}{n - n_{\alpha\alpha}}\cos\alpha - \frac{n_{x_1} - n_{x_2\alpha}}{n - n_{\alpha\alpha}}\sin\alpha \tag{16}$$

for anisotropic material. Γ^k_{ij} represent the Christoffel symbols, $n(x_1, x_2, \alpha)$ the refraction index and the index on n means partial derivative. The expression for the isotropic case is just a special case of (16).

The main objective with representing the electromagnetic wave as a wave front, as opposed to single rays, are

- The wave front is geometrically well described in caustics.
- When carachteristics depleat interpolation is easily done to increase resolution.
- The wave can be represented with a minimum of data.

5 Conclusions

We have presented three different methods to be applied when small and large scales are present at the same time in a boundary value problem. This is quite common in electromagnetic ingeneering. The methods have only been presented on a theoretical level and needs of course to be justified by numerical tests. Also while doing the tests questions, on efficiency and robustness when the methods are applied to industrial CAD-settings and user friendly interfaces, will naturally be raised.

References

[1] Unpublished communication between the Authors and J. B. Keller, UCLA, spring -99

[2] Notes for EE-148: 3d Photography, Caltech spring 2001.

[3] S. Hagdahl, Tracking wave fronts with Lagrangian markers, Royal institute of technology, Department of computer science and numerical analysis,2001,Technical report, TRITA-NA-0122.

[4] O. Runborg, PM: Derivation of ODEs from Fermat's Principle. NADA, KTH. January 24, 2002.

Optimization of broadband transition from waveguide to microstrip

Thomas Rylander[1], Pär Ingelström[2], Jan Stake[3] and Anders Bondeson[4]

[1] Dept. of Electromagnetics, Chalmers, SE-41296 Göteborg E-mail: rylander@elmagn.chalmers.se

[2] Dept. of Electromagnetics, Chalmers, SE-41296 Göteborg E-mail: pi@elmagn.chalmers.se

[3] Dept. of Microelectronics, Chalmers, SE-41296 Göteborg E-mail: stake@ep.chalmers.se

[4] Dept. of Electromagnetics, Chalmers, SE-41296 Göteborg E-mail: elfab@elmagn.chalmers.se

Abstract

An extensive parameter study of a transition from a waveguide to a microstrip line is presented. The purpose is to acheive a transition which gives low reflection for a wide frequency band and we identify two geometrical parameters which have large influence on the insertion loss, namely the distance to the termination of the waveguide and the penetration depth of the probe into the waveguide. We have used the stable FEM-FDTD hybrid method [T. Rylander and A. Bondeson, "Stable FEM-FDTD hybrid method for Maxwell's equations," *Comput. Phys. Comm.* **125**, 75 (2000)] to compute the reflection coefficient and about 200 different transitions geometries were analyzed within a week on an ordinary PC.

1 Introduction

Waveguide circuits are widely used for mixer and multiplier applications above 100 GHz. Low loss and ease of use are the main reasons for their popularity. However, active devices are usually mounted on shielded microstrip circuits. This is to simplify mounting, bias circuitry, and to allow extraction of baseband signals. Therefore, a critical part of millimeter and submillimeter wave circuits is to couple energy from a waveguide mode to a microstrip mode with as low insertion loss as possible.

The standard solution is a rectangular E-plane probe extending from the microstrip line into a waveguide. For example, Tentzeris et. al studied such a transition using the FDTD scheme [1]. However, other probe shapes than rectangular can be used to reduce the return loss over a wider frequency range. For instance, the radial probe has the property that its input impedance is purely real over a very wide range of frequencies. Kaneda et. al considered a quasi Yagi antenna arrangement for broad-band operation [2]. The waveguide can have standard height or a reduced height for wider bandwidth operation.

We analyze different probe geometries and optimize them for a large bandwidth operation by using the stable FEM-FDTD hybrid scheme presented in [3]. The hybrid can handle complex geometry without resorting to the staircase approximation. The time-step is set by the Courant condition of the FDTD and it is not limited by small cells needed to resolve the geometry of the probe. The microstrip line is terminated using an absorbing boundary condition [4]. We use the PETSc [5] for the solution of the linear system of equations associated with the FEM part of the computation.

2 Geometry

The transition from a waveguide to a microstrip line is shown in Fig. 1. Half of the perfectly electric conducting (PEC) walls of the waveguide and the shield of the microstrip line are excluded to visualize the interior structure, i.e. the substrate with the microstrip and a rectangular probe.

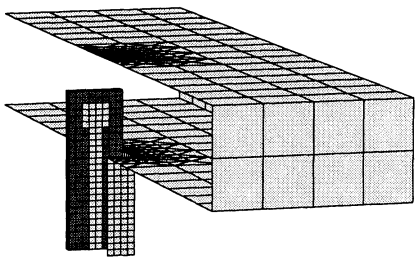


Figure 1: Parts of the discretization of a transition from waveguide to microstrip.

The cross section of the shielded microstrip transmission line is shown in Fig. 2, where the width and height of the shield are $w_m = 0.36$ mm and $h_m = 0.36$ mm, respectively. The thickness of the quartz substrate ($\epsilon_r = 3.8$) and the width of the microstrip itself are $t_s = 70$ μm and $w_l = 90$ μm, respectively.

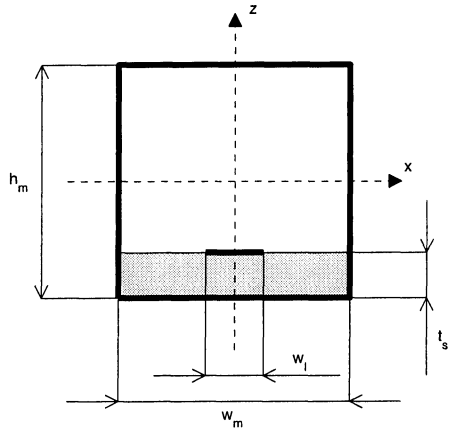


Figure 2: Cross section of the shielded microstrip line.

The cross section of the waveguide (through the microstrip line) is shown in Fig 3. The width and height of the waveguide are $w_w = 2.54$ mm and $h_w = 0.68$ mm, respectively. The substrate extends a distance $d_s = 0.68$ mm into the waveguide and the rectangular probe has the width w_p and the height h_p. The distance from the "floor" of the waveguide to the lower edge of the probe is d_p and the length of the microstrip line is $l_m = 2.04$ mm.

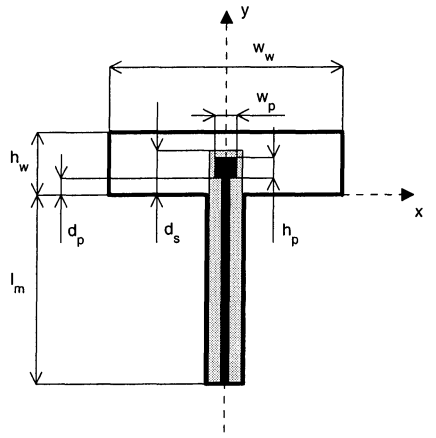


Figure 3: Cross section of the waveguide in the plane $z = 0$ (compare with Fig. 2) of the shielded microstrip line.

The incident $\vec{E}_i$, reflected $\vec{E}_r$ and transmitted $\vec{E}_t$ waves are indicated in Fig. 4. At the plane A-A shown by the dashed line to the left, the incident and the reflected fields are treated with the modal absorbing and modal source boundary conditions [6]. The transmitted fields are absorbed by a reflectionless sponge layer [4] at the end of the microstrip line, which is terminated by a PEC wall. The sponge layer occupies a length 0.68 mm of the microstrip line, which corresponds to 16 cells on the coarsest grid we used. The

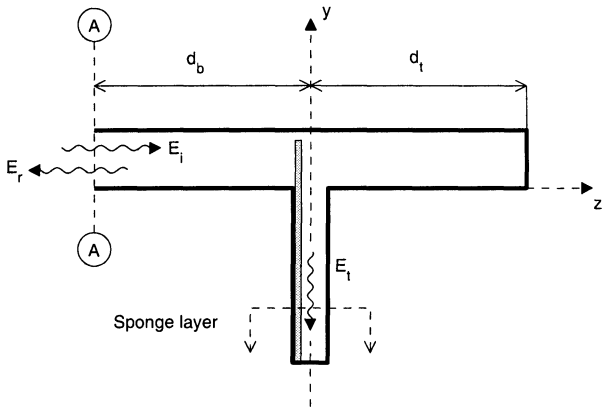


Figure 4: Cross section in the symmetry plane $x = 0$. The incidident, reflected and transmitted field are indicated by the undulating arrows.

damping parameter w'' of the sponge layer [4] is set to $0.7/\Delta t$. The waveguide is terminated by a short circuiting PEC wall at a distance d_t from the center of the microstrip line and the corresponding distance to the plane A-A is $d_b = 3.84$ mm.

3 Computation

For this setup, we impose an incident TE_{10}-mode with the amplitude $E_i(t)$ and extract the reflected TE_{10}-mode with the amplitude $E_r(t)$. The incident amplitude is given by

$$E_i(t) = E_0 \exp\left[-\left(\frac{t-t_0}{d_0}\right)^2\right] \sin(2\pi f_0(t-t_0)), \quad (1)$$

where $f_0 = 90$ GHz, $d_0 = 3/f_0$ and $t_0 = 8/f_0$. The reflection coefficient $r = E_i(f)/E_r(f)$ was computed for all feasible combinations of the parameters d_p, w_p, h_p and d_t. The range of the study is shown in discrete numbers, e.g., D_p, in Tab. 1. Here, the physical dimensions are $d_p = D_p\Delta y$, etc, and $\Delta x = 45$ μm, $\Delta y = 40$ μm and $\Delta z = 320$ μm. The combinations satisfying $D_p + H_p > 16$ were excluded since those correspond to a probe extending outside the "roof" of the waveguide.

Parameter	Integer range
$d_p = D_p\Delta y$	$D_p = 2, 4, 6, \ldots, 14$
$w_p = W_p\Delta x$	$W_p = 4, 6$
$h_p = H_p\Delta y$	$H_p = 2, 4, 6, \ldots, 14$
$d_t = D_t\Delta z$	$D_t = 2, 4, 6, 8$

Table 1: Parameter space which consists of 168 feasible geometries.

We consider the reflection coefficient S_{11} in the frequency interval $f_0 = 70$ GHz $< f < f_1 = 110$ GHz. Two different measures to quantify the reflection were used, namely the maximum value $R_\infty = \max_{f\in[f_0,f_1]} |S_{11}|$ and the root-mean-square (RMS) value $R_2 = [(f_1 - f_0)^{-1} \int_{f_0}^{f_1} |S_{11}|^2 df]^{1/2}$.

The region where $D_t = 2$ and $D_p + H_p = 10$ or 12 gives the smallest values for both R_∞ and R_2. In the sequel, we fix $D_t = 2$. A study to resolve the variation with d_t around the coarsely located optimum at 640 μm is in progress. The power reflection corresponding to R_∞ is shown in Fig. 5 for the fixed parameters $W_p = 4$ and $W_p = 6$ to the left and the right, respectively. The outer boundary of the parameter domain is shown by the black triangle.

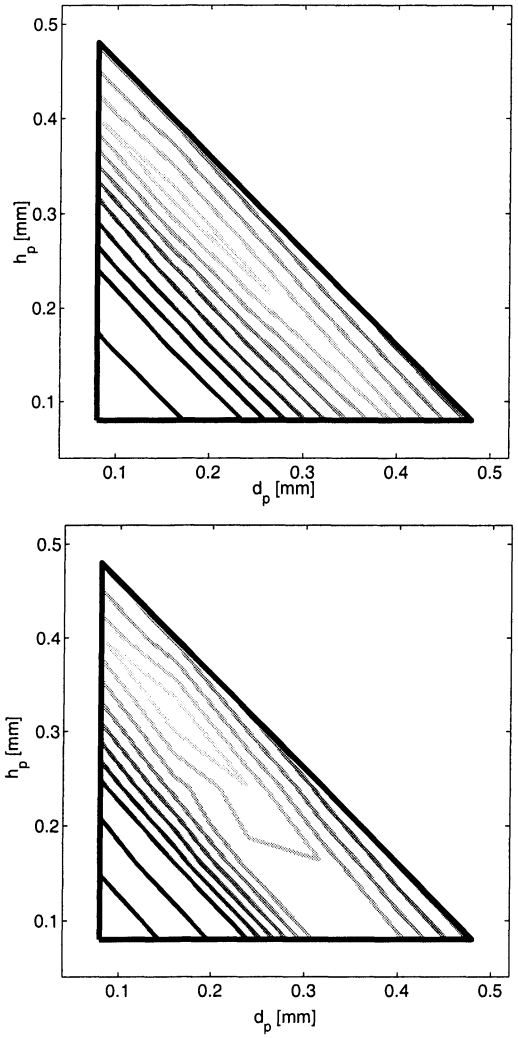


Figure 5: The power reflection corresponding to R_∞ is shown to the left and the right for $W_p = 4$ and $W_p = 6$, respectively. The contours are $R_\infty = -10, -9, \ldots, -1$ dB in consecutive order from bright to dark gray.

Furthermore, the power reflection corresponding to R_2 is shown in Fig. 6 for the fixed parameters $W_p = 4$ and $W_p = 6$ to the left and the right, respectively.

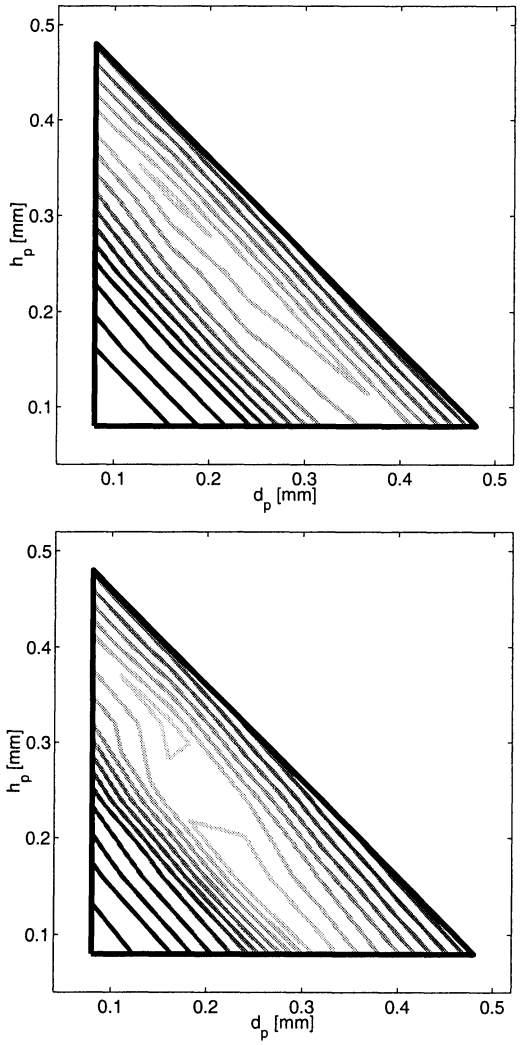


Figure 6: The power reflection corresponding to R_2 is shown to the left and the right for $W_p = 4$ and $W_p = 6$, respectively. The contours are $R_2 = -13, -9, \ldots, -2$ dB in consecutive order from bright to dark gray.

The results show clearly that the essential parameter is $d_p + h_p$, which is the total distance the probe extends into the waveguide. The width of the probe does not influence the result very much. We introduce the triplet $G_p = [D_p, W_p, H_p]$ to describe the patch geometry and position. The minimum values are R_∞ = -10.8 dB for $G_p = [4, 6, 8]$ and R_2 = -14.0 dB for $G_p = [6, 6, 4]$ for the power reflection coefficients. In Fig. 7, the power reflection coefficient is shown as a function of frequency for the two optimal configurations.

We examined $S_{11}(f)$ for each individual setup and

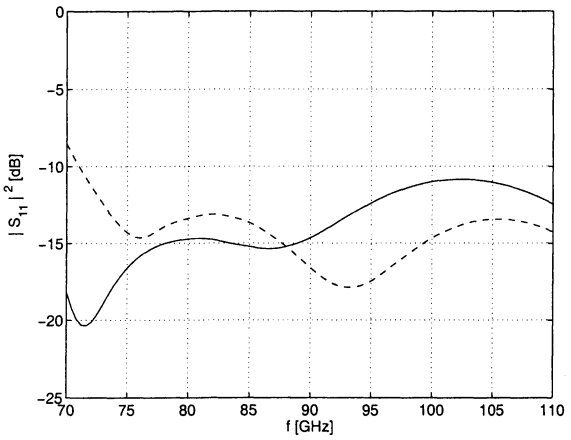


Figure 7: The power reflection corresponding to the optimal configurations $G_p = [4, 6, 8]$ and $G_p = [6, 6, 4]$ are shown by the solid and dashed line, respectively.

found that four geometries gave very deep local minima ranging from -26 dB to -47 dB. The minima were located towards the lower end of the frequency interval. The power reflection coefficient for one of the configurations, namely $G_p = [4, 4, 8]$, is shown in Fig. 8 and the corresponding Smith chart is shown in Fig 9.

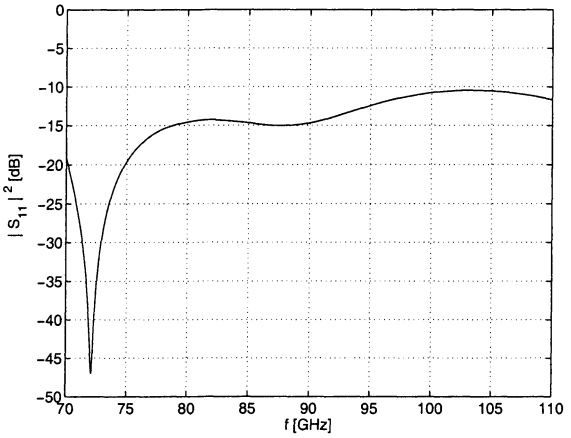


Figure 8: The power reflection coefficient corresponding to the configuration $G_p = [4, 4, 8]$.

4 Conclusion

We conclude that broadband operation with small losses from reflection is possible over a certain range of geometrical parameters. Two parameters have a major influence on the result, namely the distance to the termination of the waveguide and the separation of the probe from the opposing wall of the waveguide. The size of the probe influences the result only weakly

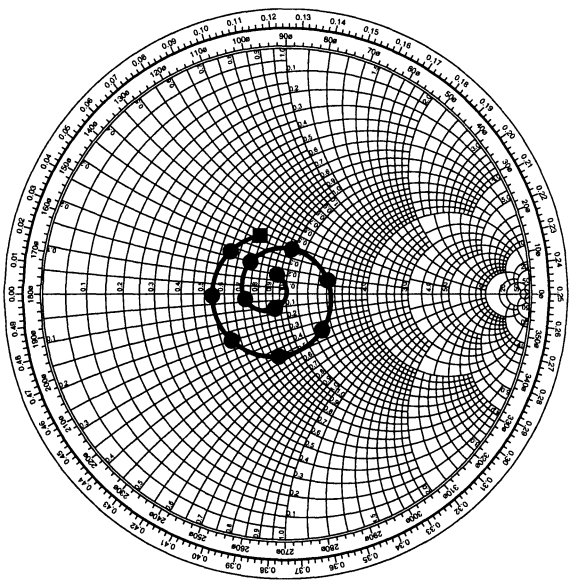


Figure 9: The Smith chart of $S_{11}(f)$ corresponding to the configuration $G_p = [4, 4, 8]$. The square indicates $f = 110$ GHz and the circles are equally distributed over freqency.

although very deep minima with low bandwidth can be found for some specific configurations.

References

[1] E. Tentzeris, M. Krumpholz, N. Dib, J.-G. Yook, and L. Katehi, "Fdtd characterization of waveguide-probe structures," *IEEE Trans. Microwave Theory Tech.*, vol. 46, pp. 1452–1460, October 1998.

[2] N. Kaneda, Y. Qian, and T. Itoh, "A broad-band microstrip-to-waveguide transition using quasi-yagi antenna," *IEEE Trans. Microwave Theory Tech.*, vol. 47, pp. 2562–2567, December 1999.

[3] T. Rylander and A. Bondeson, "Stable FEM-FDTD hybrid method for Maxwell's equations," *Comput. Phys. Comm.*, vol. 125, pp. 75–82, March 2000.

[4] P. G. Petropoulos, L. Zhao, and A. C. Cangellaris, "A Reflectionless Sponge Layer Absorbing Boundary Condition for the Solution of Maxwell's Equations with High-Order Staggered Finite Difference Schemes," *J. Comput. Phys.*, vol. 139, pp. 184–208, 1998.

[5] S. Balay, W. Gropp, L. C. McInnes, and B. Smith, "The portable, extensible toolkit for scientific computation." `http://www-fp.mcs.anl.gov/petsc/index.html`.

[6] F. Alimenti, P. Mezzanotte, L. Roselli, and R. Sorrentino, "A revised formulation of model absorbing and matched modal source boundary conditions for the efficient FDTD analysis of waveguide structures," *IEEE Trans. Microwave Theory Tech.*, vol. 48, pp. 50–59, January 2000.

Overall design of 4th generation wireless infrastructures (4GW) – A project overview & some lessons learned

Aurelian Bria*, Maxime Flament, Fredrik Gessler*, Olav Queseth*, Rickard Stridh*, Matthias Unbehaun*, Jiang Wu*, Jens Zander***

* Wireless@KTH, Royal Institute of Technology. E-mail: 4gw-pcc@s3.kth.se
** Dept of Signals and systems. Chalmers University of Technology
http://www.s3.kth.se/radio/4GW/

1 Introduction

A "hot" item, frequently debated in the wireless community these days is if there is such a thing as a "fourth generation" of wireless systems that is likely to appear after the successful deployment of the current third generation systems, say five to ten years from now. This supposedly "new" generation of wireless systems is intended to complement or replace third generation systems, as well as second generation systems that have already been in use for about a decade. A "classical" approach would design such a "system" in the same ways as the previous generations wireless systems, i.e. again focus on higher data rates (now beyond 2 Mbit/s) and finding new frequency bands for a world-wide standard (e.g. (2)). It is however not obvious that the roadmap is this straight-forward for a number of reasons discussed in this paper. An important factor contributing to this uncertainty is that we have very limited knowledge about the future environment in which a fourth generation wireless infrastructure should function. Which of today's systems will still exist when a potential 4G infrastructure is deployed? Which systems and solutions will be considered successful then? What technical bottlenecks will be apparent ten years from now? What market impact will third generation wireless systems have? How will this affect user behavior, and user demands? How much money do prospective users have available to pay for services provided over this infrastructure?

This paper presents some of the conclusion of of the Fourth Generation Wireless Infrastructure project (4GW) of the Personal Computing and Communications program (PCC), the major Swedish academic research effort on future communications systems launched in late 1997(1). The vision of the PCC project has been "Personal Multimedia to everyone at today's prices for fixed telephony". In 4GW a scenario-based approach has been used to tackle the issue of identifying reasonable research topics. In the paper we will present this method, and demonstrate how it interfaces with traditional research methods in the fields that the 4GW project incorporates. We will also give an overview of some research results from the project. Finally, we will conclude these results as a vision of what fourth generation wireless infrastructures might become.

2 The 4GW work process

Perhaps the most difficult issue in any scientific research endeavor is to identify reasonable assumptions for user behavior and telecommuncation systems in, say 10 years from now, such that relevant research topics may be identified. In fact, this is so difficult that most studies don't even try to do it. The approach chosen in the 4GW project was to work with scenarios. Scenarios are a tool to explore possible, plausible futures, by identifying key technical and social developments that are needed for them to be realized. The point of scenarios is not to predict the future, but to create an awareness of what future developments are possible. The research model used in the 4GW project is outlined in figure 1. As the figure demonstrates, the process began by the creation of techno-socio-economic scenarios based on literature studies and more informal sources of knowledge. The studied literature consisted of scenario methodology, as well as scenario work done by others, e.g. Ericsson(7), Siemens(8). The informal, experience based knowledge was gathered through Delphi interviews with academics and industry professionals. The resulting scenarios were used as an important input for the formulation of the basic assumptions that are an expression of the expectations and visions of the entire PCC program. From the basic assumptions, a number of working assumptions have been drawn. They in turn represent the more operational goals of the research program, and have been used to formulate the actual research problems of the 4GW project. By working in a multiple stage process, it has been possible to translate "fuzzy" societal developments into consequences for technologically defined research problems.

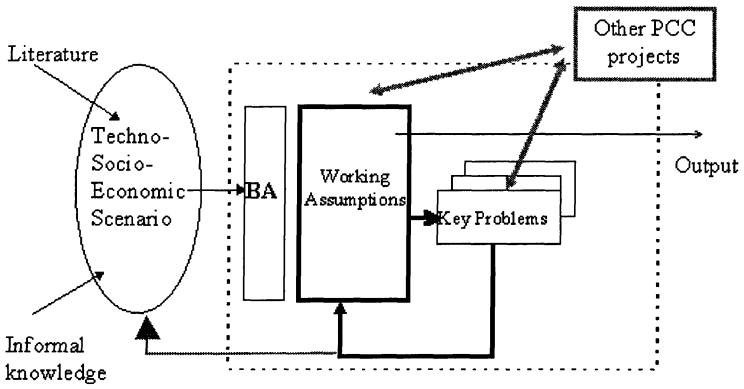


Fig 1. The 4GW work process

3 The scenarios ...

In the scenario work, three scenarios were formulated: "Pocket Computing", "Big Brother", and "Anything Goes". They address such things as user behavior and life style, telecommunications market evolution, development of supporting technologies, and evolution of values and society. The scenarios are outputs that portray the essence of what the world might become. The three scenarios are outlined below. More extensive, narrative descriptions of the scenarios can found in a full report (6) published by the project group in 1998.

Anything Goes!
The diversity of telecommunications equipment has increased dramatically, as well as the possibilities of manufacturing cheap, co-existing products. Manufacturing companies have become dominant in the telecom world. They advocate open de-facto standards, and use software solutions to create flexible, multi-standard equipment. Because of dramatic price reductions, both residential and business environments have wireless LAN solutions. They are operated by a multitude of operators, and the end-users have great freedom of choice in selecting where to purchase wireless services.

Big brother
As more and more personal information is available in the information infrastructure, personal integrity has become a major concern for the ordinary user. There is a widespread call for regulation and government intervention to ensure information integrity and secure networks. In the private sphere, most public information services use broadcasting. The complexity of products and services has increased, and thus also the cost. Service, transport and equipment providers have been reduced to a few large actors (brands) that, in the public eye, can be trusted.

Pocket computing

Technological development continues at a high pace throughout the world, but due to financial and educational differences, society is divided between those who can follow the development and those who cannot. Parts of the population have access to a multitude of advanced services, whereas others use simple services adapted to their needs. The mobile multimedia services mainly focus high-end consumer and business needs. A few operators and some very large manufacturers use standards to maintain their strategic position, and dominate the telecommunications scene.

4 ..and their implications

From the scenarios, it was possible to identify a set of key developments in the information and communication fields. They could then be expressed in terms of the following working assumptions.

Tele-presence
... is an application that will be used to create virtual meetings between individuals, and provides full stimulation of all senses required to provide the illusion of actually being somewhere else. With efficient data compression and fast sensory feedback, the bandwidth required for tele-presence is less than 100 Mbits/s. The data stream is dominated by 180-degree stereo, hi-resolution, full motion video. This type of application is likely to be the technically most demanding encounter in personal communication systems.

Information anywhere, anytime...
...with virtually seamless connection to a wide range of information services is a key feature of the future information infrastructure. Information access to large

volumes of data, pictures, video, etc., will be nearly instantaneous in small portable terminals. Compared to real-time meeting process, this application is less delay sensitive. The traffic pattern will be highly asymmetric, with 50 to 1 ratios or more favoring the system-to-terminal links.

Inter-machine communication...
...will be an important application/service. It will range from simple maintenance routines (e.g. refrigerator telling repair shop that it's broken) to sophisticated massive data exchange (e.g. camera and PC/TV exchanging video/picture information). All cars will have wireless interface as standard feature, as will household and office equipment costing as little as USD 20.

Security will be an indispensable feature...
...of the future infrastructure. Data integrity and protection against unauthorized access will be key features for providing reliable services for banking, electronic payment and handling of personal information. Schemes that reliably prevent unauthorized tracking of users and other intrusions in the private sphere will be in operation.

One-stop-shopping...
...services will be provided in a "turn-key" fashion directly to the consumer at the point of sales. The store (information provider) will take full responsibility for the service, as well as for any hardware or software that is provided.

Non-homogeneous infrastructure...
...consisting of several switching fabrics and a multitude of physical media will be the rule. All elements of significance will be digital. The fixed backbone structure will be dominated by connection-less packet switching (IP-style). The new air interfaces in wireless systems will also use packet switching technology. The wireless infrastructure will consist of a multitude of air interfaces, inherited from the wireless systems of the late 1990s, and early years of the new millennium.

Public and private access mixed.
Public Wireless access quality and bandwidth will vary. Higher data rates will be confined to dense urban areas, office environments (private/public systems) and homes (private systems). Operators and service providers will provide partial coverage for non-real-time wideband information access in most public places (info-stations).

Ad-hoc, unlicensed operation will dominate...
...and many different actors will provide parts of the infrastructure. Ad-hoc networking (spontaneous deployment, self-planning) in unlicensed bands (the 5 and 60 GHz bands) will play an important role, and compete fiercely with existing public operators, who experience dwindling market shares. Techniques for efficient multi-operator (private/public) sharing of unlicensed spectrum have been developed. Ad-hoc structures, where the equipment of the users (companies or even individuals) provide part of the infrastructure, will be adaptive to possible new communication patterns. Control of the new emerging ad-hoc networks (incl. routing, mobility etc.) will be fully distributed and highly reliable.

Multimode access ports in public systems...
...with multiple access air interfaces will be used to accommodate a wide range of terminals. Large operator systems will use advanced access ports with adaptive antennas that self-configure with non-critical installation procedures (self-configuration) to reduce cost. Access ports (wireless gateways) in ad-hoc access systems will on the other hand be simple single mode (single air interface) devices. The cost of access port hardware in these systems will be negligible in comparison with the cost of planning and physical installation.

Terminals....
...will exhibit a large range of bandwidths, from less than 10 kbits/s (simple appliances) to 100 Mbits/s (tele-presence terminals). The battery life of personal terminals will be at least one week. Battery capacity/weight/volume ratios will increase by an order of magnitude compared to today. Terminals in the 5 and 60 GHz range will use advanced adaptive antennas. Terminals will either be multi-mode, multi-function terminals, or single-purpose, cheap terminals designed solely for a specific service or function (e.g. receive only).

5 Key research areas – the WP:s

Returning to the process model in fig. 1, a number of research problems, relevant to wireless infrastructures, have been derived based on the working assumptions above. A key recurrent problem is, paraphrasing the PCC vision, to provide high data rates, everywhere *affordable* to general public. The 4GW project has conducted a number of feasibility studies focusing on techniques and architectures, given that they could be used to their full potential, could significantly change the cost or performance of wireless system. The aim is to find ways that could break the cost barrier discussed above. The project has participants from various information and communication technology research fields. Below follows brief account of these subprojects of 4GW. Details of the subprojects and their results are reported elsewhere in these proceedings.

Broadband OFDM air interface design for 60 GHz
An OFDM based modulation is a good candidate for broadband wireless systems. This scheme is well suited for transmitting high data rates in frequency-selective slow fading channels, and in the case of a properly dimensioned system the fading experienced by each sub-channel is flat. In several countries, the 60 GHz

unlicensed band provides up to 5 GHz of available bandwidth. New FCC recommendations allow us to plan for even more spectrum. In this band, we showed that coverage is not the main limitation in indoor office deployments, but rather that the short burst interference and the body shadowing create critical problems. Using a ray tracing simulator, we have studied the dynamics of the 60 GHz time varying channel in particular situations typical for office or shopping mall environments. Significant limitations, such as shadowing and fast variation of the interference characteristics in office and shopping mall environment, have been investigated. The results give a clear idea of the time variations of the channel quality for which we try to identify potential solutions. For this kind of channels, an orthogonal frequency division multiplexing (OFDM) based modulation is a good candidate. This scheme is well suited for transmitting high data rates in frequency-selective slow fading channels, and in the case of a properly dimensioned system the fading experienced by each subchannel is flat. Recently, we investigated promising system design for OFDM air interfaces. One example is based on distributed access point (AP) networks, called Virtual Cellular Networks (VCN), combined with multiple receiving antenna terminals.

Smart Antennas for 5 GHz
Using the 60 GHz band requires an increased number of access points, but may allow inexpensive radio access equipment. Systems at 5 GHz offer greater range, and have the advantage that several users can share one access point, which offers flexibility for the operator at the cost of more complex access points. Smart Antennas, i.e. antenna arrays in combination with advanced signal processing, is a tool for improving the capacity of such systems.

Our research results so far show that dual arrays at above 5 GHz, in indoor environments, fulfil the 4GW requirements of link capacity. The results indicate that operation at 5 GHz is an important alternative in the 4th generation wireless systems in all the presented scenarios. In addition to further work in this area, e.g. to map the network properties, an infrastructure study is needed in order to compare coverage and QoS vs. infrastructure cost for the proposed systems.

Wireless infrastructure architecture
The scenario work revealed three key attributes of future wireless infrastructures: very high data rates (up to 100 Mbps), simplicity and convenience in using devices and services (one-stop shopping), and flexibility (ad-hoc installation, unlicensed operation).

Our results show that user-deployment is indeed a viable alternative to traditional infrastructure installation methods. In particular, dense networks, which are typically required to satisfy the high capacity demands in for example office environments, are tolerant to arbitrary placement of the access points, as long as they are reasonably uniformly distributed over the entire area. However, our studies show that 60 GHz systems are slightly more sensitive than 17 GHz systems due to more difficult propagation characteristics. In densely populated, large buildings such as shopping malls, train stations or airports, user-deployment also achieves acceptable performance, although access point placement will require some coarse preplanning. This effort can however be limited to visual inspection of the site in compliance with a set of basic guidelines, and no expert knowledge in cellular planning is required.

Wireless resource management in multiple operator infrastructures
As noted in the "Anything Goes" and "Pocket Computing" scenarios, some feasible evolution paths will entail a large number of different services, offered over various types of access infrastructures. The number of operators and service providers will increase, as will the capacity requirements of the services they offer. In order to make future wireless services reasonably priced, we need to find more efficient methods to share frequency spectrum. Traditional licensing techniques, widely used today, are inadequate.

The aim of this subproject is to determine the feasibility of unlicensed operation for spectrum sharing. Our research results clearly demonstrate that efficient infrastructure deployments are possible in environments where multiple operators share unlicensed frequency spectrum. Unlicensed infrastructure deployments can be made as efficiently as traditional deployment, based on licensed operation. The technical feasibility of such evolutions has thus been justified. Different techniques for frequency sharing have been studied in relation to infrastructure deployment, and frequency hopping appears to be the preferred alternative. Isolation (i.e. attenuation) between operators' infrastructures improves the total available capacity. Techniques for increasing isolation will thus become key in an unlicensed environment. Examples of such techniques include smart antennas.

Seamless IP mobility support for mobile applications
In the "Anything Goes" and "Pocket Computing" scenarios, it was noted that a heterogeneous network infrastructure will play a crucial rule in building up a liberal communication market and providing a large number of innovative services. In these scenarios, the network infrastructure is not built in a well-planned way, so protocols need to be flexible and robust enough to survive despite the diverse nature of the infrastructure. This subproject investigates how seamless IP layer mobility can be supported in fourth generation wireless infrastructures. Studies are focused on both Mobile IP and IP multicast. Enhancements in both protocols for supporting IP layer seamless handover are needed.

With proper enhancements, the current Internet protocols are able to support IP layer mobility. Our research has shown that with our suggested protocol enhancements, IP layer handover latency for IP multicast is small enough to support real-time applications. In the case of Mobile IP, we believe it will be possible to achieve seamless handover in a near future. The research results demonstrate the feasibility of a heterogeneous network infrastructure, evolving through an integration of the current Internet and wireless networks. Using IP layer mobility support, wireless networks can be constructed very flexibly. Network operators will be able to provide mobile Internet services with ease, thus allowing innovative applications to be developed and deployed quickly.

Asymmetric wireless infrastructures

High demand for mobile Internet and interactive services will characterize the use of the future wireless systems. As a consequence, a large amount of traffic will be asymmetric, with user terminals requesting large amounts of data. Exploiting this fact in network design may have significant implications on infrastructure deployment costs since higher transmitter powers can be used in base stations.

6 Conclusions

The systematic process to find key research issues for fourth generation wireless infrastructures, based on the PCC vision paraphrased as "wireless high data rates - affordable everywhere", has been presented. Using a scenario based approach, the work resulted in three major scenarios describing possible "telecom futures". Key technical and business-related research issues were derived, and working assumptions for the project were formulated based on the scenarios. An overshadowing barrier in current system design, prohibiting the implementation of the vision by current techniques has been identified: the cost per transmitted bit remains almost constant as we increase the data rate. Some feasibility studies have been conducted in the project, studying key techniques or technologies that had the promise of breaking these cost / performance barrier.
The results of these ongoing studies show that user-deployed system remains as a viable candidate for short-to-moderate range wireless system with very high data rates (<100 Mbit/s). Further advances in array signal processing are shown to be practically applicable in these environments, with the potential to substantially reducing the number a required access points. Ad-hoc, multiuser/multioperator systems can be made working in these environments, but without careful system design, this can be incur a severe performance penalty. Current CDMA-type solutions do not work in this context. Mobilty management remains a big challenge in "low-hierarchical" network architectures as outlined in the working assumptions. Although, for the most demanding applications (multicasting/multiparty teleconferencing) significant progress has been reported.

The scenario work has also provided confidence that the group is working with the right problems. The scenarios have gained acceptance and created discussions within the PCC program, where they serve as a common platform for discussion of future systems and architectures.

8 References

(1) Molin, B-A, "Personal Computing and Communication -- A Swedish Strategic Research Program", Proc PIMRC 98, Boston, MA, Sept 1998.

(2) Hata, M, "Fourth generation mobile communication systems beyond IMT-2000 Communications", Proce APCC/OECC '99. Fifth Asia-Pacific Conference on Communications and Fourth Optoelectronics and Communications Conference , 1999 , Page(s): 765 -767 vol.1

(3) Pereira, J, "Fourth Generation: Now, its personal!", Proc. Personal, Indoor and Mobile Radio Communications, PIMRC 2000, Volume: 2 , 2000 , Page(s): 1009 –1016

(4) A Bria, M Flament, F Gessler, O Queseth, R Stridh, M Unbehaun, J Wu, J Zander "4th Generation wireless infrastructures - scenarios & research challenges", IEEE Personal Communications Magazine, December, 2001.

(5) M. Flament, F Gessler, F Lagergren, O Queseth, R Stridh, M Unbehaun, J Wu, J Zander "An Approach to 4th Generation Wireless Infrastructures - Scenarios and Key Issues", VTC 99, Houston, TX, May 1999

(6) M. Flament, F. Gessler, F. Lagergren, O.Queseth, R. Stridh, M. Unbehaun, J. Wu, J. Zander, "Telecom Scenarios 2010 - a wireless infrastructure perspective", http://www.s3.kth.se/radio/4GW/public/Papers/ScenarioReport.pdf

(7) Ericsson, Scenarios, 2000-12-18, Kontakten no. 1, 1997.

(8) Siemens, Scenarios, 2000-12-18,

Coverage Planning for Outdoor Wireless LAN Systems

Max Kamenetsky† and Matthias Unbehaun‡
†ISL, Department of Electrical Engineering, Stanford University, Stanford CA, USA
‡Department of Signals, Sensors and Systems, Royal Institute of Technology (KTH), Stockholm, Sweden

***Abstract*—Wireless LANs (WLANs) are becoming increasingly popular for providing high data rate network access to mobile computers. Most of the currently deployed systems operate in the 2.4 GHz unlicensed frequency band. However, increasing demand for higher data rates and network capacities has led to new system standards for the 5 GHz band.**

Unlike traditional cellular telephony systems, WLANs are deployed in an ad-hoc fashion, often based on an *educated guess* by the person installing the access points (APs). This typically results in coverage gaps or capacity loss due to misplaced APs. In this paper, we examine methods for obtaining a close-to-optimal positioning of WLAN APs and evaluate their performance in a typical downtown or campus environment.

The system performance is evaluated using an objective function which aims to maximize both the coverage area and the overall signal quality. The optimization algorithms used in this paper evaluate this objective function over a discrete search space, thereby considerably reducing the inherent complexity of the problem, while at the same time providing a reasonable approximation to the continuous optimization problem.

Numerical results show that random search algorithms, such as simulated annealing, can yield very good solutions. However, the convergence speed of simulated annealing strongly depends on the fine-tuning of simulation parameters and a good choice of the initial set of transmitter positions. Successive removal algorithms, such as pruning, though usually producing sub-optimal solutions, converge in polynomial time. We therefore propose a combination of the two approaches – using pruning for obtaining an initial set of transmitter positions and refining these by using either neighborhood search or simulated annealing.

***Keywords*—wireless LAN, coverage planning, outdoor coverage, network deployment, simulated annealing, pruning, neighborhood search**

I. Introduction

THE traditional approach to coverage planning in cellular networks aims to achieve optimal placement of the infrastructure such that the total number of base stations is minimized while maintaining a given Quality-of-Service in the entire coverage area. Hence, extensive system simulations based on terrain databases and using statistical propagation prediction tools are typically carried out. The costs for such sophisticated network optimization are justified by the large savings it yields, since the installation and operation of large-scale, often nationwide, networks requires substantial investments.

Wireless LANs (WLANs) follow a different paradigm. They are designed to provide low cost, best-effort connectivity and are usually deployed in an ad-hoc fashion, *i.e.* wherever coverage is needed and access to a wired backbone is available. Although certainly beneficial to the overall system performance, network planning for WLANs is regarded as too complex and too costly. In this paper, we investigate the performance improvements that can be obtained by optimizing the placement of WLAN access points (APs) and study ways to reduce the complexity of the optimization process.

Several methods for coverage planning in WLANs can be found in the literature, mostly based on random search [1], gradient descent [2] or genetic algorithms [3]. The performance of these methods strongly depends upon the initial starting conditions and the choice of simulation parameters. Since these parameters and a good starting set of APs are usually not known *a priori*, the above algorithms often suffer from slow convergence, and they may, in fact, not converge at all.

In this paper, we explore the use of simple heuristic algorithms such as pruning and neighborhood search to obtain a close-to-optimal solution to the AP placement problem. These algorithms do not exhibit the convergence problems described above and thus can be used either as standalone optimization routines or to obtain good starting solutions for use by more complex optimization algorithms.

The rest of this paper is organized as follows. Section II provides a brief summary of the WLAN standards considered in this work. Section III discusses the specific problems of coverage planning for such networks and introduces suitable propagation models and performance measures. In section IV, we review the different optimization approaches and select the most appropriate methods for numerical evaluation. The simulation results are then presented in section V, with special attention paid to reducing the overall complexity of the optimization. Section VI concludes with a summary.

II. Wireless LANs

Most of today's WLAN systems operate in the unlicensed 2.4 GHz band and follow the IEEE 802.11b standard, which is a wireless extension of the Ethernet standard and allows data rates up to 11 Mbps. A new generation, using the 5 GHz band and delivering up to 54 Mbps, is expected to appear on the consumer market soon. These systems are based on two harmonized standards developed by ETSI [4] and IEEE [5]. Both use advanced radio resource management schemes, such as link adaptation and dynamic channel allocation, for fine-tuning the network parameters. Although these algorithms ensure the best possible performance for a given network structure, they do not improve the topology itself, *i.e.* they cannot compensate for misplaced APs or fix coverage gaps.

In order to analyze the performance improvement achieved by optimizing AP locations, we need to consider a non-trivial network deployment, *i.e.* an environment where the propagation is difficult to estimate and hence an *educated guess* clearly results in suboptimal performance. An example of a trivial case is a completely symmetric environment, *e.g.* an open field without trees, houses or any other obstacles, or an empty room. In realistic situations it is usually not possible to guess (*e.g.* using geometric considerations) the local propagation situation.

Typical usage scenarios for WLANs are offices, campus

environments, manufacturing or storage facilities, shopping malls, conference venues, or other locations where network access for portable computers or other devices is required. A number of representative environments are proposed in [6], and it was found that typical campus or downtown locations are the most difficult to cover. Hence, we will use the campus example for examining the performance of the network planning algorithms.

III. Coverage Optimization

A. Wireless Campus Network

The campus environment has a typical urban layout with streets, small squares and complex building structures. The total area is approximately $100,000$ m^2, however only the outdoor part of $45,000$ m^2 should be considered as service area for network planning. Coverage from outdoor APs into the buildings is neglected, and only the streets and squares are used for estimating wireless coverage. All buildings are constructed with reinforced, 50 cm-thick concrete walls. APs are installed using masts or utilizing existing lamp poles, where available. The height of the APs is assumed to be 5m above street level.

The possible AP sites, from which the optimization algorithm should select a set that provides the best coverage, are limited to a grid. We discuss this approach in detail in section III-B and choose a suitable grid size in section V-A. An illustration of the campus environment and the AP placement grid is shown in Fig. 7.

The received power is sampled at $M = 450$ measurement points, regularly dispersed in the considered service area at 10 m intervals. Ray-tracing is used to obtain the propagation data, since it is a very accurate tool for modeling propagation in highly complex environments such as the one we are considering. Such detailed modeling is needed to accurately evaluate the impact of different AP locations on system coverage.

B. Grid-based Approach

The main disadvantage of ray-tracing is its computational complexity, which depends on the number of sample points M and AP locations N and is approximately of order $\mathcal{O}(MN)$. However, coverage optimization algorithms are also highly complex and combining both will most likely render the entire problem unsolvable. We therefore separate the two problems and perform propagation prediction and coverage optimization as independent, consecutive steps:

1. We define a finite set of possible AP installation sites $\mathcal{A} = \{A_1, \ldots, A_N\}$ and the propagation data are then obtained for all sites in $\mathcal{A}$. Hence, the inherently infinite search space is reduced to a finite set of size N.
2. The optimization algorithm identifies the subset of K sites that provide the best coverage.

The fundamental tradeoff is now between selecting an appropriate number of grid points, N, such that the overall complexity is manageable and providing a sufficiently fine granularity for the coverage optimization to converge to a point close to the global optimum in the continuous search space. In order to find a reasonable grid size, we first need to introduce appropriate performance measures and then select suitable optimization schemes.

C. Objective Function

Coverage planning typically comprises two objectives: improving the average signal quality in the entire service area and minimizing the areas with poor signal quality. The two objectives are not always compatible and hence a suitable tradeoff must be found. We use an approach proposed in [2], a combination of a *minisum* and a *minimax* objective function.

The first objective is achieved by evaluating and minimizing the average pathloss, expressed by the function

$$f_1 = \frac{1}{M} \sum_{i=1}^{M} \left(g_i^{(k)} + \mu \max \left\{ 0, g_i^{(k)} - g_{\max} \right\} \right) \quad (1)$$

over the entire service area. Here, M is the total number of measurement points in the service area and $g_i^{(k)}$ is the pathloss from the i-th point to AP k. Each point is assigned to the AP from which it measures the minimum pathloss, *i.e.*

$$g_i^{(k)} : k = \underset{j=1,\ldots,K}{\operatorname{argmin}} \left\{ g_i^{(j)} \right\} \quad \forall i = 1, \ldots, M \quad (2)$$

The term $g_{\max}$ defines the maximum tolerable pathloss, which causes a penalty term of $\mu \left(g_i^{(k)} - g_{\max} \right)$ to be added if the threshold of $g_{\max}$ is exceeded. By minimizing (1), we obtain a *minisum* objective function.

In order to mitigate the worst-case situation, we consider the measurement point with the maximum pathloss and try to minimize its contribution, obtaining a *minimax* objective function:

$$f_2 = \max_i \left(g_i^{(k)} + \mu \max \left\{ 0, g_i^{(k)} - g_{\max} \right\} \right) \quad (3)$$

Finally, the *total objective function* (OF) is a convex combination of (1) and (3), controlled by a balancing parameter $\psi \in [0,1]$ that defines the relative contributions of (1) and (3):

$$F = \psi f_1 + (1 - \psi) f_2 \quad (4)$$

The optimal location for only one AP can be obtained by evaluating (4) for all possible AP locations and picking the one that achieves the minimum OF. Since the number of OF evaluations grows linearly with the number of possible AP sites, N, the global optimal solution in this case can be easily obtained via exhaustive search.

The problem of placing K APs out of N possible sites is a combinatorial problem of order $\mathcal{O}\left(\frac{N!}{K!(N-K)!}\right)$, thus making exhaustive search algorithms appropriate only for cases where K and N are small. Since K is usually limited to a small range bounded by coverage (minimum number of APs to cover the area) and economic (maximum number of APs allowed by budget) constraints, one may be tempted to decrease N in order to make exhaustive search feasible. However, a small N is equivalent to a sparse quantization of the continuous search space, which increases the likelihood that none of the quantized grid points is sufficiently close to the global optimum in the continuous search space.

Unfortunately, there is no known polynomial time algorithm that can provide an exact solution to the above problem. Consequently, for realistic problems with a reasonably large N, heuristic-based and other locally optimal algorithms need to be explored.

IV. Optimization Algorithms

Different approaches to coverage optimization for wireless systems can be found in the literature. Prior work has mostly focused on gradient descent [2], random search [1] and genetic algorithms [3]. Gradient descent is an attractive approach because of existing convergence proofs that can guarantee convergence under a specific set of conditions (usually constraints on the step size in the negative direction of the gradient). Unfortunately, gradient descent depends upon knowledge of the OF surface or gradient estimate. Thus, it is restricted to cases where the OF is smooth, which is certainly not the case for the difficult propagation situation in the campus environment.

Random search and genetic algorithms do not require knowledge of the gradient. However, they lack strong convergence proofs, are highly dependent upon starting conditions and algorithm parameters, and lack a clear termination condition. Therefore, we propose a new *pruning* algorithm that is guaranteed to converge, does not require knowledge of the gradient, is independent of starting conditions, and can be used to provide a solution that can serve as a starting point for more complex search algorithms.

We first define a reference structure, which serves as a performance benchmark, and then outline the heuristic algorithms used in this paper.

A. Reference Installation

The simplest possible heuristic solution is a uniform, or regular placement. In this scheme, the environment is subdivided into K equal regions R_k, $k = 1, \ldots, K$. Then, for each k, AP A_k is placed in the Euclidean center of R_k.

This algorithm is clearly suboptimal. However, it is simple to implement, and it is reasonable to expect that a similar method will be employed in installations by users with no knowledge of the environment other than its general geometry. As such, it is a useful metric and can serve as a reference point for other more computationally expensive algorithms. Finally, it is important to point out that regular placement will be globally optimal in a completely symmetric environment.

B. Pruning

Pruning is a simple algorithm for combinatorial optimization. The pseudocode for this algorithm is shown in Table I. In the initialization stage, the algorithm starts with N APs, one for each available AP site. Then, each AP is iteratively removed, the OF is re-evaluated without the removed AP, and the removed AP is re-seeded. This proceeds until the algorithm has calculated the OF for every possible AP removal. The AP whose removal achieved the lowest OF is then permanently removed and the algorithm repeats for the remaining $N - 1$ APs. This continues until there are only K APs left.

Pruning is greedy in the sense that it tries to minimize the OF at each step, without looking ahead to how that move may affect the OF at future steps [7]. As such, it is suboptimal because it may remove an AP that, though not necessary in the early stages of pruning, may actually be part of the globally optimal solution.

TABLE I
PRUNING ALGORITHM

```
seed each of all N available APs with an AP
set number_of_placed_APs := N
while number_of_placed_APs > K
  set k := 1
  while k ≤ number_of_placed_APs
    remove AP A_k
    evaluate OF for the remaining APs, store it as J_k
    re-seed AP A_k
    set k := k + 1
  end  /* inner loop */
  set k* := argmin_k J_k
  remove AP A_k*
  set number_of_placed_APs := number_of_placed_APs − 1
end  /* outer loop */
```

However, several observations are in order. First, pruning is clearly optimal for the case where $K = N - 1$. Second, pruning is significantly less complex than exhaustive search. In general, pruning requires $\mathcal{O}(KN)$ OF evaluations compared to the $\mathcal{O}\left(\frac{N!}{K!(N-K)!}\right)$ OF evaluations for exhaustive search. Third, pruning exhibits computational complexity behavior that is, in essence, the opposite of exhaustive search. That is, for a fixed N, exhaustive search becomes more complex as K grows while pruning at the same time becomes simpler. As such, pruning provides a very attractive solution method for cases where K is very large. Fourth, pruning has a natural termination condition in that the algorithm is finished when exactly K APs remain and no further pruning is possible. And finally, pruning may be used to provide a good starting solution to other heuristic search algorithms, such as neighborhood search and simulated annealing, described below.

C. Neighborhood Search

Neighborhood search (NS) is a simple heuristic for finding a local optimum of the OF [8]. In general terms, for some initial solution S_0, the algorithm searches the neighborhood of S_0, $\aleph(S_0)$, for a solution S_1 that achieves a lower objective function. If such a solution is found, the algorithm moves to it and repeats the search. Otherwise, the algorithm is terminated with the last found solution. In practice, NS is commonly combined with steepest descent, whereby in the search step, the algorithm evaluates all the solutions in the neighborhood of the current solution and picks one that achieves the lowest OF.

Table II shows the pseudocode for one possible implementation of NS for the AP placement problem. In this implementation, the algorithm starts out with an initial solution, obtained either randomly, by pruning, or by some other algorithm. Then, for every AP, NS performs a local search in the neighborhood of that AP for a better AP site, where the neighborhood of AP number k in a solution S, $\aleph_k(S)$, is defined as the set of p grid points that are closest (in Euclidean distance) to AP A_k. If the algorithm finds more than one AP that

TABLE II

NEIGHBORHOOD SEARCH (NS)

```
start with initial solution S = {A_1, A_2, ..., A_K}
evaluate OF, store it as J_old
set J_new := J_old
while TRUE
  set k := 1
  while k ≤ K
    evaluate OF for all solutions in N_k(S)
    store the minimum value as J_k
    if J_k < J_new
      move to new solution S that achieves OF J_k
      set J_new := J_k
    endif
    set k := k + 1
  end  /* inner loop */
  if J_new = J_old
    break  /* outer loop */
  else
    set J_old := J_new
  endif
end  /* outer loop */
```

achieves a lower OF, it updates the solution with the location of the best AP. NS then iterates over all APs in the solution until it can no longer improve the OF.

NS is not guaranteed to achieve a globally optimal solution, though it is guaranteed that the solution it achieves will be at least locally optimal. Initial solutions that are "far away" from the global optimum are unlikely to converge to it. In fact, only solutions whose neighborhood includes the global optimum can be assured of convergence to it. Therefore, picking a good initial solution for NS is extremely important for achieving a low OF. For that reason, one of the main benefits of NS lies in its use as the last step for refining solutions obtained by regular placement, pruning, or other more complex optimization algorithms that can achieve good, though not necessarily locally optimal, solutions.

D. Simulated Annealing

Simulated annealing (SA) is a random search algorithm which slowly decreases the degree of randomness until it converges to a local optimum. A good example of how SA can be applied to cellular network planning is found in [1] and we will adopt the proposed approach for our purposes.

The algorithm starts with an initial solution S_0 and evaluates the OF for that solution. The AP positions are then randomly altered, resulting in a new system state S_1. If the new placement results in an improvement, *i.e.* $\mathrm{OF}(S_1) < \mathrm{OF}(S_0)$, then state S_1 is accepted as the new solution. Otherwise, if $\mathrm{OF}(S_1) > \mathrm{OF}(S_0)$, the system moves to the new state S_1 with an acceptance probability

$$P_a(T) = \min\left\{1, \exp\left(-\gamma \frac{\mathrm{OF}(S_1)/\mathrm{OF}(S_0) - 1}{(T/T_0)^2}\right)\right\} \quad (5)$$

where γ is an attenuation coefficient for limiting $P_a(T)$ within a suitable range and the parameter T is called the *system temperature* (T_0 being the starting temperature). The algorithm then repeats.

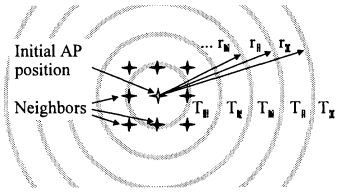


Fig. 1. Cooling strategy for simulated annealing. The circle radii are not necessarily drawn to scale.

T is a measure for the intensity of the random alterations to the system state S. To improve convergence, SA uses a *cooling strategy*, where the randomness is gradually reduced as the simulation progresses. Hence, SA tends to settle close to points with a low OF. The cooling strategy used for the AP placement problem is shown in Fig. 1. During the first phase, corresponding to start temperature T_0, the position of an AP can be changed over a large radius r_0. Since we assume a grid of available sites, an AP can move to any of the n_0 adjacent sites within the circle determined by r_0. As the temperature is lowered, the circle becomes smaller and hence the number of available neighbors decreases. In the simulations, we use five temperature levels $T = \{2\sqrt{2}R_{\mathrm{grid}}, \sqrt{5}R_{\mathrm{grid}}, 2R_{\mathrm{grid}}, \sqrt{2}R_{\mathrm{grid}}, R_{\mathrm{grid}}\}$, where R_{grid} is the distance between two adjacent grid points. On a regular grid, these temperature levels correspond to $n = \{24, 20, 12, 8, 4\}$ neighbors.

The gradual decrease of the temperature T permits a large search space in the early phases of the algorithm and restricts it as the algorithm progresses and is likely close to finding an optimum. The acceptance probability $P_a(T)$ should prevent the algorithm from becoming trapped in a local optimum. It needs to be carefully tuned by adjusting γ such that typically $0 \leq P_a(T) \leq 0.1$ to ensure convergence while permitting sufficient inertia to eventually escape a bad local optimum. The convergence speed of SA also strongly depends on the initial solution S_0, which can be obtained either randomly, by regular placement, or by pruning.

Due to the random nature of SA and the rough OF surface, it is difficult to determine convergence and decide when to stop the simulation. Therefore, we ran SA for a fixed number of iterations at each of the five temperature levels. Out of all the AP constellations which were evaluated at one temperature, the one resulting in the lowest OF was used as the start set for the next temperature level (or as the final solution when the simulation reached T_4).

Finally, it has been shown that in order to guarantee convergence to the global optimum, SA will in general require more iterations than exhaustive search [9]. Thus, like the other heuristic algorithms described above, it is only practical to use SA to attempt to find a good local optimum.

V. SIMULATION RESULTS

A. Simulation Parameters and Grid Size Selection

The objective function in (4) is based on the pathloss, thus limiting the number of relevant system parameters to a few fundamental link layer constants, which are summarized in Ta-

TABLE III

SIMULATION PARAMETERS

Target frequency	f_0	5 GHz
Channel bandwidth	B	20 MHz
Noise floor	$N_0 B$	-101 dBm
Noise figure	NF	9 dB
Transmission power	EIRP	20 dBm
Min. required signal quality	$\text{SNR}_{\min}$	15 dB
Max. pathloss	$g_{\max}$	97 dB
Outage probability	$Pr\{\text{SNR} < \text{SNR}_{\min}\}$	$\leq 10\%$
Balancing parameter	ψ	0.6
Penalty factor	μ	2
Attenuation coefficient	γ	100

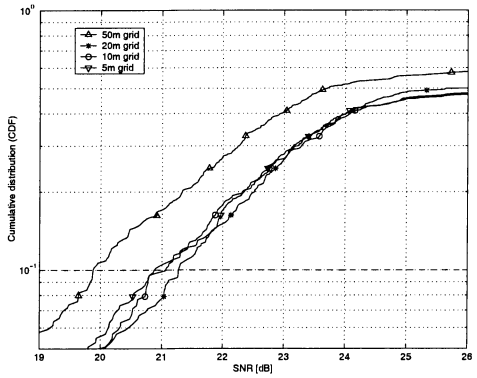


Fig. 2. SNR for different grid sizes. The dashed line denotes the maximum allowed outage probability.

ble III. The balancing parameter ψ is set to a rather low value within the range of $\psi \in [0.5, 1]$ recommended in [2]. This represents an almost even tradeoff between (1) and (3), with a slight emphasis on (1).

As discussed in section III-B, we first need to identify a suitable grid point density to be used in the optimization algorithms. The results shown in Fig. 2 are obtained by restricting the service area to a subset that is sufficiently small so as to allow exhaustive search.

We notice that a 50 m quantization of the search space is too coarse. The grid points are likely too far from the global optimum in the continuous search space and hence system performance suffers considerably compared to a finer quantization, such as 20 m. However, refining the grid further to 10 m or 5 m yields no substantial performance improvement, particularly when considering the 10% outage limit. Therefore, we select the 20 m grid as the best tradeoff between complexity and accuracy of the obtained solution.

B. Fair Comparison of the Algorithms

For comparing the convergence behavior of different algorithms, a reasonable metric is required. All of the algorithms were implemented using the same math library functions in MATLAB 6 and hence the total simulation time should be a reasonably reliable indicator for the convergence speed of the different algorithms. All simulations were run on a 700 MHz Pentium processor using Linux.

Furthermore, it is important to provide the same starting

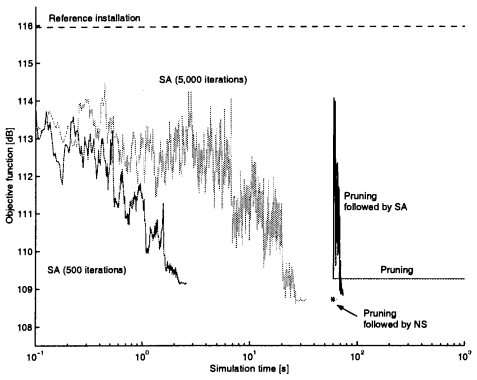


Fig. 3. Convergence of the optimization algorithms for 6 APs.

conditions for all approaches. Pruning differs considerably in that respect from random search algorithms (*e.g.* SA). It starts by evaluating the OF for all grid points, whereas SA proceeds from a set of starting points which have been chosen according to certain criteria. Selecting appropriate starting conditions for SA requires some knowledge about the underlying system geometry. For a fair comparison between pruning and SA, we need to start both with the same information about the system, that is: *no prior knowledge*. Therefore, SA will be initialized with a randomly selected set of starting points and we will use the ensemble average over a number of independent realizations for evaluating convergence and the resulting system performance.

C. Convergence Speed and System Performance

We can now compare the convergence behavior of the following algorithms:

- Pruning
- SA (100, 1000 and 10000 iterations at each temperature)
- Pruning followed by NS
- Pruning and SA (500 iterations at each temperature)

Figs. 3 and 4 show convergence and system performance for 6 APs. All algorithms achieve considerable improvement compared to the reference installation and converge in a rather short time, however none reaches the performance target of 15 dB ($\text{SNR}_{\min}$ in Table III) in 90% of the service area. Note that pruning has no convergence behavior; instead, it produces one single solution after removing all excess APs and hence appears as a straight line in the figures.

Figs. 5 and 6 show that in the 8 AP case, most algorithms achieve adequate coverage. Pruning produces a solution in about the same time as for 6 APs. Additional NS after pruning yields a further 8% performance improvement (shown as an asterisk in Fig. 5) with virtually no extra computation time (additional 260 ms). An illustration of the optimization results is given in Fig. 7, where the AP positions obtained from *pruning followed by NS* are shown as bold points (•).

By comparison, SA requires about 25 times longer than pruning to converge. If the number of iterations for SA is adjusted such that it uses approximately the same time as pruning with NS ($\approx$ 60 s), it was found that SA produces an equally good or better solution with only a 30% probability. Using SA

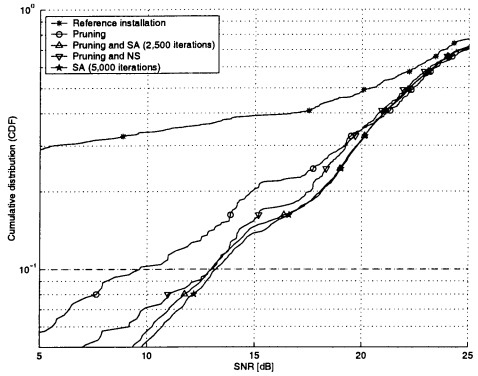


Fig. 4. SNR achieved by the optimization algorithms for 6 APs.

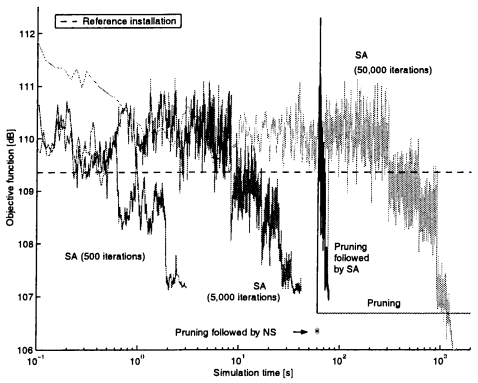


Fig. 5. Convergence of the optimization algorithms for 8 APs.

to refine the pruning solution improves performance in only about 10% of the cases.

Since SA typically does not produce unique solutions, the major difficulty in using SA by itself lies in the uncertainty of deciding whether it found a reasonably good solution or was trapped in a bad local minimum. Hence, SA needs to be repeated a number of times if a relatively reliable outcome is required. (Note that the results for SA in Figs. 3–6 are obtained from ensemble averages over 10 independent realizations.)

The trends observed for 6 (8) APs are emphasized when using fewer (more) APs. Using the reference installation, adequate coverage could not be achieved with fewer than 12 APs. Hence, using even relatively simple optimization schemes, such as pruning together with NS, translates into a 33% infrastructure cost saving. On the other hand, for the same case of 12 APs, SA alone cannot produce a reliable performance improvement even for 10000 iterations per temperature level.

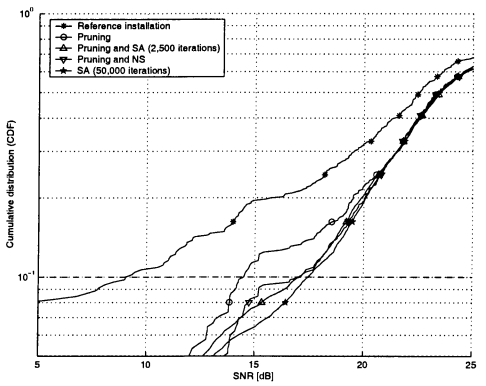


Fig. 6. SNR achieved by the optimization algorithms for 8 APs.

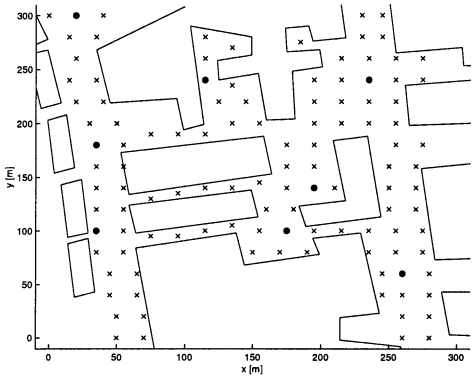


Fig. 7. Campus environment with the AP positions obtained by pruning and NS (•). The crosses (x) represent the grid of possible AP installation sites, based on a 20 m grid spacing.

VI. Summary

In this paper, we discussed different approaches to coverage planning for WLAN systems. A new optimization scheme, pruning, has been proposed, which is easy to implement, does not depend on a good starting solution, and requires no fine-tuning of simulation parameters. When combined with a comparably simple neighborhood search, it outperforms simulated annealing for realistic network sizes. Numerical simulations in a typical campus environment showed that 1/3 of the infrastructure costs (*i.e.* number of APs) could be saved compared to a regular dispersed network, where the AP positions are selected uniformly within the environment.

References

[1] H. Anderson and J. McGeehan, "Optimizing microcell base station locations using simulated annealing techniques," in *Proceedings of VTC*, Stockholm, Sweden, June 1994, vol. 2, pp. 858–862.

[2] H. D. Sherali, C. M. Pendyala, and T. S. Rappaport, "Optimal location of transmitters for micro-cellular radio communication system design," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 4, pp. 662–673, May 1996.

[3] K. Lieska, E. Laitinen, and J. Lähteenmäki, "Radio coverage optimization with genetic algorithms," in *Proceedings of PIMRC*, Boston, MA, September 1998, vol. 1, pp. 318–322.

[4] ETSI, "Broadband radio access network (BRAN); HIPERLAN type 2 technical specification; physical (PHY) layer," Tech. Rep., ETSI, 1999.

[5] IEEE Standard 802.11a, Part II, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: High speed physical layer in the 5 GHz band," Tech. Rep., IEEE, 1999.

[6] M. Unbehaun, "A feasibility study of wireless networks for 17 and 60 GHz and the impact of deployment strategies on the system performance," Licentiate thesis, KTH, Stockholm, Sweden, March 2001.

[7] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*, John Wiley, New York, NY, 1998.

[8] S. M. Sait and H. Youssef, *Iterative Computer Algorithms with Applications in Engineering: Solving Combinatorial Optimization Problems*, IEEE Computer Society, Los Alamitos, CA, 1999.

[9] V. Rayward-Smith, I. Osman, C. Reeves, and G. Smith, Eds., *Modern Heuristic Search Methods*, John Wiley, Chichester, UK, 1996.

Digital Broadcasting and Mobile Cellular Networks to Provide Asymmetric Data Services - A Survey

Aurelian Bria[1]

[1]KTH-Radiosystemteknik, Stockholm. E-mail: aurelian@radio.kth.se.

Abstract

High demand for mobile Internet services and interactive services are expected to characterize the user of the future wireless systems. For a large amount of applications the user terminal will have to receive massive amount of information while sending only a small amount of data. Exploiting this fact in the infrastructure design may have significant implications on the deployment cost of the infrastructure since higher powers can be used at the transmitting sites. Emerging examples include the integration of digital broadcasting systems with personal communication systems. Being capable of taking advantage of future digital broadcasting infrastructure (DAB/DVB-T) and the high amount of spectrum allocated for radio and TV broadcasting in VHF/UHF band, the new hybrid systems show promising performance and especially good business opportunities when deployed in wide areas with fairly low user densities. The objective of the paper is to evaluate the potential of digital broadcasting systems (mainly DVB-T) when using them as a high-speed downlink channel in a hybrid system providing large-scale asymmetric services. This investigation is part of the research on 4th Generation of Wireless Systems supported by Swedish Personal Computing and Communication program (PCC).

1 Introduction

Starting from the initial scenarios developed by the PCC-4GW project some working assumptions regarding the future generations of mobile services were developed [1]. Among others, we expect that a large amount of traffic will be of asymmetric character, where user terminals request massive amounts of information. Exploiting this fact in network design may have significant implications on infrastructure deployment costs since higher transmitter powers can be used in base stations. Emerging examples include the integration of digital broadcasting systems with personal communication systems. In the near future we expect a shift from analogue to digital TV and audio broadcasting, so that a large amount of spectrum in the band 470-862 MHz can be allocated for personal data services. The paper will consider mainly DVB-T (Digital Video Broadcasting-Terrestrial) due to its capabilities to provide a lot higher data rate and significant lower delay than DAB (Digital Audio Broadcasting).

According to PCC working assumptions and supported by estimations made by UMTS forum we expect asymmetries of 1/400 to 1/40 for high, respective medium bandwidth multimedia services, while low or even no asymmetry is assumed for interactive multimedia, switched data, simple messaging or speech services. Traffic for the asymmetric services is estimated to require around 30% (2005) to 50% (2010) of the whole bandwidth consumed for terrestrial personal communications infrastructures [2].

2 Ongoing efforts

2.1 MCP

MCP (Multimedia car Platform) was launched by European Commission in January 2000 and has as primary goal to enable the horizontal market for infotainment services in the car [3]. The project focuses on integrating the high rate downlink of digital broadcasting into the GSM and UMTS networks in order to provide unicast and multicast services to the mobile users. Integrated mobile networks supporting enhanced broadcasting and mobile positioning services are also considered. MCP will be based on open standards for broadcasting (DVB-T, DAB) and mobile radio standards (GSM) and applications are hardware platform independent. A modular and scalable approach for the terminal must include the further evolution towards 4G. On the network level MCP builds on earlier hybrid systems developed by ACTS project MEMO (DAB + GSM) [4] and on SABINA System (DVB-T + GSM) developed by Teracom [5].

2.2 DRiVE

The overall objectives of the project [6] are to enable spectrum efficient high-quality wireless IP in a heterogeneous multi-radio environment and to deliver in-vehicle multimedia-services, which ensure universally available access to information and support for education

and entertainment. To achieve this objective the DRiVE (Dynamic Radio for IP Services in Vehicular Environment) project addresses the convergence of cellular and broadcast networks to lay the foundation for innovative IP-based multimedia services. Therefore, DRiVE tackles the issues of inter-working of different radio systems (GSM, GPRS, UMTS, DAB, DVB-T) in a common frequency range with dynamic spectrum allocation and co-operation between network elements and applications in an adaptive manner.

2.3 COMCAR

The COMCAR project is an initiative of German Ministry of Education and Research and is part of the UMTSplus concept [7]. The main focus of this project is on asymmetrical and interactive mobile-IP services, based on the integration of additional downlink into UMTS and co-operation with digital broadcast systems. The key issues are high quality IP services and communication to and in vehicles. A concept demonstrator was build.

2.4 SABINA

The Swedish broadcaster Teracom AB conducted the test system and trials of the SABINA prototype (concept demonstrator). SABINA (System for Asymmetric Broadband INternet Access) is designed to carry the fundamental Internet transport protocols: TCP and UDP. Measurements from SABINA project showed data rates of 10 Mbps in highly mobile environments [5], using GSM as a return channel and a modified version of TCP for asymmetric traffic. The delay introduced by the round-trip time was around few milliseconds.

3 Digital terrestrial broadcasting

DAB – Digital Audio Broadcasting, is intended to replace the current analogue broadcasting for audio programs [8]. The modulation technique is COFDM (Coded Orthogonal Frequency Division Multiplex). Each of OFDM sub-carriers is DQPSK modulated. The system is able to provide up to 1.2 Mbps on the air interface. Currently, the success of DAB is questionable due to the slow pace of adoption and the high cost of the terminals.

DVB-T – Digital Video Broadcasting – Terrestrial is adopted as a standard for future terrestrial video broadcasting in Europe [9]. It uses the same COFDM modulation technique like DAB, but for each OFDM sub-carrier one out of several modulation techniques can be used. It can go up to 32 Mbps on the air interface. A good comparison between DAB and DVB-T standards together with a concept of a multi-standard receiver can be found in [10]. A large number of DVB-T networks are already running in most of the West European countries and the receiver penetration is increasing slowly mainly due to the high price of the devices. Already today DVB-T networks are offering e-mail and Internet browsing services, using PSTN or GSM as return channel.

3.1 SFN formation

There are many advantages of the COFDM digital modulation, but one of the major ones is its immunity against echoes generated by multipath propagation. By correctly setting the guard interval COFDM is able to constructively use the echoes, increasing the received power. In this way active echoes provided by co-channel transmitters or repeaters can reinforce the natural echoes, produced by reflection or refraction. This is the strategy used in SFN – Single Frequency Network. With COFDM, it becomes more efficient to use several low power transmitters or repeaters than using a highly powered single transmitter.
Using SFN substantially improves the frequency economy for broadcast services compared to the conventional systems which require the transmitter to use different frequencies. The only setback is loss in bandwidth due to necessary guard interval, but loss is negligible compared to the overall improvements in frequency economy.

3.2 Mobility

The efficiency of DVB-T for mobile users is demonstrated in [12]. The excellent correlation between field strength and mobile reception quality is illustrated. In [11] it is proved that the throughput per user is almost independent of the C/I as long as it is higher than a certain threshold, i.e. the breakthrough point. Under these circumstances, forward error protection methods used in DVB are so powerful that they are able to reduce transmission errors to almost zero ($< 10^{-11}$) as long as a certain carrier-to-noise ratio in the physical transmission channel is exceeded. The effect of motion is translated into a loss in signal-to-noise ratio value, but this can be compensated by an increase in transmitted power. The faster the receiver moves the more power is required [13].

3.3 Hierarchical modulation

Fixed and even mobile DVB-T receivers have the advantage of a external (roof top) antenna gain in most of the cases, while the portable receivers use small and inefficient antennas. Moreover, indoor, the portable receivers are penalized by the building penetration loss. One of the advantages offered by DAB and DVB-T is the possibility to choose protection levels and modulation methods, so there could be a lot to gain by dynamically changing some of the transmission parameters. By means of hierarchical modulation it is possible to modify the modulation parameters in order to facilitate the indoor

reception for portable receivers. Two different streams can be created: the high-priority stream (HP), intended for large coverage, and the low priority stream (LP) characterized by restricted coverage, but higher rate [14]. The HP stream is modulated as a 4-QAM and the maximum useful bit-rate is depending only on the coding rate used to protect it (it could be more than 2 times less than data rate on LP, but it is strongly protected). The LP stream is introduced as an over modulation of the HP one and is perceived by the HP receivers as an additional noise in the quadrant of the received constellation.

3.4 Delay

While providing interactive services by means of an uplink channel, the total delay of the service will be given by the round-trip-time (RTT) and the queuing time of the packets at the broadcasting transmitter. When using DAB the RTT is very large, about 1 second mainly due to interleaving time. DVB-T offers low delay values due to much lower values of the interleaving time than DAB. For the case when packets experience no queuing time at the broadcasting transmitter we have:

$$MinDelay = \frac{MessageSize}{DataRate} + 2 \cdot T_i \qquad (1)$$

where T_i represents the interleaving time. In DAB the value of T_i is 384 ms.

4 Broadcast and cellular

The big difference between the broadcasting and mobile communication networks stems from the type of service they provide. While broadcasting services aim to be received by many users in the same time and without considering any difference between users, the telecom ones address each user separately, according to the type of terminal, priority or type of service. Combining these two strategies and being able to provide multicast services in a telecom network or personal services in a broadcasting network brings an additional value and enable a large number of new applications. Overall, several reasons supporting the integration are:

- Each of the systems is optimized for a certain type of service.
- Combination can provide the same services but in a more spectrum efficient way in some cases.
- Low investments in the introduction phase and in rural areas with low terminal penetration.
- DVB-T can be seen as a complement to UMTS for providing high-speed downlink to fixed, mobile or even portable users.

At a first glance, we can formulate some requirements to be met by the hybrid broadcasting-cellular systems: network transparency, mobility, scalability of services and infrastructure, security and open architecture (based on IP).

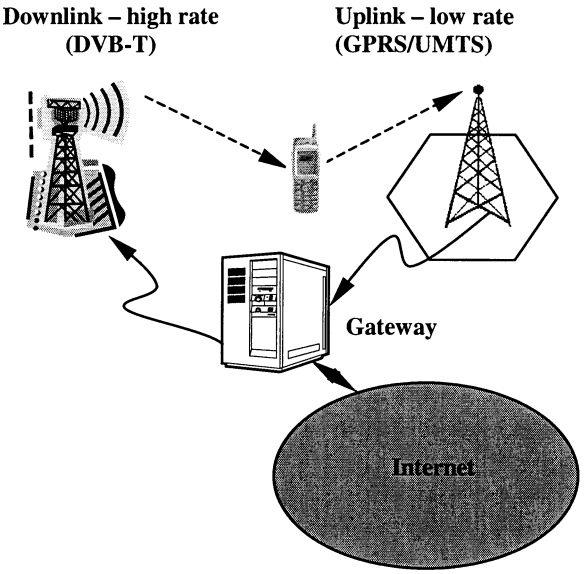


Fig. 1 – Concept of a hybrid broadcasting-cellular system

5 Key research issues

In this section we point out and discuss few emerging key issues, relevant for the success of the implementation of hybrid broadcasting-cellular systems:

5.1 Resource allocation schemes

Dynamic packet assignment was proposed in [15] as a method to increase the spectrum efficiency when providing personal services by means of DAB/DVB-T networks. The method implies packet-by-packet dynamic allocation of resources (transmitters). The concept of PARPS (Packet and Resource Plan Scheduling) is introduced. This reduces all the problems of dynamic resource management, power control, dynamic channel allocation, etc. to a scheduling problem. The approach is to have different SFN formations in each timeslot. These dynamic SFNs facilitate soft handover and efficient multicasting and broadcasting communication. The method shows an improvement in spectrum efficiency, but nowadays the implementation in a real system is questionable, due to the difficulties encountered when switching on and off a high power transmitter.

Dynamic spectrum allocation between mobile Internet and TV services is proposed in the MCP project [3]. The goal is to have full utilization of the spectrum instead of the non-uniform load we have today. Many services are delay insensitive and the content for these services can be pushed and stored in the user's receiver in the time periods when the network is less loaded. For example movies, music, or the most popular web pages can be

broadcasted over night, while during the day the spectrum will be saved for real-time communication. Solutions for dynamic spectrum allocation include automatic spectrum allocation and de-allocation allowing the system to "breathe" with respect of requested spectrum. In [16] a similar solution is proposed and a special common coordination channel (CCC) is employed for negotiations of the spectrum requests.

5.2 Receiver Storage

Using broadcasting we avoid to send the same information more than once, so the utilization of available spectrum will be improved significantly. The idea is to broadcast and store in the user terminal on-line information (web pages, etc.) or media (music, movies, maps, etc.). Users can browse files stored already locally as well as download new files when is needed or only small updates of the cached files. The user perception will be a *"virtual continuous connection"*. The concept is supported by the future availability of large storage capacities (hard-disk and on-chip memory) according to the Moore's law. Furthermore, especially for mobile and portable terminals, issues like cost efficient memory availability and power consumption have to be investigated.

5.3 Migration from macro to micro cells

When deploying isolated large cells the signal level determine the quality of service. If we have to consider much closer and smaller cells the limitations introduced by interference will lead to a worse behavior of downlink oriented channels. Local clustering in small SFNs following the model of cells in cellular systems could be a solution to provide increased capacity, but if DAB and DVB-T are suited for this migration to micro-cellular environment in a cost efficient way is a subject of future investigations. Both systems were intended for outdoor large area coverage, so migration to micro-cellular architecture will imply modifications in the current standards, at least at the levels of protocols and frame organization.

5.4 Handling Large Traffic Asymmetry

Modified TCP and selective ARQ were introduced in SABINA in order to decrease the requirements for uplink capacity on GSM network. Personal data services need error correction and retransmissions, so error control strategies to provide true interactivity for big differences between uplink and downlink have to be developed. On the other hand, media services are offered by means of the UDP protocol, so that there is no need for retransmission of the incorrectly received packets (they can be actually lost without any major impact on the service quality if the error correcting schemes can manage the situation. The reason is that voice or image is perceived in terms of quality and not bits).

5.5 Multicasting

Multicasting will enable all kind of media services in a more spectrum efficient manner, but in case of large file transfers (software updates, maps, etc.) multicasting will face the cruel reality of packet errors that implies retransmissions. Since different terminals will demand retransmissions of different packets, the advantage of spectrum efficiency introduced by multicasting can be severely affected when a large number of retransmission has to be performed. In conclusion, flexible combinations of unicast and multicast may have the ability to improve the spectrum efficiency, at least for certain services.

5.6 Co-siting

One of the issues supporting the integration of broadcasting and cellular systems is the cost minimization. On the operator side (both broadcasters and telecom) the larger investment goes into infrastructure (towers, underground cables and fiber, real estate, etc.). It seems like the price of these assets will not drop at the same pace as the price of electronic equipment does. It actually remains constant or even gets higher. From this point of view, installing the broadcasting transmitters on the GSM/UMTS towers or vice-versa will be a very good opportunity to save money. Difficulties when doing this arise from the need for agreements between different entities (operators, regulators, etc.), but from technical and economical point of view it is a promising challenge.

6 Alternative solutions

Among a large number of asymmetric wireless infrastructure architectures proposed over time by the industry or academic research, we identified two of them showing comparable potential with the combination of broadcast and cellular. First of them is the new DVB-RCT and second one is HSDPA. The technical performance of these two, say competitors, are comparable to our hybrid systems.

6.1 DVB-RCT

DVB-RCT (DVB-Return Channel Terrestrial) is a very new standard [17] and benefits from field experimentation and public demonstrations performed by the European-funded research project iTTi ("interactive Terrestrial TV integration" – 1997-1999) [18]. Two new projects have recently been set up to support DVB-RCT implementation: WITNESS ("Wireless Interactive Terrestrial Network System and Service") and

IM4DTTV ("Integrated Modem fo(u)r Digital Terrestrial TV). Among the advantages offered by DVB-RCT are:

- Multiple access uplink channel based on OFDM technique.
- It can serve large cells (up to 65 Km radius), but it can be implemented with smaller cells (3.5 Km radius) being able to carry even few megabits per second per user.
- Designed to process up to 20.000 short interactions per second.
- Does not require more than 0,5W rms transmission power from the User Terminal or Set Top Box to the base station.

The system consists of a forward channel (downlink) represented by the digital broadcasting and a return channel (uplink) based also on a wireless VHF/UHF transmission. The members of the DVB-RCT group estimate a small deployment cost of the additional return channel compared to other solutions like (PSTN, XDSL, GSM or UMTS). It is expected that the price of the conventional digital Set-top Box will increase only by few percents by adding the return channel facility [19]. Availability of mobile and portable terminals using DVB-RCT and modifications of the current regulation regarding the use of broadcasting channels can transform this solution into a big success in both, fixed and mobile, environments.

6.2 HSDPA

Asymmetry in the traffic pattern of tomorrow's wireless systems is considered very serious by 3GPP, which is already working on the development of a specialized downlink channel integrated in 3G (WCDMA) standard. High Speed Downlink Packet Access (HSDPA) [20] is based on techniques such as adaptive modulation and hybrid ARQ to achieve high throughput, reduce delay and achieve high peak rates (8 –10 Mbps). HSDPA relies on a new transport channel, the HS-DSCH (High Speed – Downlink Shared Channel), which is terminated in the Node B in the UTRAN.

Performing some comparisons of network complexity & cost between the discussed hybrid systems and UMTS, for various user populations and infrastructure densities, has to be done in order to evaluate the superiority of a system over the other.

7 Conclusions

Integration of the digital broadcasting systems (DAB/DVB-T) with the cellular ones (GSM/3G) seems to offer a cost efficient implementation of a multitude of interactive broadcasting services as well as personal data services of asymmetric character.

The article tries to build an overview on the relevant issues and challenges regarding this topic from a technical, as well as from an economical point of view. Advantages offered by the hybrid systems are in terms of spectrum efficiency, high capacity in the downlink, low investment costs and a good business opportunity

References

[1] Aurelian B., Fredrick G., Olav Q, Rikard S., Matthias U., Jiang W., Maxime F., Jens Z. *4th Generation Wireless Infrastructures – Scenarios & Research Challenges*, Special Edition IEEE Personal Communications, Dec.2001

[2] UMTS Forum, Report#6

[3] MCP Project, http://mcp.fantastic.ch/

[4] W. Klingenberg, *Using a Hybrid DAB/GSM Communication System to Provide Interactive Multimedia Services to Mobile Users,* Proceedings of the WPMC'98, Yokosuka, 1998

[5] Erik Stare, Stefan Lindgren, *Hybrid Broadcast-Telecom Systems for Spectrum Efficient Mobile Broadband Internet Access*, NRS01 Workshop Stockholm, Sweden, April 2001.

[6] DriVE Project, http://www.ist-drive.org/

[7] COMCAR Project, http://www.comcar.de/

[8] ETSI, "Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers", May 2001, EN300-401

[9] ETSI, *"Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television"*, Jan.2001, EN300-744

[10] Alexander Romanowski, *Concept of a multi-standard receiver for broadcast and communication services,* IEEE Transactions on Consumer Electronics, vol.43, nr.3, Aug.1997

[11] Anders Lindberg, *Aspects on Individual Services in a Dense Cellular Broadcasting Network*, Master Thesis at Teracom AB, March 1998

[12] U. Reimers, *Field Trial DBV-T in Northern Germany: Field Measurements,* http://www.dvb.org

[13] U. Reimers, *Digital Video Broadcasting*, IEEE Communication Magazine, June 1998

[14] Gerard Faria, *The magics of terrestrial digital TV*, http://www.itis.com

[15] M. Eriksson, H. Säterberg, *The concept of PARPS - Packet and resource plan Scheduling,* MMT'99, Venice, Italy, Oct 1999

[16] Ralf Keller & others, *Convergence of Cellular and Broadcast Networks from a Multi-radio Perspective,* IEEE Personal Communications, April 2001

[17] ETSI, *Digital Video Broadcasting (DVB); Interaction channel for Digital Terrestrial Television (RCT) incorporating Multiple Access OFDM,* March 2002, EN301-958

[18] iTTi Project, http://www.cordis.lu/infowin/acts/rus/projects/ac321.htm

[19] G. Faria, F. Scalise *DVB-RCT: A standard for interactive DVB-T*, http://www.itis.com

[20] 3GPP, *HSDPA-General Description-R5*, Dec. 2001, TS 25.308 v5.1.0

Diversity benefits in 60 GHz VCN wireless architecture

Maxime Flament and Arne Svensson
Department of Signals and Systems, Chalmers University of Technology, SE-41296 Gothenburg, Sweden
Email: {maxime.flament,arne.svensson}@s2.chalmers.se

Abstract

This paper gives comprehensive description of a broadband communication system using the 60 GHz frequency band. We describe the channel and its inherent problems such as interference situations and body shadowing. We introduce the use of Virtual Frequency Networks and multiple receiving antennas in order to solve these problems resulting in an overlapping cellular architecture. We explain how the diversity of the channel can be used to increase reliability and performance. Finally, we simulate the link performance of a VCN using two receiving antennas for a set of OFDM parameters.

1 Introduction

During the past decade, the demand for higher data rate wireless access has been growing at a dramatic pace. The forthcoming third generation networks (3G) will provide new opportunities to the wireless markets. However, the services offered by this technology barely compete with the fixed access possibilities in local or hot-spot places, which opens the door to alternative operatorless wireless technologies such as the IEEE802.11 family and HiperLAN/2 standards. The 3G actors are now feeling the pressure caused by the fast penetration of these Wireless LAN (WLAN) forcing them to solve compatibility issues and integrate them in the 3G wireless standards under the name "Beyond 3G". In the near future, we will experience a revolution going from centralized mobile phone networks to decentralized mobile computer networks thanks to the WLAN hot-spots penetration. Private and public hot-spots will cover most of the populated area and commercial centers leaving to the UMTS the small rest of the market; in the same way GSM did to the satellite mobile services such as Iridium. However, WLAN technologies are answering only a small part of the bandwidth expectation in the future. Five to ten years from now, a single wireless access will be expected to provide at least the same bandwidth as today's office computers for a lower price.

This paper is meant to give a good overview at the characteristics of the 60 GHz indoor channel and identify the different sources of diversity that we could exploit. We introduce the concept of Virtual Cellular Networks (VCN) to increase the Direction of Arrival diversity from the wireless network viewpoint and introduce channel macro-diversity. Using multiple receiving antennas, we exploit the relatively small human scale of the spatial diversity at the receiver. In addition, the diversity of the frequency-selective channel can be handled by forward error correction or using adaptive modulation across the subcarriers. For each of these types of diversity, we discuss the potential solutions and their implementation for future 60 GHz wireless infrastructure. For example, the combination of VCN and multiple antennas can yield a dramatic diversity gain in some cases and provide more stability in the network but this involves at a centralized implementation that reduce the flexibility. In the last part of this paper, we propose an OFDM modulation scheme adapted to the channel and we simulate the link performance using different receiving techniques. In the case where the channel is known both at the transmitter and at the receiver, we can show that the performance of the system can reach promising values.

2 The 60 Ghz channel

In the quest for ever-increasing bandwidth, we have chosen to investigate the unlicensed 60 GHz frequency band, offering up to 5 GHz of bandwidth. The large bandwidth offers a great capacity for wireless broadband systems. For example, a single AP in an meeting room could offer, using simple modulation, a steady throughput of 200 Mbit/s to not less than 25 users at the same time, without sharing any resources or, similarly, offer up to 5 Gbit/s to a single user.

2.1 Propagation characteristics

The propagation characteristics at this high frequency have been studied in many publications. In [1], we identified a series of inherent problems for the use of WLAN in offices and small open areas. Among others, the fast spatial change of the quality of the channel and the high attenuation due to body shadowing are the most destructive effects. The Direction of Arrival and the Time of Arrival of the rays play an important role in the rest of this paper. We illustrated it in figure 1. An extensive description is given in [2].

The propagation at 60 GHz can be explained if you imagine a cellular communication system (like GSM 1800) in which users and objects are 30 times bigger and system bandwidth is 2500 larger. The distances

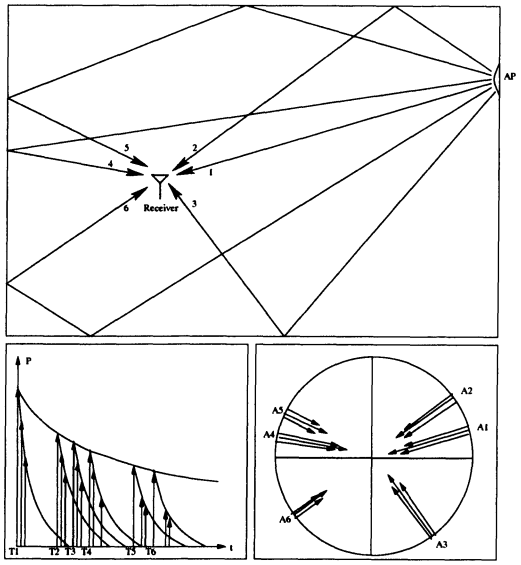


Figure 1: Illustration of the Saleh-Valenzuela model with direction information [3]. Rays arrive in cluster from the same direction due to multiple reflection in materials. The taps in the discrete-time model are likely to contain power from one single cluster i.e. originating from one particular direction.

become too small for the user scale. Hence by scaling the system to the human scale, it becomes an indoor environment (30 times smaller than GSM 1800) with larger delay spread compared to the bandwidth (about 2500 higher throughput than GSM 1800 if we don't segment the available bandwidth) and users are moving very fast compared to the 5mm wavelength. An access point would have a typical range of 25 meters but the cell shape are mainly dependent on the surrounding furniture, walls and windows (likewise the cells in GSM 1800 are dependent on the buildings and reliefs). Since typical rooms are not as big, coverage is not the main limitation. On the other hand, interference situations due to cell overlapping occur in short bursts and create unstable handover situations during which the mobile could initiate multiple handovers between different APs within a few centimeters (likewise the "manhattan grids" create problems in cities for GSM systems). This issue is worse than at lower frequencies since users are moving fast compared to the wavelength.

With this huge bandwidth, we can accept some spectral efficiency losses in order to increase the link performance or the coverage of the cells. For example, according to discussions in the FCC "ET docket 94-124", the 59.00–59.05 GHz portion of the band (50 MHz band) will be common to all users, for assistance in the coordination of real time use in the remainder of the band. Users must transmit an ID code, at least once per second, as an aid to resolving rapidly any interference problems. This can for example help the adaptive formation of virtual cells for each particular users with

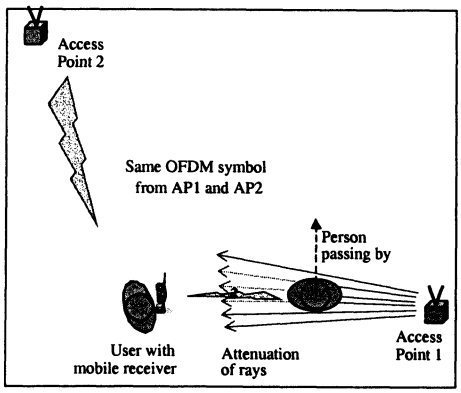


Figure 2: Typical shadowing situations for indoor Wireless LAN systems

a good spectrum efficiency, a minimum of interference and little handover traffic. If we can standardize the coordination channel and deregulate completely the rest of the frequency band it would increase dramatically the potential of the band for a great variety of applications.

2.2 Destructive Shadowing effect

The similarity with the lower frequency systems is not completely accurate. In the system we describe, people moving around the receiver destroy part of the signal by shadowing incident rays. The shadowing process becomes therefore a dynamic problem that is similar to the interference problem described above when the user itself is moving. If a user is standing still while other persons are walking inside the room, the mobile device will eventually lose the connection to the AP when a person is walking through the line-of-sight (LOS). It will then attempt to hand-over to another AP with a stronger instantaneous signal. The situation is depicted in Figure 2.

Previous studies [4] have showed that a person walking through the LOS of a communication link at 60 GHz can cause attenuation to the received power exceeding 18 dB. Achieving coverage in a crowded room is therefore a critical problem. In order to avoid that this kind of situations result in a dropped network connection, the network must be designed in a way that AP's coverage overlap each other. This effort to increase link reliability introduces even more interference problems for which we need to find a convenient solution. It could be solved using the usual frequency or code separation between users and APs. However, since in our case conventional methods would require high control traffic, it becomes increasingly interesting to solve the problem by interference mitigation or by virtual cells. In the next section we try to describe the advantage of the two methods.

3 Design alternatives in overlapping cellular network

3.1 Virtual cellular networks

Smart network infrastructures, called Virtual Cellular Networks (VCN), was first introduced in [5] but we interpret it a little bit differently. We propose this technique in order to mitigate the 60 GHz channel shadowing problems. In our case, VCN uses distributed access points (AP) and Single Frequency Networks (SFN) within one virtual cell in order to form an adaptive wireless communication architecture. The idea of VCN originates from single frequency networks, such as Digital Audio Broadcast (DAB), but VCN uses adaptive resource allocation. A VCN system intelligently selects the APs forming a virtual cell for a particular user. Firstly, in order to preserve capacity, APs only transmit if their received power contribution at the terminal exceeds a certain received threshold. Secondly, the number of APs in one virtual cell is limited in order to avoid increasingly time dispersive channels. Thirdly, in the case of good VCN channel quality, we try to combine the contributions from each VCN AP in an optimum way to increase the diversity gain. The use of VCN is particularly relevant for 60 GHz wireless LANs. At this frequency, the propagation range is dramatically limited while the 5 GHz bandwidth is large. The major motivation for VCN is the increased Direction of Arrival diversity at the receiver but it also introduce Time of Arrival diversity (frequency-selection).

The VCN basically creates a cell for a user which is going to follow him. The concept is very interesting because the handover procedure doesn't exist anymore on the terminal side. It is the network that is switching on and off APs in such a way that the signal is received by the user. It changes completely the way we think about a cellular network and new rules and adaptive techniques need to be implemented. The use of VCN requires a centralized network unit (or eventually a common control channel) that decides how the virtual cells move with the user. This solution is therefore difficult to implement in public places where many APs belong to different wireless networks. However it is the solution that will give the best results for the interference problem.

In this paper, VCN is used jointly with OFDM based modulation in order to mitigate the delay between different transmitting APs in a virtual cell. Since OFDM is already identified as an important feature in a 60 GHz wireless channel because of its time-dispersive nature, the additional complexity at the receiver due to VCN is negligible. One major problem occur when rays from different channels contribute to the same discrete tap. Two taps from two different channels could cancel each other due to phase differences. This problem is illustrated by the example in figures 3 and 4. We can notice that some subcarriers suffer greatly from the interference. This effect could be circumvented if we know the two channels before we send the signal from the transmitting APs but it would require to have a reliable feedback channel in order to provide this information at the transmitters.

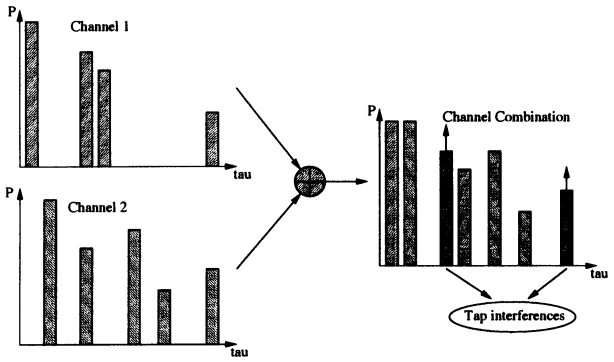


Figure 3: Channel combination in VCN results in a new time delay profile. Whenever rays from different APs contribute to the same discrete tap, they might cancel each other resulting in power loss. Knowing the channel at the transmitter we could mitigate this problem.

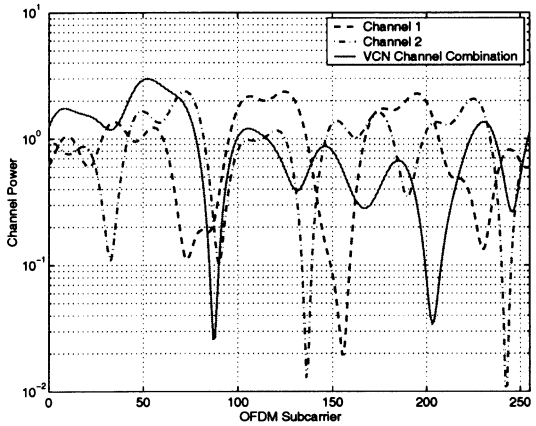


Figure 4: VCN channel combination influence on OFDM subcarriers. Due to the interfering taps, some subcarriers suffer from the VCN channel combination.

3.2 Multiple receiving antennas

In contrast with the previous section, it is interesting to consider the possible use of multiple receiving antennas at the receiver. Multiple antennas are interesting in order to select particular direction of arrival (DoA) of the desired signal. The DoA are basically detected by processing the slight differences between the estimated channels at the receiving antennas. At 60 GHz, the waves from one AP often arrive at the user from a limited number of directions and the same is true for the eventual interfering waves if the cells are overlapping. In addition, the physical implementation of multiple receiving antennas on a small terminal will be relatively easy since the wavelength is very small, i.e. the size of the antennas and the required distance between them is

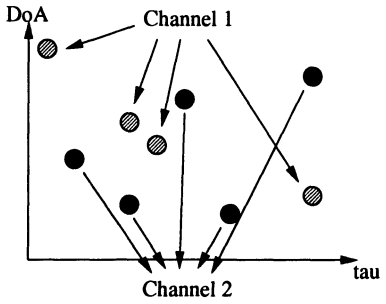


Figure 5: Direction of Arrival can be used to separate two different channels and therefore avoid intererference from different taps.

very small. It will be therefore interesting to use directional antenna techniques in order to solve the problem created by the overlapping cellular design. However, multiple receiving antennas necessitate an increased signal processing at the receiver at least proportional to the number of branches. When there is no interference, the combination of the signal exploiting the best the antenna diversity is known as maximum ratio combining but we could use also equal gain or selection combining for each subcarriers. But it is also possible to cancel the interference even if they do fall in the same discrete tap. Figure 5 shows the interesting DoA properties that multiple antennas could exploit. As soon as we know the wanted channel and the interferers, we can calculate the best way to combine the signal from the antennas and to mitigate the interference for each OFDM subcarriers. Using multiple antennas without a VCN is not extremely interesting since shadowing is still going to require temporary handovers from one APs to the others. However, due to the mitigation of the interference, it solves the problem created by overlapping cells.

3.3 Joint use of VCN and multiple receiving antennas

The VCN and multiple receiving antennas are two complimentary techniques. The system would use the multiple antennas as a safety net if the VCN has not activated one of the interfering AP in the virtual cell for a particular user. The terminal can use the 60 GHz coordination channel to ask the interfering channel to joint the virtual cell. If this AP doesn't have any resource available, the terminal will continue to cancel its signals. Figure 6 gives the procedure for the formation of the virtual cells including the option of ignoring an AP.

In addition, the system could increase dramatically the link quality as soon as the channels from different APs have a good quality without anyone shadowing it. Indeed, the multiple antennas can be used to separate the different channels coming from different APs. Even if the APs are sending exactly the same signals, it is interesting to detect them separately and combine them

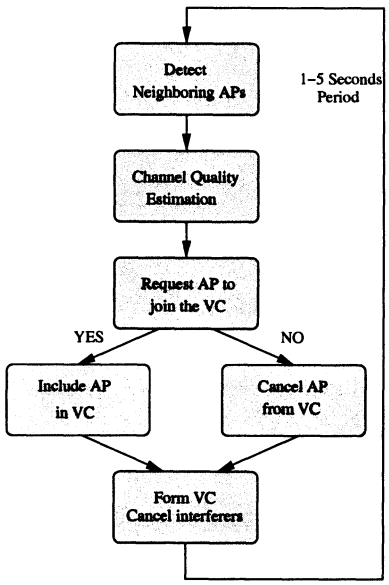


Figure 6: VCN virtual cell formation procedure when using multiple receiving antennas.

in an optimum way.

3.4 Exploiting frequency diversity

The time dispersion of the channel can be interpret as a variation of the signal quality across all the subcarriers introducing a subchannel diversity similar to the time diversity of a flat Raleigh fading channel but with a different auto-correlation function. In section 4, forward error correction know as convolution codes are used to mitigate the deep fades in particular subchannels. However, when a good knowledge of the channel is present at the transmitter, it is possible to use adaptive modulation for each of the subcarriers. The techniques to use adaptive modulation in a VCN network might become very complicated but it could increase dramatically the capacity of a channel.

4 Simple link simulation

An indoor 60 GHz WLAN link using VCN and OFDM modulation depicted in Figure 7 is simulated. Signal bandwidth is fixed to 200 MHz divided into 256 subcarriers with a Cyclic Prefix length Δ=30 i.e. a CP time period of Δ/BW=150 ns. The subcarrier bandwidth is wide enough (781 kHz) and the symbol time is short enough (1.43 μs) such that the effect of speed can be ignored in one entire frame. 16–QAM baseband modulation is used. The diversity of the frequency-selective channel is exploited by a rate R=1/2 convolutional coding with constraint length 7 using a simple block interleaver across the subcarriers. No puncturing is considered. For simplicity, hard decoding is performed at the receiver after the combiner. Perfect APs power control is assumed. The estimated channels

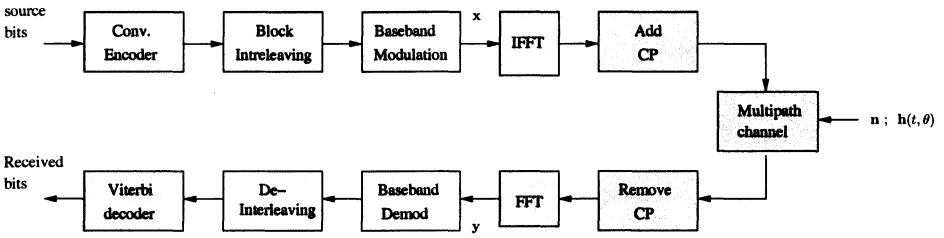


Figure 7: Overview of the OFDM communication system. Using VCN, the multipath channels are duplicated with uncorrelated noise n and channel $h(t, \theta)$.

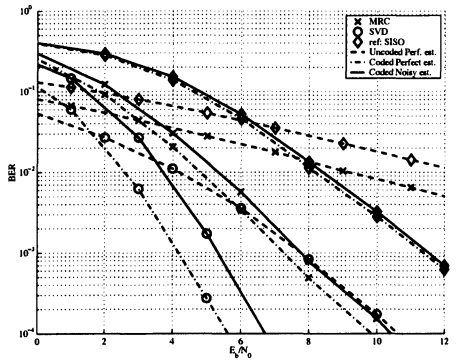


Figure 8: VCN using two antennas on each side. 16–QAM is evaluated with rate 1/2 convolutional coding (K=7). Perfect and noisy channel estimates are considered.

are generated at the beginning of a frame. Estimation error is calculated from the noise and the number of training symbols [2]. For comparison, the SISO performance are shown, and, perfect and noisy estimates are considered. Singular Value Decomposition (SVD) and Maximum Ratio Combining (MRC) are evaluated related to the case the channels are known or not at the transmitters.

In order to simulate the channels, we use a large set of Saleh-Valenzuela channel realizations represented by tap-delay models and we form the resulting time dispersive channels out of n_T randomly picked channels. VCN AP delays are uniformly distributed with maximum 30 ns. The n_R channels from one VCN antenna have a similar discrete-tap pattern but experience different tap fading due to the path differences. They are therefore generated using independent Rayleigh random values on each of the taps. The results in figure 8 show the performance of the described system for two transmitting APs and two receiving antennas.

5 Conclusions

In this paper, we described the use of OFDM, VCN and multiple receiving antennas in order to solve many problems identified in broadband wireless networks operating in the 60 GHz frequency band. The different aspects of the channel diversity suggest that it is possible to build a system that mitigate most of these problems. We identified each of the diversity degrees of freedom and explained comprehensively how the system can take advantage of it. The complexity introduced is relatively limited but simplified algorithms should be proposed. We also simulated the link performance for an 16–QAM OFDM modulation scheme for two transmitting APs and two receiving antennas. The paper show that 60 GHz wireless networks have a strong potential but many open issues still need to be investigated.

References

[1] Maxime Flament, *On 60 GHz Wireless Communication Systems*, Licenciate thesis, Chalmers University of Technology, Gothenburg, Sweden, Dec. 2000.

[2] Maxime Flament and Arne Svensson, "Performance of 60 GHz virtual cellular networks using multiple receiving antennas," in *Proc. of IEEE International Symposium on Wireless Personal Multimedia Communications*, Denmark, Sept. 2001, vol. 1, pp. 129–134.

[3] Q. H. Spencer, B. D. Jeffs, M. A. Jensen, and A. L. Swindlehurst, "Modeling the statistical time and angle of arrival characteristics of an indoor multipath channel," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 347–360, Mar. 2000.

[4] K. Sato and T. Manabe, "Estimation of propagation-path visibility for indoor wireless LAN systems under shadowing condition by human bodies," in *Proc. of IEEE Vehicular Technology Conference*. IEEE, May 1998.

[5] H. J. Kim and J.-P. Linnartz, "Virtual cellular network: a new wireless communications architecture with multiple access ports," in *Proc. of IEEE Vehicular Technology Conference*. IEEE, June 1994, vol. 2, pp. 1055–1059.

System Evaluation of Optimal Downlink Beamforming in Wireless Communication

Rickard Stridh[1], Mats Bengtsson[2] and Björn Ottersten[3]

Dept. of Signals, Sensors & Systems, Royal Institute of Technology, SE-100 44 Stockholm, SWEDEN.
[1]E-mail: Rickard.Stridh@s3.kth.se, [2]E-mail: Mats.Bengtsson@s3.kth.se, [3]E-mail: Bjorn.Ottersten@s3.kth.se

Abstract

In this paper, we investigate the use of joint optimal downlink beamforming, power control and access point allocation, in a multi-cell SDMA system. Smart antennas are used at the access points and single antennas are used at the terminals. The possibility to send messages to multiple terminals on the same frequency in the same time slot is exploited. An algorithm for removing users from a congested system, in order to avoid outage is proposed and evaluated. Results show that the proposed algorithm gives substantial increase of the system performance compared to random removal of users.

1 Introduction

Wireless communication is currently in a transition from carrying only voice communication to provide both voice and data communication and the capacity requirements are increasing. Smart antennas has been proposed in order meet the increased capacity requirements as an alternative to e.g. exploiting new frequency bands. Smart antennas utilize the fact that different mobile terminals in a wireless system have different spatial signatures, seen from the access point. Using this property, each access point is able to communicate with several mobile terminals, using the same time and frequency resource. This access scheme is denoted Space Division Multiple Access (SDMA). This paper treats the downlink of such an SDMA system. The motivation for treating the downlink is that data traffic is generally asymmetric and thus the downlink is critical for the system capacity [1].

The issue of downlink beamforming has been investigated in e.g. [2] and [3]. The problem of optimal downlink beamforming is formulated in [4] and can be solved using the algorithms in [5] or [3]. In [1], the latter joint optimal power controller and beamformer is assessed. This method uses semi-definite optimization to find the global optimum. The algorithm has been extended to perform joint optimal power control, beamforming *and* access point allocation in [6]. The algorithm described requires global knowledge of channel statistics and may be used in the search for the fundamental performance limits of a wireless system. Though for limited systems in slowly changing environments (e.g indoor systems), it may be possible to implement a system with centralized beamforming based on global channel statistics knowledge.

Using algorithms for beamforming in a system context will introduce the problem of admission control, i.e. how to admit or remove mobiles in the system. Different approaches are possible and herein the goal is to remove or not admit, the mobile that is most critical to the system, in term of the possibility to reach the target Signal-to-Interference-and-Noise Ratio (SINR) for all remaining mobile terminals.

These methods, including joint optimal beamforming, power control and access point assignment, are investigated in the context of cellular networks with traffic and radio resource management issues. Moreover, an algorithm for admission control is proposed and evaluated.

The paper is organized as follows: First the system and signal assumptions are reviewed, followed by the algorithms used for beamforming, power control, access point assignment and terminal removal. Then simulation results and a summary concludes the paper.

2 System assumptions

A cellular system, with several terminals sharing the same frequency channel is investigated. Each access point is supposed to be communicating with several terminals. We assume a controlled environment where all mobile terminals request a channel resource. When connecting, the terminals are allocated to an appropriate access point.

It is assumed that the access points are equipped with array antennas with n_T omni-directional elements and that the channel is a low rank (flat fading) statistical channel modelled with local independent scattering at the terminal. It is further assumed that the second-order statistics of the vector channels are known to the transmitter at the access point.

The user requirement on QoS is assumed to be fulfilled if the Signal-to-Interference-and-Noise Ratio (SINR) at all user terminals exceeds the SINR requirement γ_0.

2.1 Signal Model

The set $\mathcal{T}$ contains all the transmitters, i.e. access points, using the same time and frequency channel as the considered receiver. The set $\mathcal{R}$ contains all receivers, i.e. terminals in the system and the sets $\mathcal{R}_j$ contain all receivers assigned to the transmitter j.

The $n_t \times 1$ signal $\mathbf{x}_j(t)$, transmitted from access point j has the form

$$\mathbf{x}_j(t) = \sum_{i \in \mathcal{R}_j} \mathbf{w}_i s_i(t), \quad (1)$$

where $\mathbf{w}_i$ are the $n_R \times 1$ transmitter weights for terminal $i \in \mathcal{R}$. Then, the signal received at terminal i is

$$r_i(t) = \sum_{j \in \mathcal{T}} (\mathbf{v}_{i,j})^* \mathbf{x}_j(t) + n_i(t). \quad (2)$$

The channel vectors $\mathbf{v}_{i,j}$ are assumed zero mean, random variables. with channel covariance matrix defined as

$$\mathbf{R}_{i,j} = E[\mathbf{v}_{i,j}\mathbf{v}_{i,j}^*]. \quad (3)$$

The user requirement on QoS is assumed to be fulfilled if the Signal-to-Interference-and-Noise Ratio (SINR) at the user terminal i, for all terminals $i \in \mathcal{R}$ exceeds the SINR requirement γ_0.

$$\text{SINR}_i = \frac{\tilde{G}_{i,i}\tilde{P}_i}{\sum_{j \in \mathcal{T}} \sum_{\substack{n \in \mathcal{R}_j \\ n \neq i}} \tilde{G}_{i,j}\tilde{P}_n + \sigma_i^2} \geq \gamma_0, i \in \mathcal{R} \quad (4)$$

Here, $\tilde{P}_i$ is the transmitted power to user i and $\tilde{G}_{i,j}$ is the channel gain from access point j to terminal i *including* all channel effects and beamforming. The scalar σ_i^2 is the noise power received at receiver i.

3 Beamforming

The following alternatives are investigated below

- Traditional beamforming with additional decoupled power control.
- Generalized eigenvalue based beamforming [2] with additional decoupled power control.
- Joint optimal beamforming and power control w/o terminal removal.
- Joint optimal beamforming, power control and access point assignment with terminal removal.

In the first three case, the user is allocated to the access point from which is has the highest gain, while in the fourth case the allocation is included in the optimization algorithm.

3.1 Suboptimal beamforming methods

Here follows the algorithms used in the investigation as comparisons to the joint optimal beamforming and power control algorithm. These algorithms are decentralized to the single access point and thus the multi-access point notation is omitted.

Traditional beamforming is optimal, with respect to SNR in absence of interference and with spatially white noise. The optimum weight vectors when transmitting to terminal i are given by

$$\mathbf{w}_i = \arg\max \frac{\mathbf{w}_i^* \mathbf{R}_i \mathbf{w}_i}{\mathbf{w}_i^* \mathbf{w}_i} \quad (5)$$

The maximum is then found by using a scaled version of the principal eigenvector of $\mathbf{R}_i$.

The generalized eigenvalue-based beamformer is a decentralized beamformer that gives the maximum signal power at the desired terminal and keeps the total transmitted power to all not desired terminals in the set $\mathcal{R}$ below a certain value [7] [8]. The beamformer is given by

$$\mathbf{w}_i = \arg\max \frac{\mathbf{w}_i^* \mathbf{R}_i \mathbf{w}_i}{\mathbf{w}_i^* (\alpha \mathbf{I} + \sum_{\substack{n \in \mathcal{R} \\ n \neq i}} \mathbf{R}_n) \mathbf{w}_i} \quad (6)$$

where $\alpha\mathbf{I}$ may be interpreted as channels unknown to the access point or as a robustification to channel uncertainties [2, 3]. The problem can be solved as a generalized eigenvalue problem [9].

3.2 Power Control

Power control decoupled from beamforming may be performed with the following algorithm. The requirement in (4) may be rewritten as

$$\text{SINR}_i = \frac{P_i}{\sum_{j \in \mathcal{T}} \sum_{\substack{n \in \mathcal{R}_j \\ n \neq i}} Z_{ij} P_n + N_i'} \geq \gamma_0, i \in \mathcal{R} \quad (7)$$

where Z_{ij} and N_i' represent the normalized link gain and noise, respectively. It is possible to get the optimal power vector by solving the inequality above. This system of inequalities provides a SINR-balanced system i.e. all terminals will have the same target SINR, and if one mobile is outaged then all will be outaged. This is not desirable and a method to solve this problem is to remove terminals until all remaining terminals reach the target SINR, advised in [10]. In this investigation, this has not been done since the same behavior is present when the algorithm for joint optimal power control and beamforming is used without admission control. Thus, the results measured as raw outage probability will be very pessimistic, but comparable to the methods below.

3.3 Joint optimal beamforming and power control

Optimal beamforming is defined as the set of beamformers that minimize the total transmitted power in the system,

while providing sufficient received Signal-to-Interference-and-Noise Ratio (SINR) at each terminal.

The problem may be formulated

$$\min \sum_{i \in \mathcal{R}} \|\mathbf{w}_i\|^2 \tag{8}$$

$$\text{s.t. } \frac{\mathbf{w}_i^* \mathbf{R}_i \mathbf{w}_i}{\sum_{j \in \mathcal{T}} \sum_{\substack{n \in \mathcal{R}_j \\ n \neq i}} \mathbf{w}_n^* \mathbf{R}_{i,j} \mathbf{w}_n + \sigma_i^2} \geq \gamma_0, i \in \mathcal{R}.$$

where σ_i^2 is the noise variance at terminal i and γ_0 is the target SINR for fulfillment of user QoS. This problem can be solved using so-called Lagrange relaxation [11] techniques. Define the matrix $\mathbf{W}_i = \mathbf{w}_i \mathbf{w}_i^*$, then the relaxation of (8) is

$$\min \sum_{i \in \mathcal{R}} \text{Tr}[\mathbf{W}_i] \tag{9}$$

$$\text{s.t. } \text{Tr}[\mathbf{R}_i \mathbf{W}_i] - \gamma_0 \sum_{j \in \mathcal{T}} \sum_{\substack{n \in \mathcal{R}_j \\ n \neq i}} \text{Tr}[\mathbf{R}_{i,j} \mathbf{W}_n] \geq \gamma_0 \sigma_i^2$$

$$\mathbf{W}_i = \mathbf{W}_i^*$$

$$\mathbf{W}_i \succeq 0, i \in \mathcal{R},$$

where $\mathbf{W}_i \succeq 0$ denotes that $\mathbf{W}_i$ is positive semidefinite. This relaxed problem is non-linear but convex, and can be solved efficiently, using standard interior point methods for semidefinite optimization, for example with the SeDuMi toolbox [12]. In general, such a relaxation technique can only provide a lower bound on the original problem but for this specific problem formulation, it can be shown [3, 6] that if (9) has a solution, then it has at least one optimal solution where rank$[\mathbf{W_i}] = 1$ for all $i \in \mathcal{R}$, i.e. it solves 8. This formulation may be extended to include additional constraints or robustness to channel uncertainties.

3.4 Joint optimal beamforming, power control and access point assignment

Above, the assignment of mobile terminals to access points was assumed to be given. However, the optimal beamforming algorithm can be extended to determine not only the transmit weights, but also the optimal access point assignment of the mobile terminals. The trick is to first view the antenna arrays of all the access points as a single large access point and determine the optimal beamforming vectors from this combined access point to all the terminals. It can be shown that the optimal solution will always use only one of the access points for each terminal, see [6].

3.5 Terminal admission/removal

When too many users are present in the system, the optimization algorithm cannot provide a solution and all users experience outage. To fully use joint optimal beamforming and power control in a real system, an admission or scheduling algorithm must be added. When the beamforming problem is solved with SeDuMi [12] or some other convex optimization package using self-dual embedding techniques, the algorithm provides a certificate of the infeasibility in the form of a so-called Farkas-dual solution [13]. This is a direction vector for the dual problem along which the dual cost function increases infinitely. In order to find the mobile terminal that is most critical for the problem feasibility, a good heuristic is to find the largest element of the Farkas-dual solution. The terminal with the SINR constraint corresponding to this dual variable can be considered one of the most limiting for system feasibility. If two terminals are located closely together such that no feasible solution can be found, both will typically get a large Farkas-dual variable but it is often sufficient to remove one of them to reach a feasible situation. Thus, we suggest to repeatedly remove the most critical terminal, one at a time, until a feasible solution can be found. Below, we show that this heuristic method can give a significant increase in system capacity compared to random removals. Since the Farkas-dual solution is not unique in most cases, the exact result will depend on the specific algorithm.

4 Results

4.1 Simulations

A multi-cellular system, corresponding to the system assumptions above is simulated, with uniform circular arrays at all access points. Mobile terminals are arriving to the systems according to a Poisson process and they are uniformly distributed over the system area. In order to take signals from surrounding cells into account a wrapped system is used, i.e. all cells are surrounded by all other cells. Thus the whole simulated area corresponds to a torus shape.

Using the local scattering model, parameterizing the channel statistics for the channel from access point j to terminal i, by the nominal azimuth angle $\theta_{i,j}$ and the standard deviation $\sigma_{i,j}$ of the angular spread, the covariance matrix including the propagation properties in the gain matrix, will yield

$$\mathbf{R}_{i,j} = G_{i,j} \mathbf{R}(\theta_{i,j}, \sigma_{i,j}), \tag{10}$$

where $\mathbf{R}(\theta_{i,j}, \sigma_{i,j})$ is the covariance matrix. Here the gain matrix $G_{i,j}$ contains the influence of path loss and shadow fading, while the influence of beamforming and power control is included in the weight vectors given by the beamforming algorithms and power control algorithms.

For all terminals, the SINR is calculated and the probability of outage is calculated. The results are given as raw outage probability i.e. $P(\text{SINR} < \gamma_0)$ or the probability to end up in an infeasible system, when the admission scheme is not used, and probability of blocking when the admission scheme is used. The noise in the system is assumed very low, as the main issue investigated is the interference from other transmitters.

For the classical beamforming and generalized eigenvalue methods the weight vectors are created. Based on the gain matrix and the beamformed antenna diagrams, decoupled power control is applied. The joint beamforming and power control algorithm provide power controlled weights directly.

For the case with joint beamforming, power control and admission control, the optimization problem is solved and if the solution is infeasible the most critical terminal is removed and the new optimization problem is solved. The procedure is iterated until a feasible solution is reached. The positive semidefinite optimization problems are solved with the standard optimization program package SeDuMi [12].

In SeDuMi numerical errors are detected, if the solution is very close to the limit and very unstable. This kind of solutions usually require very large transmit power and are thus not of practical interest. In this case one terminal is randomly removed from the system to reach a stable solution. This means that the results may be somewhat below the possible maximum.

The throughput figures are calculated as Load $\times$ Probability of no blocking, and exclude the retransmission that would be possible in a packet switched system. In order to have a reference figure for the removing procedure, the removal of the most critical terminal is compared to the removal of a random terminal.

The parameters used in the system simulations are shown in Table 1.

Table 1: Simulation parameters.

Number of cells	9 and 16
Reuse factor	1
Noise level	-120 dB
Shadow fading standard deviation	8 dB
Propagation attenuation	$\alpha_{att} = 3.5$
Number of antenna elements	8
Angular spread variance	9°
SINR requirement	$\gamma_0 = 7, 10, 15$ dB
Uniform circular array radius	0.4λ

4.2 Numerical results

The results show that using the beamforming methods without admission gives very high outage probability, as expected. The simulations show that the use of joint optimal beamforming and power control allows five times higher load than conventional beamforming combined with optimal power control at an raw outage level of 10 % as seen in Figure 1. The corresponding figure over the generalized eigenvalue method is twice the load.

The outage probabilities for different QoS levels parameterized as SINR requirement are shown in Figure 2.

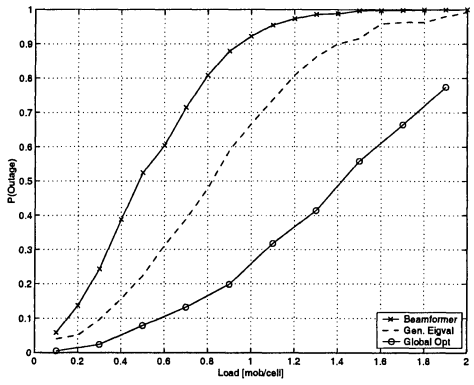


Figure 1: Raw outage probability for different downlink beamforming methods, without admission control. Required SINR is 7 dB.

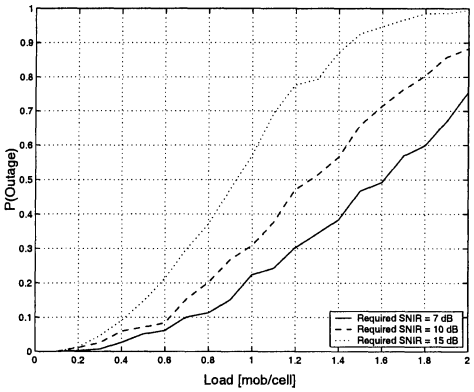


Figure 2: Raw outage probability for different QoS levels with joint optimal downlink beamforming and power control, without admission control.

For typical speech services requiring about 7 dB the double load is possible compared to services requiring 15 dB SINR. These values are valid for an raw outage level of 10 %.

Adding admission control to the system gives substantial increase of the system throughput. The blocking rate in the system is showed in Figure 3. It is shown that the proposed heuristic removal algorithm has substantially lower blocking rates than the random removal algorithm used for reference. In Figure 4 it is shown that the throughput using the random removal algorithm reaches a ceiling, while the heuristic algorithms manages to remove the critical users and thus reaches substantially higher throughput.

Adding optimal access point allocation, jointly to the optimal beamforming and power control, further decreases the blocking rate and thus increases the throughput as shown as dash-dotted in Figures 3 and 4.

At a allowed blocking rate of 10 %, it is possible to double the possible load, when using optimal access point allocation and heuristic admission, compared to the the case with maximum gain access point allocation and random admission.

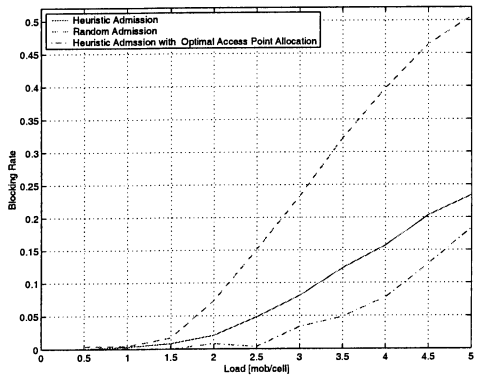


Figure 3: Blocking rate for the optimal beamformer and power control with and without access point assignment and admission control, at SNIR = 7dB

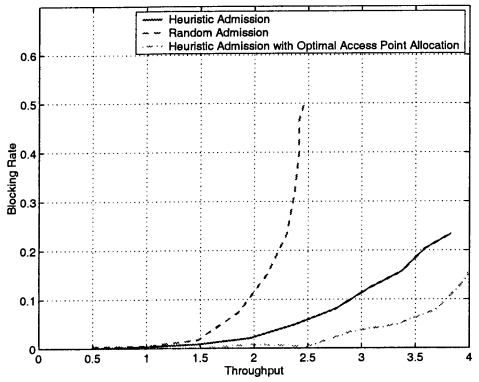


Figure 4: Relation between blocking rate and throughput for joint optimal beamforming and power control with and without access point assignment and with different admission control schemes, at SNIR = 7dB

5 Summary

In this paper we have demonstrated the use of optimal beamforming, power control and access point allocation in a wireless communication system. An heuristic algorithms for admission or removal of users/packets in the system is proposed and evaluated. The results show that the proposed algorithms substantially increases the system performance compared to a reference system with random removal of users or packets.

References

[1] R. Stridh, "Smart Antennas in Fourth Generation Wireless Systems: System Issues and Capacity" Licentiate Thesis TRITA-S3-SB-0101, Royal Institute of Technology, SE-100 44 Stockholm, Sweden, February 2001.

[2] P. Zetterberg, *Mobile Cellular Communications with Base Station Antenna Arrays. Spectrum Efficiency, Algorithms and Propagation Models.* PhD thesis, Royal Institute of Technology, S3, KTH, SE-100 44 Stockholm, Sweden, 1997.

[3] M. Bengtsson, *Antenna Array Signal Processing for High Rank Data Models.* PhD thesis, Royal Institute of Technology, S3, KTH, SE-100 44 Stockholm, Sweden, 1999.

[4] C. Farsakh and J. A. Nosek, "Channel Allocation and Downlink Beamforming in an SDMA Mobile Radio System" in *Personal, Indoor, and Mobile Radio Communications*, pp. 687–691, Sept. 1995.

[5] F. Rashid-Farrokhi, K. R. Liu, and L. Tassiulas, "Transmit Beamforming and Power Control for Cellular Wireless Systems" *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1437–1450, Oct. 1998.

[6] M. Bengtsson, "Jointly Optimal Downlink Beamforming and Base Station Assignment" in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Salt Lake City, Utah), IEEE, May 2001.

[7] P. Zetterberg and B. Ottersten, "The Spectrum Efficiency for a Base Station Antenna Array for spatially selective transmission" *IEEE Transactions on Vehicular Technology*, vol. 44, pp. 651–660, Aug. 1995.

[8] J. Goldberg and J. R. Fonollosa, "Downlink Beamforming for Spatially Distributed Sources in Cellular Mobile Communications" *Sig. Proc.*, vol. 65, pp. 181–197, Mar. 1998.

[9] G. H. Golub and C. F. van Loan, *Matrix Computations.* Baltimore: The John Hopkins University Press, third ed., 1996.

[10] J. Zander, "Performance of Optimal Transmitter Power Control in Cellular Radio Systems" *IEEE Transactions on Vehicular Technology*, vol. 41, pp. 57–62, Feb. 1992.

[11] L. Vandenberghe and S. Boyd, "Semidefinite programming" *SIAM Review*, vol. 38, pp. 49–95, Mar. 1996.

[12] J. F. Sturm, *Using SeDuMi 1.0x, a MATLAB Toolbox for Optimization over Symmetric Cones.* Department of Quantitative Economics, Maastricht University, The Netherlands, `http://www.unimaas.nl/~sturm/software/sedumi.html`, 1999.

[13] Z.-Q. Luo, J. F. Sturm, and S. Zhang, "Duality Results for Conic Convex Programming" Tech. Rep. Report 9719/A, Econometric Institute, Erasmus University Rotterdam, The Netherlands, Apr. 1997.

TCP over WLAN and GPRS with Mobile IP

Jiang Wu

KTH/IMIT, Electrum 229, S-16440, Kista. E-mail: jiang@it.kth.se

Abstract

Mobile IP(v4), as an IETF standard, has been well accepted in the Internet community. Today both WLAN (Wireless Local Area Network) and GPRS (General Packet Radio Services) networks are wide spread and publicly accessible by mobile users, which makes it practical to deploy Mobile IP in very large scale. WLAN and GPRS are different in many aspects. Besides their coverage area, they have huge differences in bandwidth and media access latency.

In this paper, I describe how Mobile IP is used over the IEEE 802.11 WLAN and GPRS networks and introduce our testbed. Using this testbed TCP behavior over WLAN and GPRS with Mobile IP is studied. The goal is to show whether it is feasible to use Mobile IP in such a hybrid wireless infrastructure. In the testbed, the MN (Mobile Node) uses Mobil IPv4 with co-located care-of address in both WLAN and GPRS. The measurements are focused on exploring handover effects on TCP throughput and TCP recovery mechanisms. TCP sessions are captured by "tcpdump (ethereal)". Another tool "tcptrace" is used for analyzing TCP behavior based on the captured sessions.

The TCP analysis shows that it is feasible to use Mobile IP over WLAN and GPRS. A TCP session runs reasonably well even during the handovers. Nevertheless, a major problem is that when an MN handover from WLAN to GPRS, the sudden increase of round trip time between the MN and the CN (Correspondent Node) causes a burst of retransmissions. During such a handover, the TCP session takes around ten seconds to be fully recovered.

1 Introduction

Today both WLAN and GPRS networks are wide spread and practically used by many users. It is of great interest that users can use both wireless networks and remain connected despite the handovers among these networks. Figure 1 shows a typical scenario in which WLANs serve as hot spots, while GPRS covers a much larger area. In this paper, Mobile IP is used for supporting handover between WLAN and GPRS.

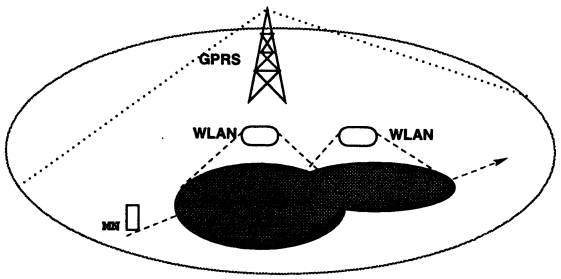


Figure 1 A hybrid wireless infrastructure

1.1 GPRS

GPRS is a new bearer service for GSM that improves wireless access to PDN (Packet Data Networks) such as X.25 networks and the Internet. Users of GPRS benefit from higher data rates and shorter access times than GSM. In practice GPRS offers session establish time below 1 second and data rates up to several tens of Kbps [4].

The GPRS system architecture is built upon the existing GSM system architecture, it is supposed to have the same coverage as GSM. To integrate GPRS into the GSM architecture, a new class of network nodes called GSN (GPRS Support Nodes), has been introduced. GSNs are responsible for interconnecting the mobile nodes and the external PDNs.

IP packet routing in GPRS is complicated. The GGSN (Gateway GSN) is a special purpose router and takes part in the routing protocol exchanges in the Internet. The SGSN (Serving GSN), however, does not exchange routing protocol messages to the other routers though it uses IP to transmit PDN packets with other GSNs. All GSNs are connected via an IP-based GPRS backbone network. Within the backbone the GSNs encapsulate the PDN packets and transmit them using the GTP (GPRS Tunneling Protocol).

An MN is required to authenticate and associate (attach) with the SGSN and GGSN before it can actually get access to the network. If the MN uses IP, it should first establish an IP PDP (Packet Data Protocol) Context with the SGSN and GGSN. The PDP Context is used for MN, SGSN and GGSN to agree on the service level and build up the context for treating the packets (including routing).

GPRS supports different QoS classes, which can be specified for each individual session. GPRS allows defining QoS profiles using the parameters service precedence, reliability, delay and throughput. The delay is defined as the end-to-end transfer time between two communicating MNs or between an MN and the GGSN. The mean delay for QoS class 1 packet is less than 0.5 second, for QoS class 2 it is less than 5 seconds, for QoS class 3 it is less than 50 second. In our testbed, the MN uses the best effort delay class.

1.2 WLAN

WLANs, like their wired counterparts, were developed to provide high bandwidth to users in a limited geographical area [5]. Usually wireless LANs use unlicensed frequency spectrum (e.g., the 2.4G ISM band) as their communication media. Unlike GPRS, a network operator is not absolutely needed in such systems. A WLAN's coverage area ranges from several meters to several hundred meters, within the coverage area users can enjoy high data rate access. Currently the widely deployed IEEE 802.11b WLANs offer a data rate of up to 11Mbps, while the soon coming IEEE 802.11a WLAN and HiperLAN2 promise a data rate up to 54Mbps.

Providing IP connectivity is very simple for IEEE 802.11 WLANs. Packet data support in IEEE 802.11 WLAN is much less complicated than in the GPRS. Usually an IEEE 802.11 AP (Access point) is attached to the IP networks directly. For an MN to use IEEE 802.11 WLANs, it should first authenticate and associate with the corresponding AP. On success the AP starts to deliver the IP packets from and to the mobile node.

1.3 Handover between WLAN and GPRS

Five different architectures for implementing handover between GPRS and IEEE 802.11 networks are described in [10]. The first two architectures involve connecting the WLAN to the GPRS network through GPRS entities such as the SGSN and GGSN, where WLAN will appear to be a GPRS cell or RA (routing area). The third architecture reverses the roles played by the GPRS and the WLAN in the first two architectures. The fourth approach introduces an MG (mobility gateway) between WLAN and GPRS. The last architecture is based on Mobile IP to handle the mobility, in which GPRS and WLAN are peer networks serving as link layer to Mobile IP. In [10], Mobile IP is claimed to be the most suitable architecture for handover in a hybrid network. It is also the chosen mechanism for implementing handover in this paper.

Mobile IP [11] is a standard Internet protocol that supports IP layer mobility. By introducing two additional routing entities: HA (Home Agent) and FA (Foreign Agent, optional), MNs are able to maintain their IP address, hence need not change their network configuration or lose ongoing communications with their CNs. In mobile IP, an MN in a foreign network registers its care-of address to the HA. The HA then sets up a tunnel between the HA and the care-of address (can be a FA or the MN itself). The packets destined to that MN are captured by the HA and delivered to the other end of the tunnel, and finally reach the MN. The packets from that MN can be delivered via normal routing, or via the same tunnel between the HA and the care-of address (bi-directional tunneling).

2 The testbed

The testbed is shown in figure 2. In the testbed, an MN has both a GPRS interface and a WLAN interface. No FA is used in the GPRS network or in the WLAN. The reason is that it is hard for the current GPRS architecture to support a FA in a suitable way [9]. Although there is a proposal from 3GPP [1] to add the FA in the GGSN, it will not be available in the near future.

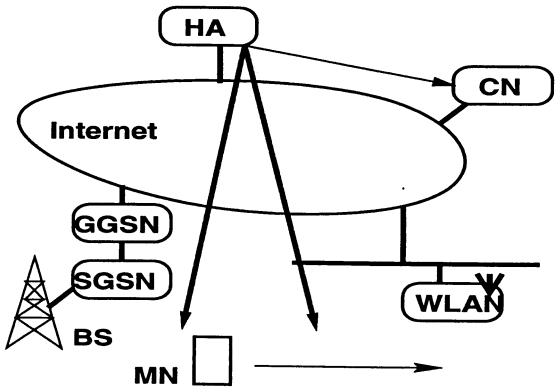


Figure 2 The Mobile IP testbed with WLAN and GPRS

Two GPRS networks are used in the testbed. One is a public GPRS network (Europolitan GPRS network in Sweden) and the other is a private owned GPRS network (Ericsson's internal GPRS network). Mobile IP over the public GPRS network needs additional mechanisms other than available from the Mobile IP standards. In public GPRS service, an MN obtains a private IP address using DHCP. A NAPT (Network Address Port Translation) gateway is used to translate the [private IP address, port] to the [public IP address, port] between the GPRS network and the public Internet. Unfortunately the NAPT can not

handle IP in IP encapsulation correctly [6]. A possible solution is to create UDP tunnels between the MN and the HA through the NAPT and run Mobile IP over the UDP tunnel [7]. Using private GPRS network means much more flexibility to set up Mobile IP tests. The experiments I have done used both public GPRS network and private GPRS network.

The MN is a HP notebook with an Ericsson R520m phone that supports GPRS. They are connected via PPP connection through a serial link. The R520m phone has 3 time slots downlink and 1 time slot uplink. The mobile node also has a Breezecom IEEE 802.11 PCMCIA card that supports up to 2Mbps. In the tests automatic movement detection mechanisms are not used. Handovers are performed by manually switching the WLAN card on/off. The GPRS connection is set up at the beginning of each test, so both GPRS and WLAN interfaces are active before and after a handover.

In the tests, the CN runs as a ftp server and the MN runs as a ftp client. A file worth of 2.5 MB size is transfered from the CN to the MN, it is the data session in the ftp session we capture for analysis. "tcpdump (ethereal)" is used for packet capture. Another tool "tcptrace" [3] is used for TCP session analysis. "tcptrace" is able to visualize the very details of a TCP session, which is proved to be very useful. The CN, HA and MN run RedHat Linux 7.2, the HA and MN use the Dynamics Mobile IP implementation version 0.8. Dynamics Mobile IP [2] is developed under the GPL license, its implementation includes HA, FA and the MN.

3 TCP behavior over GPRS and WLAN

3.1 TCP over GPRS

Since TCP and GPRS both apply their own ARQ (Automatic Repeat Request) mechanisms, it is suspected that GPRS introduces severe performance degradations for TCP traffic. However, in [8] simulation result shows that the ARQ mechanisms in GPRS and TCP coexist very well. TCP would not notice packet errors on the wireless link because the fast GPRS ARQ mechanism can hide the link error from TCP. In the tests, TCP performs worse than in the simulation results. A TCP session has quite a few retransmissions in GPRS networks. The number of retransmissions in the private GPRS network is much less than in the public GPRS network. Figure 3 shows a typical TCP session in the public GPRS network.

The explanation is that the TCP session experiences spurious timeouts, which trigger the retransmissions. The reason for these spurious timeouts is the rapid increase of the RTT (round trip time) brought by the GPRS wireless link. As figure 4 shows, there is a strong correlation between the TCP retransmissions and the rapid increase of RTT. The result shows that even though the ARQ mechanism in GPRS is fast enough to keep the wireless link error-free from TCP, the instability of the round trip time due to the GPRS network causes many retransmissions.

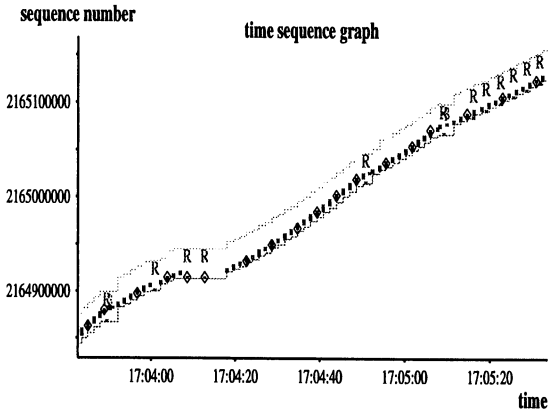


Figure 3 Time sequence graph in the public GPRS network

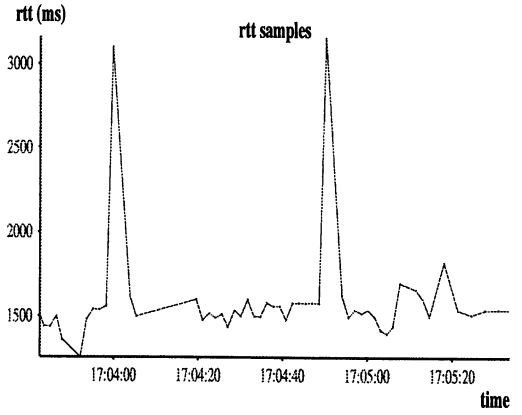


Figure 4 RTT samples in the public GPRS network

3.2 TCP over GPRS and WLAN with Mobile IP

In the tests the MN uses Mobile IP with co-located care-of address and bi-directional tunneling. Handover is done by manually switching between GPRS and WLAN interfaces, so that there is little packet loss caused by the handover. Since handover latency is very short, TCP sessions react mainly to the sudden change of the two different types of access media. Handover with automatic movement detection will complicate TCP behavior even more.

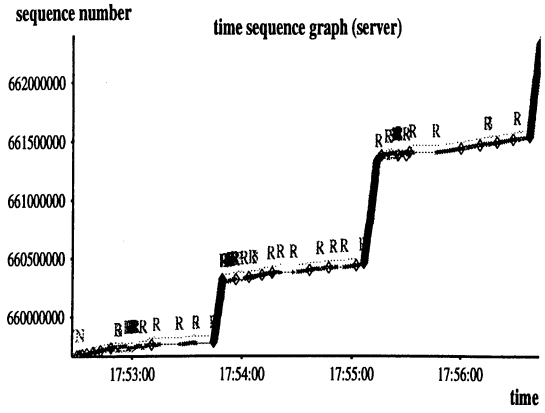


Figure 5 A TCP session over WLAN and GPRS

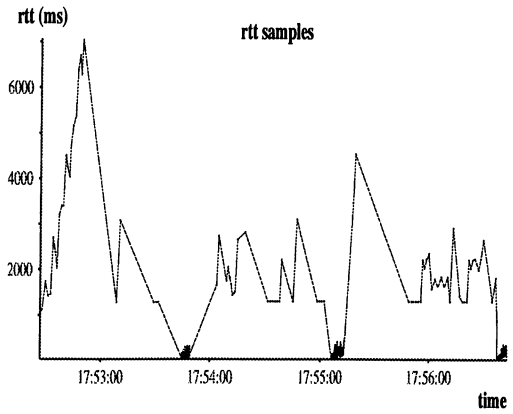


Figure 6 RTT samples over WLAN and GPRS

Figure 5 and figure 6 show a TCP session between the CN and the MN. Figure 5 shows the time sequence graph captured in the CN (server), and figure 6 shows the corresponding RTT value. The MN starts the TCP session using GPRS, and then switch to WLAN. It continues to handover between GPRS and WLAN for a couple of times until the TCP session finishes.

The measured throughput in GPRS is rather low, with the average around 16Kbps. The average RTT is around 2.5 seconds. When handover to WLAN, the average throughput increase to around 1Mbps, while the average RTT drops to 200 milliseconds. RTT caused by the fixed networks in the tests can be neglected, since the CN, HA and MN are very close to each other.

The TCP session goes very smooth when the MN uses WLAN. When using GPRS, there are much more retransmissions. The TCP session starts with slow start. During the handovers, although quite a few retransmissions occur, it does not go to slow start phase.

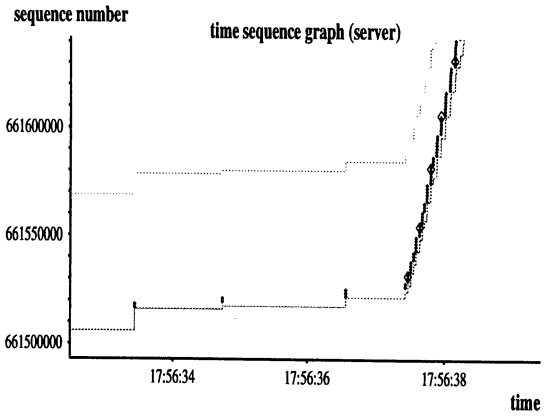


Figure 7 GPRS to WLAN handover

Figure 7 is a zoomed area from figure 5, which shows a typical TCP time sequence during the handover from GPRS to WLAN. After the MN sends a registration request message for the WLAN interface, the HA immediately sends back a registration reply. Few or no retransmissions are found during the handover. Handover latency is almost negligible.

On the opposite direction, handover from WLAN to GPRS triggers a burst of retransmissions. Figure 8 shows a zoomed area in figure 5, which describes a typical TCP time sequence (with sequence number between 660280000 and 660340000) during the WLAN to GPRS handover. Figure 9 shows the corresponding TCP session captured in the MN. Upon receiving a registration request from the MN, the HA sends back a registration reply and switches immediately the tunnel to the new co-located care-of address. During this process, due to the fast transmission speed right before the handover, there are still some TCP packets and acknowledgements on the fly between the HA and CN. In figure 8, we can see there are 19 packets with high speed from the server as if the MN is still located in the WLAN. Those packets are buffered in the GSNs and transmitted over the new tunnel through GPRS. In figure 9 we can see 13 packets are spread out in time and received by the MN in GPRS, with 6 packets missing (probably due to the lack of buffer in the GSNs). Meanwhile the CN's TCP RTO (Retransmission TimeOut) value was still based on the estimation of the previous connection with WLAN, which is much smaller than the RTT value in GPRS. The expiration of the RTO in the CN triggers a burst of retransmissions after the handover. Those retransmitted packets are received by the MN, most are discarded, while the 6 missing packets are accepted. From then the TCP session becomes stable in GPRS. From the handover moment to the stable session, it takes around 12 seconds both in the CN and MN.

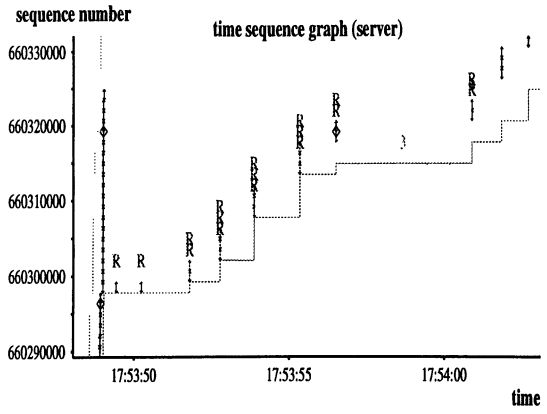


Figure 8 WLAN to GPRS handover

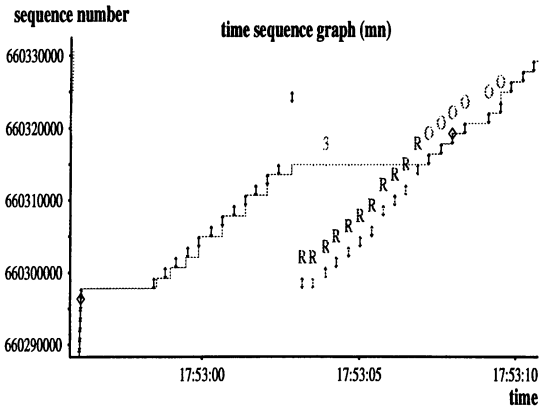


Figure 9 WLAN to GPRS handover (captured in the MN)

4 Conclusions

In this paper I described a Mobile IP testbed built on WLAN and GPRS. TCP sessions are captured and the TCP behavior during handover is analyzed. I show that the sudden increase of RTT in GPRS itself, as well as during handover from WLAN to GPRS, disturbs TCP sessions a lot, which results in bursts of retransmissions. Mobile IP with co-located care-of address is used for handover between WLAN and GPRS. The results show that it is feasible to use Mobile IP in such a hybrid wireless infrastructure. Handover from GPRS to WLAN is smooth. Handover from WLAN to GPRS, however, causes a TCP sender to retransmit a bunch of packets. It typically takes around ten seconds to have the TCP session recovered after the handover.

Due to the limitation of the testbed and the Mobile IP implementation, automatic movement detection is not used in the tests. TCP behavior will be more complicated when handover is triggered by intelligent mechanisms, rather than manually. It is remaining to be done as future work.

References

[1] http://www.3gpp.org.

[2] http://www.cs.hut.fi/Research/Dynamics.

[3] http://www.tcptrace.org.

[4] Christian Bettstetter, Hans-Jorg Vogel, and Jorg Eberspacher. GSM Phase 2+ General Packet Radio Service GPRS: Architecture, Protocols, And Aire Interface. *IEEE Communication Surveys*, 2(3):2–14, Third Quarter 1999.

[5] Brian P. Crow, Indra Widjaja, and Jeong Geun Kim; Prescott T. Sakai. IEEE 802.11 Wireless Local Area Networks. *IEEE Communications Magazine*, pages 116–126, September 1997.

[6] Juan C. Bayerri et al. Turncoat: An Investigation into Intersystem Handover. Technical report, KTH/IMIT, May 2001.

[7] H. Levkowetz and S. Vaarala. Mobile IP NAT/NAPT Traversal using UDP Tunnelling. *Internet Draft*, March 2002.

[8] Michael Meyer. TCP Performance over GPRS. In *IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1248–1252, 1999.

[9] Jonas Mohlin. Mobile IP in GPRS and IEEE802.11 Wireless LAN. Master's thesis, KTH/IMIT, April 2002.

[10] Kaven Pahlavan, Prashant Krishnamurthy, Ahmad Hatami, Mika Ylianttila, Roman Pichna Juha-Pekka Makela, and Jari Vallstrom. Handoff in Hybrid Mobile Data Networks. *IEEE Personal Communications*, pages 34–46, April 2000.

[11] C. Perkins. IP Mobility Support. *RFC2002*, October 1996.

Accurate Measurements of small Antennas and Radiation from Mobile Phones in small Reverberation Chambers

Per-Simon Kildal[1,2], Yngve Hamnerius[1]

[1]Department of Electromagnetics, Chalmers University of Technology, 41296 Gothenburg, Sweden,
E-mails: simon@kildal.se, www.kildal.se and yngve@elmagn.chalmers.se
[2]Bluetest AB, Chalmers teknikpark, 41288 Gothenburg, Sweden, per-simon.kildal@bluetest.se

Abstract

The paper gives a summary of the work that has been performed in the Antenna group at Chalmers University of Technology and the start-up company Bluetest AB on measuring radiated power from mobile phones in reverberation chambers. Reverberation chambers were originally developed for EMC measurements. We have shown that it also can be used to accurately measure performance of antennas that are designed for use in multipath propagation environment, as well as the total radiated power of complete phones. The antennas and phones can be measured with or without the presence of a head phantom or other objects. In the present paper we describe the measurements of the total radiated power and provide results for phones with different types of antennas. We will also show results of validation against measurements of the total radiated power by integration of three-dimensional far field patterns measured in anechoic chambers.

1 Introduction

A mobile phone must radiate in order to work. Thereby it is also unavoidable that part of the radiated power is absorbed in the human head. This absorption is characterized in terms of a Specific Absorption Rate (SAR), and there exist requirements to the maximum allowed value of the SAR. These values have been determined by considering the possibility of health hazards. It has therefore been an issue that the SAR value should be as low as possible, without actually mentioning the fact that a phone must radiate in order to work satisfactory. It is therefore necessary to introduce a complement to the SAR value. This is the maximum power the phone can utilize for communication in the presence of the human head. This radiated power can be measured in many ways. E.g., it can be measured in a standard anechoic chamber. The purpose of the present paper is to describe measurements of the radiated power in a reverberation chamber. The reverberation chamber has

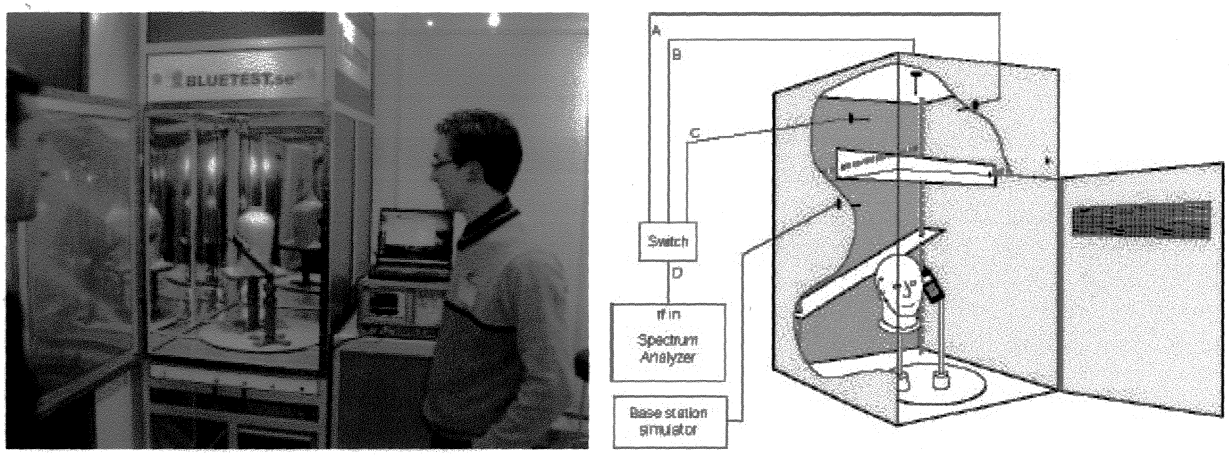


Figure 1. Photo (left) of Bluetest's reverberation chamber RC800 with open door, and drawing (right) of measurement setup with base station simulator and spectrum analyzer used in pulse power mode. The two plate shaped mechanical stirrers, the platform stirrer and the three antennas for polarization stirring are shown [19].

many advantages over an anechoic chamber. The anechoic chamber needs to be several meters in cross section, whereas the reverberation chamber we use is much smaller, only 0.8m x 1.05m x 1.6m in size (Figure 1). The measurements take also much shorter time. We will in this paper also show that the measurement accuracy is comparable with that of a good anechoic chamber. The present paper is a summary of results from [1].

2 Radiated power, head loss and SAR; three quality measures of mobile phones

The total radiated power in the presence of the head is the figure of merit of a mobile phone, when it is transmitting. The higher the radiated power, the better the phone will work in the transmit mode. On the other hand, the possible radiation hazards are characterized in terms of a Specific Absorption Rate (SAR) distribution that should be as low as possible or at least below some standardized limits, see e.g. [2]. Both the radiated power and the SAR are proportional to the maximum power that can be radiated by the phone. Therefore, a high quality phone must provide a good compromise between high radiated power and low SAR. This is possible by using an antenna that directs the radiation from the phone away from the head. In practical measurements the human head is replaced by a head phantom with similar characteristics as the human head.

The radiated power is proportional to the radiation efficiency of the antenna on the phone, measured with a head phantom present. The radiation efficiency as defined in [3] has three contributions: The reflections due to impedance mismatch between the phone antenna and the transmission line that connects the phone antenna to the receive and transmit amplifiers inside the phone, the absorption in the phone antenna itself including the chassis of the phone, and the absorption in the phantom.

The less absorption in the head, the lower is the SAR. Thus, the best compromise between high radiated power and low SAR is to make the antenna radiate away from the phantom (head). Thereby, the absorption in the phantom will decrease so that the radiation efficiency increases.

In a setup for measuring the radiated power it is not possible to separate the power absorbed in the head from the results. However, we can measure the total loss in radiated power when we move the phone from a position far from the head to one of the talk positions defined in [2]. This loss we refer to as head loss, i.e. loss due to the presence of the head phantom. It is caused by both mismatch of the antenna on the phone and absorption, and we cannot separate the two. A good phone has low head loss. A good phone has low head loss.

3 Reverberation chamber

The present tests make use of a so-called reverberation chamber. The reverberation chamber is well known within the EMC area. Its theory is described in detail in [4]-[8]. It has recently also been used to characterize antennas. The reverberation chamber is also called a mode stirred chamber, as it contains several cavity modes, which are stirred mechanically to provide several statistically independent field distributions. These field distributions correspond to what in mobile communications result from multipath propagation, as shown in [9]. The mechanical stirrers can have many forms. The reverberation chamber at Chalmers makes use of two plate stirrers. One of these can be moved across the back wall of the chamber, and the other can be moved over a complete horizontal cross section in the upper part of the chamber (see Figure 1).

The antenna group at Chalmers has shown that reverberation chambers can be used to measure radiation efficiency of antennas [10]-[11]. In order to improve accuracy we have developed platform stirring [12] and polarization stirring [13], see Figure 1. We have also shown that it can be used to measure the "free space" input reflection coefficient of small antennas in the vicinity of some object such as a head phantom [14]-[15]. With "free space" reflection coefficient we mean the reflection coefficient, as it would be seen if the antenna and the object were located in free space. We have also shown that reverberation chambers can be used to measure effective diversity gain very accurately, if the phone has more than one antenna and makes use of diversity [16]. Measurements of both radiation efficiency and radiated power in Bluetest's chamber are in [17] validated against measurements in another reverberation chamber and in two anechoic chambers. Some of the last results are given in the present report.

4 Validation

The measurements in Bluetest's reverberation chamber have been validated against results from two anechoic chambers, one at Sony Ericsson Mobile Communications AB (SEMC) in Lund, Sweden and the other at Saab Ericsson Space AB (SES) in Gothenburg, Sweden [17]. The total radiated power of each of ten different mobile phone models was measured. In addition, the antennas on eight of the phone models were fed by a cable (spare set of phones), and the radiation efficiencies were measured. Both the antenna radiation efficiency and the radiated power were measured in the presence of a head-and-shoulder phantom from ECE, Japan. The handsets were measured

in only one position - the so-called left talk, 45° user-intended position.

The results for the measured radiation efficiencies in GSM 900 and GSM 1800 bands are given in Figure 2 for the SES anechoic chamber and the Bluetest reverberation chamber. Three frequency points were measured for each GSM band. The Bluetest results were frequency stirred (averaged) over 25 MHz around each frequency point. We used 25 platform-stirring positions with 2 positions of the metal stirrers and 3 excitation antennas for polarization stirring. The SES results were at each frequency obtained by pattern integration over the whole radiation sphere. There is very good agreement between the results for all phones. This can be clearly observed from the very small difference of the corresponding bars and lines in the given charts. The measured frequency variation is seen to be very similar in the chambers as well.

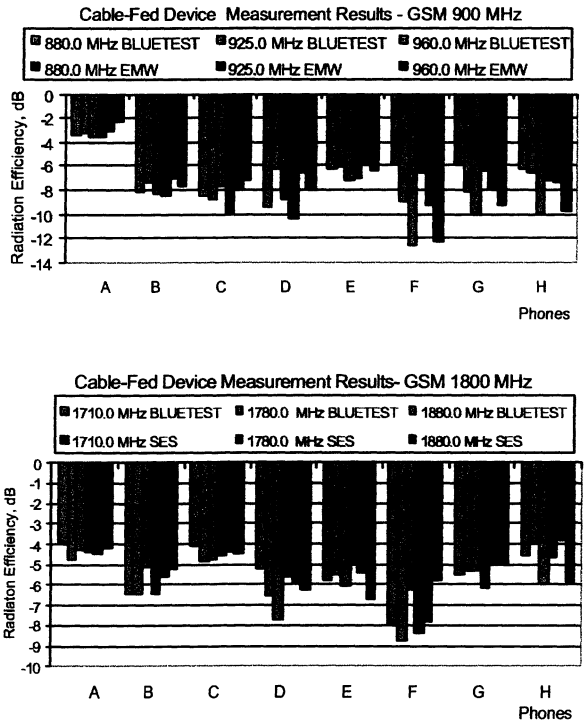


Figure 2. Comparisons of radiation efficiencies of phones A to H measured in Bluetest's reverberation chamber and anechoic chamber at SES.

The results from measuring the radiated power are shown in Figure 3 for the GSM 900 and 1800 bands. The measurements were performed in the SEMC and Bluetest chambers. The SEMC chamber did not at the time of the measurements provide absolute power results. Therefore, we normalized the power of all phones to the average radiated power of all the phones under test, in each GSM band and each chamber. The Bluetest results were obtained by 25 platform positions, 2 positions of the metal stirrers, 3 receive antennas and frequency stirring (averaging) over 5 frequency points, equally spaced, over the transmit band. The latter corresponds to 25/75/65 MHz frequency stirring for the GSM 900/1800/1900 bands, respectively. In the SEMC chamber, the output data was relative radiated power measured by pattern integration over the radiation sphere for the two frequencies defined by the transmit band edges. The results at these two frequencies were averaged before comparison with Bluetest's values. The active measurements in Bluetest's chamber gave absolute values, but they were still normalized in the same way as the relative power values measured in SEMC's chamber to obtain relevant comparison. Observing the figures we clearly see that the Bluetest and SEMC chambers provide very similar results for the different phones. They rate the phones in the same order.

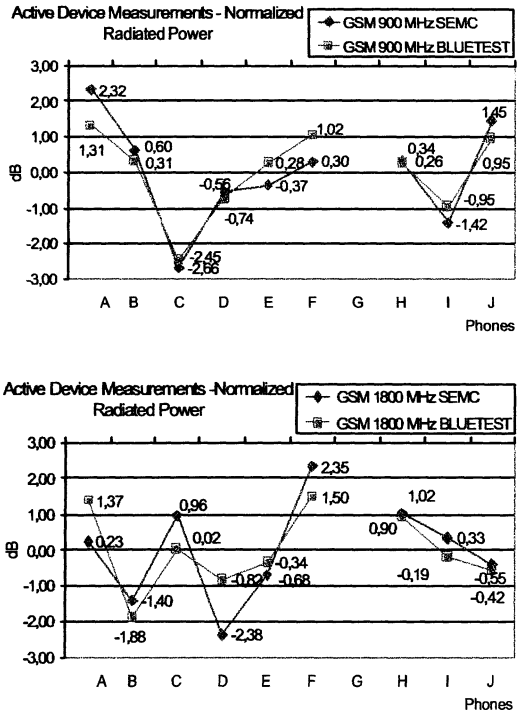


Figure 3. Comparisons of radiated power of phones A to J measured in Bluetest's reverberation chamber and anechoic chamber at SEMC.

In order to analyze statistically the difference between the measured results, we computed the mean and standard deviations of the difference between the two sets of results obtained in two different chambers, see Table 1. This comparison shows that the standard and mean deviations are very similar for the corresponding cable-fed and active phones measurements. It should be noted that these deviations are due to inaccuracies in both the measurements in the anechoic chamber and the reverberation chamber. Also,

a significant part of the deviation is due to positioning errors of the mobile phone relative to the head phantom.

	Radiation eff.		Rad. power	
GSM band	**900**	**1800**	**900**	**1800**
Mean Dev.	0.01	-0-18	0.00	0.00
Standard	0.72	0.55	0.57	0.87
Max. Dev.	1.43	-1.54	-1.0	-1.57

Table 1. Statistical deviations between measured results in reverberation chamber and anechoic chamber

5 Results for radiated power

We will here give some results for measured radiated power and head loss of 20 phones. The complete results including validation measurements are given in [1]. The stirring sequences are the same as described in the previous section. The phones are shown in Fig. 4. The radiated power at 900 MHz is shown in Fig. 5, and at 1800 MHz in Fig. 6. We have chosen to letter-code the phones because we want to study the effect of different antennas types and not the different phone models. The phone code contains even an antenna code that is an abbreviation for external (E), built-in (BI) and extractable (EL).



Figure 4. Photo of the twenty phones measured in for radiated power and head loss.

We have measured the radiated power in dBm for all the talk positions of the phones defined in the new European Standard for SAR measurements developed by CENELEC [2]; free space, cheek right, cheek left, tilt right and tilt left.

We see that all phones have free space radiated power at 900 MHz in the range 29-32.2 dBm, whereas the GSM standard is 33 dBm +/- 2 dB. At 1800 MHz all phones except two have values between 27 and 29.3 dBm when the GSM standard is 30 dBm +/- 2dB. The head loss, which is a quality measure of the antenna on the phone, is the difference in dB between the radiated power in free space and in a talk position. We see that the extractable antennas are very good at 900 MHz. There is not much difference between built-in and external antennas at 900 MHz, whereas the built-in antennas have typically 1 dB smaller head loss at 1800 MHz. The radiated powers vary much more with phone position for phones with external antennas than for phones with built-in antennas. The built-in antenna seems to be less sensitive to how the phone is used. The head loss is between 2.5 and 9 dB at 900 MHz, whereas it is between 1 and 4 dB at 1800 MHz. We must here note that the present results do not take into account the effect of a users hand.

The measurement accuracy has been investigated in detail in [1], by using the recommendations in [18]. The result is an expanded uncertainty of less than 1 dB in both the 900 and 1800 MHz bands. Expanded accuracy means two standard deviations. This value agrees well with the validation results in the previous section.

Acknowledgements

The Swedish Strategic Research Foundation (SSF) has supported the theoretical part of the research associated with the development of the reverberation chamber for measuring small antennas, under a program related to antenna technology. The measurements of the 20 phones have been supported by TCO Development AB. TCO has adopted the present measurement method for radiated power in their new quality marking of mobile phones [20]. The radiated power is there referred to as TCP (telephone Communication Power), which is a copyrighted trademark of TCO Development.

References

[1] Per-Simon Kildal and Charlie Carlsson, *TCP of 20 Mobile Phones Measured in Reverberation Chamber - Procedure, Results, Validation and Uncertainty*, Bluetest AB, Feb 2002, Chalmers Teknikpark, Gothenburg, Sweden.

[2] *EUROPEAN STANDARD EN 50361, Basic standard for the measurement of Specific Absorption Rate related to human exposure to electromagnetic fields from mobile phones (300 MHz - 3 GHz)*, CENELEC European Committee for Electrotechnical Standardization, rue de Stassart 35, B - 1050 Brussels, July 2001

[3] Per-Simon Kildal, *Foundations of Antennas - A Unified Approach*, Studentlitteratur, Feb 2000 (www.studentlitteratur.se/antennas).

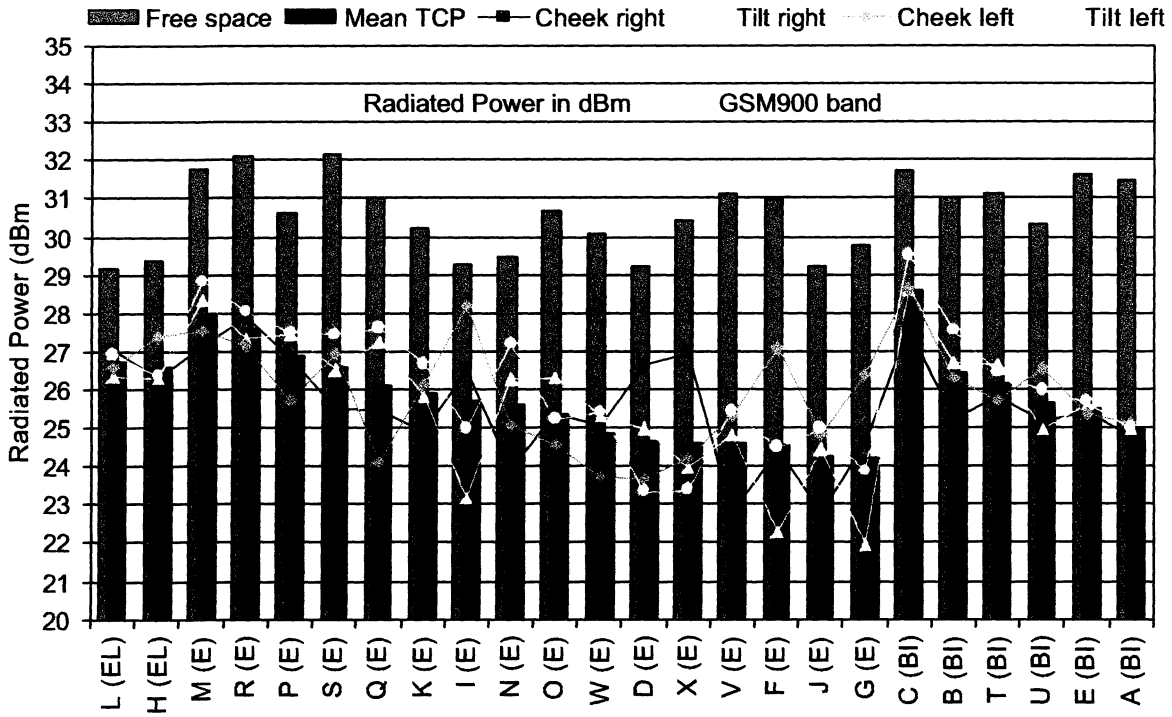


Figure 5. Measured radiated power in Bluetest's reverberation chamber in the transmit GSM 900 MHz band of 20 phones with extractable antenna (EL), external antenna (E) and built-In antenna (BI), for different positions relative to head phantom.

[4] J. G. Kostas and B. Boverie, "Statistical model for a mode-stirred chamber", *IEEE Transactions on Electromagnetic Compatibility*, Vol. 33, No. 4, pp. 366-370, Nov. 1991.

[5] D. A. Hill, M. T. Ma, A. R. Ondrejka, B. F. Riddle, M. L. Crawford and R. T. Johnk, "Aperture excitation of electrically large, lossy cavities", *IEEE Transactions on Electromagnetic Compatibility*, vol. 36, no. 3, pp. 169-178, August 1994.

[6] D. A. Hill, "Electronic mode stirring for reverberation chambers", *IEEE Transactions on Electromagnetic Compatibility*, vol. 36, no. 4, pp. 294-299, November 1994.

[7] D. A. Hill, "Plane wave integral representation for fields in reverberation chambers", *IEEE Transactions on Electromagnetic Compatibility*, vol. 40, no. 3, pp. 209-216, Aug. 1998.

[8] D. A. Hill, "Linear dipole response in a reverberation chamber", *IEEE Transactions on Electromagnetic Compatibility*, vol. 41, no. 4, pp. 365-368, November 1999.

[9] K. Rosengren and P-S. Kildal. " Study of distributions of modes and plane waves in reverberation chambers for characterization of antennas in multipath environment", *Microwave and Optical Technology Letters*, Vol. 30, No 20, pp. 386-391, Sept. 2001.

[10] K. Rosengren, P-S. Kildal, J. Carlsson, O. Lundén. "Measurement of terminal antennas performance in multimode reverberation chambers", *Antenn00, Nordic Antenna Symposium*, Lund, Sweden. 12-14 Sep. 2000.

[11] K. Rosengren, P-S. Kildal, J. Carlsson, O. Lundén "A new method to measure radiation efficiency of terminal antennas", *2000 IEEE AP-S Conference on Antennas and Propagation for Wireless Communication*, Nov. 6-8, 2000. Westin Hotel Waltham, Massachusetts

[12] K. Rosengren, P-S. Kildal, C. Carlsson and J. Carlsson, "Characterization of terminal antennas in reverberation chambers: Improved accuracy by platform stirring", *Microwave and Optical*

Technology Letters, Vol. 30, No 20, pp. 391-397, Sept. 2001.

[13] P-S. Kildal, C. Carlsson, "Detection of a polarization imbalance in reverberation chambers and how to remove it when measuring antenna efficiencies", to appear in *Microwave and Optical Technology Letters*, July 5, 2001.

[14] P-S. Kildal, J. Yang and C. Carlsson, "Measurement of free space impedances of small antennas in reverberation chambers", *Microwave and Optical Technology Letters, Vol. 32, No. 2, pp 112-115,* Jan. 20, 2002.

[15] Jian Yang, Jan Carlsson, Per-Simon Kildal and Charlie Carlsson, "Calculation of self-impedance and radiation efficiency of a dipole near a lossy cylinder with arbitrary cross section by using the moment method and a spectrum of two-dimensional solutions", *Microwave and Optical Technology Letters, Vol. 32, No. 2, pp 108-112,* Jan. 20, 2002.

[16] Per-Simon Kildal, Kent Rosengren, Joonho Byun and Juhyung Lee, "Definition of effective diversity gain and how to measure it in a reverberation chamber", to appear in *Microwave and Optical Technology Letters*, July 20, 2001.

[17] Nikolay Serafimov, *Radiation efficiencies of phone antennas and radiated power of mobile phones measured in two anechoic chambers and two reverberation chambers*, Master thesis at Department of Electromagnetics, Chalmers University of Technology (supported by Sony Ericsson Mobile Communications AB), Jan. 2002.

[18] Angivande av mätosäkerhet vid kalibrering (Examination of measurement uncertainty in calibration), Publication EAL-R2-Sv, European cooperation for Accreditation of Laboratories, Swedish translation, Jan. 1999.

[19] A method and an apparatus for measuring the performance of antennas, mobile phones and other wireless terminals, International patent application No PCT/SE01/00422, filed February 26, 2001, priority March 31, 2000.

20 TCO'01 Certification of Mobile Phones; Requirements and test methods for quality and environmental labelling, TCO Development, SE-114 94 Stockholm, Sweden, Nov. 12, 2001

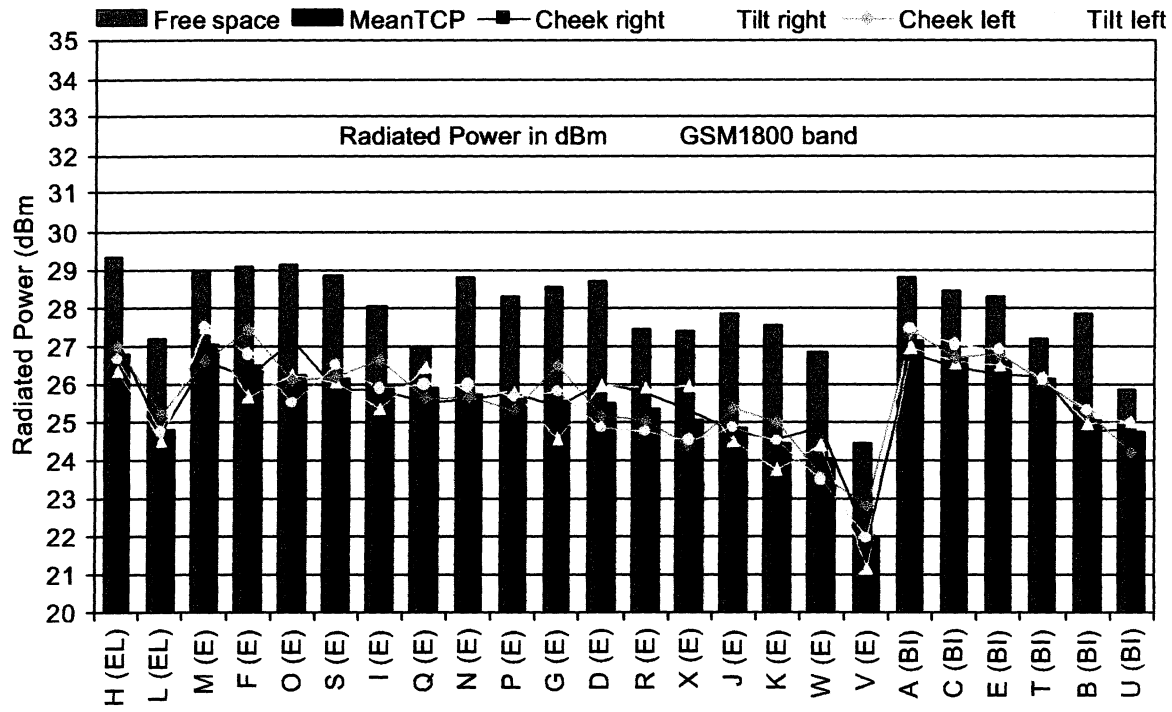


Figure 6. Measured radiated power in Bluetest's reverberation chamber in the transmit GSM 1800 MHz band of 20 phones with extractable antenna (EL), external antenna (E) and built-in antenna (BI), for different positions relative to head phantom.

Development of reverberation chamber for accurate measurements of mobile phones and mobile phone antennas

Charlie Carlsson[1], Per-Simon Kildal[2], Kent Rosengren[3] and Jan Carlsson[4]

[1]Bluetest AB, Göteborg. E-mail: charlie.carlsson@bluetest.se, http://www.bluetest.se

[2]Chalmers, Göteborg. E-mail: simon@elmagn.chalmers.se

[3]Intenna Technology AB, Kalmar. E-mail: kent.rosengren@intenna.com

[4]SP, Borås. E-mail: jan.carlsson@sp.se

Abstract

This paper deals with the development and validation of small reverberation chambers for characterization of small antennas. The reverberation chamber supports a statistical multipath environment, with a uniform and isotropic distribution of the electric field. The chamber can be utilized to measure the radiation efficiency and input impedance of antennas, as well as the total radiated power of wireless devices. Also diversity gain and effective diversity gain can be measured in the reverberation chamber.

1 Introduction

The reverberation chamber, or mode-stirred chamber, is well-known and widely used in the area of electromagnetic compatibility (EMC) testing to create a statistically uniform and isotropic field distribution. During recent years the reverberation chamber has gained a lot in interest. The basic theories are well investigated and described in journal articles [1, 2, 3, 4].

The antenna group at Chalmers University of Technology together with the spin-off company Bluetest AB have been dealing with the issue of measuring the performance of antennas in a relatively small reverberation chamber, with smallest dimension approximately three wavelengths at the lowest frequency. Several improvements have been introduced to get the desired accuracy. These improvements are described below.

2 Measurement setups in reverberation chamber

2.1 Radiation efficiency

The total received, or equivalently the transmitted, power by an antenna in a uniform and isotropic field is proportional to the radiation efficiency of the antenna. We use here the radiation efficiency that includes impedance mismatch, losses in the antenna itself and losses in the near-in environment such as a head phantom (defined in [5]).

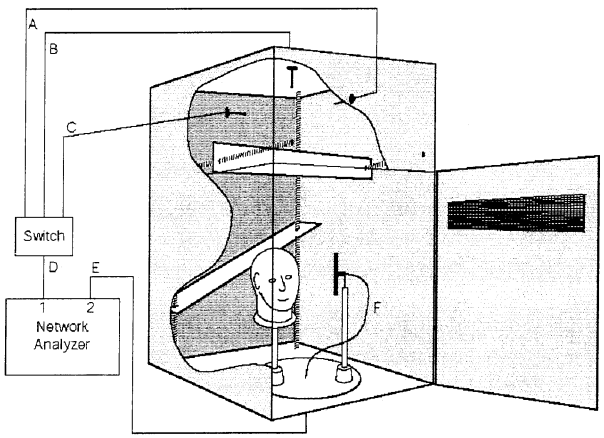


Figure 1: Setup in chamber for measurement of radiation efficiency, input impedance and possibly diversity gain.

Radiation efficiency is possible to measure for antennas with a cable connection to the antenna port, see figure 1. This setup is treated as a two-port, i.e. all S-parameters are measured. Based on several measurements of the S-parameters, we calculate the mean

power transfer function P_{21} of the chamber from [6]

$$P_{21} = \frac{\overline{|S_{21}|^2}}{(1-|\overline{S_{11}}|^2)(1-|\overline{S_{22}}|^2)} \quad (1)$$

where $\overline{|S_{21}|^2}$ is the mean of all power transmission samples and $\overline{S_{11}}$ and $\overline{S_{22}}$ are the free space reflections of the fixed and test antennas, respectively, obtained as described in the next section. The radiation efficiency of the antenna under test (AUT) is then given by

$$e_{rad} = \frac{P_{21,AUT}}{P_{21,ref}}(1-|\overline{S_{22}}|^2) \quad (2)$$

where $P_{21,AUT}$ is the mean power transfer function for the test antenna, $P_{21,ref}$ the mean power transfer function for the reference antenna and $\overline{S_{22}}$ is the free space reflection coefficient of the AUT as described in the next section. $P_{21,ref}$ characterizes the chamber properties and is obtained by measuring the relative received power of a reference antenna with unit radiation efficiency.

2.2 Free-space impedance

From the setup used above for the radiation efficiency (figure 1), also the reflection coefficient of the AUT can be extracted. The reflection from an antenna inside the chamber will consist of two contributions; one deterministic from the antenna itself and the near-in environment, and the other from reflections inside the chamber. The latter will be normally distributed in the complex plane due to the chamber statistics, and can therefore be averaged out by taking the mean of the complex reflections obtained during a measurement sequence. That is, the reflection coefficient of the antenna under test located in free space, but in the possible close vicinity of some kind of lossy object, will then be:

$$\overline{S_{11}} = \frac{1}{N}\sum_{i=1}^{N} S_{11}^{(i)} \quad (3)$$

where N is the number of samples and $S_{11}^{(i)}$ is the complex reflection sample number i.

2.3 Diversity gain

The controlled multipath environment in the reverberation chamber makes it possible to measure diversity gain in a repeatable way. One big advantage of using the reverberation chamber instead of conventional methods is that the so called effective diversity gain can be measured with high accuracy. This is all described in [7].

The setup and instrumentation used for diversity measurements is the same as in figure 1, except that two test antennas are connected and possibly a three-port network analyzer is used.

2.4 Radiated power

The setup for measuring the radiated power from active devices such as mobile phones, Blutooth units etc. is shown in figure 2. A complete description of this setup is found in [8].

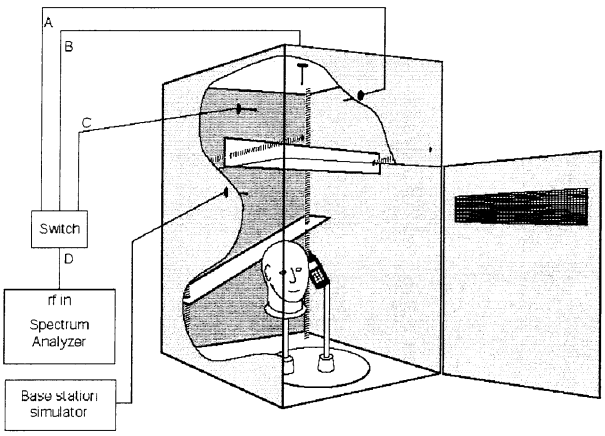


Figure 2: Setup in chamber for measurement of radiated power from wireless devices.

3 Improvements

Three major improvements of the reverberation chamber have been done in order to get the desired accuracy.

3.1 Chamber size

An important parameter for a good measurement accuracy in the reverberation chamber is the mode density $\partial M/\partial f$, since that is one of the parameters that determines the number of modes that are excited at a particular frequency. The total number of modes M_{tot} excited is also proportional to the mode bandwidth ΔF, with the relation

$$M_{tot} = \frac{\partial M}{\partial f}\Delta f \quad (4)$$

The mode density can be estimated from the derivative of Weyl's formula [9]

$$\frac{\partial M}{\partial f} = \frac{8\pi}{c^3}LWHf^2 \quad (5)$$

where the dependence to the chamber dimensions (L, W, H) are very clear. For chambers with a low mode density the frequency variation will not be as smooth as predicted by eq.(5), thus a more thorough investigation of the resonant modes is needed. In [9], the method to describe each mode with a number of plane

waves is used to get a more accurate figure of the mode density. This method is used to estimate the number of modes within a 25 MHz wide band for a chamber with the height dimension varied from 0.8 m to 2.0 m. Figure 3 shows the number of modes versus center frequency for the 25 MHz band. Other dimensions of the chamber are width 0.8 m and depth (length) 1.0 m.

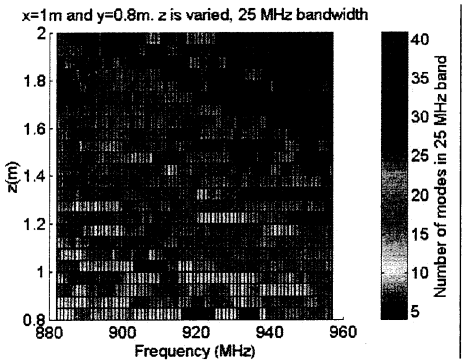


Figure 3: Calculated number of modes within 25 MHz bandwidth for chambers with varying height.

As can be seen in figure 3, the number of modes for the original size of chamber ($1.0x0.8x1.0m^3$) is very low for the GSM900 transmit band, a band with a particular interest for us at that moment. This 'mode-hole' was the reason for a poor accuracy for some measurements within that frequency band. By extending the chamber height to 1.6 m, and keeping the other dimensions, the number of modes was higher and also more smoothly distributed for the whole frequency range shown in figure 3.

3.2 Platform stirring

Figure 4 (left side) shows the measured radiation efficiency for a dipole close to a lossy cylinder in the frequency range where the dipole is tuned, i.e. the radiation efficiency is supposed to have a smooth variation with frequency. We can see that with a small amount of frequency stirring the variation is very large over the band. This inaccuracy in the measured radiation efficiency is due to a poor number of independent samples, and the solution to this was to introduce the platform stirring, i.e. the AUT is placed on a rotatable platform.

With the platform stirring introduced, the results in figure 4 (right side) for the radiation efficiency of the dipole were obtained. We can see a more smoother variation with frequency than previous, and there is no longer need for a very large amount of frequency stirring. This is a result of the increase of independent samples that the platform stirring gives, which is

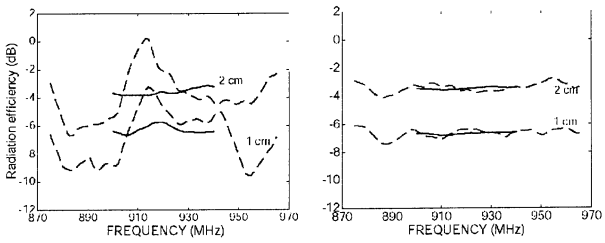


Figure 4: Measured radiation efficiency for a dipole close to a lossy cylinder without (*left*) and with (*right*) platform stirring. For the dashed lines 10 MHz frequency stirring is used and for the complete lines 60 MHz.

shown in [10] with correlation analysis. The platform stirring is therefore a powerful method that makes it possible to use the reverberation chamber for measurements even when the number of modes present are low.

3.3 Polarization stirring

When measuring the chamber transfer function, i.e. the relative received power of an ideal antenna, we were able to detect a polarization imbalance in the magnitude of several dBs. By polarization imbalance we mean the difference in power transfer function for two orthogonally polarized antennas, in this case a dipole mounted vertically and horizontally, respectively. The polarization imbalance was found in both a small reverberation chamber of 1.3 m^3 volume (the Bluetest chamber) and a large chamber of 37 m^3 volume. Figure 5 shows the measured chamber transfer function for the Bluetest chamber with a dipole as AUT and monopoles mounted on the wall as fixed antennas.

The reason for the polarization imbalance is that certain polarizations of the AUT and fixed antenna gives a preference to which modes are excited, i.e. when a single fixed antenna is used, the polarization of that antenna will determine the isotropy in the chamber. One solution to the problem is to use several fixed antennas with different polarizations. Our choice is to use three monopoles, mounted orthogonally on three different walls. This is introduced as polarization stirring. The result for the chamber transfer function with utilization of the polarization stirring is shown in the lower right part of figure 5. We can see that the polarization imbalance is removed.

A simple propagation model can predict the behavior of polarization imbalances in reverberation chambers. This is described in [11], together with more details about the polarization imbalance investigations within the research project.

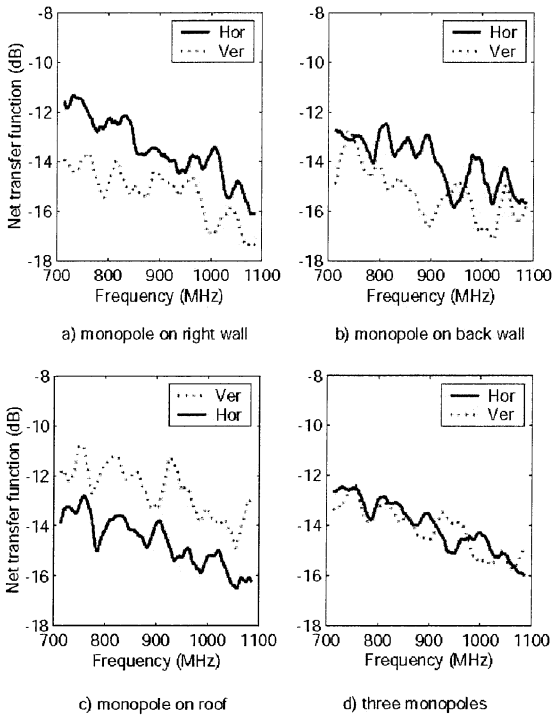


Figure 5: Measured chamber transfer function with separate monopoles used as fixed antennas.

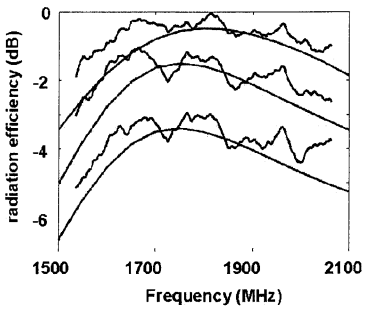


Figure 6: Measured and calculated radiation efficiency for a dipole at 55 mm, 20 mm and 10 mm distance from a lossy cylinder, respectively.

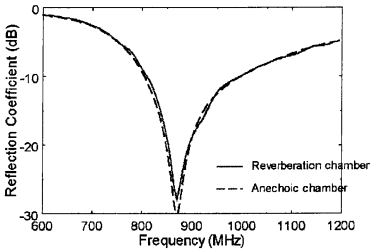


Figure 7: Measurements of reflection coefficient of a dipole at 10 mm distance from a lossy cylinder, conducted in reverberation and anechoic chambers.

4 Validations

A lot of effort has been put into validating the reverberation chamber measurement technique against other methods and numerical simulations of the measurement cases. A special test case with a dipole close to a cylinder is the main validation case for antenna measurements since the origin of the research project. This case have been chosen because it is easy to model for numerical simulations.

Figure 6 shows examples of measured and calculated radiation efficiency from [12] with a dipole placed at different distances from a lossy PVC-cylinder. The simulated results are calculated with an FDTD-code and the measurements performed in the Bluetest reverberation chamber. The agreement between measured and calculated values can be regarded as very good.

Results from reflection coefficient measurements have been reported in [13], where comparisons between reverberation chamber and anechoic chamber measurements, as well as numerical simulations, are presented. An example of these results are shown in figure 7, in which the reflection coefficient of the dipole at a 10 mm distance from the lossy PVC-cylinder is shown.

Validations of various measurement ranges, including the reverberation chamber, for radiated power measurements has been conducted in [14]. The statistical variations of some of the radiated power measurements are shown in table 1.

An analysis of measurements of active devices in two different reverberation chambers is also given in [8].

Comparison chamber	**GSM900**		**GSM1800**	
	Std.	Max.	Std.	Max.
Anechoic	0.57	1.00	0.87	1.56
Reverberation	0.62	1.10	0.81	-1.55

Table 1: Statistical deviations of measurements of active mobile phones in the Bluetest reverberation chamber to measurements in one anechoic and one other reverberation chamber. Data taken from [14].

5 Conclusion

The conclusion of the work we have done with antenna measurements in the reverberation chambers so far, is that we have shown that the reverberation chamber can be used for measuring performance of

antennas and active wireless devices. To be able to use small reverberation chambers with high accuracy, we have introduced three major improvements. These improvements include two new stirring methods, polarization and platform stirring, and changed physical dimensions of the Bluetest chamber.

Altogether the performed validations show that the reverberation chamber can be used with good accuracy for measuring the radiation efficiency and reflection coefficient of small antennas, as well as the radiated power for active units.

References

[1] J. G. Kostas and B. Boverie, *Statistical model for a mode-stirred chamber*, IEEE Transactions on Electromagnetic Compatibility, vol. 33, no. 4, pp 366-370, 1991.

[2] D.A. Hill et al, *Aperture excitation of electrically large, lossy cavities*, IEEE Transactions on Electromagnetic Compatibility, vol. 36, no. 3, pp 169-178, August 1994.

[3] D. A. Hill, *Plane wave integral representation for fields in reverberation chambers*, IEEE Transactions on Electromagnetic Compatibility, vol. 40, no. 3, pp. 209-216, Aug. 1998.

[4] D. A. Hill, *Linear dipole response in a reverberation chamber*, IEEE Transactions on Electromagnetic Compatibility, vol. 41, no. 4, pp. 365-368, November 1999.

[5] P.S. Kildal, *Foundations of Antennas - A Unified Approach*, Studentlitteratur, 2000.

[6] C. Carlsson, *Characterization of mode-stirred chamber and its set-ups for antenna measurements*, Master thesis at Department of Electromagnetics, Chalmers Univ. of Technology, Aug 2001.

[7] P-S. Kildal, K. Rosengren, J. Byun and J. Lee, *Definition of effective diversity gain and how to measure it in a reverberation chamber*, to appear in Microwave and Optical Technology Letters, July 2002.

[8] P-S. Kildal, C. Carlsson, *TCP of 20 phones measured in reverberation chamber*, Bluetest AB, Feb 2002.

[9] K. Rosengren, *Measuring radiation efficiency of terminal antennas in reverberation chambers*, Thesis of Licentiate degree, Technical Report No 388L, Chalmers University of Technology, Gothenburg, April 2001.

[10] K. Rosengren, P-S. Kildal, C. Carlsson, J. Carlsson, *Characterization of antennas for mobile and wireless terminals in reverberation chambers: Improved accuracy by platform stirring.*, Microwave and Optical Technology Letters, Vol. 30. No. 20, pp. 391-397, Sep. 2001.

[11] P-S. Kildal, C. Carlsson, *Detection of a polarization imbalance in reverberation chambers and how to remove it when measuring antenna efficiencies*, to appear in Microwave and Optical Technology Letters, July 2002.

[12] J. Carlsson, A. Wolfgang, P-S. Kildal, *Numerical Simulations of a Validation Case for Small Antenna Measurements in a Reverberation Chamber*, accepted for presentation at IEEE International Symposium on Antennas and Propagation, June 2002.

[13] J. Yang, J. Carlsson, P-S. Kildal, C. Carlsson, *Calculation of self impedance and radiation efficiency of a dipole near a lossy cylinder with arbitrary cross section by using the moment method and a spectrum of two-dimensional solutions*, Microwave and Optical Technology Letters, Vol. 32, No. 2, pp. 108-112, Jan. 2002.

[14] N. Serafimov, *Radiation efficiencies of phone antennas and radiated power of mobile phones measured in two anechoic chambers and two reverberation chambers*, Master thesis at Department of Electromagnetics, Chalmers University of Technology, Jan. 2002. With addendum: Appendix G, Feb. 2002.

Measured magnetic field distribution over a printed patch antenna compared to numerical calculations and transformation to far field distribution

Urban Lundgren†, Jonas Ekman† and Stewart Jenvey*

† EISlab, Systemteknik, Luleå Tekniska Universitet, 971 87 Luleå, Sweden

***Department of Electrical and Computer Systems Engineering,**
Monash University, Wellington Rd, Clayton, Australia 3168

Abstract
This paper describes the use of a position scanned magnetic field probe to investigate the current distributions on the surface of a probe fed patch antenna. The antenna was a probe fed patch antenna with dimensions 99 x 61 mm on a 1.6 mm thick FR4 substrate. The first resonance of the antenna was found at 1.155 GHz and the scan was performed at that resonant frequency. Measurements of the magnetic field magnitude and phase at each point of the scan were used to derive the time variant instantaneous values of the magnetic field and the instantaneous currents on the antenna.

Introduction
A patch antenna is a microstrip transmission line section that usually offers a narrow frequency band where it is in resonance. The design of the patch can vary in a wide range depending on desired far field distribution and frequency band width. To find a good design is a task that often involves some trial and error because of the narrow band characteristics of the antenna type and possible variations in material properties and tolerances in manufacturing. A near field scanning technique is used to evaluate a simple prototype. In this paper the most simple patch design is used, a rectangular patch with a probe feed. The background on patch antenna design is obtained from [1] and [2].

Patch design and model
For this work a prototype patch antenna was used that was manufactured on a FR4 laminate board. The dimension of the patch (Figure 1) is 99 x 61 mm and the FR4 substrate is 1.6 mm thick with 35 µm thick copper layers on each side. A probe feed is used to supply excitation to the patch. The feed point is positioned 15 mm from the edge in the y-direction (E-axis) and centred in the x-direction (H-axis).

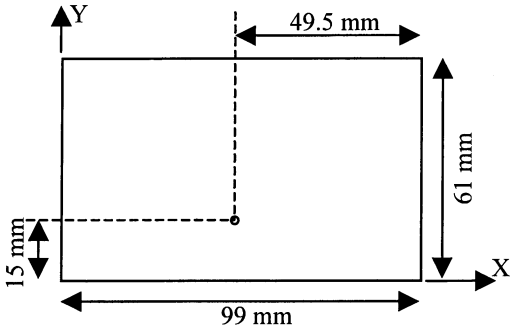


Figure 1 Patch dimensions and coordinate axes

A simple software tool was used for the patch design [2]. The chosen design has the following predicted characteristics:

SUBSTRATE HEIGHT = 0.1600 cm
SUBSTRATE RELATIVE DIELECTRIC CONSTANT = 4.50
SUBSTRATE LOSS TANGENT = 0.0100
CONDUCTOR RELATIVE CONDUCTIVITY = 1.000
PATCH LENGTH = 6.033 cm
PATCH WIDTH = 9.900 cm
FEED LOCATION = 1.500 cm
FREQUENCY = 1.1546 GHz

INPUT RESISTANCE = 47.50 Ohms
PATCH TOTAL Q = 50.813
EFFICIENCY = 94.88%
OVERALL EFFICIENCY = 40.81%
PATCH BANDWIDTH = 1.39%
FOR A 2.00:1 SWR

After reflection loss measurement the first resonance of the prototype is found at 1.155 GHz.

Numerical calculation
The patch was modelled in a software tool called PCB-MOM [3] offering a Method of Moments solver for single layer structures with an infinite ground plane as an optional second layer, the dielectric constant for the substrate in a PCB can be accounted for. The current distribution obtained can be used in post processing for predicting near field and far field radiation. Using this software the surface current distribution and the far field radiation were obtained.

Another software tool was used as a source for predicted far field radiation pattern data. The tool is called HWPATCH [2] and is a simple code that assumes a surface current distribution and reciprocity to find a far field solution. This is used as an extra source of data for comparison.

Measurements
The measurements of the current distribution was made with a technique were a coordinate table (Figure 3) was used together with a vector network analyzer. A small magnetic field probe was positioned close above the studied patch antenna and moved stepwise in a rectangular grid pattern. The center of the loop probe was about 5 mm above the surface of the antenna.

The probe used was a 1 cm diameter loop antenna shown in Figure 2, with the polarization axis normal to the plane of the loop, positioned in parallel with the plane of the antenna. The PC software for controlling the coordinate table was used with modified GPIB instruction files to enable automatic control of the vector network analyzer. This arrangement made the entire measurement procedure automatic. Between each step the magnitude or phase for the transmission from the antenna to the probe was measured. The measurement procedure was repeated to cover x and y polarization of the magnetic field.

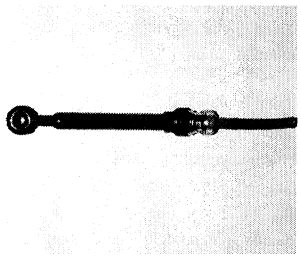


Figure 2 The magnetic field probe used had 1.0 cm diameter and was electrically shielded

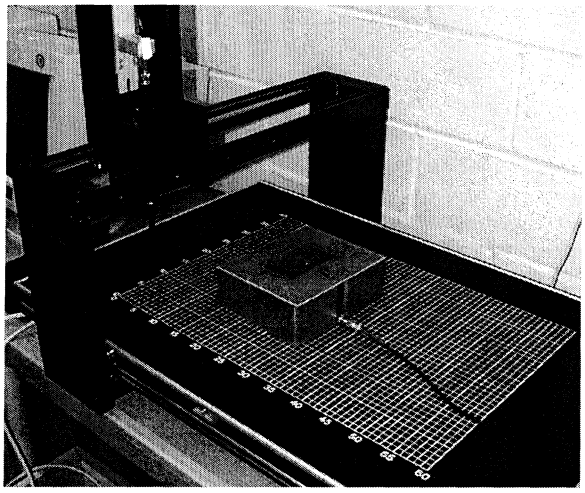


Figure 3 Scanning the patch antenna using the coordinate table

The scan consisted of 957 measurement points in a 33 x 29 point rectangle. The step size in both directions was 5.0 mm. Magnitude and phase of the magnetic field were measurements at each point. Measured magnetic field distributions are compared with current distributions obtained from numerical calculations. The measured data is transformed to predict the far field distribution from the antenna. Measured and predicted far field radiation patterns are compared.

Results

The measured magnetic field data can be used to visualize the function of an antenna (Figure 4) and when using both magnitude and phase data a topological map of the antenna can be time advanced to animate the wave movement on the antenna and around the antenna. If the assumption is made that the measured magnetic field indicates magnitude and phase of the surface currents on the patch then the far field radiation can be predicted also using the same data.

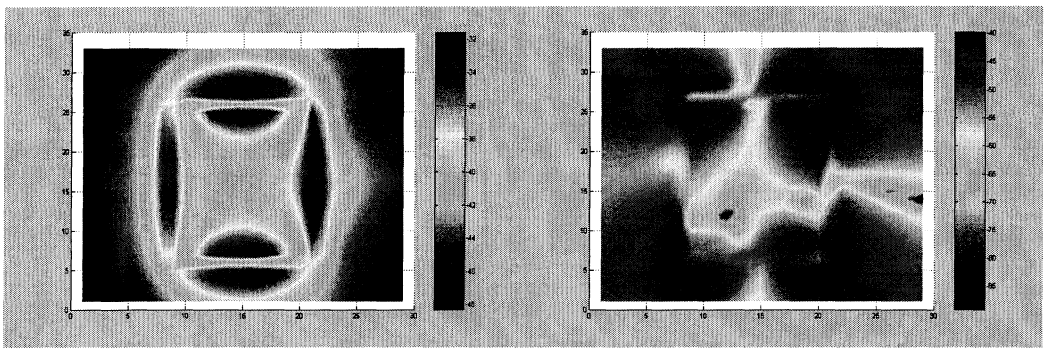


Figure 4 The magnitude-scan of the x-polarized magnetic field to the left and the magnitude-scan of the y-polarized magnetic field to the right

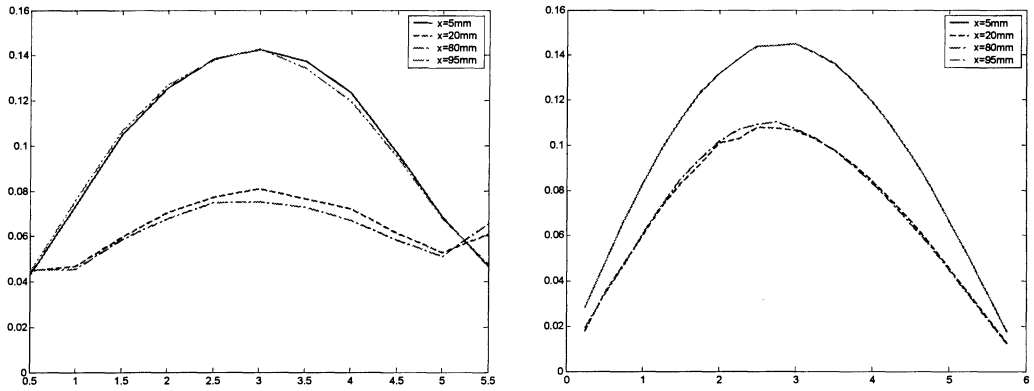


Figure 5 Left: Measured magnetic near field corresponding to currents in y-direction. Right: currents in y-direction from Method of Moments solution. X-coordinates close to edge of patch.

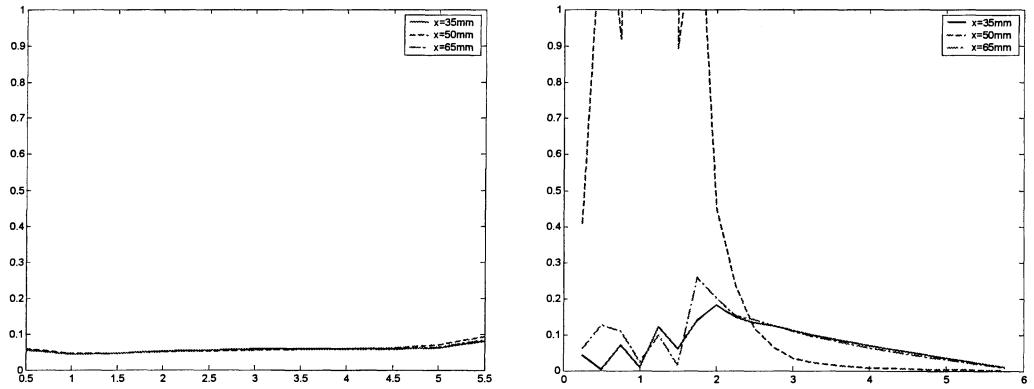


Figure 6 Left: Measured magnetic near field corresponding to currents in y-direction. Right: currents in y-direction from Method of Moments solution. X-coordinates close to centre of patch.

In Figure 5 the measured magnetic field magnitude is compared to the corresponding surface current, values are scaled so that the shape can be compared. For the magnetic field the magnitude seem to be higher closer to the edge of the patch while the current distribution are more even from the edge and 20 mm into the patch. The distribution along the y-axis is quite similar for the magnetic field and the calculated current.

When looking at current distribution closer to the centre of the patch there is no agreement with the measured magnetic field (Figure 6).

The far field radiation diagram have been obtained from four sources, the HWPATCH code, the PCB-MOM code, integration from measured near magnetic field and directly from measured far field radiation. The diagrams below are scaled in order to make the comparison clear. All diagrams are made to agree with the measured far field distribution at 0 degrees. From the measured magnetic near field the assumption is made that the surface current on the patch is known by magnitude and phase so that integration over all measured positions with the integrand:

$$E_\theta = -\frac{I \cdot dl}{4 \cdot \pi} \cdot Z_0 \cdot \beta^2 \cdot \sin(\theta) \cdot \frac{1}{j \cdot \beta \cdot r} \cdot e^{-j \cdot \beta \cdot r}$$

where

$$\beta = \frac{2 \cdot \pi}{\lambda}$$

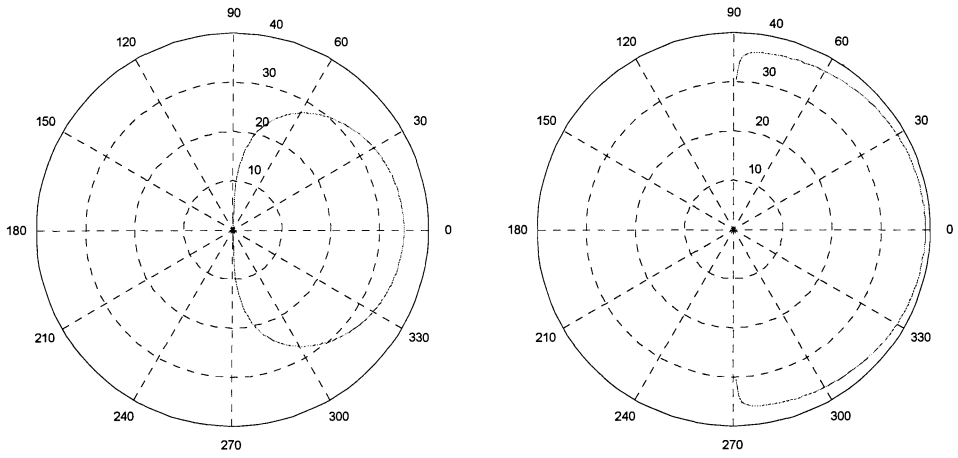


Figure 7 H-plane (left) and E-plane (right) radiation pattern from HWPATCH code [2]

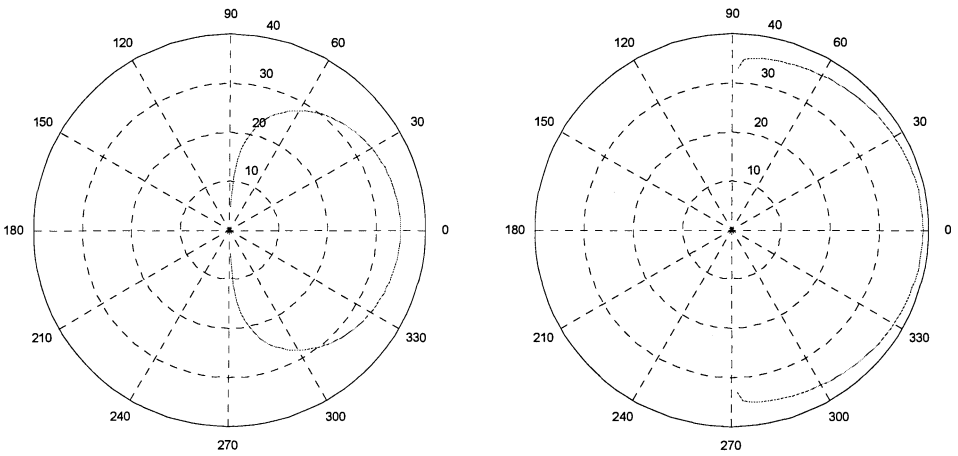


Figure 8 H-plane (left) and E-plane (right) radiation pattern from a Method of Moments code PCB-MOM [3]

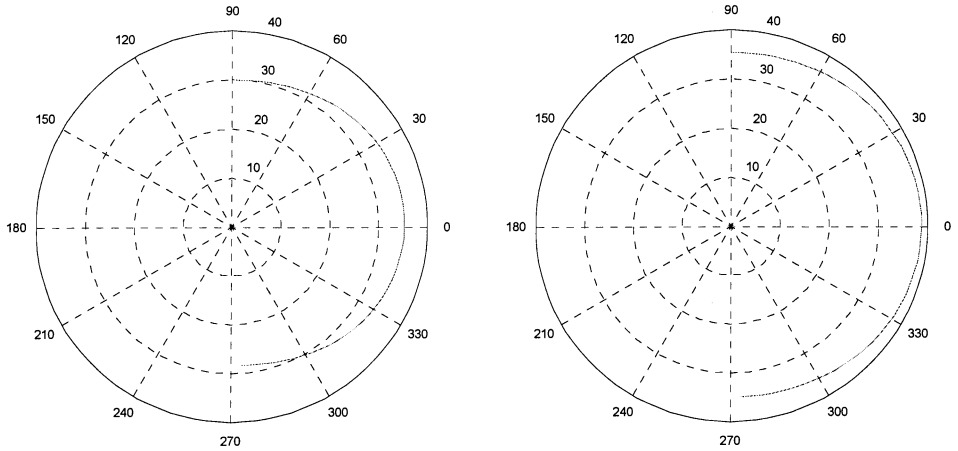


Figure 9 H-plane (left) and E-plane (right) radiation pattern resulting from integration of measured magnetic near field

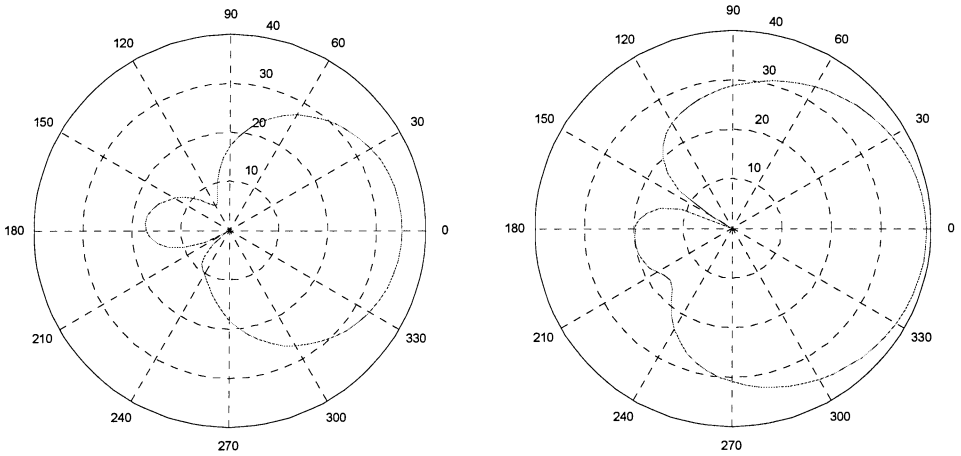


Figure 10 H-plane (left) and E-plane (right) radiation pattern measured at 3 meters distance

Conclusions

The relatively strong surface currents close to the feed probe shown in the numerical simulation can not be seen when measuring the magnetic field above the patch surface. The shape of the current distribution over the patch is indicated by the magnetic field measurements however strong local currents near the feed point can not be seen also along the edges of the patch the measured magnetic field amplitude is probably not only due to surface currents since the measured amplitudes here are stronger than expected from simulation results. When comparing radiation patterns from numerical calculations a very good agreement is found. The agreement with measured far field distribution is quite good for the H-plane, a small deviation is found for E-plane comparison. For the measured magnetic near field integration to estimate far field radiation, the agreement with simulations is good for E-plane but not as good for H-field. The ground plane on the prototype patch antenna was 20 cm x 20 cm, in both software tools the ground plane is modelled as infinite, which affects the radiation pattern at angles close to -90° and 90°. The near field scanning technique is a useful complement to numerical calculations when designing a patch antenna and evaluating a patch antenna design. The technique is likely to be used in further work regarding patch antenna design.

References

1. Pozar D., Schaubert D., "Microstrip Antennas, The Analysis and Design of Microstrip Antennas and Arrays", IEEE PRESS, 1995, Piscataway, New Jersey, USA, ISBN 0-7803-1078-0
2. Sainati R., "CAD of Microstrip Antennas for Wireless Applications", Artech House Publishers, 1996, Boston, USA, ISBN 0-89006-562-4
3. Carlsson, J., "Computation of EMC Properties of Slots and Printed Circuit Boards", Chalmers University of Technology, 1998, Gothenburg, Sweden, ISBN 91-7197-680-9

Measuring fast electromagnetic transients

Åke Wisten[1] and Per Mäkikaltio[2]

[1]EISLAB. Luleå University of Technology. E-mail: Ake.Wisten@sm.luth.se
[2]EISLAB. Luleå University of Technology. E-mail: Per.Makikaltio@luth.se

Abstract

Measuring radiated fast electromagnetic transients, occurring unpredictedly may give very different results depending on the equipment used. In this report electromagnetic transients from a passing train is measured in three different ways. Comparison was made between a measuring receiver, a spectrum analyzer and a digital oscilloscope. It was verified by laboratory experiments on electric discharges, that a measuring receiver is almost useless, a spectrum analyzer, will probably give misleading results, but the only instrument, that can be used for measuring electromagnetic transients is an oscilloscope. Still with an oscilloscope oscillations in the receiving system can give misleading results, unless great care is taken. For the oscilloscope though, the size of the memory and the upper frequency limit is crucial for the result obtained. In a situation where the frequency contents is in principle unlimited, like the electromagnetic radiation from an electric discharge, the outcome of measurements will be strongly dependent of economical resources: A memory size of 50K, used at 500 MHz, and a sampling rate of 5Gs/s, will last for only 10 us. Measuring a 1 GHz transient, at 10 GS/s for a time interval of 200 us takes a oscilloscope with 16 MB memory, where only the cost of the memory today is about 10000 SEK per MB.

1. Measuring transient electromagnetic radiation from trains.

Since the sources are transmitting a mixture of harmonics and transients the frequency spectra are changing partly at random when measurements are repeated. *Figure 1 – 5.*

2. Measuring transient electromagnetic radiation under controlled lab conditions.

Well defined electrostatic discharges, one per second, were generated in a shielded chamber. The receiving system consisted of a bilog antenna Chase 6112 connected to a) R&S ESPC measuring receiver in peak detector mode, sweeping 30 -1000 MHz, step size 50 kHz, bandwidth 120 kHz, measuring time 1 ms/step, Total sweep time for one sweep 6 min 25 sec. *Figure 6.* b) HP8591E spectrum analyzer. Sweep time 0.2 s. 100 sweeps accumulated takes in total 30.2 s. *Figure 7-8.* c) Digital oscilloscope, Tektronix TDS 724 A, 500 MHz, 1GS/s. *Figure 9.*

2 Conclusions

Measuring electromagnetic transients give different results for repeated measurements, when a measuring receiver or a spectrum analyzer is used. One reason, apart from variations of the source, is that both the measuring receiver and the spectrum analyzer take considerable time to sweep the whole frequency span, and can only pick up harmonic components if they occur when the frequency window is open. An oscilloscope, on the other hand, gives more reliable results, provided the trigger level is exceeded. But there may be unwanted oscillations in the receiving system due to mismatched antennas and cables. One way of keeping a check on oscillations is to use a four channel oscilloscope, with different antennas connected to at least two of the input channels, and also to connect a leaky cable with a matched 50 ohm load, to one of the input channels. If the 50 ohm load is well matched the signal picked up by this cable can be used as a reference.

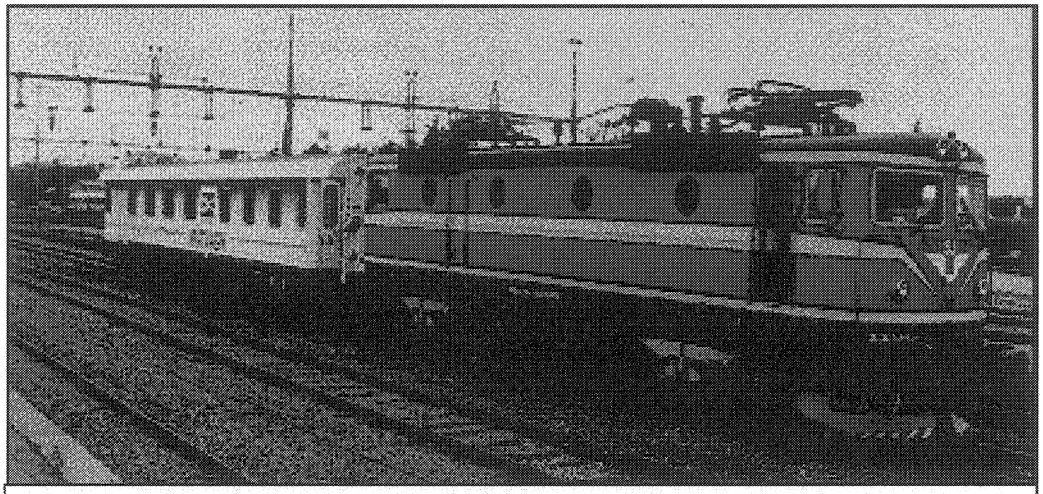


Figure 1. RC engine with measuring car. Antenna location at the front end of the car, close to the engine. In some cases the antenna was located at the rear end of the first car.

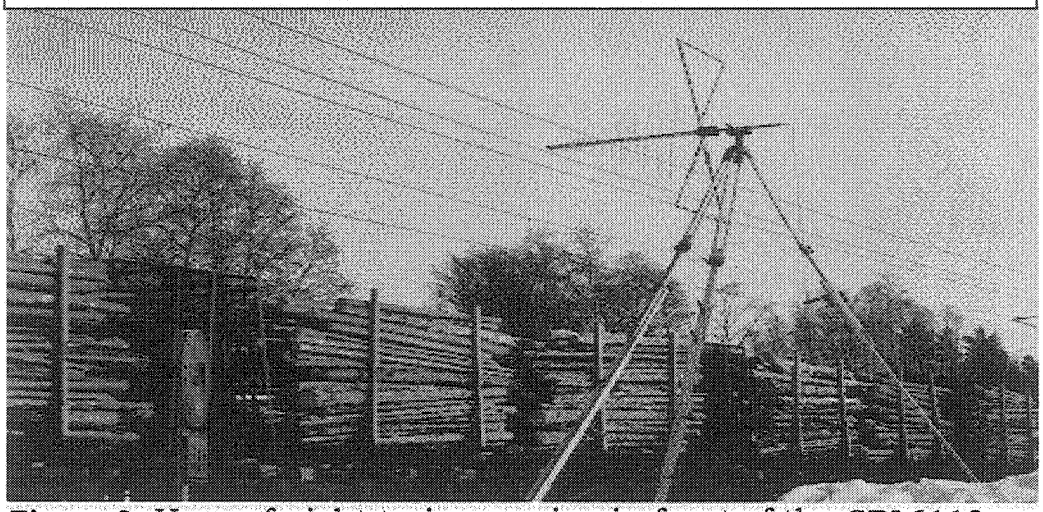


Figure 2. Heavy freight train, passing in front of the CBL6112 log periodic antenna in position for measuring the vertical electric field according to ENV 50121, at a distance of 10 m from the centre of the track and at a fixed height of about 3 m. Älvsbyn.

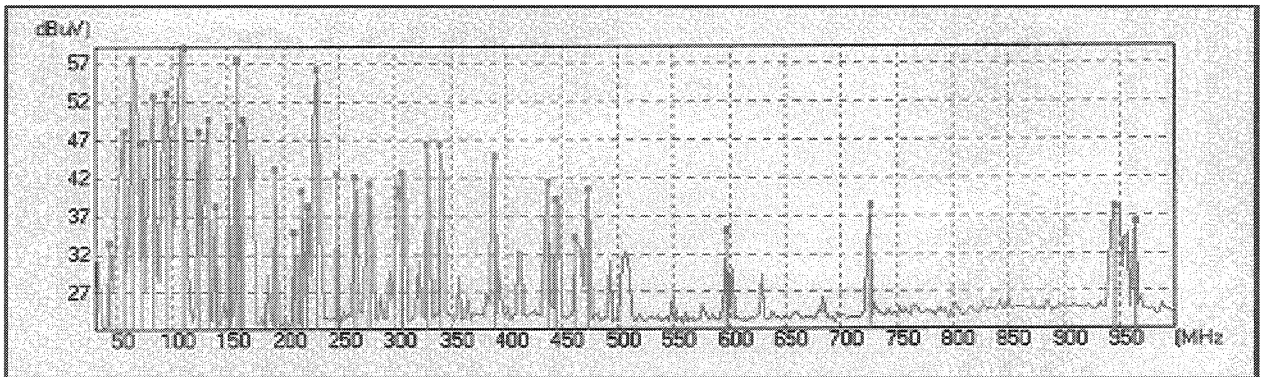


Figure 3. Vertical E field measured according to ENV 50121. Spectrum analyzer HP8195E and CBL 6112 antenna. Heavy freight train (timber) with engine RC 2 no 1128. Motor current 1700 A. Bölebyn 42 km S Älvsbyn. Repeated measurements gave other peaks, apparently distributed at random. In this diagram peaks above 550 MHz are due to radio transmitters.

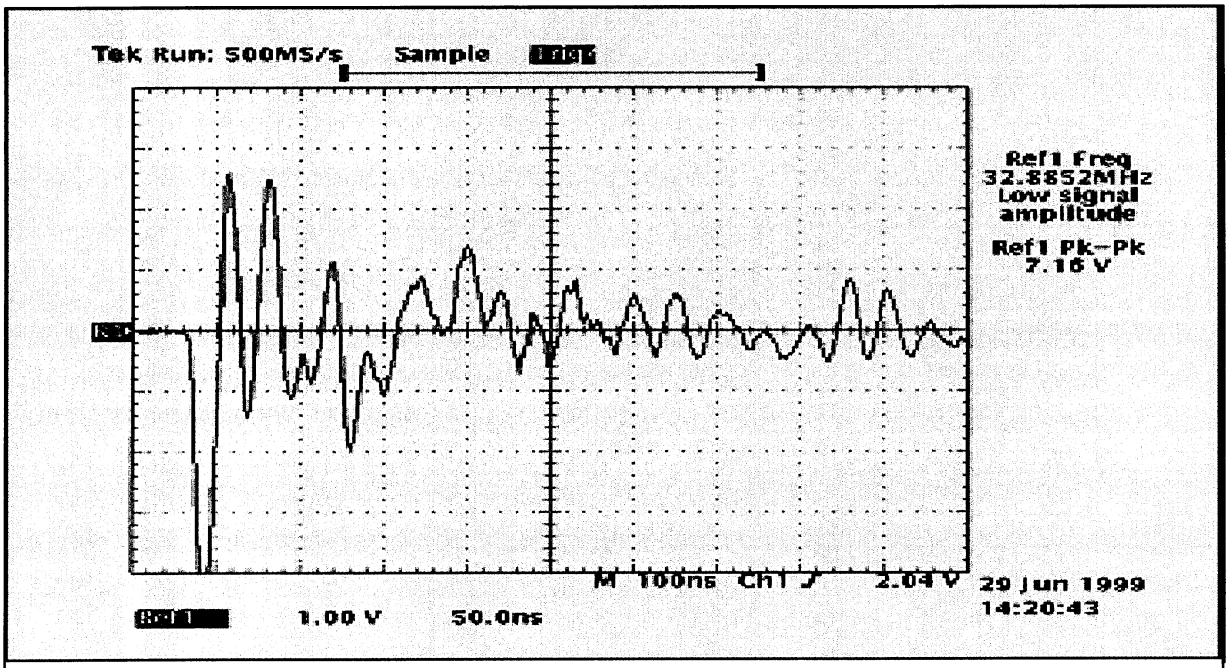


Figure 4. Transients measured on board the train, passing security zone. One single pulse. Big yagi, front end of first car. Freq 22 MHz, Trigg level 940 mV

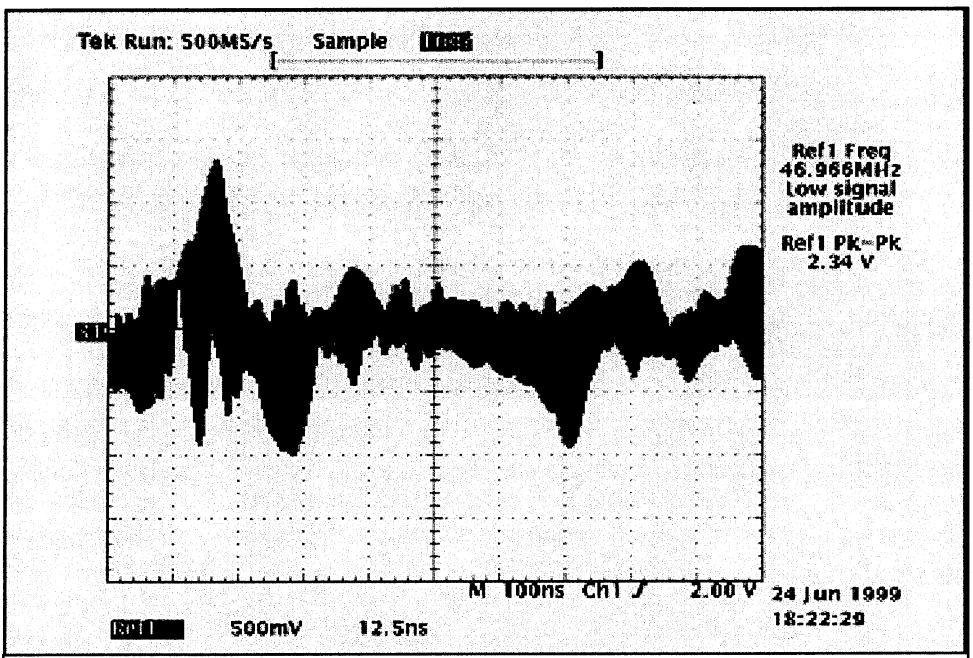


Figure 5. Horizontal electric transient pulses measured on board train. Big yagi antenna. Trigger level 400 mV. Time 1 min. Number of Pulses: 8. Constant speed. The higher trigger level, the more rare and the more equal transients.

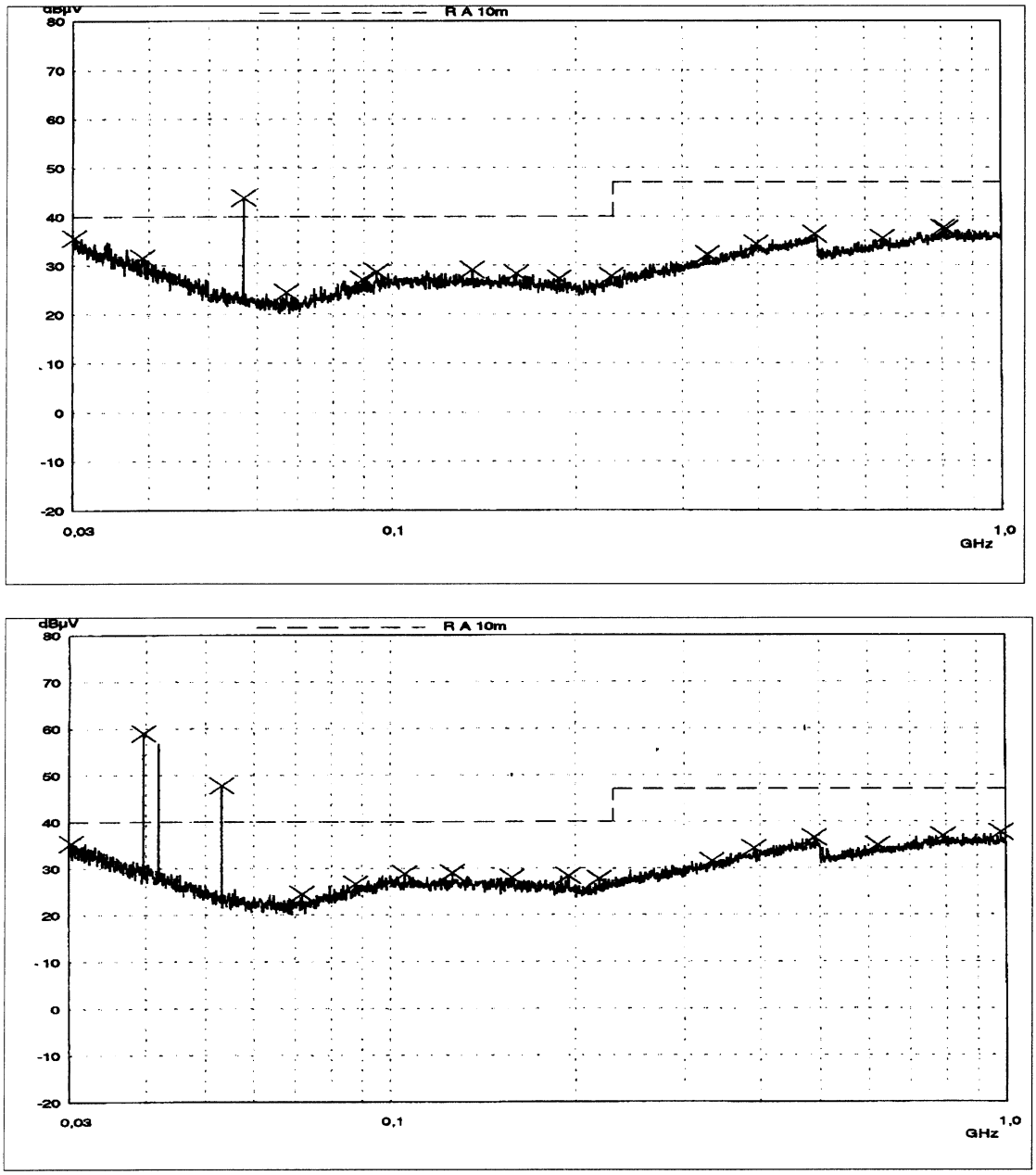


Figure 6. (Two graphs). Measuring transients under lab conditions, in a shielded chamber. One ESD pulse was transmitted every second. The receiving system consisted of a R&S ESPC measuring receiver connected to a bilog Chase 6112 antenna. One sweep 30-1000 MHz, step size 50 kHz, bandwidth 120 kHz, peak detector, 1 ms/step. Total measuring time 6 min 25 sec. In first recording only one peak was detected. In the next recording three peaks were detected, but none at the same frequency as in the first recording. In the third recording no peak at all was detected (Not shown in this report). The conclusion is that a measuring receiver in not suitable for measuring transient emissions, although it is very useful for measuring continuous harmonics.

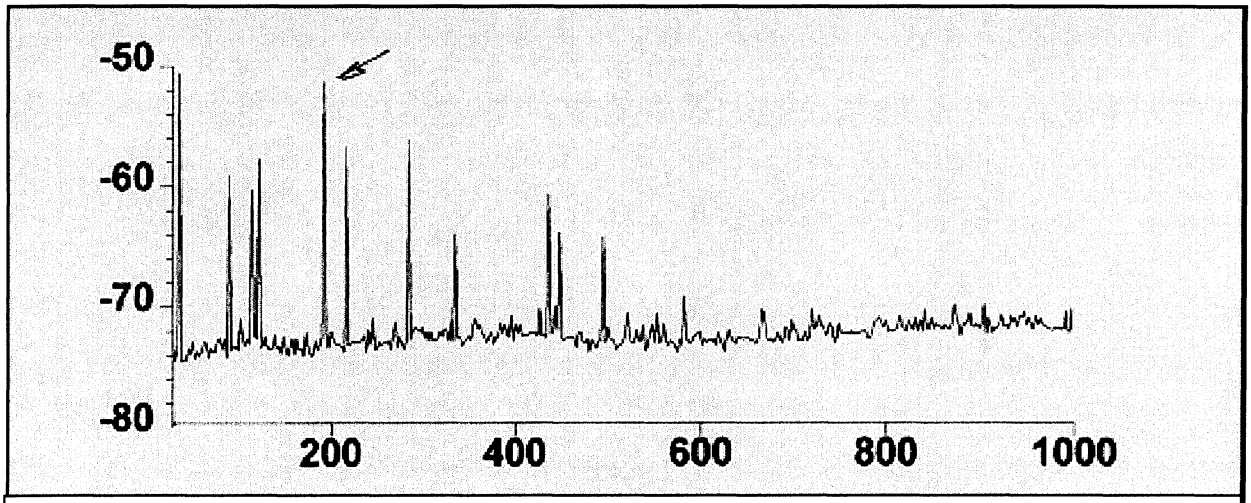


Figure 7. Lab measurements. CHASE CBL 6112 antenna. HP8591E spectrum analyzer. 100 sweeps accumulated. Sweeptime 0.2 s. total time 30.2 s. Source: ESD pulses once per second. The peaks are distributed at random: cf Figure 8.

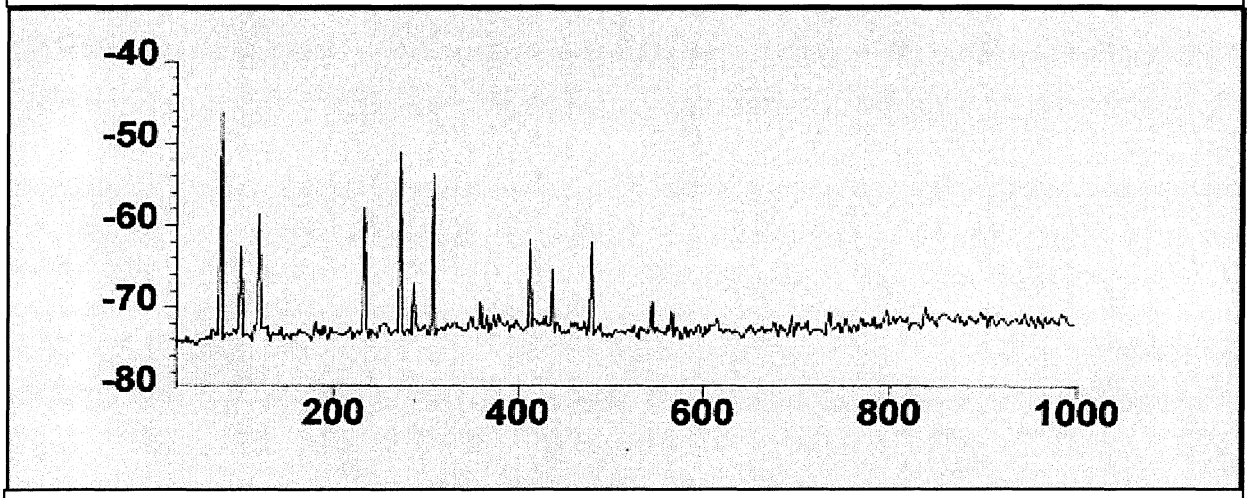


Figure 8. Lab measurements: exactly the same conditions as in Figure 7. Still the result is not the same.

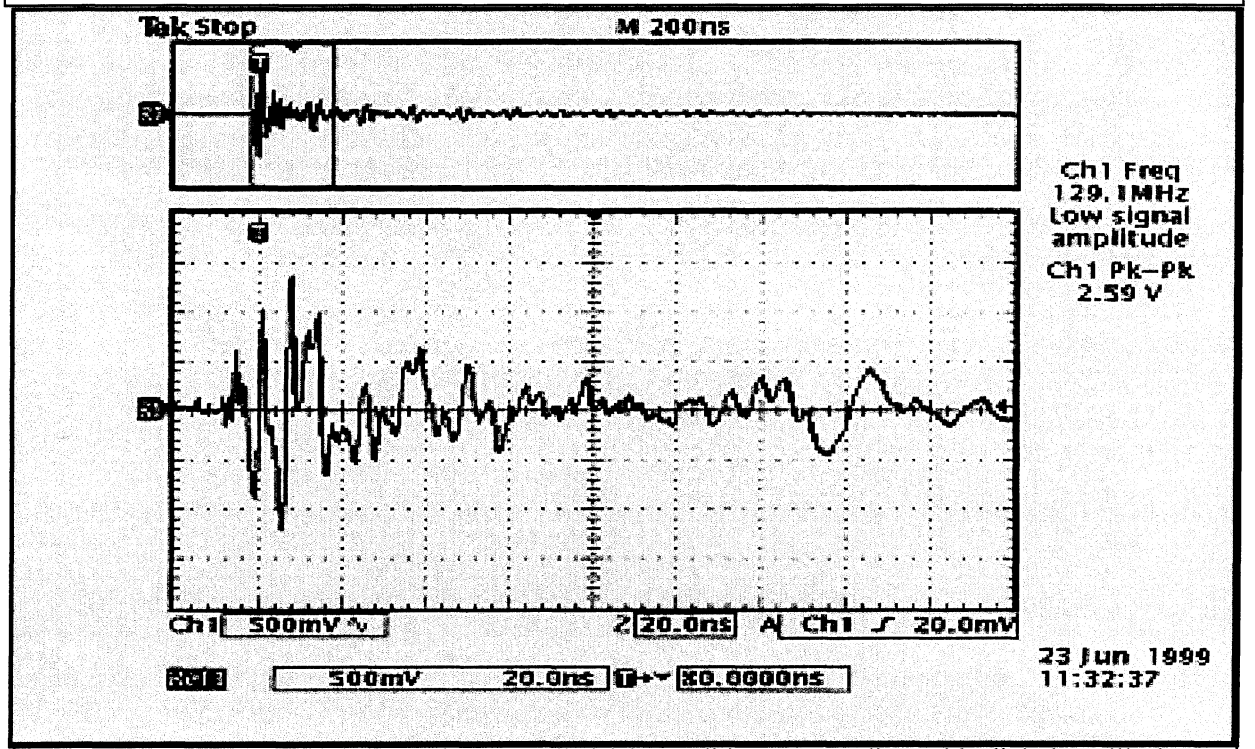


Figure 9. Measuring transients under controlled lab conditions. Recording with digital oscilloskope, Tektronix TDS 724 A, 500 MHz, 1GS/s. Chase CBL 6112 bilog antenna. Source: ESD pulses once per second. The recording never failed, ie one pulse per second was recorded. The recordings were similar for equal voltage levels.

Radar for measuring the content in storage tanks.

VP Research Olle Edvardsson

SAAB Rosemount, Box 13045, 40251 Gothenburg, Sweden, olle.edvardsson@marine.saab.se

1. Abstract

A big application for radar today is to measure the level of liquids or solids in tanks and containers. Far over 100000 sensors manufactured by more than a dozen manufacturers are installed world-wide. The first installation of a serial produced Radar Level Gauge (RLG) was done 1976 on a super-tanker and today radar is since long time the dominating methods on big ships as well as for more or less all demanding industrial applications. Radar is used over a wide range from high accuracy (<1 mm over 30 m) to measurement is complicated tanks where the biggest issue is to find the right echo. A survey of the various techniques in use, typical performance and the ongoing development is done below. The development of microwave components and low cost digital processors has been and will continue to be important driving forces for the development.

2. A few historical notes

The practical possibilities for industrial applications of microwaves started when microwave semiconductors come in use. One starting point was when the Gunn diodes were invented 1964 and started to be available a few years later. A lot of applications were investigated among them very special environments (level of molten steel) and measurements utilizing the special dielectric properties of water (moisture measurements). One application which in the late sixties was predicted to very big "soon" was anti-collision radar for cars which came into the market around 30 years later. Another application which attained less attention but which grew fast was level gauging of liquids by radar. At first glance is similar in function to an aircraft radar altimeter which had been in use since the late forties. Patents for radar level gauging were filed already in the mid sixties and some of the first experiments were directed towards very hard environments like level gauging in blast furnaces and basic oxygen furnaces. The first large scale application however was measurement of level in tanks in oil tanker ships. The big accuracy requirements (in the beginning 0.1% but later well below 0.01% to compete with older mechanical methods) and the need for low current drain (to avoid the possibility of ignitions) was big issues in the early days as well as the system cost which directed the early use to really demanding applications only. The RLG development was also much dependent on the development of digital processors as the big accuracy demand needed new methods as compared to aircraft altimeters. For the long term development in that area obviously Moores law has been a good help. Another important market change is that microwave semiconductors and manufacturing methods turned to become really low cost with satellite TV, microwave links etc.

Some notes in the market growth are:
1975: First serial produced RLG delivered (from SAAB to oil tanker)
1985: Second manufacturer turns up on the market. RLG use in oil refineries starts
1995: Nearly 10 RLG manufacturers exist. Radar is standard on ships and emerging to be standard on high precision refinery tanks. Gauges intended for process industry starts to come.
2002: RLG established standard on ships and refinery. Process industry is the fastest growing market segment.

3. Technical background

A radar level gauge in its typical application measures the plumb-line distance from the roof of a tank to the liquid surface. Thus there are certain similarities with an aircraft altimeter but for the radar level gauge (RLG) the distance is rather small (5-30 m depending on application) and the measurement can be done at a slow rate such as 1 time per second. One special detail is a big requirement for accuracy which typically is ±1 mm and another some special environmental requirements typical for the application. Very common is a certified inability to create ignition (gasoline tanks are common) and other applications may require high pressures (such as 50 bar) or special temperatures (somewhere between 300°C or -165°C) in the tanks and around the antenna.

The background to the technical solutions as compared to classical radar can be divided in the following areas:

The radar propagation is dependent on the tank environment which is very different from the radar environment around well known classical radar families.

A few different system solutions are used modified from older systems. Both pulsed and FMCW systems are used now and additional phase measurement is sometimes applied to enhance accuracy. System allowing high accuracy (0.01% or better) and systems allowing a complicated echo environment are used.

The antennas are of established types but the choice is much dependent on the tank and the general environment. The size, choice of materials etc. is dictated by the tank environment rather than the system requirements and various contamination may be the real crucial question.

Special environmental requirements may occur such as safety against explosion in tanks containing explosive material or low current consumption in order to adopt to instrument standards.

4. Different system concepts

All early systems used a FMCW radar with a linear sweep over a wide bandwidth (>1GHz) and low power (<1 mW). As most systems are used in closed metallic tanks the radiation outside of the tanks will fall below any current EMC limits. The linear modulated FMCW-technique translates the more or less complicated multitude of echoes from the tank (surface echo plus several echoes from the internal structure in the tank) to a frequency spectrum and the spectrum is many times translated to a temporal signal by a FFT procedure. High accuracy systems use the FMCW-method but for many uses pulsed systems have entered the market the last decade. Short pulses (<1ns) with a low duty cycle (<1:1000) are used and by using a technique similar to a sampling oscilloscope the pulses can be extended for a final signal processing with fairly slow methods with corresponding small current consumption. The most important feature for the pulsed systems is a low average current consumption enabled by a very low average power (≈1 μW) from the transmitter in spite of a 1 mW peak power. The low current consumption enables the transmitter to be fed from a 4-20 mA instrument bus. Another advantage is that the signal for a not too wide-band receiver appears as noise which can be made lower than the allowable EMC noise from any type of equipment. For both FMCW systems and pulsed systems the average power is the important parameter for the sensitivity and pulsed systems with low current drain typically have 20-30 dB less sensitivity than typical FMCW-systems systems intended for high accuracy application. Low sensitivity implies difficulties when small antennas are used at long distances and when extra attenuation occurs due to dirt layers at the antenna and foam layer on the liquid surface.

The sensitivity, defined as maximum two way attenuation, need to be in the range 50-90 dB to cover the full range of operating conditions. The pulsed systems with low current consumption typically have around 70 dB while the most sensitive systems on the market has 90-100 dB.

5. Current consumption

The current consumption is a crucial specification point for at least two reasons. In order to pass the most common type of explosion certification the current consumption must be low such a around 100 mW. Further restrictions is to take account to in case the feeding is via an instrument bus where the current has to be 4-20 mA at a voltage in the range of 15-5 V. In this case the feeding is done over a resistor lowering the voltage at high currents and the current may be used as an analogue output value.

6. Signal processing

An area of big importance is the signal processing. Depending on the use the focus may be to get a high accuracy with a resolution around 0.1 mm during fair signal condition or to find the real surface echo in a complicated and unstable signal environment.

At first sight the FMCW and pulsed systems appears to be very different but usually a FFT is a part of the signal processing transforming the received signal to a time like signal.

At high accuracy systems a least square calculation is often employed to utilize the available information as much as possible.

7. Frequency ranges

Four frequency ranges occur on the market. Around 10 GHz was used on the older systems but because of the big need for bandwidth (around 1-2 GHz in order to get better than a few dm resolution) it is usually necessary to restrict the use to closed metallic tanks so that the leaking power will remain below the limits allowed for any electronic equipment like a PS etc. Today many radar level gauges are located within the ISM-bands at 5.8 and 24.5 GHz. The need for bandwidth is still much bigger than the width of this two frequency bands so only very low powers can be used unless the tank is metallic and closed. A fourth "frequency band" is very short DC-pulses (1 ns) used in time domain reflectometry (TDR) using a 2-wire line or a surface waveguide. The last method has found use in silos for

solids like grain, cement etc. and the transmission line must be arranged to avoid radiation.

8. Radar propagation

The radar propagation in a tank environment is very different from a classical type of radar for surveillance, altimeters, fire control etc. Under ideal conditions the echo from the liquid surface is strong and by the use of a narrow antenna beam possible echoes from the tank structure can be suppressed. Due to a turbulent liquid surface (waves in the tanks of a ship, agitators or boiling in a fixed tank etc.) the echo may get much smaller an amplitude and be highly fluctuating and then echoes from structures inside the tank may at least temporary be much stronger than the echo from the liquid surface. A problem specific for ships is changing trim and list conditions which if a narrow antenna beam is used will cause the reflected power to fail to hit the antenna. An agitated surface will have in the order of 20 dB less reflection than a smooth surface so it will be necessary to increase the sensitivity in case a turbulent surface is the normal condition.

Other possible problems are associated with the fact that the antenna and the content of the tank are very close to each other. That may cause the antenna to be covered by product splashing up from the surface and in most cases that will deteriorate the antenna function. Not the least water will have a big influence because of the high dielectric constant of water. Just a few tenths a mm of wet dirt may degrade the transmission 20 dB. A similar degradation can occur from a foam layer on the surface. On the antenna one necessary precaution is to use a dielectric material with good water-repellent properties (like PTFE or polyethylene) in order to avoid the forming of a continuous wet layer. On PTFE the inevitable condensation droplets then will remain too small to disturb. Further more horizontal or nearly horizontal surfaces should be avoided as condensation otherwise will form growing droplets (like on a horizontal glass pane) which will have big influence when they are around one wavelength in circumference. Metal surfaces are much less sensitive as the metal electrically will "short-circuit" the water or dirt.

Atmospheric conditions have a limited influence on radar as can be guessed from the dielectric constant of air which is 1.0006. Hydrocarbon gases, hot water vapor etc. can have a higher dielectric constant and in for instance pressurized propane the measured distance will be around 1% longer than in air. Accurate RLG systems employ a pressure dependent correction for that.

9. Antennas for RLG

The antenna is a key component of each RTG-system because it is in close contact with the liquid in the tank. Mainly broadside radiators like horn antennas, parabolas and array antennas are used and typical for the RTG-application is that the antennas may not be chosen after their optimal size but after the installation conditions. For instance the diameter of the antenna may be limited of the hole through the antenna is to be mounted. Unlike classical antennas the antenna and not the least the feeding must be designed to stand and isolate from difficult environment like high pressures (40-100 bar) and special temperatures (-175 to 250 degrees).

Parabolas are used extensively on big ships requiring a rather narrow antenna beam at 30 m distance. By their nature the essentially metallic construction can be made to resist a harsh environment.

Horns are the most common antennas for smaller tanks. Horns up to 200 mm are used and the bigger ones may be shaped to be shorter or be fitted with a lens for the same reason.

Array antennas are used made from PCBs. The obvious geometrical advantages as compared to parabolas and horns have to be valued against the practical problems with good sealing and the condensation problems possibly occurring with horizontal surfaces. For all level gauging antennas incl. arrays it is very important to have the same electrical length to all radiation elements which sometimes calls for non-conventional array configurations.

End-fire antennas made of PTFE in shape of a rod antenna are sometimes used to be able to install the antenna through a fairly small hole (30-40 mm) and still have 10-12 dB antenna gain.

10. RLGs measuring through a pipe

A number of RLG applications use pipes in order to adopt to tank constructions or to avoid echoes from tank structures. A pipe of 50-300 mm can be used and in many cases an old pipe, already installed in the tank, is used. In such cases the pipe does not look as a conventional waveguide but is very rusty, has an oil lining and a lot of holes to ensure the same liquid level inside and outside of the pipe. Pipes are used both in tanks with floating roofs and in tanks for liquid gas.

For use in existing pipes of very poor quality (seen as waveguides) the propagation mode H_{01} is usually utilized. Due to its low currents in the wall (if the pipe is big) this mode is less sensitive for heavy rust layers and

oil clogging than the basic propagation mode H_{11} in a circular guide.

In order to ensure high accuracy to get sub-mm accuracy the mode purity must be very high such as >25 dB to the next strongest propagating mode. Special mode converters have been developed for this and a long nicely shaped cone is one solution transforming H_{01} in a small diameter (just allowing H_{01} to pass) to the real pipe diameter which is 4-8 times bigger. An array antenna with radiating elements arranged to create H_{01} is another solution which is used.

H_{11} is also used in applications where the pipe can be kept clean such as liquid gas measurements. The mode generator is simpler and there is also a possibility to use the polarization to measure on calibration marks in order to enhance accuracy.

11. Conducted radar

There are quite a number of RLGs using one or two wires to conduct the radar beam to the surface. In this case a low frequency is used or rather DC-pulses around 1 ns long. The electronics and echo logic is rather simple in this case and this type of RLGs is usually a low cost version. It is typically used for solids (grain etc.) where the shape of the surface makes the reflection of a free beam unreliable.

12. Safety requirements

A very common requirement, not the least within the petroleum industry, is that all electronic equipment must be guaranteed to be incapable to create ignition or explosion. Typical for this is a very high degree of safety typically requiring that even after two serious failures there should remain a margin.

For explosion safety three basic methods bare used. At "intrinsic safety" currents and voltages are limited by a combination of a fuse, 3 zeener diodes and a resistor to such an extent that even at the worst material combination no ignition can occur in any gas mixture. Fairly small powers are available but with modern components most semiconductor functions can be established. Another method is "explosion safety" where a possible internal explosio9n is allowed but will be contained within a strong enclosure. This can allow motors and other power consuming items but in not allowed in areas close to the tank interior. A third method is to keep the possible explosive mixture away from the electric circuits by various casting, oil filling or sand filling. A stringent official testing is an important part of this.

Another kind of safety is that many applications concern beverages, pharmaceuticals etc where special requirements for cleanness, freedom of voids where bacteria could hide etc.

13. Future development

There has been a very interesting development since 1975 with 10 generations of consecutive RLG equipment launched from SAAB alone. With regard to ongoing component development the following 25 years can be expected to be even more exiting.

The favorable cost development for microwave semiconductors and MMICs is one important trend and the increased capacity of low cost processors is another.

The Radio-Aeronomy Receiver Systems at Onsala Space Observatory

Peter Forkman and Anders Winnberg
Onsala Space Observatory

Introduction

During the last 6 years a station for measurements of gases in the middle atmosphere (stratosphere: 10 – 50 km; mesosphere: 50 – 100 km) has been developed in an experimental process. The driving rationale has been to apply passive microwave and millimetre-wave technique to study the rotational spectral lines from middle atmospheric molecules in much the same way as is done in radio astronomy. The station now consists of two receiver systems: one for the water vapour line at 22 GHz and one for the ozone line at 111 GHz and for the carbon monoxide line at 115 GHz. Both receivers are served by a common digital spectrometer and an online steering computer.

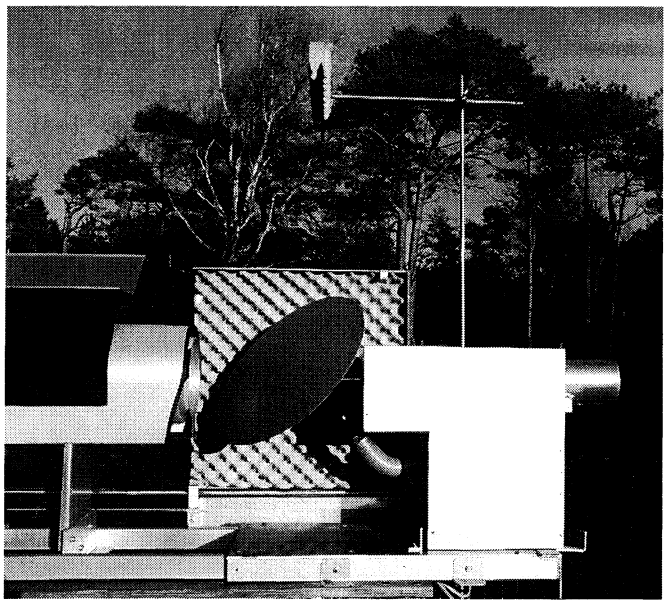


Fig. 1 Part of the 22-GHz receiver with horn, plane mirror and calibration load

The 22-GHz receiver

In order to achieve a maximum in reliability, stability and ease of operation this receiver system has been designed as an uncooled system based on a HEMT preamplifier and with a minimum of components (Figures 1 and 2). The antenna is a corrugated aluminium horn and the direction of measurement is determined by a plane, rotatable mirror. Standing waves are eliminated by driving the mirror unit in a reciprocating motion ('path length modulation').

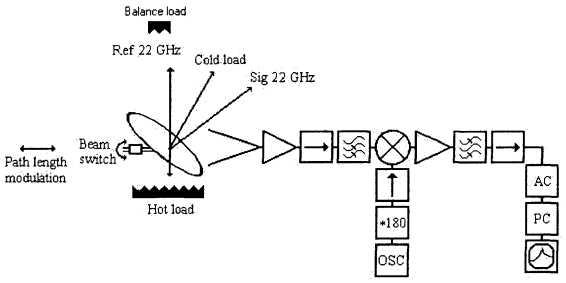


Fig. 2 Block diagram of the 22-GHz receiver system. OSC is the local oscillator, AC is the autocorrelator spectrometer, and PC is the steering computer

The receiver noise temperature is about 250 K and the contribution from the atmosphere is of the order of 50 K. However, the line amplitude of strato/mesospheric water vapour is less than one Kelvin only. Therefore a measurement is made at an elevation of about 20° ('sig' in Figure 2) and another one close to 90° ('ref') and then the difference between them is formed. In the 'sig' position the line strength is much larger than in the 'ref' position. The error made due to the weak line in the 'ref' position is corrected in the calibration process. In addition, by changing between 'sig' and 'ref' as quickly as practically possible, effects of gain variations on timescales longer than the measurement cycle are eliminated.

The dominating line broadening mechanism in the atmosphere below an altitude of about 75 km is pressure broadening. Therefore there is information about the water vapour abundance as a function of altitude in the shape of the line profile (Figure 3) and it is possible to retrieve this 'height profile' (Figure 4) through a mathematical algorithm.

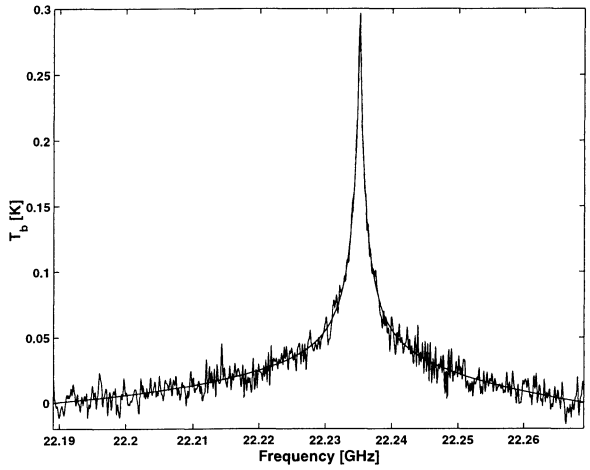


Fig. 3 A spectrum of the 22-GHz water vapour line in the strato/mesosphere. The intensity scale is in brightness temperature; the blackbody equivalent temperature.

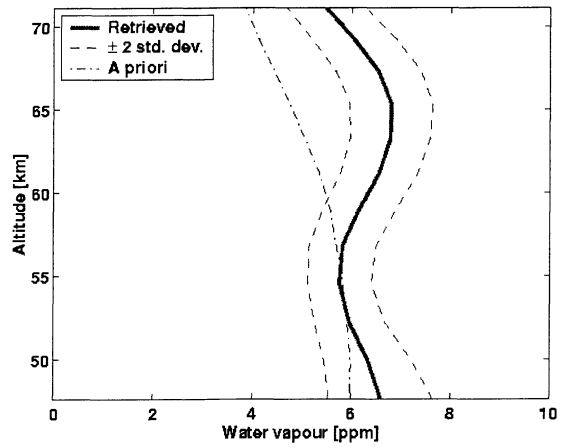


Fig. 4 A 'height profile' (solid line) of water vapour corresponding to the line profile in Figure 3. The unit of concentration is 'parts per million' (ppm) in volume, i.e. no. of water molecules per million 'air molecules'.

The 3-mm receiver

Only recently HEMT amplifiers have been successfully manufactured for these high frequencies. Since the 3-mm receiver was designed some five years ago it has no preamplifier. The first system component therefore is a mixer of Schottky type cooled to 20 K together with the feed horn and IF amplifier by a helium gas cooling machine (Figure 5).

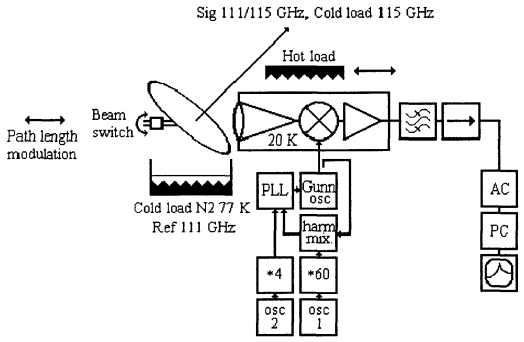


Fig. 5 Block diagram of the 3-mm receiver system. The spectrometer (AC) and computer (PC) are the same as for the 22-GHz receiver. The computer is programmed to be able to switch between the two receiver systems. PLL means Phase Lock Loop.

Since the height distribution of ozone (O_3) and carbon monoxide (CO) is quite different, the CO being situated mainly above the O_3, the line widths are quite different too. Because of the narrowness of the CO line it is most easily detected by switching the receiver pass band between two frequencies while the mirror is 'looking' at the same position of the sky ('frequency switching'). The much broader O_3 line, however, is detected by switching between the sky and a container with absorbing bottom and walls and filled with liquid nitrogen ('load switching').

Some Issues in Low Power Arithmetic for Fixed-Function DSP

Oscar Gustafsson and Lars Wanhammar

Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden
E-mail: {oscarg, larsw}@isy.liu.se

Abstract

In this paper we discuss some aspects of designing arithmetic circuits with low power consumption. We focus on digit-serial processing techniques, as well as minimum adder multipliers, and multiple constant multiplication. A review of power dissipation sources in CMOS is given. Further, we discuss some methods to decrease the power dissipation of the algorithmic, arithmetic, and architecture levels.

1 Introduction

Digital signal processing algorithms are key components in communication systems. In this paper we discuss some of the key issues involved in design and implementation of the arithmetic circuits for the inner loop in such algorithms. Generally the throughput of the systems of concern here are fixed and the design problem is viewed as a resource adequate problem. The throughput is therefore a fixed constraint within which we will make a trade-off between chip area and power consumption. Further we assume that the required flexibility is low, i.e., the circuitry will only perform one single algorithm that is not required to be easily modified. Note that high flexibility in terms of modifiable algorithms is contradictory to highly efficient implementation. For example, a programmable signal processor is flexible, but highly inefficient in terms of throughput, chip area and power consumption.

In this restricted context we regard design efficiency and power consumption as the two most important aspects in this type of applications. In this paper, however, we will focus on the later.

In Section 2 we review the different sources for power consumption in CMOS circuits. In Section 3 we discuss various techniques to reduce the power consumption. We concentrate on digit-serial processing, minimum adder multipliers, and multiple constant multiplication techniques. Finally, in Section 4 we summarize our conclusions.

2 Energy Consumption

Minimization the energy consumption in portable and battery powered systems has become an important aspect of System-om-Chip design. In battery powered systems the peak power that can be delivered is limited by the batteries and we are interested in maximizing the runtime.

2.1 Power Dissipation in CMOS Circuits

CMOS technology is currently and will continue to be the dominating technology for VLSI and ULSI circuits. The feature size is rapidly shrinking towards 0.1 μm and below. Circuits implemented using modern CMOS processes are no longer characterized by the number of transistors on the chip. Instead the energy or power consumption is a more appropriate metric.

Power dissipated in CMOS circuits consists of several components

$$P_{total} = P_{sw} + P_{sc} + P_{static} + P_{leakage} \quad (1)$$

where P_{sw} is the power required to charge and discharge a capacitive load, P_{sc} is the short circuit power consumed during transitions, P_{static} is the static power consumed by the device, and is $P_{leakage}$ the leakage power consumed by the device. P_{sw} and P_{sc} are present when a gate is changing state, while P_{static} and $P_{leakage}$ are present regardless of state changes. P_{sw} is [1]

$$P_{sw} = f_{CL}\alpha C_L V_{dd}\Delta V \quad (2)$$

where f_{CL} is the frequency of operation (often the same as the clock frequency), C_L is the equivalent switched capacitance of the whole circuit, V_{dd} is the power supply voltage, ΔV is the voltage swing across the switched capacitance, α represents a switching activity factor based on the probability of a transition. The product αC_L is also referred to as the effective switched capacitance, or C_{eff}. In most cases we use full swing, i.e., $\Delta V = V_{dd}$.

$$P_{sw} = f_{CL}C_{eff}V_{dd}^2 \quad (3)$$

Hence, the power consumption can be reduced by using a lower power supply voltage. Unfortunately this has the effect, that the speed of the circuit is also reduced [1]–[3] according to

$$f_{CL} = \frac{k(V_{dd} - V_{th})^{\alpha}}{C_L V_{dd}} \quad (4)$$

For properly sized gates, P_{sc} is less than 10% of P_{total}, although this may increase with increased device scaling. P_{static} is not present in pure CMOS circuits, but occur in sense amplifiers, voltage references, and constant current sources. $P_{leakage}$ is due to leakage currents from reversed biased PN junctions at the source and drain of MOS transistors and subthreshold currents. The leakage is proportional to device area and temperature while the subthreshold leakage is strongly dependent on device

threshold voltages, and becomes an important factor as power supply voltage scaling is used to lower power. For systems with a high ratio of stand-by operation to active operation, $P_{leakage}$ may be the dominant factor in determining overall battery life.

Using a special low-power CMOS process can, of course, reduce the power consumption of digital circuits. Here, however, we will assume that the digital filters will be implemented using a standard digital CMOS process. Reducing any of the factors in (1) can decrease the power consumption.

3 Design Techniques for Power Reduction

Opportunities for power consumption minimization are available and must be explored at all levels of the design hierarchy [1], [2], [6]–[10]. Here we will only discuss some possibilities at the algorithm-arithmetic-hardware architecture levels.

3.1 Algorithmic Level

At the algorithmic level energy reduction techniques typically focus on minimizing the number of operations weighted by the energy cost of those operations. Multirate techniques are efficient in terms of operations per sample [11]–[15]. Common to these techniques is that the effective number of arithmetic operations per sample is reduced.

Wave digital filter structures [5], [16] have, besides their robustness, very low coefficient sensitivity which reduces the required coefficient word length but also the data word length. Hence, they are good low- power candidates. Other techniques try to reduce and minimize the arithmetic complexity, e.g., strength reduction [17], or minimize memory traffic.

In some algorithms, the power consumption can be reduced by eliminating operations with known results altogether.

3.1.1 Architecture-Driven Power Supply Voltage Scaling

Architecture-driven power supply voltage scaling techniques exploit the quadratic dependency on V_{dd} in (3) to reduce the power consumption [1]. In order to compensate for the associated reduction in circuit speed, (4), the inherent computational parallelism in the algorithm should be exploited. We have therefore proposed the following way to reduce the power consumption [5], [18].

- Design a maximally fast and resource minimal realization of the basic DSP algorithm.
- Convert any excess circuit speed to reduced power consumption by reducing the supply voltage.

From (4) it is evident that from a speed point of view a CMOS process with a low threshold voltage is to be preferred. In practice, many circuits are therefore manufactured with a process with two threshold voltages. Devices with a low threshold voltage are used for high-speed circuits and devices with a high threshold voltage are used for low speed circuits since the latter have smaller leakage currents. Typically, the power supply voltage is in the range $2V_{th}$ to $3V_{th}$.

3.1.2 Maximally Fast Algorithms

DSP algorithms can be divided into iterative processing and block processing algorithms [5]. The maximal sample frequency for an iterative recursive algorithm, described by a fully specified signal-flow graph, is [5],[19]

$$f_{max} = \max_i \left\{ \frac{N_i}{T_{opi}} \right\} \tag{5}$$

where T_{opi} is the total latency due to the arithmetic operations, and N_i is the number of delay elements in the directed loop i [19]. The minimum sample period is also referred to as the iteration period bound.

The iteration period bound is of course not possible to improve for a given algorithm. However, a new algorithm with a higher bound can often be derived form the original algorithm. We recognize from (5) that there are two possibilities to improve the bound.

- Reduce the operation latency in the critical loop.
- Introduce additional delay elements into the loop.

Removing superfluous arithmetic or logic operations from the loop can often reduce the operation latency [5]. Using low-sensitive filter structures, e.g., wave digital filters can also minimize the latency of multipliers. Another efficient approach is using cascade form structures [12], [20].

A new algorithm with more delay elements can be obtained in several ways, for example, scattered look-ahead [21] and frequency masking techniques, which was first proposed by Lim. We do not advocate the former since some finite word length problems may occur [22].

3.2 Arithmetic

At the arithmetic level with have several options. We can divide the data in smaller digits requiring smaller chip area as will be discussed in Section 3.2.1, exploit redundancies inside and among operations as will be discussed in Section 3.2.2 and Section 3.2.3, use distributed arithmetic for complex multiplication and sum-of-products [5], and simplified/truncated multipliers [23]. Further we almost always select fixed-point over floating-point arithmetic.

The number representations is another area that offer power trade-offs [4], [5], [10]. Selection of sign-magnitude versus two's-complement representation result in significant power reduction if the input samples are uncorrelated and dynamic range is minimized [1]. Data and coefficient word length minimization are also important in order to minimize the power consumption.

3.2.1 Digit-Serial Processing

Bit- and digit-serial processing techniques has received considerable attention during the last decade [5], [24], [25]. In bit-serial processing one bit of the input word is processed each clock cycle. A bit-serial adder computing $c = a + b$ is shown in Fig. 1 (a) while a bit-serial serial/parallel multiplier computing $b = \alpha a$ is shown in Fig. 1 (b).

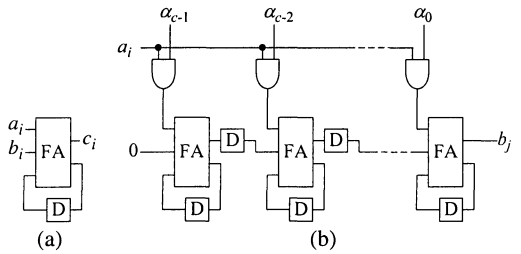


Figure 1: *Bit-serial (a) adder and (b) serial/parallel multiplier.*

For digit-serial processing a number of bits of the input word, a digit, is processed in parallel. If the digit-size, i.e., the number of bits processed concurrently, is one the digit-serial system reduces to a bit-serial system, while for a digit-size equal to the word length the system reduces to a bit-parallel system. The motivation for digit-serial processing is to find an optimum trade-off between area and processing capacity.

Traditionally, digit-serial multipliers has been obtained either via unfolding of a bit-serial multiplier [26] or via folding of a bit-parallel multiplier [27]. The problem with these approaches is that the obtained circuits have not been pipelinable at the bit-level. The reason for this is shown in Fig. 2 where a digit-serial adder obtained via unfolding of a bit-serial adder is shown. The recursive loop prohibits the insertion of pipelining to reduce the critical path to less than d full-adders but solutions to this problem has been proposed in [28]–[31].

The latency of digit-serial processing elements are conveniently described in terms of number of clock cycles. An adder has a latency of zero clock cycles while a serial/parallel multiplier has a latency of $\lceil W_f/d \rceil$ clock cycles, where W_f is the number of fractional bits of the coefficient. The clock frequency will be determined by the longest critical path, which basically is the number of adjacent operations. Introducing a pipelining stage will increase the latency with one clock cycle, but the critical (electrical) path will be decreased, and, thus, the clock frequency increased. The optimal level of pipelining in maximally fast, digit-serial, implementations is presented in [32].

3.2.2 Minimum Adder Multipliers

Multiplication with a constant fixed-point number is a common operation in many DSP algorithms. The multiplier can be expressed as a network of shifts, adders and subtractors. For bit-parallel arithmetics the shifts can be realized without any gates using hard wiring. To obtain an efficient implementation with a small switched capacitance it is therefore of interest to minimize the number of additions and subtractions. As the area cost is similar for additions and subtractions we will refer to both as additions.

Representing a number using canonic signed digit (CSD) representation is a common method when an area efficient multiplier required. A B-bit integer coefficient c is represented as

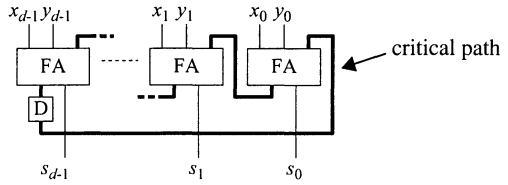


Figure 2: *Digit-serial adder with digit-size d obtained from unfolding an bit-serial adder.*

$$c = \sum_{i=0}^{B-1} b_i 2^i \tag{6}$$

where $b_i \in \{-1, 0, 1\}$ and $b_i b_{i+1} = 0, 0 \le i \le B-2$. The average number of non-zero bits is about $B/3$ for $0 \le c \le 2^{B-1}$ [33].

CSD multipliers are based on sums of powers of two of the input word. However, other network topologies, not restricted to CSD representation, are available that require an even smaller number of adders. An optimal method was described in [34]. The method is based on the graph representation of the multipliers introduced in [35]. The search method involved was simplified and extended in [36]. In Fig. 3 a graph representation of a multiplication with 341 is shown. The vertices corresponds to additions while the edges corresponds to multiplications with $\pm 2^n$, i.e. shifts and possible replacement of the subsequent addition by a subtraction. The value obtained in each vertex is known as a fundamental. Using CSD requires four adders while only three adders are required using the graph representation. This concept has also been extended in [37] to a redundant type of adders known as carry-save adders [38].

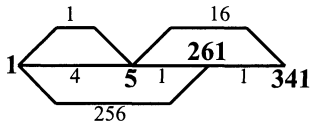


Figure 3: *Graph representation of multiplication with 341. Fundamentals are in bold.*

3.2.3 Multiple Constant Multiplication

In many DSP algorithms one data value is multiplied by several constants. One typical example is the transposed form FIR filter, where one input data is multiplied with the filter coefficients as shown in Fig. 4.

The hardware cost can be decreased by utilizing redundancy between the coefficients [35], [39]–[45]. This is known as multiple constant multiplication (MCM). The previous work in this area can be divided in two techniques. The first technique is based on pattern matching techniques [40]–[45] and the result depends on the initial representation of the constants. This type of algorithms are often referred to as using subexpression sharing or subexpression elimination. The second technique is independent of the representation and only considers the values after each addition [35], [39]. The algorithm presented in [39] utilizes a graph representation and is shown to be optimal in terms of minimum number of additions. Most of the al-

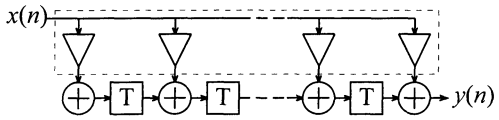


Figure 4: *Transposed form FIR filter.*

gorithms using the first technique can be applied to other subexpression sharing applications than multiplications, such as Hadamard matrix evaluation. Note that by transposing the signal flow graph a sum-of-product network is obtained, i.e., the problem of sum-of-products, e.g., for a direct form FIR filter, can be solved in the same way.

3.2.4 Number Representation

In older CMOS processes the equivalent switched capacitance is due to devices and wires and the dominant wire capacitance is between the wire and substrate. In deep submicron CMOS technologies the capacitance associated with the devices are relatively small and the dominant wire capacitance is instead dominated by interwire capacitance as illustrated in Fig. 5. For long buses it is therefore important to order the transmission of successive data on the bus so that the charging/discharging of the interwire capacitors is minimized. A useful measure to minimize is the sum of the Hamming distances of the transmitted words. In speech signals, for example, it may be advantageous to use sign-magnitude representation on long buses, since most samples have small magnitudes which causes only the least significant bits to vary while in two's-complement representation almost all bits change when a sample value changes sign.

3.3 Architectural

The architecture design starts from an operation schedule that is isomorphically mapped to a corresponding hardware structure. It is important that the schedule is maximally fast so that any excess speed can be traded for lower power consumption by power supply voltage scaling.

3.3.1 Cyclic Scheduling

To achieve a maximally fast schedule it is necessary to schedule over several sample intervals by connecting several computation graphs [5], which will yield a periodic schedule. Each computation graph represents the operations and the shimming delays within one sample interval. It may help to imagine the computation graphs drawn on a cylinder of circumference equal to the scheduling period, as shown in Fig. 6. Hence, the schedule has no beginning or ending which may restrict the scheduling the operations. We denote this scheduling formulation as cyclic. This is the most general scheduling formulation and it can be used to find maximally fast, resource-optimal schedules.

To attain the minimum sample period, it is necessary to perform cyclic scheduling of the operations belonging to several sample successive intervals if

- the execution time for a processing element (PE) is longer than T_{min} or
- the critical loop(s) contains more than one delay element.

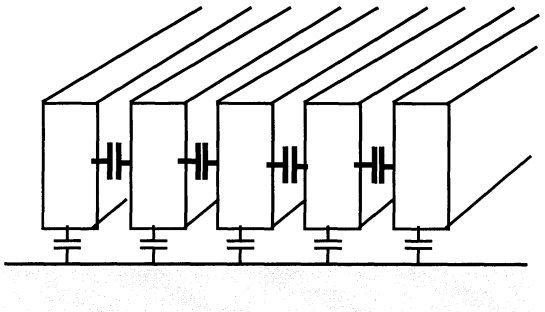


Figure 5: *Interwire and wire-to-substrate capacitance model.*

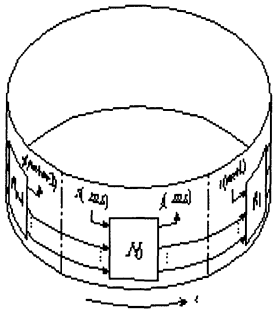


Figure 6: *Illustration of the cyclic operation schedule.*

Generally, the critical loop should be at least as long as the longest execution time for any of the PEs in the loop. Bit-serial and digit-serial operations with execution times longer than the minimal sample period can thereby be completed within the scheduling period. The minimum number of sample periods to schedule for is

$$m = \left\lceil \frac{\max_i (T_{exe,i})}{T_{min}} \right\rceil \tag{7}$$

where $T_{exe,i}$ is the execution time of operation i.
The execution time is $\lceil W_d/d \rceil$ clock cycles for a digit-serial adder. For a serial/parallel multiplier it is $\lceil W_d/d \rceil + \lceil W_f/d \rceil$ clock cycles, where W_d is the data word length.

4 Conclusion

In this paper we have discussed some aspects of designing arithmetic circuits with low power consumption. We focused on digit-serial processing techniques to be able to find an optimal trade-off between area and processing capacity. Further, minimum adder multipliers and multiple constant multiplication was discussed to minimize the switched capacitance and area. A review of power dissipation sources in CMOS was also given. Further, discussions on decreasing the power dissipation were given for the algorithmic, arithmetic, and architecture levels. As noted, the power optimization must be performed at all levels in the design hierarchy. We proposed using maximally fast implementations and trade any excess speed for lower power consumption by lowering the power supply voltage.

References

[1] A.P. Chandrakasan and R.W. Brodersen, *Low Power Digital CMOS Design*, Kluwer Academic, 1995.

[2] A. Bellaur and M. I. Elmasry, *Low Power Digital CMOS Design: Circuits and Systems*, Kluwer, 1996.

[3] T. Njølstad and E. J. Aas, "Validation of an accurate and simple delay model and its application to voltage scaling," *Proc. IEEE Intern. Symp. Circuits Syst.*, Monterey, CA, Vol. II, May 31-June 3, 1998, pp. 101–104.

[4] I. Koren, *Computer Arithmetic Algorithms*, Prentice Hall, Englewoods Cliff, New Jersey, 1993.

[5] L. Wanhammar, *DSP Integrated Circuits*, Academic Press, 1999.

[6] J. M. Rabaey and M. Pedram (Eds.), *Low Power Design Methodologies*, Kluwer, 1996.

[7] M. S. Elrabaa, I. S. Abu-Khater, and M. I. Elmasry, *Advanced Low-Power Digital Circuit Techniques*, Kluwer, 1997.

[8] L. Benini and G. De Micheli, *Dynamic Power Management: Design Techniques and CAD Tools*, Norwell, MA: Kluwer, 1998.

[9] B. Moyer, "Low-power design for embedded processors," *Proc. IEEE*, Vol. 89, No. 11, pp. 1576–1587, Nov. 2001.

[10] M. Mehendale and S.D. Sherlekar, *VLSI Synthesis of DSP Kernels, Algorithmic and Architectural Transformations*, Kluwer Academic, 2001.

[11] H. Johansson, "Multirate single-stage and multistage structures for high-speed recursive digital filtering," *Proc. IEEE Intern. Symp. Circuits Systems*, Orlando, Florida, Vol. 3, May 30-June 2, 1999, pp. 291–294.

[12] H. Johansson and L. Wanhammar, "Filter structures composed of allpass and FIR filters for interpolation and decimation with factors of two," *IEEE Trans. Circuits Syst. II*, Vol. 46, No. 7, pp. 896-905, July 1999.

[13] T. Njølstad and T. A. Ramstad, "Low-power narrow transition-band FIR filters," *Proc. IEEE Nordic. Symp. Signal Processing*, Kolmården, Sweden, June 13-15, 2000, pp. 181–185.

[14] H. Johansson, "On high-speed recursive digital filters," *Proc. X European Signal Processing Conf.*, Vol. 3, Tampere, Finland, Sept. 5-8, 2000, pp. 1701–1704.

[15] G. Jovanovic-Dolecek (Ed.), *Multirate Systems: Design and Applications*, Idea Group Publ., Hersey, USA, 2001.

[16] A. Fettweis, "Wave Digital Filters: Theory and Practice", Proc. IEEE, Vol. 74, No. 2, pp. 270–327, Feb. 1986.

[17] K. K. Parhi, *VLSI Digital Signal Processing Systems: Design and Implementation*, John Wiley & Sons, 1999.

[18] M. Vesterbacka, *On Implementation of Maximally Fast Wave Digital Filters*, Diss. No. 487, Linköping University, June 1997.

[19] M. Renfors and Y. Neuvo, "The maximal sampling rate of digital filters under hardware speed constraints", *IEEE Trans. Circuits Syst.*, CAS-28, No. 3, March, pp. 196–202, 1981.

[20] H. Johansson and L. Wanhammar, "High-speed recursive filter structures composed of identical sub-filters and QMF banks with perfect magnitude reconstruction," *IEEE Trans. Circuits Syst.*, Vol. 46, No. 1, pp. 16–28, Jan. 1999.

[21] K.K. Parhi and D.G. Messerschmitt, "Pipeline interleaving and parallelism in recursive digital filters—Part I: Pipelining using scattered look-ahead and decomposition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, no. 7, pp. 1099–1117, July 1989.

[22] K. S. Arun, and D. R. Wagner, "High-speed digital filtering: Structures and finite wordlength effects," *J. VLSI Signal Processing*, vol. 4, no. 4, pp. 355–370, Nov. 1992.

[23] M. J. Schulte, J. E. Stine, and J. G. Jansen, "Reduced power dissipation through truncated multiplication," *Proc. IEEE Alessandro Volta Memorial Workshop Low-Power Design*, Como, Italy, Mar. 1999, pp. 61–69.

[24] S. G. Smith and P. B. Denyer, *Serial Data Computation*, Kluwer, 1988.

[25] R. I. Hartley and K. K. Parhi, *Digit-Serial Computation*, Kluwer, 1995.

[26] K. K. Parhi, "A systematic approach for design of digit-serial signal processing architectures," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 358–375, Apr. 1991.

[27] K. K. Parhi, C.-Y. Wang, and A. P. Brown, "Synthesis of control circuits in folded pipelined DSP architectures," *IEEE J. Solid-State Circuits*, vol. 27, pp. 29–43, Jan. 1992.

[28] A. S. Ashur, M. K. Ibrahim, and A. Aggoun, "Systolic digit-serial multiplier," *IEE Proc. Circuits, Devices, Syst.*, vol. 142, no. 1, pp. 14–20, Feb. 1996.

[29] Y.-N. Chang, J. H. Satyanarayana, and K. K. Parhi, "Systematic design of high-speed and low-power digit-serial multipliers," *IEEE Trans. Circuits Syst. II*, vol. 45, no. 12, pp. 1585–1596, Dec. 1998.

[30] O. Nibouche, A. Bouridane, M. Nibouche, and D. Crookes, "A new pipelined digit-serial multiplier," *Proc. IEEE Int. Symp. Circuits Syst.*, Geneva, Switzerland, May 28–31, 2000, vol. 1, pp. 12–15.

[31] O. Gustafsson and L. Wanhammar, "Bit-level pipelinable general and fixed coefficient digit-serial/parallel multipliers based on shift-accumulation," in preparation.

[32] O. Gustafsson and L. Wanhammar, "Optimal logic level pipelining for digit-serial implementation of maximally fast recursive digital filters," *National Conf. Radio Science (RVK)*, Kista, Sweden, June 10–13, 2002.

[33] A. Avizienis, "Signed-digit number representation for fast parallel arithmetic," *IRE Trans. Electronic Comp.*, vol. 10, pp. 389–400, 1961.

[34] A. G. Dempster and M. D. Macleod, "Constant integer multiplication using minimum adders," *IEE Proc. Circuits Devices Syst.*, vol. 141, no. 6, pp. 407–413, Oct. 1994.

[35] D. R. Bull and D. H. Horrocks, "Primitive operator digital filters," *IEE Proc. G*, vol. 138, pp. 401–412, June 1991.

[36] O. Gustafsson, A. G. Dempster, and L. Wanhammar, "Extended results for minimum-adder constant integer multipliers," *IEEE Int. Symp. Circuits Syst.*, Phoenix, AZ, May 26–29, 2002.

[37] O. Gustafsson, H. Ohlsson, and L. Wanhammar, "Minimum-adder integer multipliers using carry-save adders," *IEEE Int. Symp. Circuits Syst.*, Sydney, Australia, May 6–9, 2001, vol. 2, pp. 709–712.

[38] T. G. Noll, "Carry-save architectures for high-speed digital signal processing," *J. VLSI Signal Processing*, vol. 3, pp. 121–140, 1991.

[39] A. G. Dempster and M. D. Macleod, "Use of minimum-adder multiplier blocks in FIR digital filters," *IEEE Trans. Circuits Syst. II*, vol. 42, pp. 569–577, Sept. 1995.

[40] M. Mehendale, S. D. Sherlakar, and G. Vekantesh, "Synthesis of multiplierless FIR filters with minimum number of additions," *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, Los Alamitos, CA, 1995, pp. 668–671.

[41] M. Potkonjak, M. B. Shrivasta, and A. P. Chandrakasan, "Multiple constant multiplication: Efficient and versatile framework and algorithms for exploring common subexpression elimination," *IEEE Trans. Computer-Aided Design*, vol. 15, pp. 151–161, Feb. 1996.

[42] R. I. Hartley, "Subexpression sharing in filters using canonic signed digit multipliers," *IEEE Trans. Circuits Syst. II*, vol. 43, pp. 677–688, Oct. 1996.

[43] R. Pasko, P. Schaumont, V. Derudder, S. Vernalde, and D. Durackova, "A new algorithm for elimination of common subexpressions," *IEEE Trans. Computer-Aided Design*, vol. 18, pp. 58–68, Jan. 1999.

[44] A. Yurdakul and G. Dündar, "Multiplierless realization of linear DSP transforms by using common two-term expressions," *J. VLSI Signal Processing*, vol. 22, pp. 163–172, Sept. 1999.

[45] O. Gustafsson and L. Wanhammar, "ILP modelling of the common subexpression sharing problem," in preparation.

A Digital Down Converter for a Wideband Radar Receiver

Henrik Ohlsson and Lars Wanhammar

Department of Electrical Engineering, Linköping University SE-581 83 Linköping, Sweden
email: {henriko, larsw}@isy.liu.se

Abstract

In a conventional radar receiver errors are introduced due to mismatch between the ADCs and gain and phase mismatch in the I/Q demodulation. By performing the I/Q demodulation in the digital domain some of these errors are alleviated. In this paper several different filter structures suitable for digital I/Q demodulation are evaluated. Three different structures for such a digital down converter has been considered and their implementation properties has been examined. A solution combining an FIR filter with a wave digital filter is found to be the most efficient solution.

1 Introduction

The increased performance of analog-to-digital converters (ADCs) allows the digital signal processing to be moved closer to the antenna in transmitters and receivers. An example of a task suitable for digital signal processing is the I/Q demodulation in a radar receiver. Such a radar receiver is considered in this paper.

The radar receiver considered in this paper is to be used in an array antenna radar system and it is developed in cooperation between FOI and Linköping University. A receiver based on the same concept has previously been designed and implemented at FOI in Linköping [1].

1.1 Motivation

In conventional radar receivers the I/Q demodulation is performed in the analog domain, as shown in Figure 1. The mixers for the I and Q channel works at the same frequency, LO2, but with a phase difference of 90 degrees between them. The two channels should have matched gains between them as well. If these two requirements is not fulfilled for all frequencies within the radar bandwidth the dynamic range of the receiver is reduced. Also, false targets may appear due to phase and gain mismatch. Another drawback with the conventional receiver is that two ADCs, one in each channel, is required. Mismatch between these ADCs will reduce the receiver performance further.

By performing the I/Q demodulation in the digital domain these mixers are not needed and phase and gain mismatch between the channels can be avoided. Also only one ADC is needed. Thus, the errors due to ADC mismatch are avoided. However, this ADC works at twice the sample rate of the ADCs for the conventional structure, This makes the ADC a critical component. The requirements on the digital filters are severe with respect to power consumption and throughput when the sample rate is high.

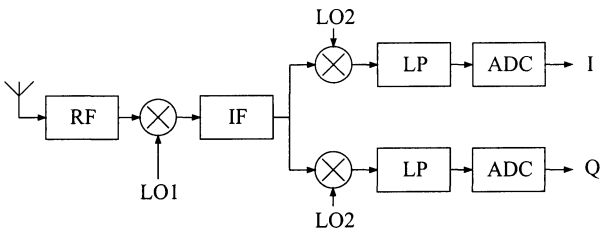


Figure 1: *Conventional receiver structure.*

1.2 Considered Receiver Structure

The digital I/Q demodulation result in a receiver structure with a RF part with only one mixer stage. This mixer down converts the signal to IF. The signal is then bandpass sampled by the ADC at IF and the I/Q demodulation is performed in the digital domain [2]. Such receiver structure is shown in Figure 2. The RF and IF blocks in the figure consists of amplifiers and filters.

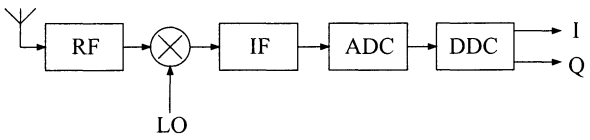


Figure 2: *Receiver structure with digital I/Q demodulation.*

The input signal to the receiver is in the range 8-12 GHz. The IF is 360 MHz and the sample frequency of the ADC is 160 MHz. After the I/Q demodulation the sample frequency is 40 MHz with a signal bandwidth of 36 MHz in each channel.

1.3 Software Receiver Model

A software model of the receiver structure has been developed in cooperation between Saab Bofors Dynamics, Eric-

sson Microwave, FOI, and Linköping University [3]. The purpose with the model is to evaluate the performance of the receiver as well as for identification of possible design trade-offs between the RF part, the ADC, and the digital filters. For the evaluation of the receiver a chirp pulse generator has been implemented in MATLAB and incooperated in the model [4]. The chirp pulse is a typical radar signal which is used in high resolution radars.

The receiver model has been implemented using the Agilent tool ADS™. For the RF part ADS components has been used while the ADC and the digital filters are modelled in MATLAB.

2 The Digital Down Converter

The digital down converter (DDC) considered here is based on a Hilbert transform for I/Q demodulation [5]. The quadrature component is obtained by applying a Hilbert transform on the input signal. This corresponds to a frequency shift of $\pi/2$ [6] [7]. The inphase component is obtained by delaying the input signal. This delay is matched to the delay of the Hilbert transform. In this stage a decimation of the signal sample frequency with a factor two is performed as well.

For the frequency shift of the signal to baseband a highpass filter and a modulator is used. This includes a decimation of the sample rate with a further factor of two. The structure of the considered DDC is shown in Figure 3. A DDC using FIR filters has previously been implemented in full custom layout, but with more relaxed requirements [8] [9].

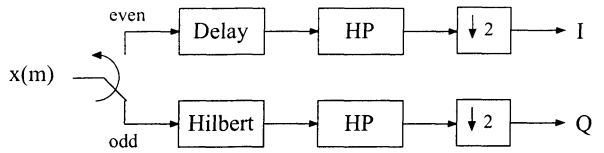


Figure 3: *Structure of the DDC.*

An important issue for the DDC implementation is the power consumption. In a multiple antenna radar system a large number of antennas and receivers are integrated in a small space. To avoid problems due to heat dissipation components with low power consumption should be used.

2.1 Lattice Wave Digital Filters

Lattice wave digital filters (LWDFs) is a class of digital filters that have several advantages compared to FIR filters [10]. For narrow transition band specifications the arithmetic complexity for a LWDF is reduced compared to an FIR filter meeting the same specification. Also, the group delay for a LWDF is lower than for a corresponding FIR filter. However, LWDFs can only be realised with approximately linear phase as opposed to FIR filters which can be designed with exact linear phase. Also, LWDFs are regular structures which makes them suitable for VLSI implementation.

Since LWDFs are recursive structures there is a bound on the maximal sample frequency [11]. This bound is given by

$$f_{s,max} = \frac{1}{T_{min}} = \min_i \left\{ \frac{N_i}{T_i} \right\} \quad (1)$$

where $f_{s,max}$ is the maximal sample rate, T_{min} is the iteration period bound, N_i is the number of delay elements in loop i, and T_i is the total latency in loop i. The loop yielding $f_{s,max}$ is called the critical loop.

By increasing the maximal sample rate for an algorithm the power consumption can be reduced since any excess speed can be traded for reduced power consumption through power supply voltage scaling [12].

When implementing interpolators and decimators with factors of two bireciprocal LWDFs, or half-band filters, are useful. Such filters are anti-symmetric around $\pi/2$ and has a reduced arithmetic complexity compared to LWDFs since every second adaptor coefficient is zero. A bireciprocal LWDF can be implemented using first-order allpass sections only. Such a filter section is shown in Figure 4 with the critical loop indicated. The sample rate bound for the first-order allpass section is

$$f_{s,max} = \frac{1}{T_{min}} = \frac{1}{2T_{add} + T_{mult}} \quad (2)$$

where T_{add} is the latency of an addition and T_{mult} is the latency of the multiplication.

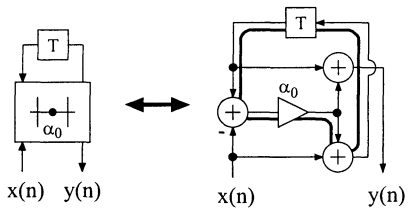


Figure 4: *A first order allpass section.*

3 Design of the DDC

Three different approaches for realization of the DDC has been considered [13]. The first case uses FIR filters for both stages. For the second case an FIR is considered for the Hilbert transform and a bireciprocal LWDF is used for the highpass filter. Finally, for the third case, bireciprocal LWDFs is considered for both stages. For all three cases the data word length is 12 bit.

The timing requirements differ between the Hilbert transform and the highpass filter. The former has a sample frequency of 80 MHz while the sample frequency for the later is 40 MHz.

3.1 FIR-FIR Solution

For the first case, using two FIR filters, we require a sixteenth-order filter for the Hilbert transform. The higpass filter is a 66:th-order filter. To obtain an efficient imple-

mentation of these filters they has been designed for a minimal number of nonzero bits in the coefficients using the method presented in [14].

3.2 FIR-WDF Solution

For the second case the highpass filter is implemented with an eleventh-order bireciprocal LWDF instead of an FIR filter. The Hilbert transform is implemented using the same FIR filter as in the first case. The coefficients for the bireciprocal LWDF has been rounded to 11 bit to meet the specification.

3.3 WDF-WDF Solution

For the third case bireciprocal LWDFs are used for both filter stages. The Hilbert transform require a seventh-order filter and the highpass filter is the same as for the second case. The coefficients for the Hilbert transform has been rounded to 9 bit to meet the specification.

4 Evaluation of the DDC

An estimation of the arithmetic complexity of the three solutions can be obtained by comparing the number of arithmetic operations required. Such comparison is shown in Table 1. These measures are, however, not good estimates of the resources needed for a fixed, algorithm-specific implementation of the filters. A more accurate estimation is possible if the hardware requirements for each multiplication is considered as well. Then the total number of adders needed can be computed. To do this we need to consider the algorithms used for the filters as well as the architecture to be used for the implementation.

The WDFs are implemented by an isomorphic mapping of the structures to bit-parallel, carry save arithmetic [15] [16]. The FIR filters are linear phase filters and are implemented using the direct form algorithm. The multiplications in the FIR filter is merged to one adder tree and mapped to carry save adders.

Case	Multiplications	Additions
FIR - FIR	41	73
FIR - WDF	15	39
WDF - WDF	13	41

Table 1: *Arithmetic operations required.*

Case	CSA	CPA	Memory elements
FIR-FIR	84	39	108
FIR-WDF	102	15	30
WDF-WDF	122	13	13

Table 2: *Hardware components required.*

Case	Area (μm^2)
FIR-FIR	442896
FIR-WDF	331337
WDF-WDF	395562

Table 3: *Area results from logic synthesis.*

Case	$f_{sHil,max}$ (MHz)	$f_{sHP,max}$ (MHz)
FIR-FIR	128	81
FIR-WDF	128	75
WDF-WDF	86	75

Table 4: *Timing results from logic synthesis.*

The result of the comparison when the architecture is considered is shown in Table 2, where CSA stands for carry-save adders and CPA stands for carry propagation adders. As can be seen the case with two FIR filters requires many memory elements as well as a large number of CPAs. On the other hand the realizations which includes LWDFs requires a larger amount of carry save adders. However, the complexity of a CSA is always the same, or lower than for a CPA. Thus, to evaluate the complexity of the different realizations, implementations of all three cases should be considered.

5 DDC Implementation

The three different solutions has been implemented in structural VHDL. The filters has been mapped to a 0.18 µm CMOS process from STMicroelectronics using the Synopsys tool Design Compiler™. In Table 3 the area requirement for the three solutions after the logical synthesis is shown. Table 4 shows the maximal sample frequency for the three cases using a power supply voltage of 1.55 V. As can be seen the maximal sample frequency for the FIR filters and the WDF highpass filter is significantly larger than the required sample frequency. Thus, the power supply voltage may be scaled to reduce the power consumption for these two cases.

The result from the logic synthesis is that the FIR-WDF case is the most area efficient solution with an area of about 75% of the area for the worst case, the FIR-FIR solution. Also, all three solutions meets the timing requirements. However, the WDF-WDF does not have as large margin for power supply voltage scaling as the two other cases. Thus, the results from the logic synthesis is that the FIR-WDF solution is the most efficient for implementation.

However, it is possible to improve the maximum sample frequency of the WDF and thereby reduce the power consumption further. This can be done by arithmetic transformations of the algorithm [17] [18] [19]. For a bitparal-

lel, carry save arithmetic implementation the maximal sample frequency can be increased by 20% to 30%, depending on the number of nonzero bits in the coefficients. This is at the expense of a small increase of the area required for the transformed allpass section.

6 Conclusions

In this paper we discuss the implementation of a digital down converter for a radar receiver. The purpose with the digital down converter is to remove errors due to phase and gain mismatch in the I/Q demodulation as well as mismatch between the ADCs. By performing the I/Q demodulation in the digital domain some of these errors are alleviated. We propose three different filter structures and compare their implementation properties. The results from the logic synthesis give that the FIR-WDF case is the most efficient for implementation.

References

[1] L. Petersson, M. Danestig, and U. Sjöström, "An experimental S-band digital beamforming antenna," in Proc. Symp. Phased Array Syst. and Tech., Boston, USA, Oct 1996, pp. 93–98.

[2] W. M. Waters and B. R. Jarret, "Bandpass signal sampling and coherent detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol 18, pp. 731-736, 1982.

[3] A. Gustafsson, K. Folkesson, and H. Ohlsson, "A simulation environment for integrated frequency and time domain simulations of a radar receiver," in Proc. Gigahertz 2001 Symposium, Lund, Sweden, Nov. 26–27, 2001.

[4] R. Standert, Software model of a radar receiver, Royal Institute of Technology IR-SB-EX-0202, Feb. 2002.

[5] D. W. Rice and K. H. Wu, "Quadrature sampling with high dynamic range," *IEEE Trans. Aerospace Elec. Syst.*, vol 18, pp. 736-739, 1982.

[6] S. K. Mitra and J. F. Kaiser, *Handbook for Digital Signal Processing*, John Wiley and Sons, 1993

[7] H. Johansson and L. Wanhammar, "Digital Hilbert transformers composed of identical allpass subfilters," in Proc. IEE Int. Symp. Circuits Syst., Monetrey, USA, May 31–June 3, 1998.

[8] U. Sjöström, M. Karlsson, and M. Hörlin, "A digital down converter chip," in Proc. European Signal Processing Conf., Vol 1, pp. 284–287, Trieste, Italy, Sept., 1996.

[9] Karlsson M.: *Distributed Arithmetic: Design and Applications*, Thesis No. 696, Linköping University, Sweden, 1998

[10] A. Fettweis, "Wave digital filters: Theory and practice," *Proc. IEEE*, vol. 74, no2, pp. 270–327, Feb. 1986.

[11] M. Renfors and Y. Neuvo, "The maximum sampling rate of digital filters under hardware speed constraints," *IEEE Trans. Circuits Syst.–II*, vol. 28, pp. 196–202, 1981.

[12] A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE*, vol. 83, no. 4, pp. 498–523, Apr. 1995.

[13] H. Ohlsson, H. Johansson, and L. Wanhammar, "Design of a digital down converter using high speed digital filters," in Proc Gigahertz 2000 Symposium, Gothenburg, Sweden, March 13–14, 2000, pp. 309–312.

[14] O. Gustafsson, H. Johansson, and Lars Wanhammar, "An MILP approach for the design of linear-phase FIR filters with minimum number of signed-power-of-two terms," in Proc. European Conf. Circuit Theory Design, Espoo, Finland, Aug. 28-31, 2001.

[15] L. Wanhammar, *DSP Integrated Circuits*, Academic Press, 1999.

[16] T. G. Noll, "Carry save architectures for high-speed digital signal processing," *J. VLSI Signal Processing*, vol. 3, pp. 121–140, 1991.

[17] H. Ohlsson, O. Gustafsson, and L. Wanhammar, "Arithmetic transformations for increased maximal sample rate of bit-parallel bireciprocal lattice wave digital filters," Proc. IEEE Symp. Circuits Syst., Sydney, Australia, May 6–9, 2001, vol. 2, pp. 825–828.

[18] H. Ohlsson, O. Gustafsson, H. Johansson, and Lars Wanhammar, "Implementation of bit-parallel wave digital filters with increased maximal sample rate," in Proc. IEEE Int. Conf. Elec. Circuits Syst., Malta, Sep. 6-9, 2001, vol. 1, pp. 71–74.

[19] K. Palmkvist, M. Vesterbacka, and L. Wanhammar, "Arithmetic transformations for fast bit-serial VLSI implementations of recursive algorithms," *Proc. Nordic Signal Processing Symp.*, Espoo, Finland, Sept. 24-27, 1996, pp. 391–394.

AN EFFICIENT PIPELINED COMPLEX MULTIPLIER

Weidong Li, Shengxian Zhuang, and Lars Wanhammar

Department of Electrical Engineering
SE-581 83 Linköping, Sweden
E-mail: {weidongl, zhuangsx, larsw}@isy.liu.se

ABSTRACT

In this paper, we propose a pipeline scheme for efficient realization of a complex multiplier using distributed arithmetic. The pipelined multiplier consists of one conventional multiplier that is multiplexed and some small additional circuitry on the boundary. The proposed scheme reduces the chip area as well as the interconnections by nearly half compared to a conventional complex multiplier. The pipelined multiplier is efficient in terms of chip area and interconnect, which reduces the power consumption.

1. INTRODUCTION

Arithmetic operations with complex numbers are required in many DSP algorithms, e.g., in digital filters and FFT. The complex multiplication is often the most expensive arithmetic operation and one of the dominating factors in determining the performance in terms of speed and power consumption. Therefore, the design of an efficient complex multiplier is important.

Straightforward implementation of complex multiplication requires four real multiplications, one addition and one subtraction. However, the number of multiplications can be reduced to three with a transformation that uses extra pre and post additions [1]. A more efficient way to reduce the cost of multiplication is to use distributed arithmetic [2]. Distributed arithmetic is based on addition of precomputed sums of coefficient values. With this algorithm, the complexity of complex multiplication corresponds to two real multiplications.

In this paper, we present an efficient pipelined complex multiplier and its implementation. By adding some logical control circuitry on the boundary, the partial product generator is shared by both the real part and the imaginary part. This leads to area efficiency and has little interconnect, which tends to reduce the power consumption.

In the following section, we first review the principles of distributed arithmetic using offset binary coding and show how it can be used to realize a complex multiplier [3]. Then in section 3, we discuss the pipelined complex multiplier. In section 4, the detailed implementation is presented. Finally, conclusions are given in section 5.

2. DISTRIBUTED ARITHMETIC

Distributed arithmetic uses precomputed partial sums for an efficient computation of inner products of a constant vector and a variable vector.

Let $C_R + jC_I$ and $X_R + jX_I$ be two complex numbers of which $C_R + jC_I$ is the coefficient and $X_R + jX_I$ is the variable complex number. In the case of a complex multiplication, we have

$$Z_R + jZ_I = (C_R + jC_I)(X_R + jX_I) = \\ = (C_R X_R - C_I X_I) + j(C_R X_I + C_I X_R) \quad (1)$$

Hence, a complex multiplication can be considered as two inner products of two vectors with two elements. We will realize the real and imaginary parts separately.

For the sake of simplicity we will only consider the first inner product equation (1), i.e., the real part. The complex coefficient, $C_R + jC_I$, is assumed to be fixed and two's-complement representation is used for both coefficient and data. The data is scaled so that $|Z_R + jZ_I| < 1$. The inner product can be rewritten

$$Z_R = C_R\left[-x_{R0} + \sum_{k=1}^{W_d-1} x_{Rk}2^{-k}\right] - \\ -C_I\left[-x_{I0} + \sum_{k=1}^{W_d-1} x_{Ik}2^{-k}\right]$$

where x_{Rk} and x_{Ik} are the k:th bits in the real and imaginary parts, respectively. By interchanging the order of the two summations we get

$$Z_R = -C_R x_{R0} + C_I x_{I0} + \\ + \sum_{k=1}^{W_d-1} (C_R x_{Rk} - C_I x_{Ik})2^{-k}$$

which can be written

$$Z_R = -F_k(x_{R0}, x_{I0}) + \sum_{k=1}^{W_d-1} F_k(x_{Rk}, x_{Ik})2^{-k} \quad (2)$$

where

$$F_k(x_{Rk}, x_{Ik}) = C_R x_{Rk} - C_I x_{Ik} \quad (3)$$

F_k is a function of two binary variables, i.e., the k:th bits in X_R and X_I. Since F_k can take on only four values, it can be computed and stored in a look-up table.

In the same way we get the corresponding binary function for the imaginary part as

$$G_k(x_{Rk}, x_{Ik}) = C_R x_{Ik} + C_I x_{Rk}$$

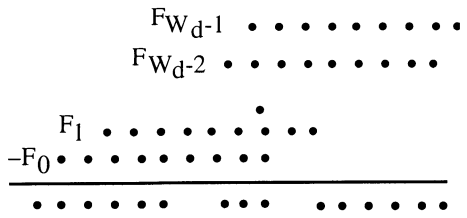


Figure 1. Bit-products in Eq. (2) to be added.

Equation (2) can be illustrated as shown below.Note the similarity between equation (2) and the corresponding expression for a conventional real multiplier in which the functions F_k (G_k) are bit-products between the coefficient bits and the k:th data bit. The bit-products can be stored in a look-up table with only four words as illustrated in Figure 2. Hence, equation (2) and its imaginary counterpart can be realized with any real multiplication algorithm where the bit-products are replaced with the corresponding bits from the functions F_k and G_k. For example, bit-parallel array or tree multiplier structures [1], or bit-serial multipliers [3 - 4] can be used. Each inner product can be implemented directly with an addition tree and precomputed partial products.

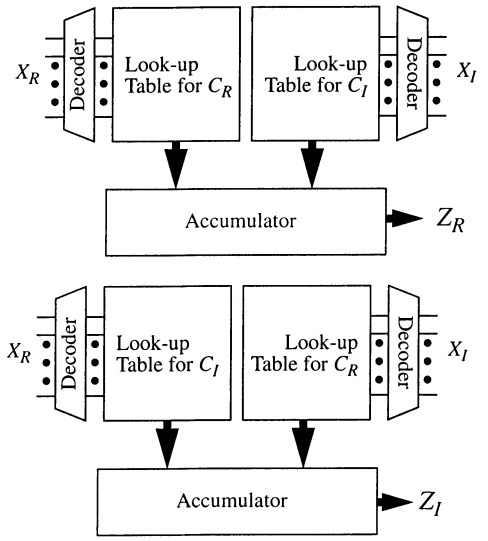


Figure 2. Complex multiplier using distributed arithmetic.

2.1 Complex Multiplier with Offset Binary Coding

The offset binary coding can be applied to complex multiplication. Without any loss of generality, we assume that the magnitudes of C_R, C_I, X_R, and X_I all are less than 1. The word length of X_R, and X_I is W_d. Then, the complex multiplication can be written

$$Z_R + jZ_I = (C_R X_R - C_I X_I) + j(C_R X_I + C_I X_R)$$

$$= \left\{ -C_R(x_{R0} - \overline{x_{R0}})2^{-1} + \sum_{i=1}^{W_d-1} C_R(x_{Ri} - \overline{x_{Ri}})2^{-i-1} - C_R 2^{-W_d} \right\}$$

$$- \left\{ -C_I(x_{I0} - \overline{x_{I0}})2^{-1} + \sum_{i=1}^{W_d-1} C_I(x_{Ii} - \overline{x_{Ii}})2^{-i-1} - C_I 2^{-W_d} \right\}$$

$$+ j\left\{ -C_R(x_{I0} - \overline{x_{I0}})2^{-1} + \sum_{i=1}^{W_d-1} C_R(x_{Ii} - \overline{x_{Ii}})2^{-i-1} - C_R 2^{-W_d} \right\}$$

$$+ j\left\{ -C_I(x_{R0} - \overline{x_{R0}})2^{-1} + \sum_{i=1}^{W_d-1} C_I(x_{Ri} - \overline{x_{Ri}})2^{-i-1} - C_I 2^{-W_d} \right\}$$

$$= \left\{ -F(x_{R0}, x_{I0}) + \sum_{i=1}^{W_d-1} F(x_{Ri}, x_{Ii})2^{-i-1} + F(0,0)2^{-W_d} \right\}$$

$$+ j\left\{ -G(x_{R0}, x_{I0}) + \sum_{i=1}^{W_d-1} G(x_{Ri}, x_{Ii})2^{-i-1} + G(0,0)2^{-W_d} \right\} \quad (4)$$

The functions F and G can be expressed as follows.

$$F(x_{Ri}, x_{Ri}) = C_R(x_{Ri} - \overline{x_{Ri}}) - C_I(x_{Ii} - \overline{x_{Ii}}) \quad (5)$$

$$G(x_{Ri}, x_{Ri}) = C_I(x_{Ri} - \overline{x_{Ri}}) + C_R(x_{Ii} - \overline{x_{Ii}}) \quad (6)$$

In equation (6), $(x_i - \overline{x_i})/2$ is either 1 or –1. Hence the partial product, i.e., function F_k (G_k), for each bit is of the form $C_R \pm C_I$. All possible partial products are tabulated in Table 1. Obviously, only two coefficients, i.e., $-(C_R - C_I)$ and $-(C_R + C_I)$, are sufficient to store since $(C_R - C_I)$ and $(C_R + C_I)$ can easily be generated from the two former coefficients by inverting all bits and adding 1 in the least-significant position.

x_{Ri}	x_{Ii}	$F(x_{Ri}, x_{Ii})$	$G(x_{Ri}, x_{Ii})$
0	0	$-(C_R - C_I)$	$-(C_R + C_I)$
0	1	$-(C_R + C_I)$	$(C_R - C_I)$
1	0	$(C_R + C_I)$	$-(C_R - C_I)$
1	1	$(C_R - C_I)$	$(C_R + C_I)$

Table 1: Partial Products.

The complex multiplier with distributed arithmetic is illustrated in Figure 3. The accumulators, which adds the

partial products, are the same as in a real multiplication. The partial product generation is only slightly more complicated than for a real multiplier. Hence, the complexity of the complex multiplier in term of chip area corresponds to approximately two real multipliers.

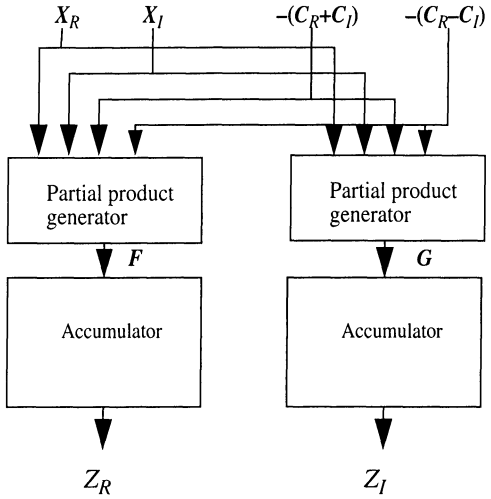


Figure 3. Complex multiplier using DA.

The complex multiplier requires that the coefficients, $-(C_R - C_I)$ and $-(C_R + C_I)$, are precomputed. However, for many applications, like the FFT, the coefficients are known in advance and can be stored in ROM, which simply changes the content of the coefficient ROM from storing C_R and C_I to storing $-(C_R - C_I)$ and $-(C_R + C_I)$.

3. PIPELINED COMPLEX MULTIPLIER

As discussed in section 2, a complex multiplier with distributed arithmetic has approximately the same area complexity as two real multipliers. However, the implementation of the complex multiplier (accumulator) is still a difficult task due to the complex routing.

In order to reduce the complexity of the wire routing, we propose the use of a pipelined multiplier. The pipelined multiplier is based on the fact that the real part and imaginary part are identical in circuit realization except the partial product generation.

3.1 Pipelining of the Multiplier

To obtain a pipelining scheme for the complex multiplier, the partial product generation for the real and imaginary parts has to be merged. For the sake of simplicity, we define $S = -(C_R - C_I)$ and $T = -(C_R + C_I)$ and for the j:th bit, $S_j = -(C_{Rj} - C_{Ij})$ and $T_j = -(C_{Rj} + C_{Ij})$.

Consider now the bit partial product for x_{Ri}, x_{Ii}, s_j, and t_j. Let $pr_{i,j}$ be the real partial product and $pi_{i,j}$ be the imaginary partial product. The partial products can be obtained from the following equations.

$$pr_{i,j} = C_{Rj}(x_{Ri} - \overline{x_{Ri}}) - C_{Ij}(x_{Ii} - \overline{x_{Ii}})$$

$$pi_{i,j} = C_{Ij}(x_{Ri} - \overline{x_{Ri}}) + C_{Rj}(x_{Ii} - \overline{x_{Ii}})$$

From Table 1, we see that all of possible partial products can be expressed with

$$pr_{i,j}(x_{Ri}, x_{Ii}, s_j, t_j) = x_{Ri} \oplus ((x_{Ri} \oplus x_{Ii})t_j + \overline{(x_{Ri} \oplus x_{Ii})}s_j)$$

$$pi_{i,j}(x_{Ri}, x_{Ii}, s_j, t_j) = x_{Ii} \oplus ((x_{Ri} \oplus x_{Ii})s_j + \overline{(x_{Ri} \oplus x_{Ii})}t_j)$$

Defining $mux(a, b, m) = ma + \bar{m}b$, the above equations can be rewritten as

$$pr_{i,j}(X_{Ri}, X_{Ii}, s_j, t_j) = X_{Ri} \oplus mux(t_j, s_j, X_{Ri} \oplus X_{Ii})$$

$$pi_{i,j}(X_{Ri}, X_{Ii}, s_j, t_j) = X_{Ii} \oplus mux(s_j, t_j, X_{Ri} \oplus X_{Ii})$$

By introducing an extra control signal m, the partial product generation can be expressed by

$$pp_{i,j} = mux(x_{Ri}, X_{Ii}, m) \oplus mux(mux(t_j, s_j, m), mux(s_j, t_j, m), x_{Ri} \oplus x_{Ii})$$

where $pr_{i,j} = pp_{i,j}$ when $m = 1$, and $pi_{i,j} = pp_{i,j}$ when $m = 0$. Hence the computation of the real product can be changed to the computation of the imaginary product by a simple multiplexing scheme.

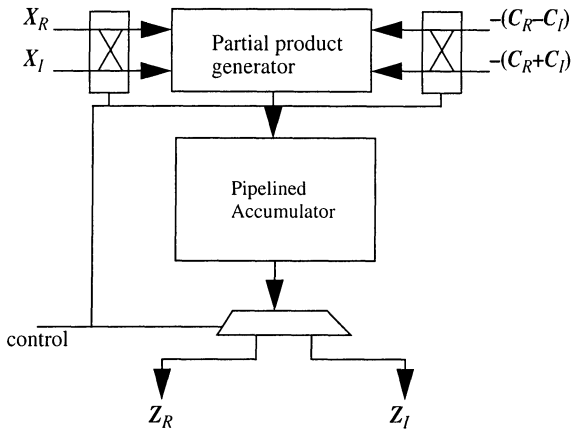


Figure 4. Pipelined complex multiplier.

3.2 Multiplexed Accumulator

The realization of the pipelined complex multiplier is illustrated in Figure 4. The pipelined complex multiplier uses a multiplexed pipelined accumulator and requires only multiplexers and demultiplexers at the inputs and outputs. The cost in terms of chip area for this complex multiplier is therefore about half of a direct implementation with distributed arithmetic. We can simply use the clock as control signal and add pipeline stages with

latches.

The routing is also simplified since there is only one of two parts to be implemented. The disadvantage of the pipelined multiplier is the computation time which is twice as long as the non-pipelined one.

4. IMPLEMENTATION

In this section, we discuss the parallel implementation for the complex multipliers. There are two main blocks for the complex multiplier: the partial product generation block, and the accumulation block. These two blocks will be studied in more detail.

4.1 Partial Product Generation Block

For the partial product generation, the number of partial products has not only direct impact to the partial product generation block but also the accumulation block. The sign-extension is a key issue to reduce the number of partial products. Hence it is important to use an efficient sign-extension scheme.

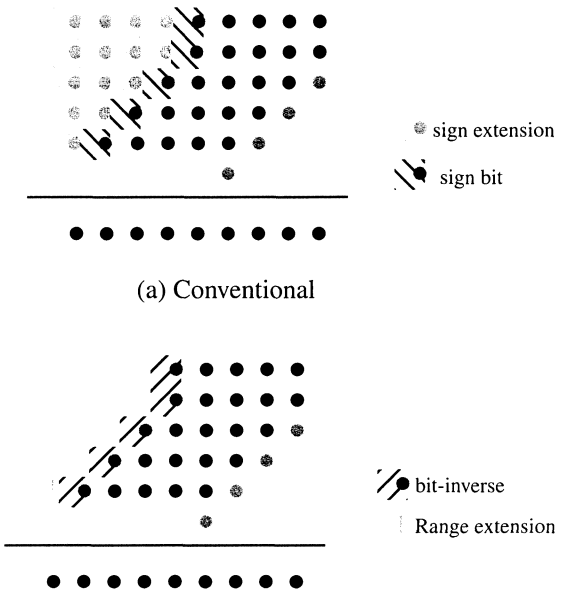


Figure 5. Sign extension schemes.

For the conventional approach, the sign-extension is done by extending the sign-bit to the most-significant bit (MSB) of the complex multiplier as shown in Figure 5 (a). For an $m \times n$ complex multiplier, the number of partial products is $2(m + 1)(n + m) - (m - 2)(m - 1) + 2$ with the conventional sign-extension scheme.

The sign-extension method in our implementation is similar to the method used in real multiplication, which is proposed in [5]. The sign-extension is accomplished by inverting the sign bit of each coefficient and adding 1 in the MSB to handle the range extension. This is illustrated in Figure 5 (b). The total number of partial product is $2(m + 1)(n + 1)$ for an $m \times n$ complex multiplier. A reduction with $m^2 + 3m - 2$ partial products is achieved

4.2 Accumulator

The accumulator adds all generated partial products. The partial product addition can be implemented with various types, like array [5], Wallace tree [6], etc. The realisation with array structure has regular in layout but is slow. The realization with Wallace tree structure is fast, but lack regularity. The realization with 4:2 tree [7 - 8] has a good trade-off for layout regularity and speed. Therefore we select the 4:2 tree for the accumulator implementation.

In full custom design, the complex multiplier are often implemented with real-imaginary-alternative (RIA) structure [9 - 11]. The RIA structure merges the real and imaginary computation in a block. However, it also introduces extra feedthroughs between two real or imaginary parts, which in turn increases the parasitic capacitances. With proposed pipelined complex multiplier, the real and imaginary part are reduced to only one part and hence there is no extra feedthroughs and the parasitic capacitances can be kept low.

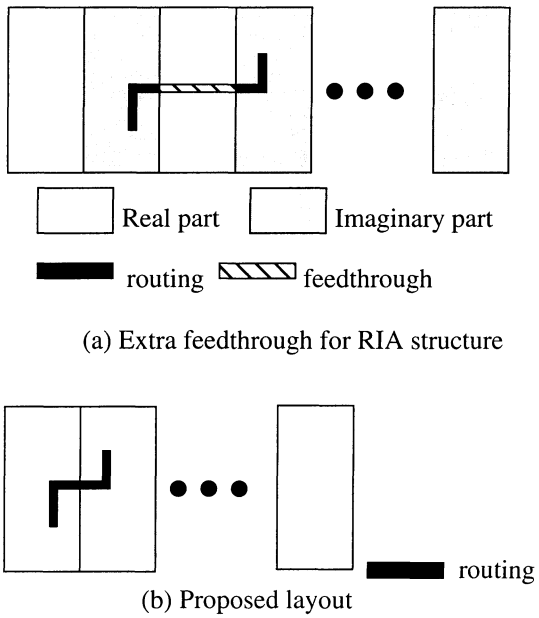


Figure 6. Parasitic capacitance reduction for full custom layout.

4.3 Evaluation

In order to confirm the proposed principle, an 17 × 13 complex multiplier has been implemented with semi-custom design method. We implemented the complex multipliers with both the proposed structure and the conventional stucture. The proposed multiplier uses time-multiplexing to compute real and imaginary part without changing of the accumulator structure. The accumulators for both multipliers use a 4:2 tree to summarize the partial

products. Both multipliers are written with structural VHDL and synthesized using standard triple metal 0.35 μm CMOS technology. The place and route is performed with *Silicon Ensemble* from Cadence. We consider only the interconnections within the multiplier cores and exclude the I/O interconnections. The result is shown in the Table 2. Even though the automatic place and route tool cannot retain the structure, we still find that the average parasitic capacitance is reduced with the proposed structure.

We cannot simulate the power consumption due to the lack of backannotation. However, there are two factors which contribute to the reduction of power cunsumption in the proposed complex multiplier. One is the reduction on parasitic capacitances as shown in Table 2. The other is the reduction of glitches due to the proposed scheme applies more pipelining to the complex multiplier.

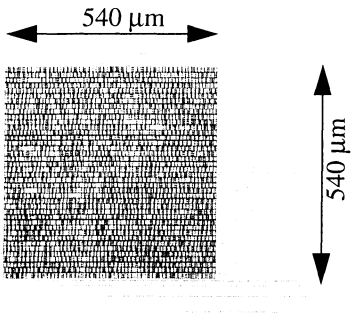


Figure 7. Layout of the pipelined complex multiplier.

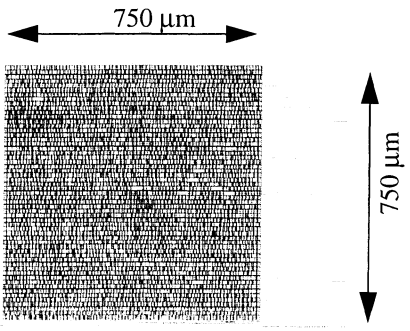


Figure 8. Layout of the conventional multiplier.

	Conventional Real Multiplier	Pipelined Complex Multiplier
Area	0.56 mm^2	0.28 mm^2
Average interconnection parasitic capacitance	5.80 fF	5.61 fF

Table 2: Comparison of conventional complex multiplier and pipelined complex multiplier.

5. CONCLUSIONS

In this paper we present an efficient pipelined complex multiplier based on distributed arithmetic and compare it with conventional complex multiplier. The proposed pipelined complex multiplier has less area, routing complexity and power consumption than that of the conventional one.

Full custom implementation and evaluation of the proposed pipelined complex multiplier would be the future works.

ACCKNOWLEDGEMENT

The authors would like to thank Henrik Ohlsson for supports with the logic synthesis. The project is financially supported by SSF, the Swedish Foundation for Strategic Research.

REFERENCES

[1] I. Koren, *Computer Arithmetic Algorithms*, Prentice Hall, 1993.

[2] A. Croisier, D. J. Esteban, M. E. Levilion, and V. Rizo, *Digital Filter for PCM Encoded Signals*, U. S. Patent 3777130, Dec., 1973.

[3] L. Wanhammar, *DSP Integrated Circuits*, Academic Press, 1999.

[4] J. Melander, T. Widhe, P. Sandberg, K. Palmkvist, M. Vesterbacka, and L. Wanhammar, "Implementation of a Bit-Serial FFT Processor with a Hierachical Control Structure," *Proceedings of European Conf. on Circuit Theory and Design (ECCTD)*, Vol. 1, pp. 423-426, Turkey, Aug. 1995.

[5] C. R. Baugh and B. A. Wooley, "A two's complement parallel array multiplication algorithm," *IEEE Trans. on Computers*, C-22, pp. 1045-1047, Dec., 1973.

[6] C. S. Wallace, "A suggestion for a fast multiplier," *IEEE trans. on Electronic Computers*, EC-13, pp. 14-17, Feb., 1964.

[7] W. Li, J. B. Burr, A. Peterson, "A fully parallel VLSI implementation of distributed arithmetic," *Proceedings of IEEE Int. Conf. on Circuits and Systems*, pp. 1511-1515, Finland, 1988.

[8] M. Santoro and M. Horowitz, "SPIM: Apipelined 64x64-bit iterative multiplier," *IEEE J. Solid-State Circuits*, pp. 487-493, April, 1989.

[9] Y. Chang and K. K. Parhi, "High-performance digit-serial complex-number multiplier-accumulator," *Proc. of Int. Conf. on Computer Design (ICCD)*, pp. 211-213, USA, Oct., 1998.

[10] T. Aoki, Y. Ohi, and T. Higuchi, "Redundant complex number arithmetic for high-speed signal processing," *Proc. of IEEE Workshop on VLSI Signal Processing*, pp. 523-531, Japan, Oct., 1995.

[11] T. Aoki, H. Amada, and T. Higuchi, "Real/complex reconfigurable arithmetic using redundant complex number systems," *Proc. of IEEE Symp. on Computer Arithmetic*, pp. 200-207, USA, Jul., 1997.

Analog-to-digital Converter Error Correction using Frequency Selective Tables

Henrik Lundin[1], Tomas Andersson[2], Mikael Skoglund and Peter Händel

Department of Signals, Sensors and Systems, Royal Institute of Technology, Stockholm, Sweden

[1]E-mail: henrik.lundin@s3.kth.se

[2]E-mail: tomas.andersson@s3.kth.se

Abstract

Error correction for analog-to-digital converters (ADCs) is considered. The frequency-dependent nature of ADC errors motivates the proposal of a novel scheme, incorporating look-up table correction and fast frequency estimation. The method is evaluated using experimental converter data, and the performance, measured in SFDR and SINAD, is found to be superior to that of non frequency-dependent correction methods.

1 Introduction

The demand for highly linear analog-to-digital converters (ADC) is ever increasing. It is a well-known fact that practical ADCs suffer from various errors, e.g., gain, offset and linearity errors. These errors stem from numerous sources such as non-ideal spacing of transition levels and timing jitter, to mention a few, and they contribute to deterioration of the linearity of the converter. Several methods have been proposed to *externally* compensate for such errors, e.g., [1–4]. External in this case implies that digital signal processing methods which operate *outside* of the actual converter are used in the calibration and compensation schemes.

Throughout this paper, we will consider a b-bit ADC. The ADC will be modeled as an ideal sample-and-hold circuit followed by an imperfect quantizer, depicted in Figure 1. The sample-and-hold circuit samples the continuous-time input signal $s(t)$ at the sampling rate f_s, resulting in a discrete-time signal

$$s(n) = s(t)\big|_{t=n/f_s}. \tag{1}$$

The Q-block of Figure 1 then quantizes $s(n)$ into one of the $M = 2^b$ output states $\{x_j\}$, $j = 0, \ldots, M-1$, and produces the corresponding output $x(n) = x_j$.

One frequently used method to correct ADCs is the look-up table correction. In classic look-up table correction, the correction table (containing the corrected output $\hat{s}_j$ associated with each ADC output state x_j) is addressed using the present ADC output sample, $x(n)$. Obviously, this addressing yields the same correction for a given ADC output sample, regardless of the dynamic properties of the input signal. This is referred to as *static* correction. However, the errors of an ADC are in general frequency dependent. This often results in a severe performance loss for table look-up correction methods when applied at frequencies other than the calibration frequency. In this paper a *frequency-selective correction* scheme is presented and evaluated. The method is based on two key components: a *fast frequency region estimator*, based on [5], and a *correction table*. These are described in the following two sections. Results obtained using experimental ADC data are presented in Section 4.

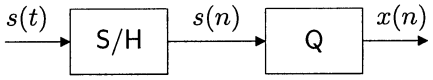


Figure 1: The ADC model. The first block is an ideal sample-and-hold circuit and the second block is an imperfect quantizer.

2 Frequency region estimator

Tone frequency estimation from an N-sequence of noise corrupted data is a well known problem which can almost be considered solved. However, with "almost solved" we refer to the case where the number of data N is large, the signal to noise ratio (SNR) is high, and there exists an infinite amount of computer resources. In the ADC error correcting application this is not the case. The number of data N can not be made large, since we then will miss the information about the instantaneous frequency. In the considered application the SNR is usually high. In terms of computer resources this is an application where almost none are present.

A traditional way of constructing frequency estimators is by optimizing some criterion related to the

frequency. The perhaps most commonly used method is the method of maximum likelihood, or approximate variants thereof [6]. That is, choosing an estimate of the frequency in such a way that the model in use is the most likely given some data samples. In common for most frequency estimation methods is that the output frequency estimate is a continuous variable. Here, on the other hand, we consider the problem of finding the most probable region to hold the unknown frequency, out of a finite (small) set of regions.

Consider the input $s(n)$ to the quantizer in Figure 1 to be modelled as a sine wave and additive Gaussian noise. The input is then given by,

$$s(n) = A\sin(2\pi f_0 n + \phi) + w(n) \tag{2}$$

where $A > 0$ is the real valued amplitude, ϕ is the initial phase, f_0 is the unknown normalized frequency, $0 < f_0 < 1/2$, and $w(k)$ is the noise with variance σ^2. The output from the quantizer $x(n)$ is a b-bit quantized version of $s(n)$. It has been shown [5] that there exists a high-performance frequency estimator of low complexity employing only 1-bit of the input signal. The use of 1-bit data also has the advantage that the estimator does not depend on the power of the input signal, that is no gain control is needed. Here, we are not limited to use 1-bit data but the resulting structure with a table look-up procedure is tractable since it supports the demand of a fast estimator of low complexity.

2.1 Operation

The frequency estimator input $y(n)$ is given by the most significant bit (msb) of $x(n)$,

$$y(n) = \text{sign}(s(n)), \tag{3}$$

where $\text{sign}(x) = 1$ for $x \geq 0$ and $\text{sign}(x) = -1$ for $x < 0$. By collecting N successive binary samples at each time instant n, that is

$$\{y(n), \ldots, y(n-N+1)\}, \tag{4}$$

we can uniquely map one input sequence onto an integer $i \in \{0, \ldots, 2^N - 1\}$. The index i is then used as a pointer to an entry in a *frequency region estimation table*, see Figures 2–3. Finally, the i-th table entry contains a region estimate $\hat{F}(n) \in \{F_1, \ldots, F_K\}$, indicating that the instantaneous signal frequency is within the k-th frequency region. The frequency regions F_k are defined as,

$$F_k = \{f \in [0, 1/2) : \\ |f - f_k| \leq |f - f_l|,\ l = 1, \ldots, K\} \tag{5}$$

where $k = 1, \ldots, K$. In this paper, the frequencies f_l have been chosen equally spaced over the region $[0, 1/2)$, but could be chosen arbitrary over the space of possible input frequencies.

2.2 Design

As a frequency region estimate we choose the region that maximizes the probability of including the unknown frequency f_0, that is

$$\hat{F}(n) = \arg\max_{\forall F_k} \Pr\{f_0 \in F_k | y(n), \ldots, y(n-N+1)\} \tag{6}$$

A straightforward way to obtain the table is to use a training approach [5]. Given a set of data samples $x(n)$ based on different frequencies, within the regions $F_1, \ldots, F_K$, it is possible to build a training set $\mathcal{T} = \{i_l\}_{l=1}^{L}$, where each i_l corresponds to a block of N samples of the msb in $x(n)$. The samples $x(n)$ are generated using an input of a single sinusoid, at a known frequency, disturbed by noise. Hence, to each block i_l there is a corresponding true frequency belonging to one of the regions $f_1, \ldots, f_K$. Now, given a training set $\mathcal{T}$ the i -th table entry can be computed as the index of the most probable frequency region over those i_l corresponding to the index i. For completeness, we let $\hat{f}_k(i) = \lceil K/2 \rceil$ for those i that are not in the training set $\mathcal{T}$.

3 Correction table

Static ADC correction yields the same corrected value $\hat{s}_j$ given the ADC output x_j, regardless of the signal frequency, while the errors sought to mitigate for in general are frequency dependent. The correction scheme presented here utilizes a frequency selective correction table. This is accomplished by extending the usual one-dimensional correction table of classical look-up table compensation to a two-dimensional table, using *both* the present ADC output $x(n) = x_j$ *and* the present frequency region estimate $\hat{F}(n) = F_k$ for addressing. This method can also be interpreted as selecting a specific one-dimensional correction table for each frequency estimate $F_k \in \{F_1, \ldots, F_K\}$. Thus, the corrected output $\hat{s}(n)$ is the table entry $\hat{s}_{j,k}$ associated with x_j and F_k. The correction system has two operation modes, *compensation* and *correction*, which are described in detail below.

3.1 Compensation

In compensation mode, i.e. normal ADC operation with correction engaged, the ADC output sample, $x(n)$, is mapped through the correction table to a compensated output value $\hat{s}(n)$. The correction is determined by the present ADC output together with the current frequency region estimate, as depicted in Figure 2. Thus, the compensation becomes

$$s(t) \rightarrow (x_j,\ F_k) \rightarrow \hat{s}_{j,k} = \hat{s}(n) \tag{7}$$
$$\hat{s}_{j,k} \in \{\hat{s}_{i,\ell}\}_{(i,\ell)=(0,1)}^{(M-1,K)}.$$

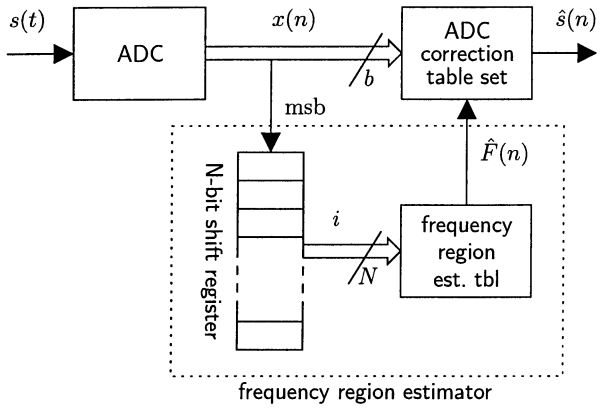


Figure 2: Compensation system outline. The frequency region estimator selects the appropriate ADC correction table.

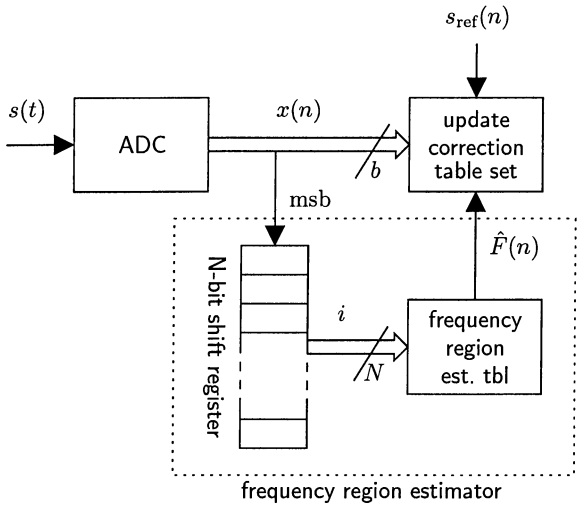


Figure 3: ADC correction table calibration system. The reference signal $s_{\text{ref}}(n)$ is an estimate of the input signal samples $s(n)$.

With this structure, the compensation is made dynamic, with table addressing depending on the frequency contents of the signal.

3.2 Calibration

Prior to using the correction table for compensation, it must be calibrated. Generally, calibration is performed with a calibration signal $s(t)$ applied to the ADC input. The calibration system outline is shown in Figure 3. The table entries $\hat{s}_{j,k}$ should be selected such that the resulting conversion $s(t) \rightarrow (x_j, F_k) \rightarrow \hat{s}_{j,k} = \hat{s}(n)$ is "better" than without correction. The employed design criterion is to minimize the mean squared error, $\mathrm{E}[(\hat{s}(n) - s(n))^2]$, where $\mathrm{E}[\cdot]$ denotes the expected value. Since the selection of $\hat{s}(n) = \hat{s}_{j,k}$ depends on the ADC output $x(n)$ and the frequency region estimate $\hat{F}(n)$, the criterion becomes

$$\hat{s}_{j,k} = \arg\min_{\hat{s}} \mathrm{E}\big[(\hat{s} - s(n))^2 \,\big|\, x(n) = x_j,\, \hat{F}(n) = F_k\big] \tag{8}$$

It can be shown [1, 7] that in order to minimize the criterion (8), $\hat{s}_{j,k}$ should be set to the mean value of all input samples, $s(n)$, that were quantized into the value x_j while the frequency region estimate was equal to F_k. Under the interpretation that the frequency region estimate $\hat{F}(n)$ selects which table, out of a set of one-dimensional correction tables, to use, the result above is equivalent to saying that the correction value s_j in the k-th table should be set to the mean of all samples, $s(n)$, that produced the ADC output $x(n) = x_j$ while the k-th table was selected.

We see that in order to calibrate the correction table, the discrete time versions $s(n)$ of the analog calibration signal must be known. However, these are in general not available and must therefore be estimated with some estimate $s_{\text{ref}}(n)$. This estimate can be obtained in several ways; a "better" ADC in parallel with the ADC under test, a digitally generated calibration signal fed to the ADC through a digital-to-analog converter [8], or signal reconstruction using optimal filtering [1] are all feasible methods for producing $s_{\text{ref}}(n)$.

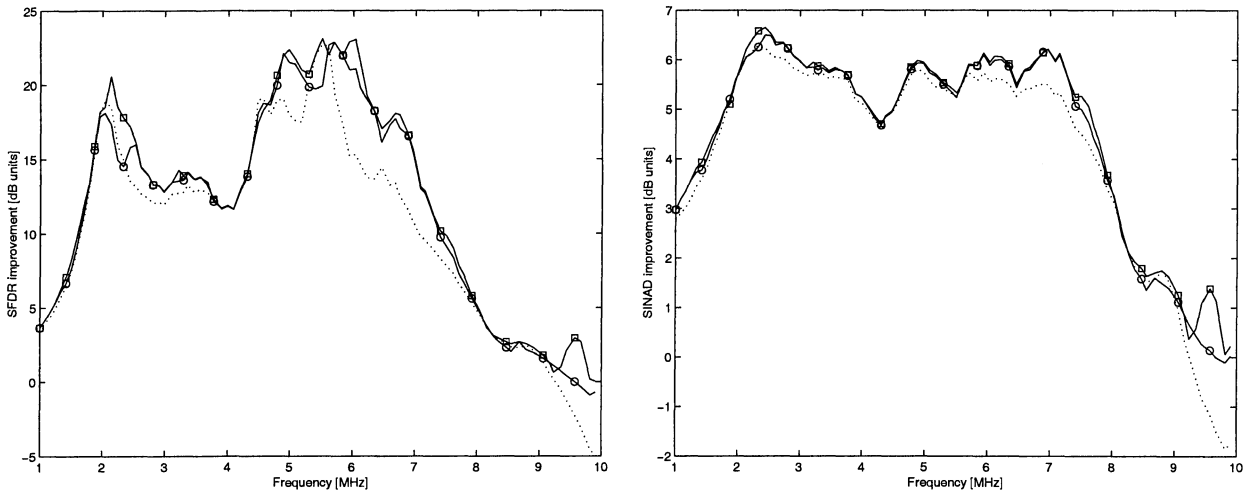
4 Performance

The proposed method has been evaluated with experimental ADC data from an Analog Devices AD876 10-bit converter, running at 20 MHz sampling frequency. The ADC correction table was calibrated using sinusoid calibration signals at several different frequencies. The calibration signal estimate $\hat{s}(n)$ was obtained using the optimal filtering method proposed in [1].

Spurious-free dynamic range (SFDR) and signal-to-noise and distortion ratio (SINAD) [9] are used to evaluate the method. The performance is presented as SFDR and SINAD improvements compared to the uncompensated case, and is shown in Figures 4(a) and 4(b), respectively. The performance for a static correction scheme is also plotted in Figure 4 for comparison.

The frequency-selective correction was evaluated for two test cases: the first case having 8 frequency regions ($K = 8$) and the second case having 16 regions ($K = 16$). Both cases comprise a 16-bit shift-register ($N = 16$). The results indicate that the frequency-selective correction method is superior to the frequency-static method in general, but also that increasing the number of frequency ranges K from 8 to 16 does not give any significant improvement.

We see from the results in Figure 4 that the SFDR is improved with between zero and 7 dB, while the improvement in SINAD is below 1 dB except at fre-



(a) SFDR (b) SINAD

Figure 4: SFDR and SINAD improvements for frequency-selective correction (solid lines) compared with static correction (dotted line). Circles represent a frequency estimator with 8 regions ($K = 8$) and squares represent 16 regions ($K = 16$).

quencies near the Nyquist frequency (10 MHz), where the SINAD improvement is approximately 2 dB. As opposed to the case of static correction, the improvement for frequency selective correction never fall below zero, i.e., the performance of the corrected ADC will never be inferior to that of the uncompensated ADC.

5 Conclusions

We have in this paper derived an extension of classic look-up table ADC correction. The extension comprises a frequency selective look-up table method, using a fast frequency region estimator together with a two-dimensional correction table. The motivation for using a frequency selective correction method was that the errors of AD converters in general vary with frequency. The proposed method was evaluated using experimental ADC data, and results showed that the ADC performance, measured as SFDR and SINAD, improved compared with the performance of a static correction scheme.

References

[1] P. Händel, M. Skoglund, and M. Pettersson, "A calibration scheme for imperfect quantizers," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, pp. 1063–1068, Oct. 2000.

[2] H. Lundin, M. Skoglund, and P. Händel, "On external calibration of analog-to-digital converters," in *IEEE Workshop on Statistical Signal Processing*, pp. 377–380, Aug. 2001.

[3] F. H. Irons, D. M. Hummels, and S. P. Kennedy, "Improved compensation for analog-to-digital converters," *IEEE Transactions on Circuits and Systems*, vol. 38, pp. 958–961, Aug. 1991.

[4] D. Moulin, "Real-time equalization of A/D converter nonlinearities," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, vol. 1, pp. 262–267, IEEE, 1989.

[5] T. Andersson, M. Skoglund, and P. Händel, "Frequency estimation by 1-bit quantization and table look-up processing," in *Proceedings European Signal Processing Conference*, (Tampere, Finland), pp. 1807–1810, EURASIP, Sept. 2000.

[6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[7] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. IT-28, pp. 129–137, Mar. 1982.

[8] J. Tsimbinos and K. V. Lever, "Improved error-table compensation of A/D converters," *IEE Proceedings - Circuits, Devices and Systems*, vol. 144, pp. 343–349, Dec. 1997.

[9] *IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters*. IEEE Std. 1241.

Estimation Error in Blind Time Error Estimation Algorithm for Time Interleaved A/D Converter System

Jonas Elbornsson[1], Fredrik Gustafsson[1] and Jan-Erik Eklund[2]

[1]Linköpings universitet. E-mail: {jonas,fredrik}@isy.liu.se

[2]Ericsson Microelectronics E-mail: jan-erik.eklund@mic.ericsson.se

March 25, 2002

Abstract

Time interleaving is an efficient way of increasing the sample rate of an A/D converter systems. However, time interleaving introduces mismatch errors in time, amplitude and gain, between the different A/D converters. In this paper a method for blind estimation of static time errors in time interleaved A/D converters is investigated. The method only requires that the input signal is band limited to the Nyquist frequency, but the estimation accuracy is dependent on the frequency content of the input signal. We will in this paper investigate how the estimation error depends on the input signal frequency content for some different signal types. The theory is also verified with simulations.

1 Introduction

In recent years the development of high bandwidth wireless communications applications have accelerated. Since most communications applications today are digital while the transmission itself is analog, the requirements on the A/D converters are rapidly increasing. With a fast A/D converter, the A/D conversion can be done closer to the antenna and more of the signal processing can be done digitally. To achieve higher sample rates, an array of M A/D converters, interleaved in time, can be used. Each ADC should work at $1/M$th of the desired sample rate [1], see Figure 1. This structure, however, introduces three types of matching errors since the A/D converters are not identical:

- **Time errors (static jitter)**
  The delay time of the clock to the different A/D converters is not equal. This means that the signal will be periodically but non-uniformly sampled.
- **Amplitude offset errors**
  The ground level can be slightly different in the different A/D converters. This means that there is a constant amplitude offset in each A/D converter.
- **Gain error**
  The gain, from analog input to digital output, can be different for the different A/D converters.

We will in this paper focus on the time errors and assume that the amplitude and gain errors are removed. The time errors give a periodically, but irregularly, sampled signal with unknown sample instances. If the time errors are known the uniformly sampled signal can be reconstructed [2], assuming that the input signal is band limited.

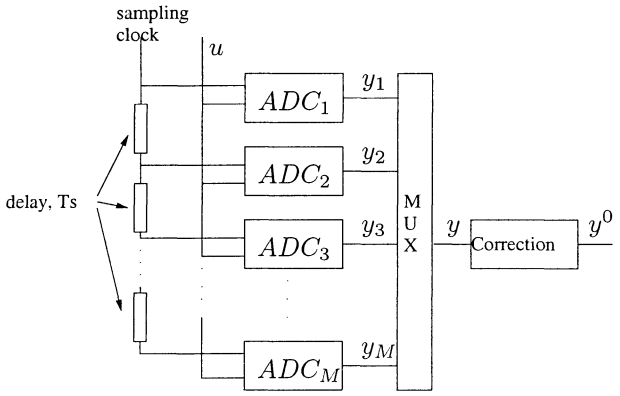


Figure 1: M parallel ADCs with the same master clock. M times higher sample rate is acheived by this structure. Due to mismatch errors in delay elements, ground levels, and gain, in the different ADCs, the sampled signal is distorted.

Methods for estimation of errors in A/D converters have been presented in for instance [3], [4] and [5] but those methods require a known calibration signal. Calibration of A/D converters is time-consuming and expensive. Therefore a lot of costs can be saved if the errors in the ADC can be automatically estimated and compensated for under drift. We will in

this paper investigate an estimation method for timing errors in interleaved ADCs, previously presented in [6, 7], that does not require knowledge about the input signal. The method has also been verified on measurement in [8] and influence from gain and amplitude errors is investigated in [9]. The only assumption that is made on the input signal is that it is band limited to the Nyquist frequency, but the estimation accuracy increases with further lowered input signal bandwidth. We will in this paper investigate how the the estimation error of the algorithm depends on the frequency content for different types of input signals.

2 Notation and Definitions

The nominal sampling interval, that we would have without time errors, is denoted T_s. M denotes the number of A/D converters in the time interleaved array. The time errors are denoted Δ_{t_i}. The estimates of these errors are denoted $\hat{\Delta}_{t_i}$. We use the following notation for the signals involved:

- $u(t)$ is the analog input signal.
- $u[k]$ denotes an artificial signal, sampled without time errors.
- $u_i[k]$ denotes the M subsequences of $u[k]$.
- $y_i[k]$ are the output subsequences from the M A/D converters, sampled with time errors.

$$y_i[k] = u((kM + i)T_s + \Delta_{t_i})$$

- $y[m]$ is the multiplexed output signal from all the A/D converters.

$$y[m] = y_{(m \mod M)}[\lfloor \frac{m}{M} \rfloor]$$

We assume that each A/D converter has a constant sample interval, without errors, so that the sample interval for each subsequence, $y_i[k]$, is exactly MT_s. We assume throughout this paper that $u(t)$ is band limited to the Nyquist frequency.

2.1 Definitions

We will in this section establish a few definitions which will be used in the following sections.

Definition 1 (Quasi-stationary signal [10]) *A signal $u(t)$ is quasi-stationary if*

$$m = \lim_{N\to\infty} \frac{1}{N} \int_0^N E(u(t))dt$$

$$R_u(\tau) = \lim_{N\to\infty} \frac{1}{N} \int_0^N E(u(t)u(t+\tau))dt$$

exist, where the expectation is taken over possible stochastic parts.

A stationary stochastic process is quasi-stationary, with m and $R_u(\tau)$ being the mean value and covariance function respectively.

Definition 2 (Quasi-stationary spectrum [10])

Assume that $u(t)$ is quasi-stationary. Then the power spectrum of $s(t)$ is defined as:

$$\Phi_u(\omega) = \int_{-\infty}^{\infty} R_u(\tau) e^{-j\omega\tau} d\tau$$

Definition 3 (Modulo M quasi-stationary) *Assume*

$$\bar{g}_{y_{i_1}, y_{i_2}, \cdots} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} g(y_{i_1}[t], y_{i_2}[t], \dots)$$
$$i_1, i_2, \cdots = 0, \dots, M-1$$

exists for a function $g(\cdot, \cdot, \cdots)$. Then y is modulo M quasi-stationary with respect to g if

$$\bar{g}_{i_1, i_2, \cdots} = \bar{g}_{\{(i_1+l) mod M, (i_2+l) mod M, \cdots\}}$$
$$\forall l \in \{\dots, -1, 0, 1, \dots\}$$

The modulo M quasi-stationarity property guarantees that the signal has the same statistical properties in all the A/D converters in the time interleaved array.

Definition 4

$$R_{y_i, y_j}[l] = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} E\left\{ \left(y_{(i mod M)}[k + \lfloor \frac{i}{M} \rfloor + l] - y_{(j mod M)}[k + \lfloor \frac{j}{M} \rfloor] \right)^2 \right\}$$

where $\lfloor \cdot \rfloor$ denotes integer part.

3 Time Error Estimation Algorithm

The time error estimation algorithm is here briefly reviewed. The algorithm is presented in more detail in [6, 7].

The basic idea of the estimation method is that the signal changes more on average if the sampling interval is longer than the nominal sampling interval and vice versa, see Figure 2. We assume, for this estimation algorithm, that the input signal is band limited. We use the first time error as reference, i.e., we assume that $\Delta_{t_0} = 0$.

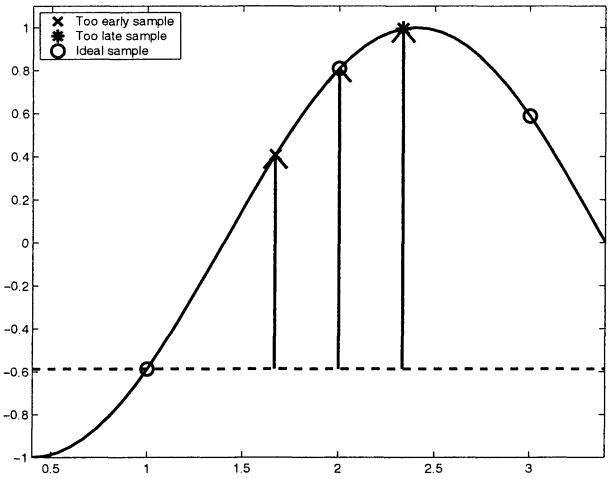


Figure 2: The basic idea for the timing error estimation method. If the sample is taken too early the average difference between adjacent sample values is smaller, if the sample is taken too late the average difference between adjacent samples is larger.

A Taylor expansion of $\Delta y_i[k]$ is calculated around the nominal sampling time of A/D converter $i-1$, mod M is here omitted on the indices for convenience

$$\begin{aligned}\Delta y_i[k] &= y_i[k] - y_{i-1}[k] \qquad (1)\\ &\approx (T_s + \Delta_{t_i} - \Delta_{t_{i-1}})u'(kMT_s + (i-1)T_s)\\ &\quad i = 0, 1, \ldots, M-1\end{aligned}$$

Calculating the mean square value of (1) we get

$$\begin{aligned}R_{y_i,y_{i-1}}[0] &= \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\{\Delta y_i[k]\}^2 \qquad (2)\\ &= (T_s + \Delta_{t_i} - \Delta_{t_{i-1}})^2\bar{E}\{(u'(t))^2\}\end{aligned}$$

If we sum over the square roots of $R_{y_i,y_{i-1}}[0]$ the unknown time errors will cancel, if $u[k]$ is mod M quasistationary with respect to $R_{u_i,u_{i-1}}[0]$, and we have

$$\begin{aligned}&\frac{1}{M}\sum_{i=0}^{M-1}\sqrt{R_{y_i,y_{i-1}}[0]} \qquad (3)\\ &= \frac{1}{M}\sum_{i=0}^{M-1}(T_s + \Delta_{t_i} - \Delta_{t_{i-1}})\sqrt{E\{(u'(t))^2\}}\\ &= T_s\sqrt{E\{(u'(t))^2\}}\end{aligned}$$

If we take the square root of (2) and insert into (3) we get M equations, linear in Δ_{t_i}

$$\begin{aligned}&\sqrt{R_{y_i,y_{i-1}}[0]} \qquad (4)\\ &= (T_s + \Delta_{t_i} - \Delta_{t_{i-1}})\frac{1}{MT_s}\sum_{i=0}^{M-1}\sqrt{R_{y_i,y_{i-1}}[0]}\end{aligned}$$

Using the first time error as reference ($\Delta_{t_0} = 0$) we get the time error estimates by solving the equation system

$$\begin{aligned}&\begin{pmatrix}1 & 0 & \cdots & 0\\ -1 & 1 & \cdots & 0\\ \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}\begin{pmatrix}\hat{\Delta}_{t_1}\\ \hat{\Delta}_{t_2}\\ \vdots\\ \hat{\Delta}_{t_{M-1}}\end{pmatrix}\\ &= \begin{pmatrix}\sqrt{R_{y_1,y_0}[0]}\\ \sqrt{R_{y_2,y_1}[0]}\\ \vdots\\ \sqrt{R_{y_{M-1},y_{M-2}}[0]}\end{pmatrix}\frac{1}{\hat{\sigma}_{u'}} - T_s\\ &\hat{\sigma}_{u'} = \frac{1}{MT_s}\sum_{j=1}^{M}\sqrt{R_{y_j,y_{j-1}}[0]} \qquad (5)\end{aligned}$$

4 Estimation Error

In this section we will investigate how the estimation errors depends on the input signal frequency content. The investigation is first done in detail for a dual A/D converter system with three different input signal types. The results are then extended to a general M.

4.1 Dual A/D Converter System

For a dual A/D converter system we have

$$\begin{aligned}R_{y_1,y_0} &= \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\{y_1[k] - y_0[k]\}^2\\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\{u((2k+1)T_s + \Delta_{t_1}) - u(2kT_s)\}^2\\ &= 2R_u(0) - 2R_u(T_s + \Delta_{t_1}) \qquad (6)\\ R_{y_1,y_0} &= 2R_u(0) - 2R_u(T_s - \Delta_{t_1}) \qquad (7)\end{aligned}$$

For $M = 2$ the equation system (5) consists of only one equation, and the solution is

$$\begin{aligned}\hat{\Delta}_{t_1} &= T_s\left(\frac{\sqrt{R_{y_1,y_0}[0]}}{\frac{1}{2}(\sqrt{R_{y_1,y_0}[0]} + \sqrt{R_{y_0,y_1}[0]})} - 1\right) \qquad (8)\\ &= T_s\left(\frac{\sqrt{R_u(0) - R_u(T_s + \Delta_{t_1})}}{\bar{\sigma}_u T_s} - 1\right)\\ \bar{\sigma}_u &= \frac{1}{2T_s}\left(\sqrt{R_u(0) - R_u(T_s + \Delta_{t_1})}\right.\\ &\quad \left.+\sqrt{R_u(0) - R_u(T_s - \Delta_{t_1})}\right)\end{aligned}$$

4.1.1 Sinusoidal input

For a sinusoidal input signal, $u(t) = \sin(\omega t)$, we have

$$R_u(\tau) = \frac{\cos(\omega\tau)}{2} \qquad (9)$$

Inserting this into (8) we get

$$\hat{\Delta}_{t_1} = T_s \left(\frac{2\sqrt{1-\cos(\omega(T_s+\Delta_{t_1}))}}{\sum_{i=0}^{1}\sqrt{1-\cos(\omega(T_s+(-1)^i\Delta_{t_1}))}} - 1 \right) \tag{10}$$

If we make a Taylor expansion of (10) with respect to ω and Δ_{t_1} we get

$$\frac{\hat{\Delta}_{t_1}-\Delta_{t_1}}{\Delta_{t_1}} = -\frac{(\omega T_s)^2}{12} \tag{11}$$

The exact value of (10) for $\omega = \pi/T_s$ is $\hat{\Delta}_{t_1} = 0$. This means that the estimate will always be between 0 and Δ_{t_1} for a sinusoidal signal below the Nyquist frequency, i.e., we always get improvement after estimation.

4.1.2 Band limited white noise input

Assume that $u(t)$ is a band limited white noise signal, so that

$$\Phi_{uu}(\omega) = \begin{cases} \frac{1}{\omega_{max}(1-\alpha)} & \alpha\omega_{max} \le \omega \le \omega_{max} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

From this we get the covariance function

$$R_u(\tau) = \frac{\sin(\omega_{max}\tau) - \sin(\alpha\omega_{max}\tau)}{\pi\omega_{max}(1-\alpha)\tau} \tag{13}$$

If we put (13) into (8) and make a Taylor expansion with respect to Δ_{t_1} and ω_{max} we get

$$\frac{\hat{\Delta}_{t_1}-\Delta_{t_1}}{\Delta_{t_1}} = -\left(\frac{1}{20} + \frac{\alpha^4+\alpha^3}{20(\alpha^2+\alpha+1)}\right)(\omega_{max}T_s)^2 \tag{14}$$

If $\alpha \to 0$, i.e., low pass filtered white noise, we get

$$\frac{\hat{\Delta}_{t_1}-\Delta_{t_1}}{\Delta_{t_1}} = -\frac{(\omega_{max}T_s)^2}{20} \tag{15}$$

and if $\alpha \to 1$ we get the same result as for a sinusoidal signal with frequency ω_{max}.

4.1.3 Multisine input

For a multisine input signal with L equally spaced frequencies

$$u(t) = \sum_{l=1}^{L} \sin(\frac{l\omega_{max}}{L}t) \tag{16}$$

we have, after Taylor expansion with respect to ω_{max} and Δ_{t_1}

$$\frac{\hat{\Delta}_{t_1}-\Delta_{t_1}}{\Delta_{t_1}} = -\frac{(\omega_{max}T_s)^2(3L^2+3L-1)}{60L^2} \tag{17}$$

Here we can see that when $L \to \infty$ we get, as expected, the same result as for low pass filtered white noise.

4.2 General Time Interleaved A/D Converter System

For a general time interleaved A/D converter system, with M A/D converters, we have

$$\hat{\Delta}_{t_1} - \hat{\Delta}_{t_{i-1}} = \tag{18}$$
$$T_s \left(\frac{\sqrt{R_u(0) - R_u(T_s + \Delta_{t_i} - \Delta_{t_{i-1}})}}{\frac{1}{M}\sum_{j=0}^{M-1}\sqrt{R_u(0) - R_u(T_s + \Delta_{t_j} - \Delta_{t_{j-1}})}} - 1 \right)$$

Following the same calculations as for the dual A/D converter case we find that the terms linear in Δ_{t_k}, $k \neq i, i-1$ disappear and we get the same error approximation as for the case $M = 2$, but expressed as a difference between adjacent A/D converter errors. With a sinusoidal input, for instance, we have

$$\frac{\hat{\Delta}_{t_1} - \hat{\Delta}_{t_{i-1}} - \Delta_{t_1} - \Delta_{t_{i-1}}}{\Delta_{t_1} - \Delta_{t_{i-1}}} = -\frac{(\omega T_s)^2}{12} \tag{19}$$

5 Simulations

The simulations in this section are done for a dual A/D converter system, i.e. $M = 2$. The number of data from each A/D converter $N = 2^{19}$. The simulation results are here shown for two types of input signals:

- Sinusoidal input, see Figure 3: The plot shows good accuracy between theoretical and simulated estimation errors. The degradation for low frequencies in the simulation results is caused by finite amount of data. The theoretical results are based on infinitely many data, and the simulations are done with $N = 2^{19}$ samples.
- Multisine input, see Figure 4: Also for the multisine input signal the resemblance between theoretical and simulated results is good. The simulations have been done with $L = 256$ tones, and a single tone simulation is shown for comparison.

6 Conclusion

An estimation algorithm for time errors in a time interleaved A/D converter has been investigated. The estimation method requires only that the input signal is band limited to the Nyquist frequency to improve the signal quality after estimation and correction. But the estimation accuracy depends on the frequency content of the input signal. In this paper we have analyzed this dependence for three different input signal types, sinusoidal, multisine and band limited white noise. The input signal types have been chosen to cover the two extremes: the flat spectrum

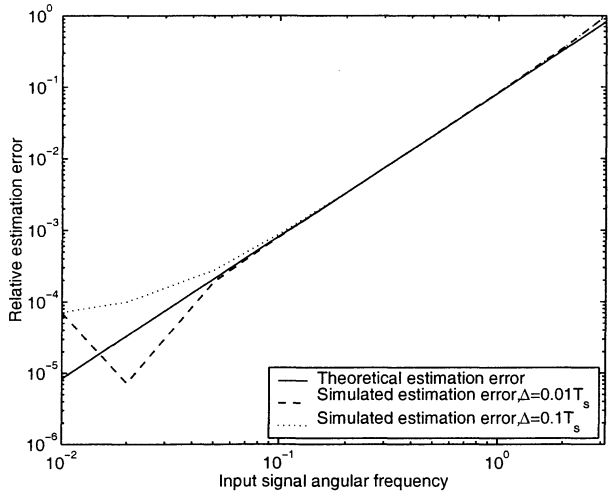


Figure 3: Relative time error estimation error with sinusoidal input for a dual A/D converter system. The estimation errors are calculated from simulations with $\Delta_{t_1} = 0.1T_s$ and $\Delta_{t_1} = 0.01T_s$.

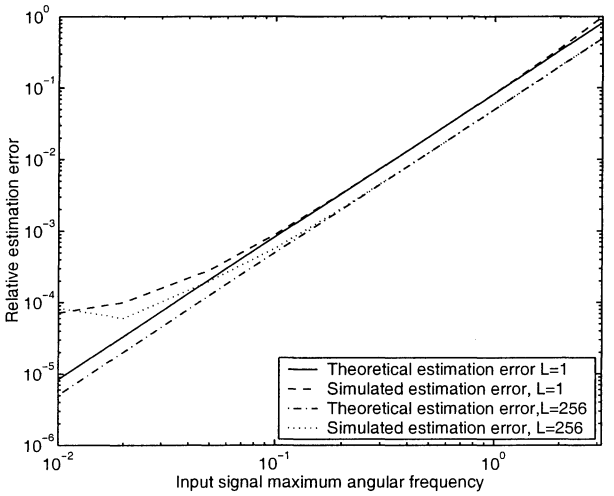


Figure 4: Relative time error estimation error with multisine input for a dual A/D converter system. The estimation errors are calculated from simulations with $\Delta_{t_1} = 0.1T_s$ and with $L = 1$ and $L = 256$ respectively.

of a bandlimited white noise signal, and the δ-spike of a sinusoidal signal. The theoretical results have been verified with simulations of a dual A/D converter system and the simulation results show good resemblance with the theoretical results.

References

[1] Y-C Jenq. Digital spectra of nonuniformly sampled signals: A robust sampling time offset estimation algorithm for ultra high-speed waveform digitizers using interleaving. *IEEE Transactions on Instrumentation and Measurement*, 39(1):71–75, February 1990.

[2] A. Papoulis. *Signal Analysis.* McGraw-Hill, 1977.

[3] J.J. Corcoran. Timing and amplitude error estimation for time-interleaved analog-to-digital converters. US Patent nr. 5,294,926, October 1992.

[4] H. Jin and E.K. Lee. A digital-background calibration technique for minimizing timing-error effects in time-interleaved ADC's. *IEEE Transactions on Cicuits and Systems*, 47(7):603–613, July 2000.

[5] H. Lundin, M. Skoglund, and P. Händel. On external calibration of analog-to-digital converters. In *Proc. IEEE Signal Processing Workshop 2001*, pages 377–380. IEEE, 2001.

[6] J. Elbornsson and J.-E. Eklund. Blind estimation of timing errors in interleaved AD converters. In *Proc. ICASSP 2001*, volume 6, pages 3913–3916. IEEE, 2001.

[7] J. Elbornsson. *Equalization of Distortion in A/D Converters.* Lic. thesis 883, Department of Electrical Engineering, Linköping University, Linköping, Sweden, April 2001.

[8] J. Elbornsson, K. Folkesson, and J.-E. Eklund. Measurement verification of estimation method for time errors in a time-interleaved a/d converter system. In *To appear in Proc. ISCAS 2002.* IEEE, May 2002.

[9] J. Elbornsson, F. Gustafsson, and J.-E. Eklund. Amplitude and gain error influence on time error estimation algorithm for time interleaved A/D converter system. In *To appear in Proc. ICASSP 2002.* IEEE, May 2002.

[10] L. Ljung. *System Identification, Theory for the user.* Prentice-Hall, 2 edition, 1999.

Narrow-Band and Wide-Band Frequency Masking FIR Filters with Short Delay

Linnéa Svensson and Håkan Johansson

Department of Electrical Engineering, Linköping University
SE-581 83 Linköping, Sweden
linneas@isy.liu.se and hakanj@isy.liu.se

ABSTRACT

In this paper, we introduce frequency masking (FM) FIR filters with short delay. The overall filter has approximately linear phase, and its delay can be chosen arbitrarily. The paper considers how to design both narrow- and wide-band filters.

1. INTRODUCTION

Finite-impulse response (FIR) filters have many good properties [1]. However, FIR filters need a high order to fulfil a specification with a narrow transition band. The order of an FIR filter is inversely proportional to the width of the transition band [1] and the arithmetic complexity grows with the filter order. An FIR filter with order N takes $N+1$ multipliers (around $N/2$ for linear phase) and N adders to implement. To come around this problem occurring for narrow transitions bands, one can use an approach proposed by Neuvo, Cheng-Yu, and Mitra in 1984 [2]. We refer to this approach as frequency masking FIR filter design and it gives a filter with lower complexity compared to a linear phase equiripple FIR filter [1]–[3] for a given specification with narrow transition band and passband. The filter design method in [2] reduces the complexity, but the delay is increased compared to a linear-phase FIR filter [2]. In this paper, we introduce FM FIR filters with short delay. If it is important to have a short delay in an application, the design introduced in this paper is preferable. However, shortening the delay one has to pay with some higher complexity of the design, compared to the original method [2]. Also, the exact linear phase is no longer maintained in this approach.

2. NARROW-BAND FM FIR

For a narrow-band FM FIR filter, the transfer function is expressed as

$$H(z) = G(z^M)F(z) \tag{1}$$

where M is some positive integer. The corresponding structure is shown in Fig. 1. The filter $G(z)$ works as a *model filter* and $F(z)$ as *masking filter* which extracts the desired passband and eliminates the unwanted images of

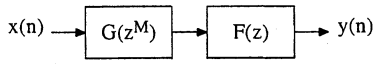


Figure 1. Structure for narrow-band FM filters.

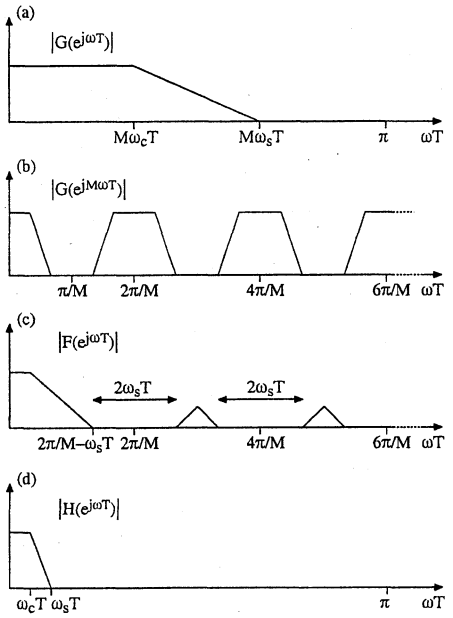


Figure 2. Illustration of magnitude responses for narrow-band FM lowpass filters. (a) model filter, (b) periodic model filter (c) masking filter, and (d) overall filter.

the *periodic model filter* $G(z^M)$. Typical magnitude responses for the model, masking, and overall filters are as shown in Fig. 2. Further, we let $\omega_c T$, $\omega_s T$, δ_c, and δ_s denote the passband edge, stopband edge, passband ripple, and stopband ripple, respectively, for the overall filter $H(z)$.

2.1 Proposed narrow-band filter design

This section introduces the new approach of designing narrow-band FM FIR filters with short delay. The delay $K = M \cdot K_G + K_F$ for the overall filter, where K_F and K_G are the delays of the masking filters $F(z)$ and $G(z)$ respectively, is fixed. The delays K_F and K_G are not restricted to be integers, but we only consider this case in this paper in order to limit the number of possible ways to

distribute the total delay K. Equation (2) shows the specification of $H(z)$ where δ_c and δ_s are the ripples in the passband and stopband, and $\phi_e(\omega T) = \phi_H(\omega T) + K\omega T$ is the deviation from linear phase in the pass band.

$$\begin{aligned} 1-\delta_c \le |H(e^{j\omega T})| &\le 1+\delta_c && \omega T \in [0, \omega_c T] \\ |H(e^{j\omega T})| &\le \delta_s && \omega T \in [\omega_s T, \pi] \\ |\phi_e(\omega T)| &\le \phi_{e,max} && \omega T \in [0, \omega_c T] \end{aligned} \quad (2)$$

The above requirements can be met by separately optimizing the subfilters $F(z)$ and $G(z)$ according to Section 2.1.1. Such an approach leads however to an overdesigned overall filter $H(z)$. To reduce the complexity and/or improve the results one should consider simultaneous optimization of the subfilters (Section 2.1.3).

2.1.1 Separate optimization of F(z) and G(z)

The specification of $H(z)$ must here be translated into specifications for the two subfilters so that they together fulfil the specification of $H(z)$. If second-order effects are neglected, we get the following results.

The passband ripple of the overall filter is in the worst case the sum of the ripples of $F(z)$ and $G(z)$ and the maximum deviation from linear phase is the sum of the deviation of the subfilters. That is,

$$\begin{aligned} \delta_c^{(F)} + \delta_c^{(G)} &= \delta_c \\ \phi_{e,max}^{(F)} + \phi_{e,max}^{(G)} &= \phi_{e,max} \end{aligned} \quad (3)$$

In the stop band, if $F(z)$ and $G(z)$ have certain ripples, then $H(z)$ is ensured to have the same or smaller ripple. This requires that none of the subfilters have a peak in their transition bands which may occur for short delay filters. If this is the case, the ripple requirements must be sharpened slightly to compensate for this. Assuming no peaks in the transition bands the stopband ripples can be chosen as

$$\delta_s^{(F)} = \delta_s^{(G)} = \delta_s \quad (4)$$

This will give us the specifications according to

$$\begin{aligned} 1-\delta_c^{(F)} \le |F(e^{j\omega T})| &\le 1+\delta_c^{(F)} && \omega T \in [0, \omega_c T] \\ |F(e^{j\omega T})| &\le \delta_s && \omega T \in \Omega_s \\ |\phi_e^{(F)}(\omega T)| &\le \phi_{e,max}^{(F)} && \omega T \in [0, \omega_c T] \end{aligned} \quad (5)$$

$$\begin{aligned} 1-\delta_c^{(G)} \le |G(e^{j\omega T})| &\le 1+\delta_c^{(G)} && \omega T \in [0, M\omega_c T] \\ |G(e^{j\omega T})| &\le \delta_s && \omega T \in [M\omega_s T, \pi] \\ |\phi_e^{(G)}(\omega T)| &\le \phi_{e,max}^{(G)} && \omega T \in [0, M\omega_c T] \end{aligned} \quad (6)$$

where Ω_s is the union of the stopbands for the masking filter according to Fig. 2. A simple way to design the filters to meet the above requirements is to use linear programming for the real and imaginary parts, in a similar way as that described in [4].

2.1.2 Finding M, K_G and K_F

For linear-phase FIR filters there are several formulas giving the approximate filter order needed for a certain specification [5], [6]. For non-linear phase FIR filters such formulas do not exist, thus one has to search for the lowest order for which the specification is fulfilled. There is no analytic way to find neither the optimal M nor the optimal distribution of K to K_G and K_F. Therefore, an exhaustive search over M and K_G (K_F) must be done. In principle, there is an infinite number of different distributions, but in practice this has to be limited (e.g. restricting K_G and K_F to be integers). Sometimes the optimal filter has a delay less than the specified K. In this case, extra delays are added in the design to achieve the wanted delay. Depending on the application, these extra delay elements can be removed and we get a smaller K for the exact same specification.

2.1.3 Simultaneous optimization of F(z) and G(z)

To reduce the complexity and/or improve the results, the overall filter is further optimized by simultaneously optimizing the subfilters instead of separately. To this end, a standard nonlinear optimization routine may be used. The solution obtained in Section 2.1.1 is a good initial solution for further optimization. To find the minimum-complexity overall filter that meets the requirements in (2), one should start with a relaxed specification, then design the subfilters separately, and finally further optimize the overall filter.

2.2 Design examples

As a means of demonstrating the proposed version of narrow-band FM FIR filter design with short delay, a filter with $K = 75$ is designed. The specification is the following: $\omega_c T = 0.08\pi$ rad, $\omega_s T = 0.1\pi$ rad, $\delta_c = 0.01$, $\delta_s = 0.01$, and $\delta_{e,max} = 0.01$.

An equiripple linear-phase FIR filter with this specification (disregarding K) gives a delay of half the filter order, in this case $195/2 = 97.5$. This delay is too long, so another method must be used to design the filter.

To this end, we use the proposed approach given in Section 2.1. The optimal M and distribution of K is found by an exhaustive search with M from 3 to 9 and K_G from 5 to $(K-5)/M$ for every M. The design with the lowest complexity is found for $M = 5$, $K_G = 14$ and $K_F = K - MK_G = 5$. The filter orders are $N_G = 48$, $N_F = 24$. This corresponds to 74 multipliers in an implementation. The resulting passband and stopband attenuation is 0.053 dB and 40.0 dB respectively, and the phase error is less than 0.0066. The magnitude response and the phase error is plotted in Figs. 3 – 5. Because of a peak in the transition band of $F(z)$, $\delta_s^{(G)}$ had to be lowered somewhat to compensate for this.

The requirements on the filter in some regions are even without the final optimization largely fulfilled. The margin of over-design can be used to lower the filter

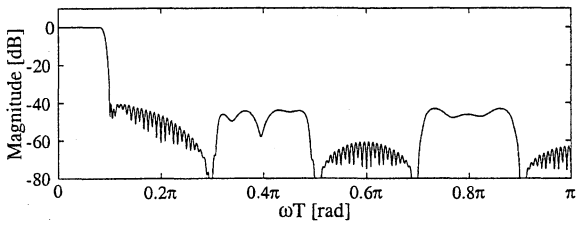


Figure 3. Magnitude response of $H(z)$ before further optimization.

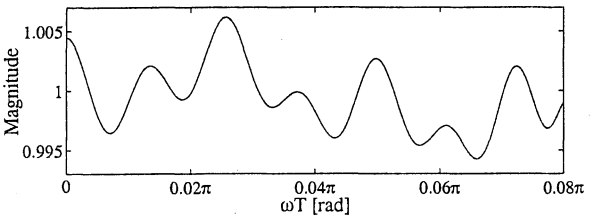


Figure 4. Magnitude response of $H(z)$ in the passband before further optimization.

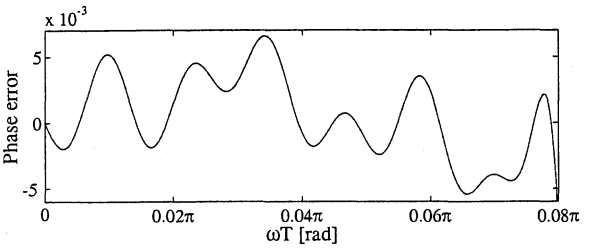


Figure 5. Phase error of $H(z)$ before further optimization.

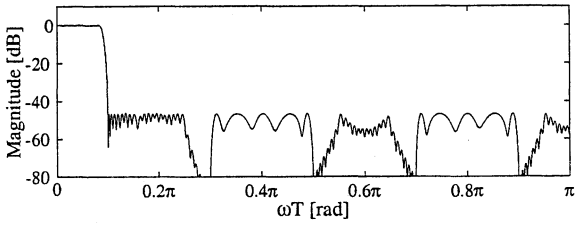


Figure 6. Magnitude response of $H(z)$ after further optimization

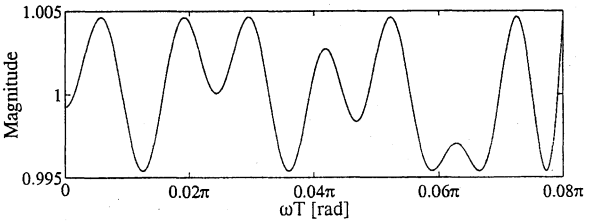


Figure 7. Magnitude response of $H(z)$ in the passband after further optimization.

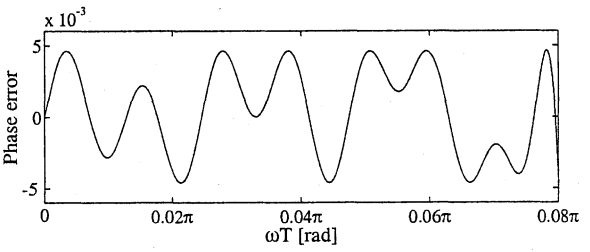


Figure 8. Phase error of $H(z)$ after further optimization.

orders to further reduce the complexity of the filter or it can be left to minimize the ripples and the phase deviation. Here the latter is demonstrated. The overall filter $H(z)$ is optimized with non-linear programming using *fminimax.m* in MATLAB. The results can be seen in Figs. 6 – 8. The passband and stopband attenuations are now 0.040 dB and 46.0 dB and the phase deviation is less then 0.0047.

As a comparison, one FIR filter with the same specification is designed with linear optimization. To fulfil the specification where the real and imaginary parts of the error were considered separately, an order of 211 was needed. This corresponds to 212 multiplications in an implementation. With further optimization this number could be lowered somewhat, but not so much that it can compete with the proposed design method.

3. WIDE-BAND FM FIR

To obtain a wide-band LP filter, a narrow-band HP filter is subtracted from a pure delay filter. The transfer function is expressed as

$$H(z) = z^{-K} - G(z^M)F(z) \tag{7}$$

and the structure is shown in Fig. 9.

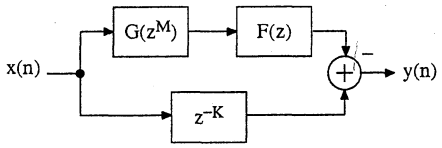


Figure 9. Structure for wide-band FM filters.

The masking filter $F(z)$ is an HP filter and the model filter $G(z)$ must be an LP filter for even M and an HP filter for odd M. Typical magnitude responses of the different filters can be seen in Figs. 10 and 11.

3.1 Proposed wide-band filter design

This section introduces the new approach of designing wide-band FM FIR filters with short delay. The specification of $H(z)$ is translated into specifications for the two subfilters so that they together fulfil the specification of the overall filter. The second order effects are neglected and we get the following specifications.

$$\delta_s^{(F)} = \delta_s^{(G)} = \min(\delta_c, \phi_{e,max}) \tag{8}$$

$$(\phi_{e,max}^{(F)} + \phi_{e,max}^{(G)})^2 + (\delta_c^{(F)} + \delta_c^{(G)})^2 = \delta_s^2 \tag{9}$$

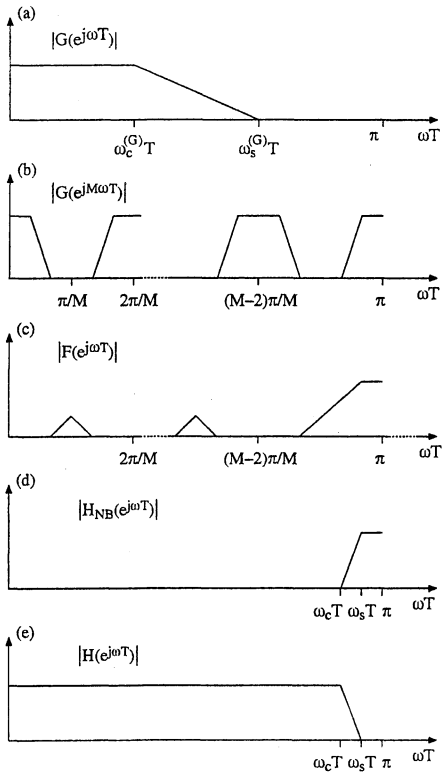


Figure 10. Illustration of magnitude responses for wide-band FM lowpass filters with even M. (a) model filter, (b) periodic model filter, (c) masking filter, and (e) overall filter.

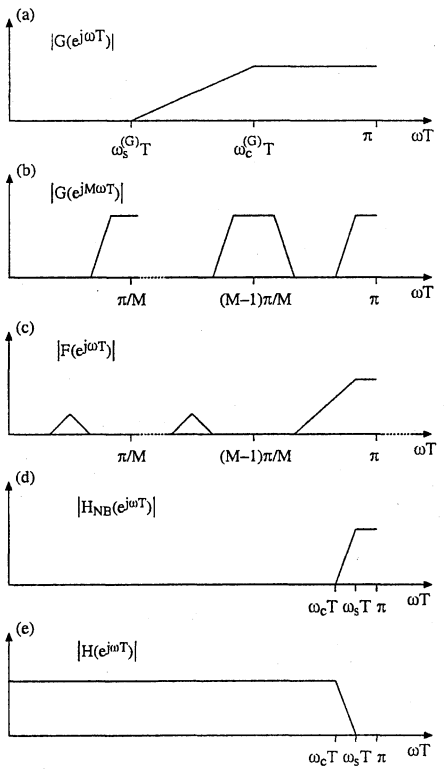


Figure 11. Illustration of magnitude responses for wide-band FM lowpass filters with odd M. (a) model filter, (b) periodic model filter, (c) masking filter, and (e) overall filter.

$$
\begin{aligned}
1-\delta_c^{(F)} \le |F(e^{j\omega T})| &\le 1+\delta_c^{(F)} && \omega T \in [\omega_c T, \pi] \\
|F(e^{j\omega T})| &\le \delta_s && \omega T \in \Omega_s \\
|\phi_e^{(F)}(\omega T)| &\le \phi_{e,max}^{(F)} && \omega T \in [\omega_c T, \pi]
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
1-\delta_c^{(G)} \le |G(e^{j\omega T})| &\le 1+\delta_c^{(G)}, && \omega T \in [0, \omega_c^{(G)} T] \\
|G(e^{j\omega T})| &\le \delta_s && \omega T \in [\omega_s^{(G)} T, \pi] \\
|\phi_e^{(G)}(\omega T)| &\le \phi_{e,max}^{(G)} && \omega T \in [0, \omega_c^{(G)} T]
\end{aligned} \tag{11}
$$

where Ω_s is the union of the stopbands for the masking filter according to Figs. 10 and 11. For even values of M, $G(z)$ is a lowpass filter with

$$
\begin{aligned}
\omega_c^{(G)} T &= M(\pi - \omega_s T) \\
\omega_s^{(G)} T &= M(\pi - \omega_c T)
\end{aligned} \tag{12}
$$

and for odd values, the edges are

$$
\begin{aligned}
\omega_c^{(G)} T &= M(\omega_s T - \pi) + \pi \\
\omega_s^{(G)} T &= M(\omega_c T - \pi) + \pi
\end{aligned} \tag{13}
$$

As for the narrow-band filters, linear programming for the real and imaginary parts can be used for designing the filters. Also, further optimization should be used in order to reduce the complexity and/or improve the result.

3.2 Design examples

The proposed wide-band FM FIR filter design is demonstrated by means of some examples. The total delay K and the ripple specification are the same as in the previous examples while $\omega_c T = 0.9\pi$ rad and $\omega_s T = 0.92\pi$ rad.

Since the only difference between the specifications is the location of the transition band, the equiripple solution would give us the same delay of 97.5 which is not short enough. To reduce the delay, we use the approach in Section 3.1. An exhaustive search is done for M and K_G and the design with the lowest complexity fulfilling the specifications (8) – (11) is found with $M = 6$, $K_G = 10$, and $K_F = 15$. The filter orders are $N_G = 55$ and $N_F = 31$. Combining them according to Fig. 9 results in passband and stopband ripples of 0.064 dB and 48.2 dB, respectively and a phase error less than 0.007. See Figs. 12– 14. This design results in 88 multiplications per sample which can be compared with the 212 multiplications needed for a single FIR filter.

The design margin is used to minimize the ripples and the phase error further in a non-linear optimization of $H(z)$. The phase error is now less than 0.0030 and the passband and stopband ripples are 0.025 dB and 53.2 dB respectively. The magnitude of $H(z)$ and its phase error in the passband is plotted in Figs. 15 – 17.

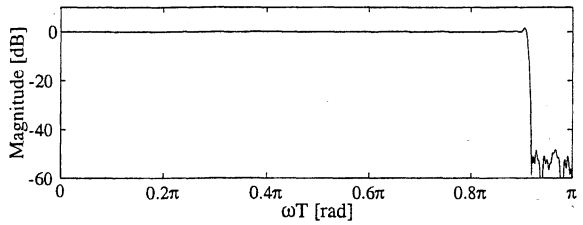


Figure 12. Magnitude response of $H(z)$ before further optimization.

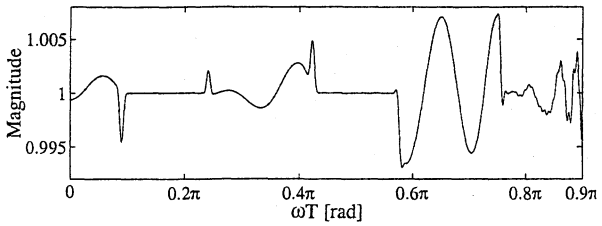


Figure 13. Magnitude response of $H(z)$ in the passband before further optimization.

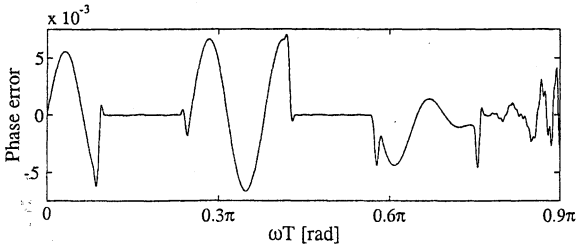


Figure 14. Phase error of $H(z)$ before further optimization.

There is a peak in the transition band of 1.76 dB, but this is typical for short-delay filters and not specific for our approach. Adding constraints in the transition band to keep down this peak the passband and stopband attenuation and maximal phase error deviation becomes 0.036 dB, 48.1 dB, and 0.0039 respectively. The magnitude response without the peak in the transition band is shown in Fig. 18. With a shorter overall delay, it might not be possible to flatten out the peak and still find a feasible solution.

4. CONCLUSION

This paper introduced a new version of the FM technique. Using the proposed method it is possible to design narrow- and wide-band FIR filters with short delay and a lower complexity compared to an ordinary nonlinear-phase FIR filter.

REFERENCES

[1] T. Saramäki, "Finite impulse response filter design," in *Handbook for Digital Signal Processing*, eds. S. K. Mitra and J. F. Kaiser, New York: Wiley, 1993, ch. 4, pp. 155–277.

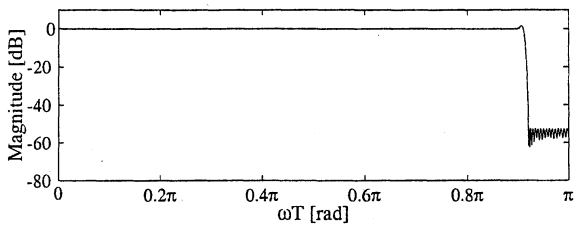


Figure 15. Magnitude response of $H(z)$ after further optimization.

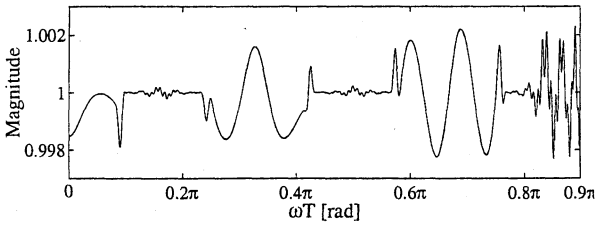


Figure 16. Magnitude response of $H(z)$ in the passband after further optimization.

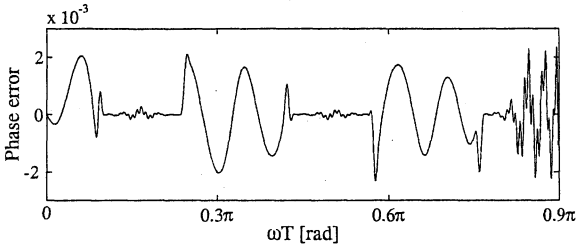


Figure 17. Phase error of $H(z)$ after further optimization.

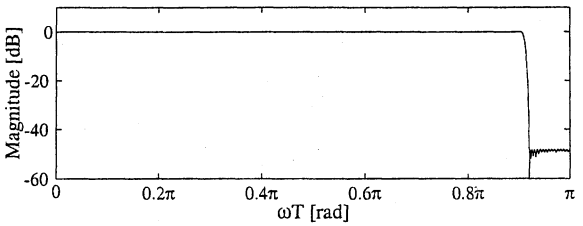


Figure 18. Magnitude response of further optimized $H(z)$ with flattened peak.

[2] Y. Neuvo, D. Cheng-Yu, S. K. Mitra, "Interpolated Finit Impulse Response Filters," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 3, pp. 563–570, June 1984.

[3] T. Saramäki, "Design of computationally efficient FIR filters using periodic subfilters as building blocks," in *The Circuits and Filters Handbook*, ed. W. K. Chen, CRC Press, Inc., 1995, pp. 2578-2601.

[4] L. Svensson and H. Johansson, "Frequency-response masking FIR filters with short delay," in *Proc. IEEE Int. Symp. Circuits Syst.*, Phoenix, USA, May 26–29, 2002.

[5] O. Herrmann, L. R. Rabiner, and D. S. K. Chan, "Practical design rules for optimum finit impuls response digital filters", Bell Tech. J., vol. 52 no. 6, pp. 769–799, 1973.

[6] K. Ichige, M Iwaki, R. Rokuya, "Accurate estimation of minimum filter length for optimum FIR digital filters," *IEEE Trans. Circuits Syst.*, vol. 47, no. 10, Oct. 2000.

Reducing Power Consumption with Selected Orthogonal Signals

Niklas Petersson[1], Albin Johansson[2], Per Ödling[3] and Per Ola Börjesson[4]

[1]Lund University, Dept. of Electroscience, Box 118, SE-221 00 Lund, Sweden. Niklas.Petersson@es.lth.se
[2]Ericsson Telecom AB, SE-126 25 Stockholm, Sweden. Albin.Johansson@etx.ericsson.se
[3]Lund University, Dept. of Electroscience, Box 118, SE-221 00 Lund, Sweden. Per.Odling@es.lth.se
[4]Lund University, Dept. of Electroscience, Box 118, SE-221 00 Lund, Sweden. Per.Ola.Borjesson@es.lth.se

Abstract

The use of existing twisted-pair lines for high speed data communication with various DSL standards provides a low-cost internet access to a wide public. In order to limit the total power consumption in the transceiver, reducing the Peak-to-Average Power ratio, PAR, is important.

In this work we use the tone reservation approach to PAR reduction, where a subset of the data-carrying tones instead are used for PAR reduction.

We derive a bound on the achievable performance under realistic system constraints. As long as a moderate reduction performance is needed, this could be achieved with an arbitrary selection of the PAR reduction tones. In an ADSL-like system, the tone reservation approach could be used for PAR reduction down to about 11.5 dB, without sacrificing considerable bandwidth.

1 Introduction

High-speed data transmission over existing telephone lines provides a low-cost access for a large number of people. The existing twisted-pair lines can be used for a much wider frequency spectrum than originally intended. The connection between the internet and the twisted pair cable that belongs to a specific customer is located in the central office. It is here important that the total power dissipation for all the transceivers is kept low. Then the temperature will be lower, and it is also possible to pack equipment for more customers in the available space.

Looking at the total power consumed by the transceiver, typically 50% of the power is used by the line driver, which is the power amplifier closest to the cable, see Figure 1. A/D and D/A converters consumes about 25% and the last 25% goes to the digital processing parts. This makes it obvious that by reducing the power in the line driver, we can improve the total power consumption of the system. The power in the line driver will be dependent on its supply voltage. This has to be big enough to admit a linear transfer function between the desired amplitude span. Thus, if the amplitude span of the line driver can be kept low, the total power dissipation will also be low.

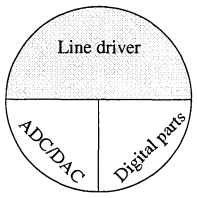


Figure 1: Power distribution in a DMT transceiver

We here focus on DMT baseband wireline systems, e.g. ADSL and VDSL, where the channel is a twisted pair copper wire. These systems, along with other multicarrier systems, have high Peak-to-Average Power Ratios, *PAR*. Therefore the reduction of signal peaks is important here. Several proposals have been made to reduce the amplitude of multicarrier signals, [1–3]. The one used here is tone reservation, which has been presented in [1, 4, 5]. A subset of the available tones are reserved for PAR reduction, and used for creating a reduction signal that is added to the data signal in order to reduce the signal peaks, see Figure 2.

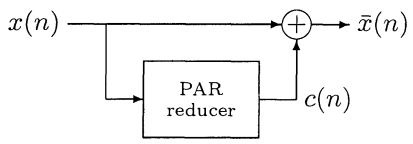


Figure 2: Addition of a PAR reduction signal, $c(n)$, that counteracts the peaks in $x(n)$.

The presented work is a cooperation between Lund University and Ericsson Telecom AB, Access Signal Processing Laboratory.

In Section 2, we define the data model and look at how the number of clips of an unreduced signal will depend on the chosen clip level. Depending on the used set of PAR reduction tones, and what system constraints that has to be met, the reduction performance will differ. In Section 3, we look closer on the tone selection and a bound on the achievable performance under system constraints is derived.

2 System model

A measure of the level of variation in a signal is the Peak-to-Average power ratio, *PAR*. Since we want to keep the peak value as low as possible, and in the same time keep the average power high, thereby maintaining a high SNR, we look at the ratio of these two numbers.

To reduce the PAR level, a subset of the available tones is reserved for PAR reduction and thereby not available for data transmission. The PAR reduction signal, $c(n)$ in Figure 2, is created with this subset. The reduction signal is then added to the data signal, leading to a signal with lower peaks, $\bar{x}(n) = x(n) + c(n)$. Based on the data signal $x(n)$, the reduction signal $c(n)$ can be constructed using for example linear programming algorithms.

2.1 Definition of PAR

Here we have worked with only the time-discrete PAR. On a data signal $x(n)$, this is often defined [1] as

$$\mathrm{PAR}\{x(n)\} = 10\log_{10}\frac{\max|x(n)|^2}{E|x(n)|^2}. \qquad (1)$$

According to this definition of PAR, the expression for the PAR of the reduced signal $\bar{x}(n)$ should include the reduction signal in the denominator. However, if this was the case, a decrease in PAR level could be caused by a increase in the total power, without any reduction of the peak power. Thus, we define the PAR after reduction as

$$\mathrm{PAR} = 10\log_{10}\frac{\max|x(n)+c(n)|^2}{E|x(n)|^2}. \qquad (2)$$

This is *not* a true Peak-to-Average ratio of a signal, since it is a function of the signal both before and after reduction. This definition will, however, describe how much average signal power it is possible to use under a certain peak power constraint. If this PAR level can be lowered, we can use a higher average signal to a fixed peak level. Then the SNR on each frequency will increase, and thereby also the number of data bits per subcarrier.

2.2 Signal model

An ADSL system [6], which we consider in particular, uses symbol length $N = 512$ in the downstream direction. The information values are modulated on several subcarriers in a multi-carrier system. Each subcarrier can be regarded as having a random amplitude, since this is data dependent. The summation of this large number of random sequences, with data values often drawn from a large constellation, will give a resulting sequence that is approximately Gaussian, due to the Central Limit Theorem. This means that there is a probability of having very high peaks in the signal.

We have modeled the signal as Gaussian with variance σ^2, although all frequencies not always are used in the data signal. The probability that a single sample exceeds a given level $\gamma\sigma$, the *sample clip probability*, is

$$p = P(|x(n)| > \gamma\sigma) = 2Q(\gamma). \qquad (3)$$

The number of clips in one symbol follows the binomial distribution. In particular, the probability that a symbol has at least one sample exceeding this level, the *symbol clip probability*, can be expressed as

$$p_{\mathrm{sym}} = 1 - \binom{N}{k}p^0(1-p)^{N-0} = 1-(1-p)^N. \qquad (4)$$

Figure 3 shows the symbol clip probability as a function of the clip level γ i dB, as well as the theoretical calculations, based on the binomial distribution.

3 Tone selection

In order to increase the the bitrate of the system, a trade-off has to be made between reserving tones for PAR reduction and keeping valuable data tones. This means that if there are certain frequencies in the spectrum where the channel attenuation is high, these could be used for PAR reduction without discarding data capacity. In particular, the highest frequencies have bad characteristics on long wire lengths. Therefore, it could be attractive to use a number of the highest tones for PAR reduction.

From [1] follows that a choice of contiguous tones performs worse than just selecting a tone set by random. In Figure 4, simulations are shown for 12 reduction tones. The difference between the random selection and the contiguous tones is very big in the unconstrained case, where the added reduction signal is unlimited in size. However, when putting a constraint on the maximum allowed energy per tone, the PAR reduction performance for different tone sets changes abruptly. In the figure the difference between the tone placement classes is now much smaller. According to [7], constrained energy on the tones means that for a small number of tones, the placement of

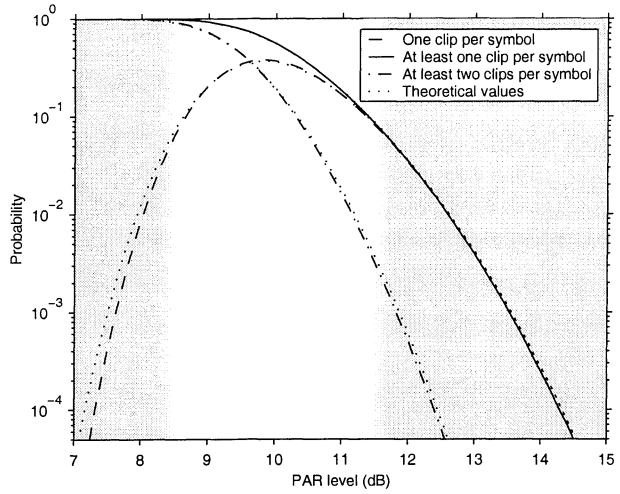


Figure 3: Probabilities of different number of clips. In the shaded area to the left, below about 8.5 dB PAR level, we will almost always have at least one clip. In the rightmost shaded area, over 11.5 dB, we will most often have maximum one clip in the signal. Simulations made with symbol size $N = 512$ and 223 data tones used. As comparison, the theoretical values for a white Gaussian symbol are also shown as dotted lines.

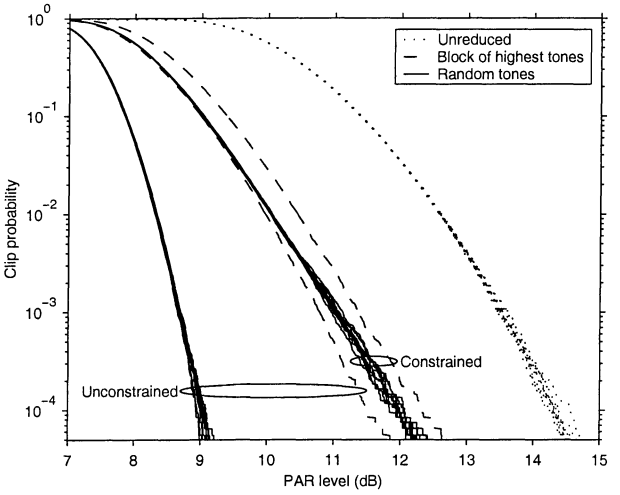


Figure 4: Symbol clip probability (one symbol equals 512 samples) is shown for 12 reduction tones. One class with a random selection of tones is shown, as well as the tone set using the 12 highest tones. Both simulations for unconstrained and PSD-constrained reduction are shown.

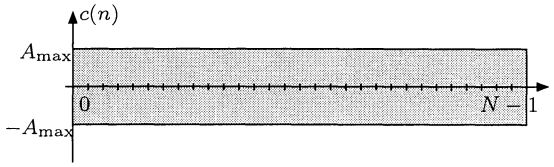


Figure 5: Limits for reduction signal in bound for PSD-constrained tone placement.

the reduction tones is not important. We will now see what is possible to do with a constrained energy per tone at our disposal, and also what constraints are imposed by a selection of only odd or even tones.

3.1 Bound on achievable PAR reduction for constrained tone power

Having a small number of PAR reduction tones, and each of these tones has a limitation in power, a bound could be derived on the achievable PAR level. The largest reduction signal that is possible to create is when all tones are placed in phase at the same sample. If the difference between any data sample and the desired PAR level is bigger than what we at maximum can get by placing all tones in phase, it will be impossible to reduce the symbol down to this desired PAR level. We now will derive a bound on what results can be achieved when operating under a PSD constraint.

We assume the symbol to be a white Gaussian signal with variance σ^2, a model that according to Figure 3 quite closely describes our data symbol. Then the symbol clip probability after reduction can be expressed as

$$\begin{aligned} p_{\text{sym}} = P(\text{clip}) &= P(\max(|x(n) + c(n)|) > \gamma\sigma) \\ &= 1 - P(\text{all } |x(n) + c(n)| \leq \gamma\sigma). \end{aligned} \tag{5}$$

Let $A(i)$ denote the maximum amplitude on reduction tone i. We now assume that we can create an arbitrary reduction signal $c(n)$, with the only limitation being that it nowhere can have a value that is higher than the sum of the amplitudes of the PAR reduction tones,

$$|c(n)| \leq A_{\max} = \sum_{i=1}^{T} A(i), \tag{6}$$

see Figure 5. T is the number of tones reserved for PAR reduction, This means that we can reduce all samples in $x(n)$ where $|x(n)| \leq \gamma\sigma + A_{\max}$ down to the clip level $\gamma\sigma$. Since this model gives more degrees of freedom than a PSD constraint on each tone, and we make optimal use of this freedom, we will then get a lower bound on the achievable PAR level for the

constrained PSD case. The clip probability can be bounded as

$$p_{\text{sym}} \geq 1 - P(\text{all } |x(n)| \leq \gamma\sigma + A_{\max}) = 1 - \left(1 - 2Q\left(\gamma + \frac{1}{\sigma}\sum_{i=1}^{T} A(i)\right)\right)^N, \quad (7)$$

where we have used the assumption that the different $x(n)$ are uncorrelated. Solving this for γ, we get the lower bound on the achievable PAR level γ as a function of the desired clip probability:

$$\gamma \geq Q^{-1}\left(\frac{1-(1-p_{\text{sym}})^{1/N}}{2}\right) - \frac{1}{\sigma}\sum_{i=1}^{T} A(i). \quad (8)$$

This is the analytical lower bound, and we will now compare this with the simulation results.

3.2 Numerical evaluation of the bound

The bound for PSD-constrained reduction in (8) is a result dependent on the tone amplitudes $A(i)$, and in order to compare this with the simulations, we can include the implemented constraints. The constraint used in the simulations can be written as

$$A(i) = \frac{\sigma\sqrt{2}}{\sqrt{T_{d0} - T}} \cdot \frac{1}{\cos\frac{\pi}{s}}. \quad (9)$$

The expression is independent of i, i.e., we use a constant-valued PSD constraint. When minimizing the PAR level using a limited energy per tone, the algorithms often choose to use the maximum allowed energy on most tones and the average added energy will be approximately the same as the maximum allowed energy. The constraint implemented therefore constrains the maximum energy per reduction tone to the average energy per data tone[1].

The $\sigma/\sqrt{T_{d0} - T}$ part of (9) divides the total energy with the number of used data tones. T_{d0} is the number of available transmitting tones, so $T_{d0} - T$ will be the number of used data tones. The cosine expression on the end comes from the linear approximation of the power constraint, see Figure 6.

The bound can now be evaluated numerically using (8) and (9). Figure 7 shows how the PSD constraint affects the reduction performance for a system with 12 reduction tones. For this number of tones, the bound becomes a tight limit somewhere below a clip probability of 10^{-3}. Figure 8 shows what reduction performance can be achieved as a function of number of PAR tones. Since the bounds limits the average performance, some individual simulations can cross this level. The results are shown for a symbol

[1]PSD masks, as often defined in standards, imply a constraint on the average power per tone [6].

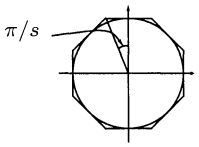


Figure 6: Constraints implemented in the linear programming algorithm. The circle represents the ideal power constraint, but the octagon is implemented as a linear approximation. This corresponds to $s = 8$ in (9).

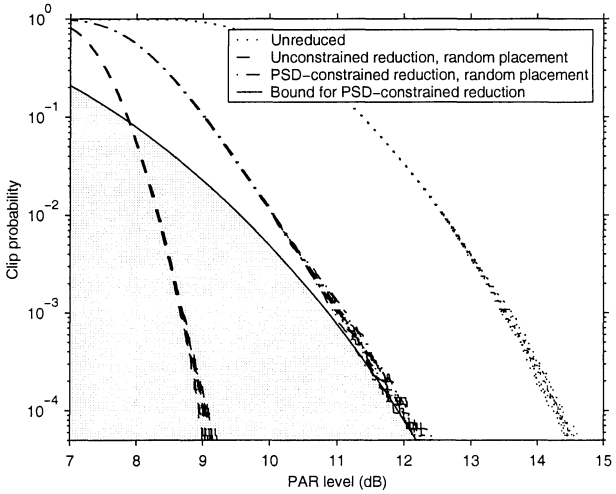


Figure 7: Symbol clip probability (one symbol equals 512 samples) for tone sets with 12 randomly selected reduction tones. The bound for constrained power is shown as a solid line.

clip probability of 10^{-3}, but this could be chosen otherwise. In Figure 7, the bound seems to be tighter with lower clip probabilities. It can be noted that up to about 10 PAR reduction tones, both the random placement and the block placement closely follow the bound. This means, that up to this number of tones, we could choose arbitrary placement.

4 Summary

A reduction system using the tone reservation approach would have to operate under a PSD constraint, which we have seen strongly limits the performance. In practice, it would be convenient to use for example a block of the highest tones in the spectrum, or a block at some other place where the channel characteristics are bad. From the bound on the reduction performance derived for a white Gaussian symbol, it can be seen that up to about 10 PAR tones out of 256, any tone set is sufficiently good to use. Then we could as well use the tones which carry the lowest number of bits.

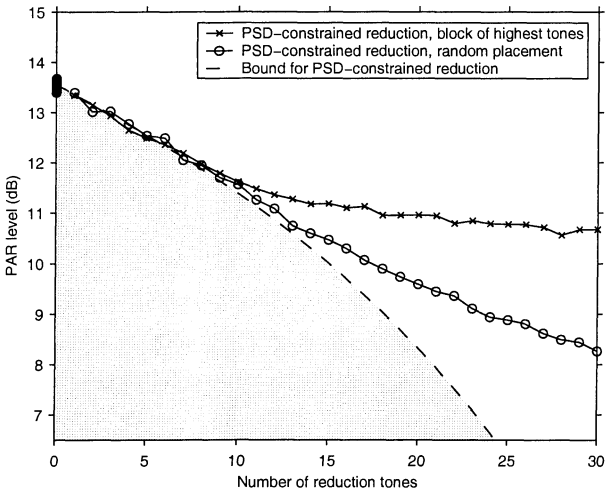


Figure 8: PAR level as a function of number of PAR reduction tones. The results are shown for a symbol clip probability of 10^{-3}. Up to 10 reduction tones, the difference between the placement methods is negligible.

From Figure 8 we can also see that with a reduction signal of 10 tones, the reachable PAR level will be about 11.5 dB. In section 2.2, it was seen that over this clip level, a clipped symbol almost always only had one clipped sample. We can interpret this result as if you have a signal that is unlikely to create more than one sample over the desired maximum limit, any tone set will do it. Then it does not matter that the specific tone set is incapable of controlling the position of a second peak.

From the system-design point of view, the tone reservation approach could be used without sacrificing considerable bandwidth, for PAR reduction down to about 11.5 dB. This is because a choice of the subcarriers carrying the least amount of bits could be used, with results as good as with any other selection.

References

[1] J. Tellado-Mourelo, *Peak to Average Power Reduction for Multicarrier Modulation.* PhD thesis, Stanford University, Sept. 1999.

[2] M. Friese, "Multitone signals with low crest factor," *IEEE Transactions on Communications*, vol. 45, pp. 1338–1344, Oct. 1997.

[3] A. Gatherer and M. Polley, "Controlling clipping probability in DMT transmission," in *Proc. Asilomar conf. on Signals, Systems and Computers*, 1997.

[4] P. O. Börjesson, H. G. Feichtinger, N. Grip, M. Isaksson, N. Kaiblinger, P. Ödling, and L.-E. Persson, "A low-complexity PAR-reduction method for DMT-VDSL," in *Proc. 5th International Symposium on Digital Signal Processing for Communication Systems*, (Perth, Australia), pp. 164–199, Feb. 1999.

[5] P. O. Börjesson, H. G. Feichtinger, N. Grip, M. Isaksson, N. Kaiblinger, P. Ödling, and L.-E. Persson, "DMT PAR-reduction by weighted cancellation waveforms," in *Proc. Radiovetenskaplig Konferens*, (Karlskrona, Sweden), June 1999.

[6] ITU-T Recommendation G.992.1.

[7] N. Petersson, A. Johansson, P. Ödling, and P. O. Börjesson, "Analysis of tone selection for PAR reduction," in *Proc. International Conference on Information, Communications & Signal Processing*, (Singapore), Oct. 2001.

Noise radar using binary ADC

Sune R.J. Axelsson

Saab Bofors Dynamics
Swedish Defence Research Agency
S-581 88 Linköping, Sweden
Email: Sune.Axelsson@dynamics.saab.se

Abstract

Presently, there is an increasing interest in noise modulated radar. Wide-band signal processing with high range resolution, reduced ambiguities and high ECCM performance can be achieved. The processing involved correlates the received signal and the delayed reference. Range/Doppler processing and digital beam forming can be performed with surprising good results even for binary or a few bits ADC, if a second noise signal is added.

1 Introduction

Noise modulated radar can be used in a wide range of applications, such as surveillance, tracking, and sub-surface profiling. Wide bandwidth gives high range resolution, and a suppression of range ambiguities is achieved.

Already in 1959, a range measuring radar was suggested, transmitting modulated noise and estimating the range by correlation processing [1]. In [2] and [3], implementations of varying complexity were discussed. See also [4] for a survey.

In a recent paper [5], the performance of noise modulated radar was analysed in more theoretical terms. Both ultra wide-band noise modulation and narrow-band around a carrier was studied. Phase modulating a sine wave by random noise was treated as well.

In this paper, the use of noise modulated radar in digital beam forming is also discussed. The performance of binary ADC is analysed, and the significant improvement achieved by adding a second noise signal before the binary or few bits ADC is pointed out and displayed.

2 Principles of noise radar

A simple type of noise radar can be constructed as follows. A random noise signal x(t) is transmitted and a delayed version defines the LO-signal of the mixer.

The signal received from the target is delayed $T=2R/c$ and that from the delay line T_f. The output from the mixer passes a low-pass filter, and for $T=T_f$, a strong correlation peak is obtained.

The delay T_f defines a resolution cell in range at $R_0=cT_f/2$, and when the target is moved through the resolution cell, a high-resolution range mapping of the target occurs. The range resolution of the cell is given by $\Delta R=c/(2B)$, where B is the noise bandwidth.

The output of the LP-filter is close to the noise correlation function $R_X(T-T_f)$ when the filter bandwidth (B_{LF}) is much less than the HF-bandwidth of the noise (B).

The time-bandwidth product $N=2BT_m$ defines the number of independent noise samples during the time of measurement (T_m).

If digital signal processing is applied, a true correlation with unit weight at the integration can be performed over the time of measurement:

$$S_{km} = \sum x_{n-m} x_{n-k} \qquad (1)$$

where n=1 to N, x_n is a sequence of random numbers with zero mean and variance σ^2. The integers m, k correspond to T and T_f.

When $T_f=T$ (k=m), a strong correlation peak is received, and due to N>>1, the peak amplitude is close to its mean value: $S_0 \approx \langle S_{kk} \rangle = N\langle x_n^2 \rangle = N\sigma^2$. The variance of S_{km} is $N\sigma^4$. As a result, the ratio of the squared peak amplitude and the side-lobe variance becomes

$$S_0^2 / \sigma_S^2 = N \qquad (2)$$

Figure 1 displays simulation results for $N=10^4$. For wide band low-frequency radar waveforms, there is no carrier frequency, and the Doppler analysis cannot be implemented as usual. Still the received signal is compressed or extended as a result of the relative velocity between the target and the radar, however. The compression factor can be estimated and used for velocity estimation [5].

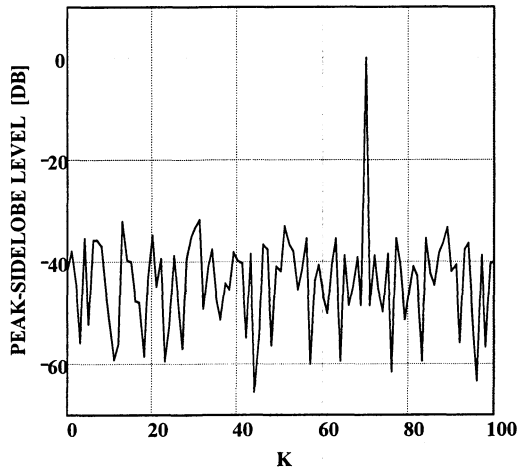


Figure 1: The function $S_{km}^2 = (\Sigma x_{n-m}x_{n-k})^2$ normalised with respect to its peak value (N=10 000 and m=70).

3 Narrow-band noise

More accurate velocity estimations are possible using narrow-band noise by detecting the Doppler shift of the carrier frequency. The noise process can then be modelled as

$$s(t) = X(t)\cos(\omega_0 t) - Y(t)\sin(\omega_0 t) \quad (3)$$

where ω_0 is the angular frequency of the carrier, and X(t), Y(t) are stationary stochastic processes, statistically independent with Gaussian distribution and zero means.

The received signal from a point reflector at range R is then s(t-T) with the delay T=2R/c. After down-converting, the received signal can be written in complex form as $V = V_1 + jV_2$, where

$$V_1 = X(t-T)\cos(\omega_0 T+\phi) + Y(t-T)\sin(\omega_0 T+\phi) \quad (4)$$

$$V_2 = X(t-T)\sin(\omega_0 T+\phi) - Y(t-T)\cos(\omega_0 T+\phi) \quad (5)$$

The reference signal $s(t- T_r)$ defines similarly $V_r = V_{1r} + jV_{2r}$. The correlation is a time-average of the product $V V_r^*$:

$$R_C(T, T_r) = \int \{V_1V_{1r}+V_2V_{2r}+j[V_2V_{1r}-V_1V_{2r}]\}dt \quad (6)$$

where the integration runs from t=0 to $t=T_{int}$.

If the correlation function of X(t), Y(t) is substituted: $R_X(\tau)=R_Y(\tau)=R(\tau)$, the real and imaginary parts of the correlation integral $R_C(T, T_r)= R_r(T, T_r)+j R_r(T, T_r)$ can be rewritten as

$$R_r(T, T_r)= 2\int R(T-T_r)\cos[\omega_0(T-T_r)]dt \quad (7)$$

$$R_i(T, T_r)= 2\int R(T-T_r)\sin[\omega_0(T-T_r)]dt \quad (8)$$

4 Range and doppler processing

The range/velocity estimation above was based on finding the delay and Doppler shift of the reference signal, giving a correlation peak.

The steps in Δv and ΔT must be chosen so short that the phase term is near-constant during the integration time. Also time compression effects make this solution complex to implement efficiently for long codes.

A possible way, similar to conventional Doppler radar processing, is to divide the velocity/range processing into two steps. First, a large number of correlation computations are performed with different delays, forming K range cells.

By repeating the range processing under M time steps, we get in each cell a signal vector, whose phase time variation reflects the Doppler frequency shift. After FFT-processing, the Doppler/range matrix output is derived.

A great advantage is that the efficient FFT-algorithm can be applied in the Doppler processing. Also the range correlation procedure might be implemented by FFT, making use of the fact that the convolution of two signals can be computed from the inverse Fourier transform of the product of their Fourier transforms.

5 Noise modulated phase

In the analysis above, the transmitted signal was Gaussian noise. Let us include the related case, when the transmitted sine signal is phase modulated by Gaussian noise

$$s(t) = \cos[\omega_0 t+\theta(t)] \quad (9)$$

By defining $X(t)=\cos[\theta(t)]$ and $Y(t)=\sin[\theta(t)]$, we can rewrite (9) in the same form as (3):

$$s(t) = X(t)\cos(\omega_0 t) - Y(t)\sin(\omega_0 t) \quad (10)$$

The mathematical formalism can then be reused, and the complex correlation functions are given by (6)-(8).

We will now look further at the modified correlation functions of X(t) and Y(t), which have to be used in (6)-(8). The normalised correlation function with R(0)=1 is given by

$$R(\tau) = Av\{2\cos[\theta(t+\tau)]\cos[\theta(t)]\}=$$

$$= Av\{\cos[\theta(t+\tau)+\theta(t)]\}+Av\{\cos[\theta(t+\tau)-\theta(t)]\} \quad (11)$$

By introducing characteristic functions [6] and the correlation function of the phase noise, (11) is rewritten as

$$R(\tau)=\exp[-R_\theta(0)- R_\theta(\tau)]+ \exp[-R_\theta(0)+ R_\theta(\tau)] \quad (12)$$

Usually $R_\theta(0)>>1$, giving $R(\tau) = \exp[-R_\theta(0)+ R_\theta(\tau)]$.

6 Binary correlation

6.1 Wide-band noise

If we detect only the sign (+1 or –1) of the received signal and the reference one, the correlation sum is

$$S_{km} = \Sigma \mathrm{Sign}(x_{n-k})\mathrm{Sign}(x_{n-m}) \quad (13)$$

and the correlation peak to side lobe level becomes

$$S_0^2/\sigma_{sid}^2 = N \quad (14)$$

Let us study the sum $S = \Sigma \mathrm{Sign}(x_n)\mathrm{Sign}(y_n)$, where x and y are two correlated Gaussian variables with zero means and the two-dimensional probability density function

$$p(x,y) = (2\pi\sigma_x\sigma_y)^{-1}(1-\rho^2)^{-1/2}\exp\{-0.5(1-\rho^2)^{-1}[(x/\sigma_x)^2 + (y/\sigma_y)^2 - 2\rho xy/(\sigma_x\sigma_y)]\} \quad (15)$$

Because N is a very large integer, we can approximate the sum by its mean: $S = N \langle \mathrm{Sign}(x)\mathrm{Sign}(y) \rangle$. Hence, S is estimated from

$$S = N(p_1 - p_2) \quad (16)$$

where p_1 is the probability of x and y having the same sign, and p_2 defines the probability that they have different signs.

The probability p_1 is derived by integrating p(x,y) over the first and third quadrants, and similarly p_2 is obtained from integration over the second and fourth quadrants.

We solve the double integral by changing to polar co-ordinates, followed by integration with respect to r and φ. If we denote $\rho(\tau) = R(\tau)/R(0)$, the result is

$$S = N \langle \mathrm{Sign}(x)\mathrm{Sign}(y) \rangle = N\,(2/\pi)\arcsin\{\rho(\tau)\} \quad (17)$$

which is similar to that of [7], describing the correlation function of clipped noise.

6.2 Narrow-band noise

The output signals from the binary AD-converters are $U_1 = \mathrm{sign}(V_1)$, $U_2 = \mathrm{sign}(V_2)$, $U_{1r} = \mathrm{sign}(V_{1r})$, and $U_{2r} = \mathrm{sign}(V_{2r})$, which can be rewritten as $U = U_1 + jU_2$ and $U_r = U_{1r} + jU_{2r}$.

The real and imaginary parts of the complex correlation function $R_C(T, T_r)$ then are

$$R_r(T, T_r) = \int [U_1U_{1r} + U_2U_{2r}]dt \quad (18a)$$

$$R_i(T, T_r) = \int [U_2U_{1r} - U_1U_{2r}]dt \quad (18b)$$

with integration from t=0 to $t=T_{int}$.

As the time-bandwidth product is very large, we can approximate

$$R_r(T, T_r) = \int [\mathrm{Av}(U_1U_{1r}) + \mathrm{Av}(U_2U_{2r})]dt \quad (19a)$$

$$R_i(T, T_r) = \int [\mathrm{Av}\{U_2U_{1r}\} - \mathrm{Av}\{U_1U_{2r}\}] \quad (19b)$$

where Av{ } represent ensemble averages or local time averages formed over time intervals much shorter than the period time of $\cos[\omega_0(T-T_r)]$.

If V_1, V_{1r} are Gaussian with zero means and correlation coefficient ρ_v: $\mathrm{Av}\{U_1U_{1r}\} = (2/\pi)\arcsin(\rho_v)$.

From $\mathrm{Av}\{V_1V_{1r}\} = R(T-T_r)\cos[\omega_0(T-T_r)] = \rho_v\sigma^2$ and the normalised correlation function of X(t) and Y(t): $\rho(\tau) = R(\tau)/\sigma^2$, we find $\rho_v = \rho(T-T_r)\cos[\omega_0(T-T_r)]$.

Insertion into (19) yields corresponding to (7)-(8)

$$R_r(T, T_r) = (4/\pi)\int \arcsin\{\rho(T-T_r)\cos[\omega_0(T-T_r)]\}dt \quad (20a)$$

$$R_i(T, T_r) = (4/\pi)\int \arcsin\{\rho(T-T_r)\sin[\omega_0(T-T_r)]\}dt \quad (20b)$$

For narrow-band noise, the phase term is varying much more rapidly than $\rho(T-T_r)$. Close to the peak, $\omega_0(T-T_r) \ll \pi$ is fulfilled during the integration time, giving

$$R_r(T, T_r) = T_{int}(4/\pi)\arcsin\{\rho(T-T_r)\cos[\omega_0(T-T_r)]\} \quad (21a)$$

$$R_i(T, T_r) = T_{int}(4/\pi)\arcsin\{\rho(T-T_r)\sin[\omega_0(T-T_r)]\} \quad (21b)$$

Due to the non-linearity of the arcsine function, the Doppler signal may be distorted with high side-lobes. This effect is maximal when $\rho=1$ but is reduced for lower values of ρ.

For $\rho=1$, the function $f(t) = (2/\pi)\arcsin\{\rho\sin[\omega_d t]\}$ becomes a triangular wave, generating high Doppler side-lobes. For reduced ρ-values, f(t) is gradually transformed into a sine signal with low side-lobes in Doppler (Figure 2).

We can reduce the fluctuations of $\rho\sin[\omega_d t]$ in (20)-(21) to the linear range of the arcsine by adding a second noise signal to the received one before AD conversion. It is also possible to add the extra noise to the transmitted signal.

Figure 3 shows, as an example, the simulated Doppler spectrum for two point targets. Binary ADC without noise addition gives high Doppler side lobes due to signal interference. After adding extra noise before ADC, the interference is removed and the sidelobes are highly suppressed.

The cost to be paid compared with high-resolution ADC is lower signal-to-noise ratio. In case of low-bits ADC, the extra noise required becomes more marginal, however. In some ECM-environments, the extra noise is already there and has not to be added.

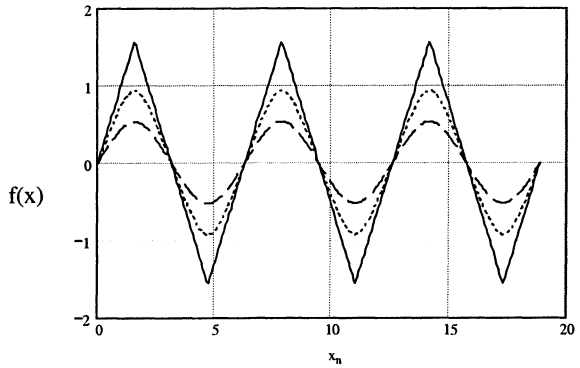


Figure 2: The function f(x)=arcsin(ρsin(x)) for ρ=1 (—), 0.8 (- - -) and 0.5 (— —) with f(x) gradually transformed from a triangular wave to a sine.

7 Digital antenna beam forming

Noise radar modules can be used in array antennas with digital beam forming after ADC. Digital beam forming is also possible using binary or low-bits ADC, if we add noise to reduce the correlation peak of (21), which makes it possible to approximate the arcsine function by its argument.

As an example, we consider a linear array of radar modules (m=0 to M). The target delay T_m will then vary between the antenna elements depending on the angle of incidence (φ). By noise addition before ADC, we can approximately describe the complex correlation function (21) as

$$R_C(m,T,T_r) \approx A\exp[j\omega_0(T_m-T_r)] \qquad (22)$$

Using m=0 as a reference, we can rewrite

$$R_C(m,T,T_r) \approx A\exp[j\omega_0(T_0-T_r)+j\delta_m] \qquad (23)$$

For a linear array with element separation d, the phase addition becomes $\delta_m = 4\pi dm\sin(\varphi)/\lambda$. At the beam forming, we multiply the detected signals by $\exp[-j\delta_m]$ and add, giving

$$R_c(\varphi,T,T_r)=\sum R_C(m,T,T_r)\exp[-j4\pi dm\sin(\varphi)/\lambda] \qquad (24)$$

Define $\varphi_n=n\Delta\varphi$ for n=0,1,...,M and $\Delta\delta = 4\pi d\sin(n\Delta\varphi)/\lambda$, which can be approximated for $n\Delta\varphi<1$ by $\Delta\delta = 4\pi dn\Delta\varphi/\lambda$.

Insertion of $d=\lambda/2$ gives $\Delta\delta = 2\pi n\Delta\varphi$ and (24) is rewritten

$$R_c(n\Delta\varphi,T,T_r) \approx \sum R_C(m,T,T_r)\exp[-j2\pi nm\Delta\varphi] \qquad (25)$$

From (25), M beams are formed if we FFT-process the vector $\{R_C(m,T,T_r)\}$. By similar technique, we can perform a two-dimensional scanning using a rectangular antenna array combined with two-dimensional FFT.

Figure 4 shows as an illustration the angular response of a linear array with 32 antenna elements, 128 range cells, and 128 Doppler cells. When extra noise is added before the binary ADC in each antenna channel (SNR=-7 dB), the side lobe suppression improves from 10 dB to 30 dB.

Simulation results showed that the noise adding method before low-resolution ADC also works for other pulse compression waveforms than noise, such as chirp, step frequency, and phase code modulation.

8 Conclusions

The theory behind noise radar for Doppler/range indication and digital beam forming was described. The use of binary or low-bits ADC highly improves the signal processing rate and reduces the costs. The spectacular improvement of sidelobe suppression, when extra noise is added before ADC, was explained theoretically and confirmed by simulation results.

Acknowledgements

This research was performed at the Sensor Department of Saab Bofors Dynamics, Linköping.

References

[1] B.M. Horton: Noise-modulated distance measuring system, Proc. IRE, vol47, pp.821-828, May 1959.

[2] G.R. Cooper and C.D. McGillem: Random signal radar, School Electr. Eng., Purdue Univ., Final Report, TR-EE67-11, June 1967.

[3] R.M. Naryanan et al.: Design and performance of a polarimetric random noise radar for detection of shallow buried targets, in Proc. SPIE Meeting on Detection Techn. Mines, Orlando, April 1995, vol. 2496, pp.20-30.

[4] I.P. Theron et al: Ultra-Band Noise Radar in the VHF/ UHF Band, IEEE AP-47, June 1999, pp. 1080-1084.

[5] S.R.J. Axelsson: On the Theory of Noise Doppler Radar, Proc. IGARSS 2000, Honolulu, 24-28 July 2000, pp. 856-860.

[6] N.U. Prabhu: *Stochastic Processes*. MacMillan, New York, 1956.

[7] J.H. van Vleck: The spectrum of clipped noise. RRL Report 51, July 21, 1943.

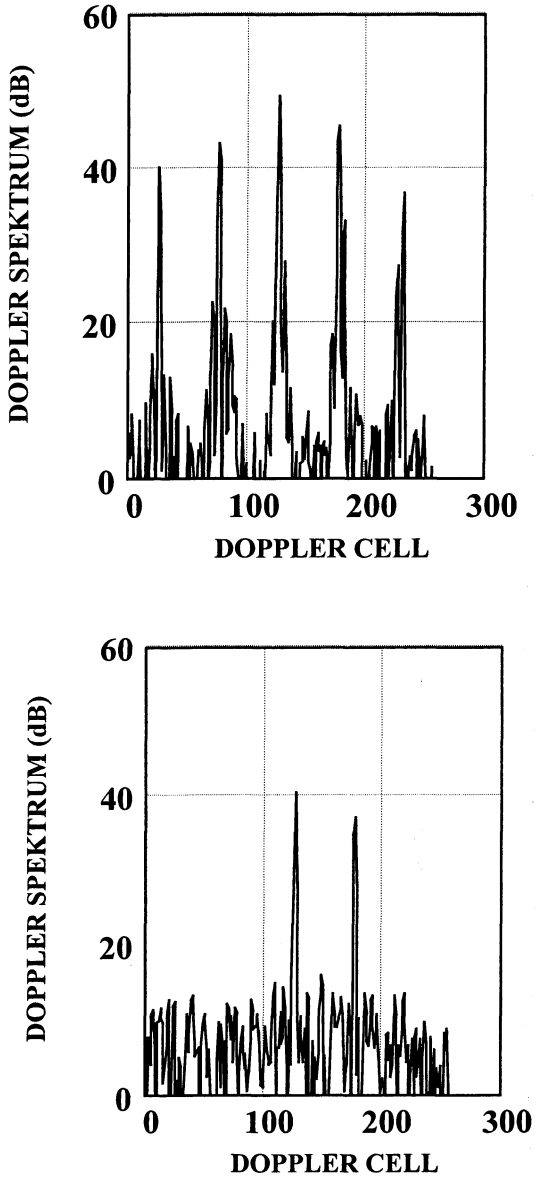


Figure 3: Simulated Doppler spectrum for two point targets moving with different velocities through the same range cell. Upper graph binary ADC with no noise added. Below with extra noise added before ADC (SNR=-7 dB). Computed for 1024 range cells and 256 Doppler cells.

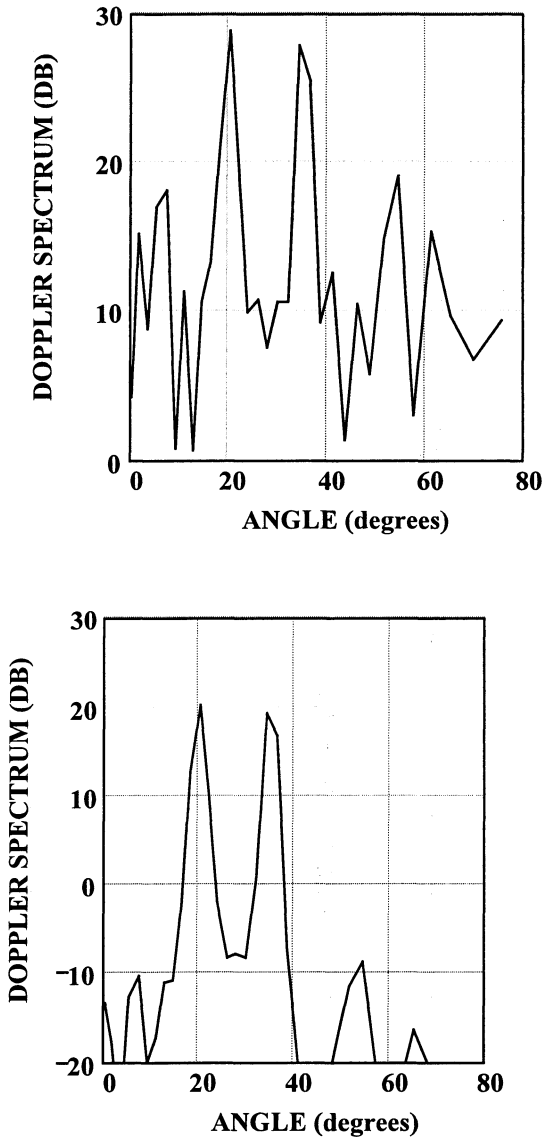


Figure 4: Simulated beam forming with binary ADC for a linear antenna array with 32 elements, 128x128 range and Doppler cells. Upper graph: no noise added before ADC. Lower graph: after noise has been added. Two targets are moving at 20 and 35 degrees through the same range/Doppler cell.

Modeling Inter-Domain Routing

Johan Nykvist and Lenka Carr-Motyčková
{johan,lenka}@sm.luth.se
Luleå University of Technology 97187 Luleå, Sweden

Abstract

Routing is one of several crucial components for securing the quality of end-user traffic. Characteristics of routing convergence and stability has a high impact on the ability of the network to perform quick route repairs and provide a good performance even in the presence of network faults.

In our on-going project, we are interested in convergence properties of inter-domain routing. In the first phase, we developed a discrete-event driven simulator with a high level abstraction model of the BGP protocol. The purpose is to perform simulations to examine the complexity of routing information exchange and to test different possible reasons for delayed convergence of BGP routing.

As a part of the verification of the simulation model, we have run simulations and compared the result with simulations done by others, and also compared the result with some theoretical findings. Our simulator showed theoretically predicted characteristics.

1 Introduction

The Internet is an important information and communication resource in many fields and societies around the world. Social infrastructures and economical interests depend on the efficiency and reliability of the network technology used in the Internet. The prospects of services such as voice over IP (VoIP) is tremendous if quality of service (QoS) and reliable connectivity for traffic on the Internet could be guaranteed to some extent. On a global scale the Internet of today only gives best-effort guarantees.

From a technical viewpoint, routing is one of several crucial components for securing the quality of end-user traffic. Characteristics of routing convergence and stability have a high impact on the ability of the network to perform quick route repairs and provide a good performance even in the presence of network faults. Inter-domain route repair convergence on Internet today might have a latency in order of several minutes, during which the quality of end-user traffic is highly degraded[1].

The Border Gateway Protocol version 4 (BGP)[2, 3, 4] is the de facto inter-domain routing protocol used in Internet today. BGP is a policy based path-vector protocol, distributing route information between Autonomous Systems (ASes). Since introduction of BGP several issues regarding its convergence properties have been arisen.

In our on-going project, we are interested in convergence properties of inter-domain routing. In the first phase, we developed a discrete-event driven simulator[5] with a high level abstraction model of the BGP protocol. The purpose is to perform simulations to examine the complexity of routing information exchange and to test different possible reasons for delayed convergence of BGP routing.

Several network simulators already exists, such as ns-2[6], JavaSim[7], SSFNet[8] and MaRS[9]. With the exception of MaRS, these simulators have a model granularity that includes the IP protocol. The model used in ns-2 even comprises link layer protocols. As we want avoid interaction with underlying protocols and stay on a control plane in first stage of our work, we built our own simulator.

Our simulation model includes some fundamental mechanisms of the BGP protocol, but excludes underlying mechanisms such as TCP/IP[10, 11]. Controlling simulations and analysing results with a high-level model is appropriate as a first approach to the problem. Low-level models are more complex with respect to the number of system variables and may hide important routing features. It also implies that the implementation of the model is less complex, which makes it easier to verify, extend and change. Furthermore, simulations with less detail scale better with respect to the size of the system model in terms of computational processing and memory usage.

In general, the selected abstraction level and the purpose of the simulations dictates how well a model represents its target system. We believe that our model possesses the characteristics of the BGP protocol that are relevant for the type of analysis we intend to perform, although extensive verification of the implementation of the protocol semantics yet has to be done.

In the remainder of the paper, an overview of related work and motivation for our work is given in section 2. Section 3 describes the simulator and the BGP simulation model. Section 4 presents some experiments performed with the simulator. Finally, in section 5 our conclusions and direction for future work are presented.

2 Related Work and Motivation

2.1 Experiments and Measurements

Major work on the field of Inter-domain Routing (IDR) convergence has been done by Labovitz et al. They conducted a large experimental study during 1996 – 1998 when they measured effects of route failures and updates on routing table changes and end-to-end traffic performance. They were primarily interested in routing instability, informally defined as the rapid change of network reachability and topology information [12]. Reasons for instability discovered here were: router misconfigurations, transient link problems and poor implementations of BGP protocol. Most BGP withdrawal messages have been duplicates generated due to a stateless implementation of BGP. The number of BGP announcements on the other hand grows with the decrease in the number of default-free routing-table entries. Instability is distributed across both autonomous systems and prefix space which suggests that there is a general problem that is not caused by particular routers. More serious problems - persistent oscillations (high frequency oscillations lasting for several hours or more) constitute about 10% of observed oscillations. Another major observation suggests that instability is closely tied to network usage patterns (congestion, failures) plus topological and policy changes.

Delayed convergence (the temporary routing table fluctuation generated during the path selection process on Internet backbone after a fault) as a result of Route Advertisement Messages timers setting is studied in [13]. Convergence time of BGP is dominated by an implicit time setting (30s) and the length of the longest alternative ASPath between the source a destination. Path selection depends not only on individual timer settings but also on the interaction of timers on multiple alternative paths in a topology.

In [13] Labovitz et al examine the nature of path failure, failover and repair due to convergence properties of BGP by both actively measuring and developing a theoretical upper and lower bounds for the convergence latencies. For routing table changes, it is shown that all types of routing updates, exhibit small but noticeable percentage of events lasting for as long as fifteen minutes. More than twenty percent of route fail-overs and forty percent of route withdrawals lasted for more than three minutes. Announcements of new routes and replacement of existing routes with shorter routes, converged faster and generated fewer update messages than route withdrawals and fail-overs. A strong correlation between the convergence latency and the amount of update messages was found.

The delays of routing convergence resulted in packet losses and an increased packet delivery latency for the end-to-end traffic. Immediately after a failover to a longer route, the traffic experienced average packet losses of up to 32 percent during a three minute period. A failover to a shorter route gave losses up to 17 percent during a one-minute period. The packet latency after a failure of any type more than tripled during a period of three minutes. Labovitz et al. also propose that sender-side loop detection be used, as this speeds up convergence time considerably. Griffin's experiments [14] on the other hand do not confirm that.

2.2 Theoretical Work and Simulations

Griffin et al. published several papers [15, 16, 17] that analyze BGP as a distributed graph algorithm. The problem with BGP is that it does not use shortest path routing (Dijkstra's algorithm). For algorithms that find a shortest path between routers, it is easy to prove that the computation will stabilize in finite time to a state that will not change until the network itself changes. Shortest path protocols like RIP converge in $O(n^3)$ [18] but these protocols do not allow any policy setting. Once BGP allows network operators to prefer a path without considering its length (policy decisions) the proof of convergence does not work any more. When we look closer to BGP routing, not only that we cannot prove its convergence but in practice we can observe changes of routing tables going on for several minutes even if the network topology itself does not change. It means that the routing policies of autonomous systems conflict, while with a different setting of the policies no oscillations would occur. Theoretical proof of BGP routing divergence was published in [19], where Varadhan et al. show that independent route selection protocols like BGP may exhibit persistent oscillations and never converge on a single route.

The question is, how can we decide whether the system is set up so that it would converge and how to identify the policy decisions that collide with others. Theoretically we could analyze all policy decisions together with the current topology at the time and decide about stability. This has two drawbacks: first, the policy decisions are not disclosed by most of autonomous systems and second, such a computation would take unreasonably long time. Griffin proposes a dynamic solution to the problem: he suggests to

record a *path history* and to check whether there is a cycle in it. A path history is a sequence of paths between a source and a destination node that are generated in a source node. A cycle in the path history is detected when a path is changed several times and then the same path is chosen again. If so, such a path should be supressed.

The same topology may behave differently with different policy settings. With consistent policy decisions, we may have a system that always converges, it means it behaves as a distant-vector routing scheme. Other policy settings may cause oscillations but at the same time there is a possibility that the system will converge nicely. This ambiguity stems from timing of routing message exchange. One sequence of exchanges leads to a stable state while with a different timing the system may oscillate. The third possibility is that the policy settings are so bad that it is impossible to converge to a stable state.

Griffin presents a formal specification of interdomain routing protocol which adds new attributes to BGP. The attributes are used to identify cycles in path histories. Once the oscillating paths are detected, they can be supressed (not advertised any more). A simulator described in [14] was used to verify optimal values for Minimum Route Advertisement Interval (MRAI), and effects of Sender Side Loop Detection (SSLD) and Withdrawal Rate Limiting (WRATE) on convergence. While a proper setting of MRAI seems to be effective, experiments with SSLD and WRATE did not show a substantial influence on the convergence time and the number of updates.

3 Simulator and System Model

3.1 Overview of the Simulator

Our discrete-event driven simulator, *RouteSim*, is implemented in the Java programming language and is composed of a simulation engine, model entities and events. The entities model components such as routers, links and messages, and the events represent changes of the system state (such as message arrivals). The simulation engine keeps all pending events chronologically ordered and dispatches the most imminent event in a cyclic fashion. The entities, and thus the system model, are declared in a text file and loaded before the simulation is commenced. Events are either created implicitly by an entity at some system state change, or explicitly injected. Injected events are typically events from outside the simulation model, for example a change in the routing policy for a certain router. Injected events are declared in a text file and loaded before the simulation is commenced.

3.2 Overview of the System Model

The system model forms a graph, describing a network topology of BGP speaking *routers* as vertices and communication *links* between the routers as edges.

A link entity models a full duplex point-to-point communication link, delivering *packet* entities between a pair of routers. A packet pending for transmission is enqueued in an unbounded FIFO queue and imposed with a transmission latency. The delivery delay (i.e. the delay between the packet being enqueued by one router and delivered to another) for one packet is the sum of its own transmission latency and all latencies of the packets in front of it in the queue. The transmission latency does not dependent on the packet's byte size or type. The value of the transmission latency and its uniformly randomized maximum deviation (a jitter) is configurable on a per link basis. As the latencies and the jitter are bounded, packet delivery is guaranteed and no packet loss is modeled. The characteristics of a link entity is comparable to a reliable transport protocol.

A BGP router entity models a *BGP speaker* that exchanges external BGP (EBGP) routing information with peer routers in neighboring ASes. Every router represents one particular AS, every AS consists thus of exactly one router, and Internal BGP (IBGP) is not modeled.

The simulation is performed on the *routing control-plane*, i.e. the only modeled communication is the exchange of routing information between peering routers. Forwarding of user traffic is not supported. Elaboration of link latency deviation may be used for modeling jitter on the link as caused by user traffic.

The model only supports the advertisement of one single network destination from a single router in any given topology. As route selection for a particular destination is not affected by decisions made for routing to other destinations, modeling a single destination is sufficient for most types of analysis regarding routing convergence and stability. For analysis involving route and/or prefix aggregation we would have to upgrade the model. The advantage of this approach is that the implementation of the system model and setup of simulations are simple. A disadvantage is that the model does not consequently consider the influence of concurrent workload of more than one destination imposed on the system in terms of CPU workload in the routers and traffic delay and jitter on the links. Even if the stochastic latencies of processing several destinations simultaneously can readily be modeled, effects of any possible global synchronization of processing many destinations concurrently is probably not feasible without supporting several destinations in the model.

CPU workload latency for every router is modeled

in the same way as described in [14]. In the same fashion as done in a link entity, every packet received by a router is imposed with a uniformly randomized latency before it is processed.

The routing information exchange and route selection is based on the specification of the BGP protocol according to [2]. The BGP protocol communication relies upon a reliable transport protocol. In Internet TCP is used for this purpose. To avoid the complexity and processing overhead of the TCP protocol in the model, BGP messages are sent directly over the links as these offer adequate reliability. As all communication in the system model is on a peering basis and no forwarding is supported, there is no need to model the IP protocol.

The only BGP message type modeled is the UPDATE message. A consequence of this is that peering between the routers is assumed to be in *established* state at all times during a simulation and no type of failure that would normally cause a NOTIFY message and terminate the connection can occur.

The only explicit addressing scheme used in the model is the labeling of the routers with AS numbers (ASN).

Without IP addressing and thus BGP *identifiers*, tie-breaking at route selection between routes with the same degree of preference is done by choosing the route offered by the peer with the lowest ASN.

The model incorporates the use of MinRouteAdvertisementInterval (MRAI) timers with optional jitter. The MRAI value and the timer jitter is configurable on a per router basis.

The model supports sender-side loop detection as proposed in [1], which can optionally be enabled on a per router basis.

A router can be configurated to heed routing policies. These policies have the form of an enumerated set of preferred AS-paths which overrides shorter, unpreferred paths.

4 Experiments and Results

As a part of the verification of the simulation model, we have run simulations and compared the result with simulations done by others, and also compared the result with some theoretical findings.

The behavior of BGP deteriorates in the case of a route withdrawal [1]. The path-vector protocol prevents routing-loops, but not the selection of infeasible routes. All learned routes to a certain destination becomes invalid when the destination is withdrawn, but the dynamics of the routing system impose a latency before this fact is propagated throughout the network. When facing a withdrawal, routers typically fall back to another route learned before the withdrawal. These newly selected routes are advertised and the prolonged behavior of enumerating routes with increasing lengths results in an exchange of a large number of routing messages and delayed convergence time. This behavior is somewhat similar to the count-to-infinity problem faced by Distant-Vector (DV) routing protocols [20]. For this problem a clique topology is the worst case for connectivity, and thus the freedom of enumeration routes, is greater than in any other type of topology.

In [1], Labovitz et al. presents theoretical bounds for convergence complexity and delay with respect to the number of ASes in a clique topology after a withdrawal. The theoretical upper bound for the number of computational states is factorial, $O(n!)$, assuming that no MRAI timers are used and message propagation delay is unbounded. It is further shown that the usage of MRAI timers plays an important role to limit the excessive enumerating of routes, as it forces synchronization of path selection into "computational rounds" where the AS path length increases monotonically for every round. In a clique topology consisting of n ASes, the usage of MRAI timers imposes a convergence delay with a linear lower bound of $\Omega((n-3) * MRAI)$. We note however that this bound holds for a routing system with synchronized MRAI timers, but not necessarily when the timers are temporally skewed.

We have performed 3000 simulations for studying what effect MRAI timers have on the convergence complexity in a clique topology at the event of a withdrawal. This was done by configurating a designated router to advertise, and later withdraw a route to the destination. The time between the advertisement and the withdrawal was kept long enough for the system to converge into an idle state where all routers had learned and selected a feasible route to the destination. After the withdrawal, the total number of routing messages and the time for the system to converge was measured. The system consisted of 15 routers and 105 links, configurated as follows; CPU processing latency between 0.01 and 1.0 seconds, no MRAI timer jitter, and routing decisions based on shortest-path route selection. The link latency was set to 0.01 seconds with no deviation.

To examine the impact of different MRAI timer values, 50 simulations with different random seeds where run for every MRAI integer value between 0 and 30 seconds. The graphs in figure 1 shows the result from the 1500 simulation runs. In figure 1(a) and 1(b), *the symptom of excessive route enumeration for low values of MRAI* is clearly shown in the form of large amounts of routing messages and long convergence delays. As the MRAI increases, this behavior is dampened as the MRAI timers impose synchronization of the route selections.

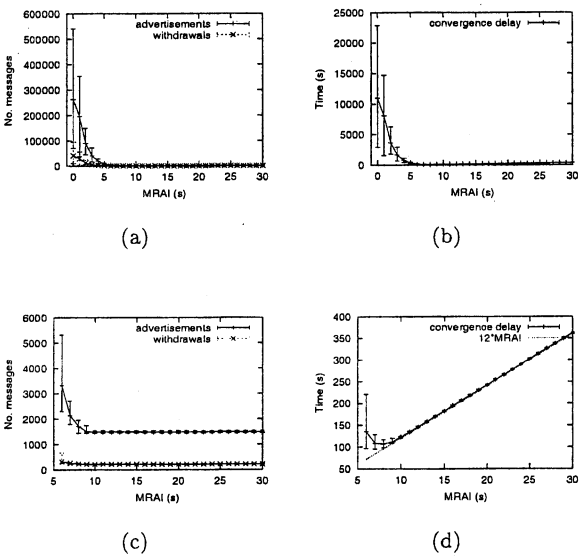


Figure 1: Graphs showing average and maximum deviation for traffic complexity and convergence time. The MRAI timers are synchronized.

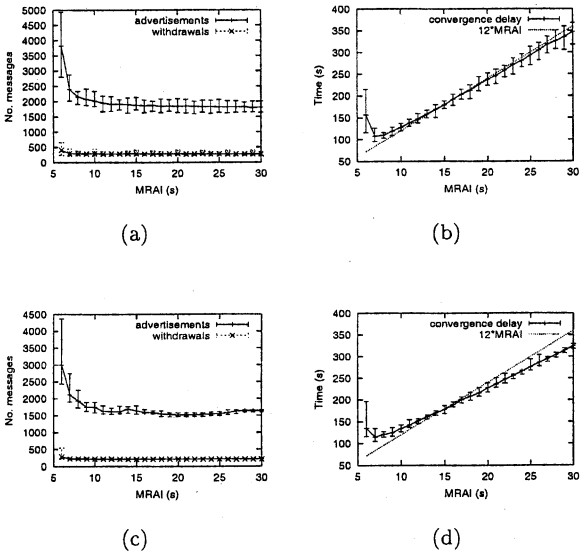


Figure 2: Graphs showing average and maximum deviation for traffic complexity and convergence time. The MRAI timers are skewed.

At a MRAI value of approx. 8 seconds, the system experiences the lowest amount of routing messages and the shortest convergence latency. At this point the MRAI is long enough to impose the route selection synchronization. For MRAI values larger than 8 seconds, the amount of routing messages is constant (figure 1(c)), but the convergence delay increases linearly. This follows from the fact that the MRAI timers impose excessive delay between the computational rounds in the route selection. Figure 1(d) shows how the convergence delay is confined by the lower bound, $\Omega((n-3) * MRAI)$. This result should however not be interpreted as that 8 seconds is the absolute optimal value for the MRAI, this is only true for this particular system, other configurations give different results[14].

In a real BGP routing network it is unlikely that all MRAI timers are temporally synchronized. First, the BGP protocol specification suggests the usage of timer jitter to avoid peaks in the distribution of routing messages. Second, even if it would be optimal to use one timer per peer and destination, most BGP routing implementations typically uses one timer per peer only, in order to save host system resources. We have performed additional simulations where we elaborated with skewed timers. The result is shown in figure 2. For the result shown in in figure 2(a) and 2(b) the MRAI timers were uniformly randomized between 0 and MRAI. In figure 2(c) and 2(d) the MRAI timers where evenly skewed between 0 and MRAI.

Our results are comparable to the results for the corresponding experiments done in [14].

5 Conclusions and Future Work

Our model implements a simplified version of BGP. To verify the correctness of our simulator we have run several tests and compared them to results obtained by others from BGP traffic measurements or simulations. The model exhibits properties corresponding to the behavior of Internet implementations of BGP in the tested areas of our interest. We were mostly interested in the influence of timer setting and their mutual synchronization on the convergence time and a number of update messages.

As a future work we plan to look at the possibility of adding more information to routing messages that could be used to suppress oscillations. On the other hand this additional mechanism should not increase the routing overhead. One possibility is to trigger the mechanism for recording a path history only if a node experiences a conflict. Implementation details, possibilities of recording a path history and suppressing oscillating paths, have to be solved and tested. It can hardly be expected that we can reach stable routing over the Internet. The point is to improve local stability and to decrease delays observed so far.

We are also interested in effects of a position of a router in a topology (relative to a backbone) and its timer setting on convergence. We perform extensive tests to assure that the properties of model correspond to real implementations.

Our plan for the future includes an extension of the simulation model with more BGP mechanisms such as route flap dampening[21] and inbound/outbound route filtering. Furthermore, we intend to perform simulations with more realistic topologies and routing policies.

References

[1] Craig Labovitz, Abha Ahuja, Abhijit Bose, and Farnam Jahanian. Delayed Internet Routing Convergence. In *Proc. of ACM SIGCOMM*, volume 30, pages 175–187, October 2000.

[2] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). Request For Comments RFC 1771, Internet Engineering Task Force, March 1995.

[3] John W. Stewart III. *BGP4: Inter-Domain Routing in the Internet.* Addison-Wesley, 1999. ISBN 0-201-37951-1.

[4] Sam Halabi and Danny McPherson. *Internet Routing Architectures.* Cisco Press, second edition, 2000. ISBN 1-57870-233-X.

[5] Jerry Banks, John S. Carson II, Barry L. Nelson, and David M. Nicol. *Discrete-Event System Simulation.* Prentice Hall, third edition, 2001. ISBN 0-13-088702-1.

[6] ns-2, `http://www.isi.edu/nsnam/ns/` .

[7] JavaSim, `http://javasim.cs.uiuc.edu/` .

[8] SSFNet, `http://www.ssfnet.org/` .

[9] MaRS, `http://www.cs.umd.edu/projects/netcalliper/software.html` .

[10] Jon Postel. Internet Protocol. Request For Comments RFC 791, Internet Engineering Task Force, September 1981.

[11] Jon Postel. Transmission Control Protocol. Request For Comments RFC 793, Internet Engineering Task Force, September 1981.

[12] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. Origins of Internet Routing Instability. In *Proc. of INFOCOM.* IEEE, June 1999.

[13] Craig Labovitz, Ahba Ahuja, Roger Wattenhofer, and Srinivasan Venkatachary. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proc. of INFOCOM.* IEEE, April 2001.

[14] Timothy G. Griffin and Brian J. Premore. An Experimental Analysis of BGP Convergence Time. In *Proc. of ICNP*, November 2001.

[15] Timothy G. Griffin and Gordon Wilfong. An Analysis of BGP Convergence Properties. In *Proc. of ACM SIGCOMM*, pages 277–288, September 1999.

[16] Timothy G. Griffin, F. Bruce Shepherd, and Gordon Wilfong. Policy Disputes in Path Vector Protocols. In *Proc. of ICNP*, November 1999.

[17] Timothy G. Griffin and Gordon Wilfong. A Safe Path Vector Protocol. In *Proc. of INFOCOM*, pages 490–499. IEEE, March 2000.

[18] Karthikeyan Bhargavan, Davor Obradovic, and Carl A. Gunter. Formal Verification of Distance-Vector Routing Protocols. In *International Conference on Theorem Proving and Higher Order Logics*, August 2000.

[19] Kannan Varadhan, Ramesh Govindan, and Deborah Estrin. Persistent Route Oscillations in Inter-Domain Routing. *Computer Networks*, 32(1):1–16, 2000.

[20] G. Malkin. RIP Version 2. Request For Comments RFC 2439, Internet Engineering Task Force, November 1998.

[21] C. Villamizar, R. Chandra, and R. Govindan. BGP Route Flap Damping. Request For Comments RFC 2439, Internet Engineering Task Force, November 1998.

Hybrid ARQ Based on Serially Concatenated Block Codes Using Iterative Decoding for Real-Time Communication*

Elisabeth Uhlemann[1,2], Tor M. Aulin[2], Lars K. Rasmussen[2,3] and Per-Arne Wiberg[1]

[1]Halmstad University, Box 823, SE-301 18 Halmstad, Sweden. E-mail: {bettan, per-arne.wiberg}@ide.hh.se
[2]Chalmers University of Tech., Dept. of Computer Eng., SE-412 96 Göteborg, Sweden. E-mail: tor@ce.chalmers.se
[3]Univ. of South Australia, Inst. for Telecom. Research, Mawson Lakes, Australia. E-mail: larsr@ce.chalmers.se

Abstract

A protocol for real-time communication over a wireless channel, based on concatenated codes using iterative decoding is proposed. The concept of deadline dependent coding (DDC), previously suggested by the authors, to maximize the probability of delivering the information before a given deadline, is further developed to include concatenated codes. The strategy of DDC is to combine different coding and decoding methods with automatic repeat request (ARQ) in order to fulfill the application requirements. These requirements are formulated as two Quality of Service (QoS) parameters: deadline (t_{DL}) and probability of correct delivery before the deadline (P_d), leading to a probabilistic view of real-time communication. An application can negotiate these QoS parameters with the DDC protocol, thus creating a flexible and reliable scheme.

1 Introduction

The tremendous development in wireless communication has provided opportunities in many related fields, such as remote sensing, monitoring, and control. New technologies and products are introduced into the market at an ever-increasing rate. This wireless evolution offers improvements for industrial applications, where traditional wireline solutions have prohibitive problems in terms of cost and feasibility. Wireline implementations are for example not cost efficient for large, temporary production lines, and may not be feasible at all for systems including rotating or high mobility machinery such as measurement and control of moving objects. Another sample application, requiring wireless access, is communication to and from different kinds of vehicles in factory automation situations.

Monitoring and control of industrial applications often introduce strict real-time constraints in order to guarantee proper operation. A time-critical system or a real-time system is characterized by the fact that it has deadlines to meet. In this work, we are interested in real-time communication systems where the timeliness of the delivered data is equally important as the correctness of the delivered data. This provides a challenge for *wireless* real-time communication due to the harsh communication environment encountered when transmitting over a wireless channel. The inherent consequence is a relatively high error rate, making the wireless channel unreliable in comparison to optical wireline channels. This has limited the extensive use of wireless access in real-time systems. Implementations of wireless communication systems for industrial use, guaranteeing real-time delivery, have been attempted, however, no general framework has been suggested. As a consequence, it is not straightforward to evaluate the general dependability of such systems with respect to real-time delivery and reliability.

1.1 QoS in Real-Time Communication

The literature in the real-time field often discusses two different classes of systems, *hard* and *soft* real-time systems. In a hard real-time system, late delivery cannot be tolerated. In contrast, in a soft real-time system a specified low probability of late delivery is tolerated, while permitting performance degradation by relaxing the real-time constraints. In this work we use a probabilistic view of the real-time constraints. This means it is no longer meaningful to talk about hard or soft real-time systems. We rather talk about deadline for delivery and the probability of succeeding in delivering correct information before this deadline. We therefore use two parameters: deadline (t_{DL}) and probability of correct delivery before deadline (P_d). They can be viewed as Quality of Service (QoS) parameters of the real-time communication system. It follows that a protocol can negotiate these parameters with an upper or a lower layer. The protocol can then either reject or accept a request, guaranteeing the delivery based on the given values of the QoS parameters. If the request cannot be accepted the application can possibly re-negotiate. One of the objectives of the real-time communication protocol is to maximize the probability that the communication system will be able to accept the request. Note that a hard real-time system defined using these QoS parameters will have P_d=1, a situation that cannot be achieved in any physical system due to noisy channel conditions, [1].

* This work is funded by the national Swedish Real-Time Systems research initiative ARTES (www.artes.uu.se), supported by the Swedish Foundation for Strategic Research.

1.2 Deadline Dependent Coding

The main idea behind the concept of Deadline Dependent Coding (DDC) [2] is to make all aspects of the communication protocol deadline dependent. The protocol should also attempt to minimize the required bandwidth, the transmitted energy and the time required to successfully deliver the information. The QoS parameters t_{DL} and P_d are mapped onto a retransmission protocol, which plays the role of maximizing the probability of correct delivery before deadline, using a minimum of resources. In a multiple access radio system, multiple access interference (MAI) is encountered. Hence, it is important to limit the interference created by the acting nodes. Consequently, another advantage of a carefully designed retransmission scheme is that we only use the required amount of redundancy suitable for the current channel condition and thus the amount of MAI is reduced.

The DDC scheme differs from existing communication protocols in a number of ways. Most importantly, DDC explicitly uses the deadline to control the transmission suite. There are a number of real-time communication protocols, e.g., [3] and [4] that are best effort protocols and consequently do not give any guarantees or explicit predictions on the probability of delivery. Other protocols have been developed, which guarantee hard deadlines [5], [6]. They depend, however, on a reliable channel, while DDC in contrast strives to maximize the probability of correct delivery over an unreliable channel.

In this paper the concepts of DDC are further developed to incorporate serially concatenated block codes using iterative decoding.

2 Iterative Decoding

Concatenated codes using iterative decoding is a way of providing long codes with manageable decoding complexity, [7],[8]. The iterative decoder is sub-optimal and hence less complex, but for certain channels, it approaches the performance of the optimal decoder in an iterative fashion.

A time and safety critical application benefits from the long powerful code yielding reliable communication, while the processing time of the decoder is kept low. Within a DDC protocol, the iterative decoding algorithm also gives the opportunity to always deliver something to the receiver just before the deadline. We can offer a fast tentative response and progressively provide iterative refinements. This last-minute delivery can be complemented by a measure of reliability of the delivered data based on *a posteriori* probability (APP) information [8].

A common approach for iterative decoding is to allow for a fixed number of iterations in the process. This may lead to unnecessary iterations, or to performance degradation if the process is terminated prematurely. Applying a performance based stopping criterion these problems can be addressed. Such stopping rules are connected to the convergence behavior of the iterative decoder. The convergence behavior of concatenated codes using iterative APP decoding algorithms has been paid considerable attention recently, e.g. [9], where a stopping rule based on thresholding the average log-likelihood ratio (LLR) of the so-called extrinsic APPs over a packet is found to be efficient. In this work, we consider this approach for serially concatenated codes. It should be noted that other methods, such as using an additional error detection code or using the error detection capabilities of the constituent codes might also be used. We have chosen the average LLR because it is simple and provides information about the current error rate.

The stopping criterion is intended to stop the iterative process once the mean LLR is above the threshold, however, if the decoding algorithm does not converge for a specific packet, this may never happen. As we are dealing with real-time communication we must have an upper limit on the time to decode a packet. Since the decoding complexity is directly related to the number of iterations, we consequently need to have an upper limit on the number of iterations. Examining the convergence behavior in a concatenated system, it is generally noticed that for a majority of packets, convergence is observed after a fixed number of iterations. Consequently, in this work the maximum allowed number of iterations is set accordingly. A non-negligible number of packets may, however, converge after less iterations.

3 Concatenated Hybrid ARQ

By concatenated hybrid automatic repeat request (CHARQ) we mean a HARQ scheme using concatenated codes as the error control code, but also a concatenation between retransmissions. Now, soft information may also be passed between retransmissions to be used in the iterative decoding process.

The convergence behavior of the iterative detector yielding the stopping criterion is also used to define the retransmission criterion that will be applied in this work. Consequently, we will not use an additional code for the sole purpose of error detection, but instead request a retransmission if the iteration-stopping criterion is not fulfilled after the maximum allowed number of iterations.

Since the retransmission scheme is intended for wireless real-time applications, the number of retransmissions is limited and the ARQ scheme becomes truncated [10]. Note that, once a request is accepted, the final decoder results are always delivered prior to the deadline, regardless of quality.

In a pure ARQ scheme, rejected packets are discarded. Previously received packets may, however, be used for so-called packet combining, in order to improve performance. There are two major types of packet combining, diversity combining, [11], and code combining, [12]. Diversity combining is in general done before the decoder whereas code combining is generally done within the decoder.

Since the CHARQ scheme is based on serially concatenated block codes using iterative decoding, some of the packet combining techniques may alternatively be viewed as different decoding strategies. The extrinsic APP information used in the iterative decoding process can be saved and used when decoding after a retransmission, similar to the work in [13]. The strategy of saving the extrinsic information can be seen as a packet combining technique, i.e. some sort of turbo code combining or concatenated code combining. The process of saving the extrinsic information requires no extra memory since this information is already saved to be passed and updated between the decoders. It is simply not reset when a new copy of a packet arrives. Saving the extrinsic information may also be seen as a doping criterion yielding faster convergence [14]. The concept of doping is based on providing side information to the decoding process. The extrinsic information from the decoding process of the previous transmission constitutes useful side information. The use of prior knowledge leads to a constructive bias, speeding up convergence. Applying this point of view, the extrinsic information can be used in conjunction with traditional diversity combining schemes. The basic scheme used in this work is a type-I HARQ scheme, hence we can directly apply the diversity combining technique equal gain combining [15]. This implies that the receiver just averages over the demodulator output metrics from all received copies of a packet to produce a combined packet for decoding. Hence, the code rate will remain the same.

3.1 Performance Results

We have used binary phase shift keying (BPSK) modulation [1] throughout this work, as it is a commonly used method in many existing hardware platforms. Further, we consider serially concatenated block codes based on two binary Reed Solomon RS(7,3) codes with a pseudo-random interleaver of size 945 used in conjunction with the ARQ protocol [16]. The communication channel is the additive white Gaussian noise (AWGN) channel [1], which models thermal noise present in all electronic equipment. An error free feedback channel is assumed throughout this work. The maximum number of iterations was chosen, based on simulation results, to be seven and the maximum number of retransmissions allowed was chosen to be four. Consequently, the maximum number of iterations that can ever occur is $7 \cdot 5=35$, yielding an upper bound on the execution time or communication time of each task. The LLR threshold was set to $\pm$ 10.

We consider the following procedure:

1. Use iterative decoding on the received packet, checking the LLR threshold after each iteration. If the stopping criterion is fulfilled go to step 3. If seven iterations are reached go to step 2.
2. If four retransmissions are made, go to step 3, else request a retransmission. Do packet combining, if any. Go to step 1
3. Output hard quantized values.

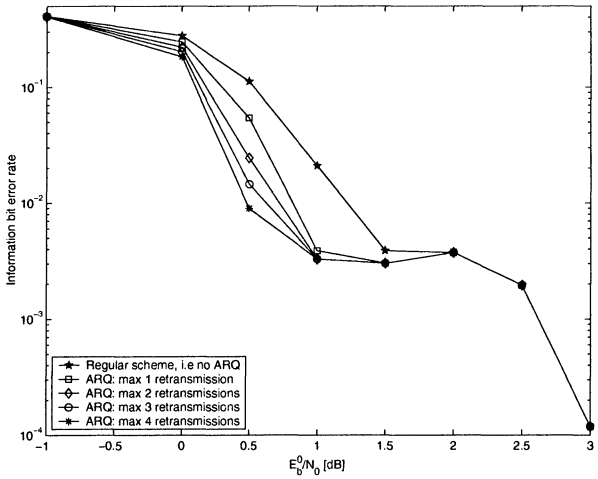


Figure 1: *BER as a function of E_b^0/N_0 for different ARQ schemes. Erroneous packets are discarded.*

In Figure 1, schemes allowing from 0 to 4 retransmissions are considered and compared in terms of bit error rate (BER) plotted against E_b^0 / N_0, where E_b^0 is the bit energy for the regular scheme with no retransmissions. For high signal to noise ratio, SNR= E_b^0 / N_0, all the curves tend towards the regular scheme without ARQ, since the probability of a retransmission tends to zero. For very low SNR, the curves of the different schemes are also on top of each other, since the probability of a retransmission tends to one. Virtually all transmissions are rejected and thus there is nothing to gain from a retransmission since the final packet must be accepted. The region between these two extreme values, however, is the working area of the ARQ schemes and a noticeable gain can be observed. The reason for showing the different truncation lengths in one figure is to see the improvements when more time is allocated to allow for additional retransmissions.

At high SNR, a loss in performance due to the stopping criterion can be observed. This is a result of the fact that the threshold is fixed for all values of SNR. The performance in terms of BER for a particular stopping criterion should, however, be compared to the number of iterations actually needed. It can be concluded that, although we have a considerable performance loss for high SNR, the average number of iterations needed to fulfill the stopping criterion is low [16]. It follows that we gain real-time benefits at the expense of BER performance. A truncated ARQ system offers most improvement in the region where the initial probability of retransmission is relatively high at the same time as the probability for accepting a packet within the allowed number of retransmissions is high. This is the case for the region 0-1.5 dB [16]. Consequently, in this region, the stopping criterion is good since we obtain faster convergence at virtually no performance loss. It follows that even though the chosen stopping criterion has an

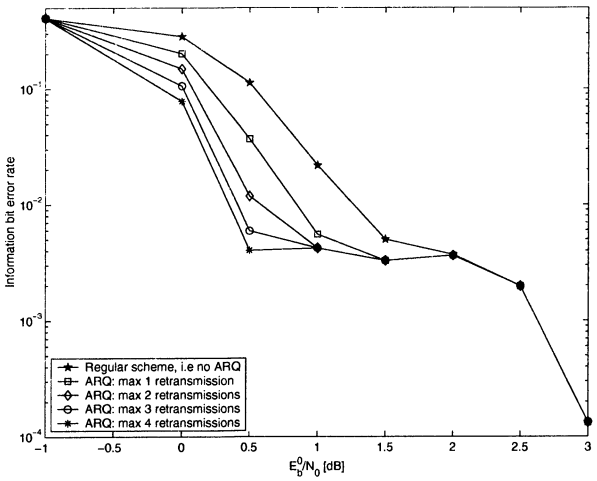


Figure 2: *BER as a function of E_b^0/N_0 for different ARQ schemes. The extrinsic information is used for doping.*

error plateau at 1.5-2 dB, it is still relevant in the SNR area where we expect to operate.

In Figure 2 the BER for the same schemes are shown, but this time the extrinsic information is saved and used when decoding after a retransmission. It can be seen that the BER for low SNR is improved as compared to the case with no packet combining in Figure 1. Note that the curves representing a regular scheme with no ARQ in both Figure 1 and Figure 2 are the same since no packet combining is made for these curves and the same stopping criterion is used.

When equal gain combining is applied, additional memory is required to store erroneous copies used in the combining process, but the gain in terms of BER is considerable. Saving the extrinsic information yields minor additional improvements to the resulting BER performance for the equal gain combining technique. Due to space limitations, no BER plots of equal gain combining are included. Results can be found in [16].

Whenever an ARQ scheme is considered its performance is usually measured both in terms of BER and throughput. In our case we are not only concerned with the number of retransmissions, but also the number of iterations made for each retransmission. We want to know if the packet combining technique considered improves both the throughput and the convergence speed. We will compare the scheme saving the extrinsic information to that where no packet combining technique is used. The relative frequency of occurrence of each iteration is plotted in Figure 3. Note that after 7 iterations a retransmission occurs. Consequently, iteration number 8 in Figure 3 corresponds to the first iteration of a newly received retransmission. The black columns in Figure 3 correspond to the scheme discarding erroneous copies of a packet. Hence, when a new packet arrives the iterative process is restarted. This can be seen as a staircase characteristic with a step every seventh iteration, since each new packet requires a few iterations to converge. The lighter columns correspond to the scheme saving the

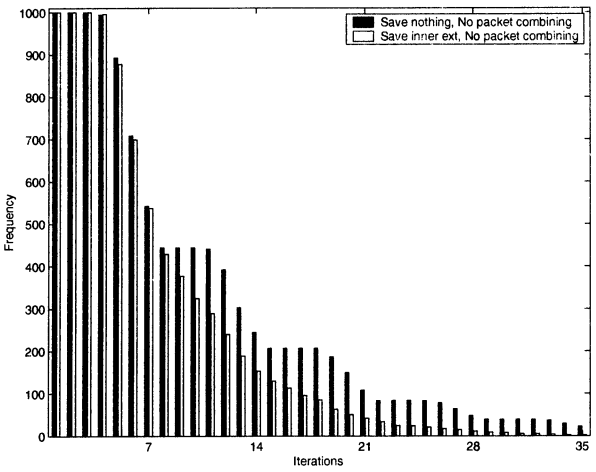


Figure 3: *Relative frequency of occurrence for each iteration. A scheme without packet combining is compared to a scheme saving the extrinsic information and is plotted at E_b^0/N_0 =0.5 dB.*

extrinsic information. The number of iterations decreases smoothly due to the fact that the decoder does not have to restart every time a retransmission arrives. Hence, convergence is faster. Note that the first seven iterations are the same for the two different schemes, as no retransmissions have yet occurred

In Figure 4 the equal gain combining schemes are compared in a similar way. Here, one scheme is using the saved extrinsic information and the other scheme is not. The most noticeable difference between Figure 3 and Figure 4 is that equal gain combining significantly reduces the number of iterations. For the first seven iterations in Figure 4 the two schemes are the same, since no retransmissions have been made and hence no packet combining is performed. However, we can still see an improvement in the convergence speed when the

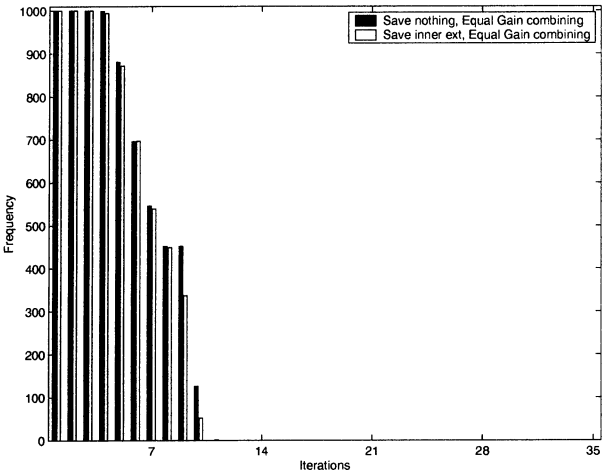


Figure 4: *Relative frequency of occurrence for each iteration, plotted at E_b^0/N_0 =0.5 dB. Both schemes employ equal gain combining, but one is saving the extrinsic information whereas the other is not.*

extrinsic information is saved and used, since the columns for iteration nine and ten have a lower frequency for this scheme. Consequently, since saving the extrinsic information does not require any major changes in the decoder it is reasonable to include it.

Additional results, together with a more detailed analysis can be found in [16].

When a real-time scheme is considered, the transmission propagation time has to be taken into account. Consequently, an iteration cannot be compared directly to a retransmission. In the bar charts on relative frequency of occurrence for each iteration, one may use one time unit to make one iteration, while a retransmission with a following iteration takes, for example, five time units.

4 Conclusions

The resulting DDC protocol presented here is based on concatenated hybrid ARQ techniques, creating a flexible and reliable scheme to meet real-time constraints. Each packet has a certain t_{DL} and P_d required by the user or the application. These two parameters will be translated into a maximum number of retransmissions allowed and a maximum number of iterations allowed for the sub-optimal iterative decoding algorithm per transmission.

If simple, cheap transmitters and receivers are required, e.g., a mobile sensor with limited battery supply, the mapping of the QoS parameters onto the CHARQ-DDC protocol may be done using a look up table. If the transmitter and receiver can be more costly, the mapping can be done adaptively based on the current estimated channel conditions.

Two different packet combining techniques for the CHARQ-DDC scheme were presented and evaluated. The diversity combining technique equal gain combining was applied. It is concluded that equal gain combining gives substantial performance improvements. However, since the hybrid ARQ system is based on serially concatenated codes using iterative decoding, a different category of packet combining techniques can be used. The iterative decoder passes extrinsic information between its constituent component decoders, which can be saved from one transmission and then used in the decoding process of a retransmission. This technique can be viewed as a code combining technique, and using it leads to performance improvement in terms of BER and convergence speed at very low additional cost in decoder complexity. Saving the extrinsic information for use in the decoding process of a retransmission can also be viewed as a doping operation of the iterative decoder. With this observation in mind, equal gain combining can be used in conjunction with the saved extrinsic information. It is concluded that the inclusion of the extrinsic information obtained from a previous transmission does speed up the convergence process of the iterative decoder.

References

[1] J. G. Proakis, *Digital Communications*, McGraw-Hill Book Co., New York, 1995, 3rd edition.

[2] H. Bengtsson, E. Uhlemann and P-A. Wiberg, "Protocol for wireless real-time systems", *Proc. of the 11th Euromicro Conference on Real Time Systems*, York, England, UK, June 9-11, 1999.

[3] W. Zhao, J. A. Stankovic and K Ramamritham, "A window protocol for transmission of time-constrained messages," *IEEE Transactions on Computers*, vol. 39, no. 9, pp. 1186-1203, September 1990.

[4] S.-K. Kweon, K. G. Shin and Q. Zheng, "Statistical real-time communication over Ethernet for manufacturing automation systems," *Proceedings of the IEEE Real-Time Technology and Applications Symposium*, Phoenix, Arizona, pp. 192-202, May 1999.

[5] T. F. Abdelzaher, E. M. Atkins and K. G. Shin, "QoS negotiation in real-time systems and its application to automated flight control," *IEEE Transactions on Computers*, pp. 1170-1183, Nov. 2000.

[6] J. Kim and K. G. Shin, "Performance evaluation of dependable real-time communication with elastic QoS," *Proceedings of the International Conference on Dependable Systems and Networks*, pp. 295-303, 2001.

[7] G. D. Forney, Jr., *Concatenated Codes*, M.I.T. Press, Cambridge, MA, 1966.

[8] C. Berrou, A. Glavieux and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: turbo codes," *Proceedings of the International Conference on Communications*, Geneva, Switzerland, pp. 1064-1070, May 1993.

[9] A. C. Reid, T. A. Gulliver and D. P. Taylor, "Convergence and errors in turbo-decoding," *IEEE Transactions on Communication*, pp. 2045-2051, December 2001.

[10] E. Malkamäki, *Performance of Error Control over Block Fading Channels with ARQ Applications*, Ph.D. dissertation, Helsinki University of Technology, October 1998.

[11] P. Sindhu, "Retransmission error control with memory," *IEEE Transactions of Communications*, vol. 25, no. 5, pp. 423-429, May 1977.

[12] D. Chase, "Code combining - a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Transactions of Communications*, vol. 33, no. 5, pp. 385-393, May 1985.

[13] K. R. Narayanan and G. L. Stüber, "A novel ARQ technique using the turbo coding principle," *IEEE Communications Letters*, vol. 1, no. 2, pp. 49-51, March 1997.

[14] S. ten Brink, "Code doping for triggering iterative decoding convergence," *Proceedings of the IEEE International Symposium on Information Theory*, ISIT 2001, Washington D.C., USA, p. 235, June 2001.

[15] M. Schwartz and W. R. Bennett, *Communication Systems and Techniques*, McGraw-Hill, Inc. 1966.

[16] E. Uhlemann, *Hybrid ARQ using serially concatenated block codes for real-time communication - an iterative decoding approach*, Licentiate Thesis, Chalmers University of Technology, Göteborg, Sweden, October, 2001. Available at http://www.hh.se/staff/bettan.

IP-based Distribution System for IEEE 802.11 Wireless LAN

Héctor Velayos, Gunnar Karlsson

Department of Microelectronics and Information Technology, KTH, Royal Institute of Technology, Electrum 229, 164 40 Kista, Sweden. E-mail: {hvelayos,gk}@imit.kth.se.

Abstract

We have designed an IP-based distribution system for IEEE 802.11 wireless LAN as an alternative to the predominant design based on Ethernet. Our distribution system allows the terminals to roam among access points connected to different IP networks. The terminals are not modified, they use the original wireless LAN link-layer mobility protocol. The access points implement a new Inter Access Point Protocol based on the exchange of IP datagrams to support the terminal mobility. The main advantage is the flexibility in the deployment of wireless LANs, while the drawbacks compared to the Ethernet-based design are additional overhead per packet and less efficient broadcast among the access points. We analyze the severity of these drawbacks.

1 Introduction

The increasing use of laptops and handheld devices equipped with IEEE 802.11 wireless LAN interfaces is demanding the wireless LAN networks to cover all the premises of an organization. In order to achieve this coverage, several access points must be deployed, since each one can cover only a few hundred meters in radius. The set of access points should behave like a single access point with increased coverage. This feature was already considered in the IEEE 802.11 standard, which included a distribution system to cope with this requirement. The distribution system connects the access points, providing link layer support for terminal roaming and frame exchange between access points. Figure 1 illustrates this architecture.

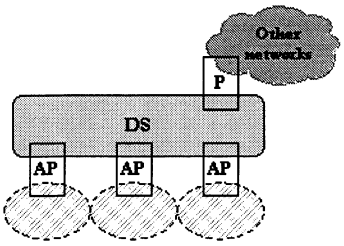


Figure 1: IEEE 802.11 architecture

There are several commercial access points that incorporate the distribution system functionality. However, they only provide link level distribution systems, despite the fact that the IEEE standard explicitly contains provisions to implement the distribution system at either the network or the link level. These products typically require an Ethernet network connecting the access points, which behave just like Ethernet bridges. They run the Inter Access Point Protocol (IAPP) to perform the distribution system services, transmitting the protocol information in Ethernet frames. Ethernet was chosen to connect the access points because it is cost-effective, widely available and its broadcast feature facilitates the IAPP implementation. This approach works well, but it limits the maximum size of the distribution system to the maximum size of an Ethernet network. While this limit is not a concern for most of the commercial usage scenarios, it is a problem for some large organizations because a single Ethernet cannot be run through all their premises. These organizations would need access points capable of intercommunicating across different IP networks, and thus enabling a single wireless LAN across the different networks of the company.

The next section summarizes proposed solutions to the deployment of this type of wireless LAN. Then we present our solution based on the redesign of the distribution system to span several IP networks, followed by the analysis of its features. The paper closes with some remarks and conclusions.

2 Related work

Different solutions have been proposed for the deployment of wireless LANs in large areas, which are typically served by different IP networks. The more established solution consists of installing an independent wireless LAN in each IP network and then using Mobile IP [1] to support roaming among them. When Mobile IP is introduced, terminal mobility is supported at two different layers: at the link layer by the distribution

system when roaming among access points belonging to the same wireless LAN, and at the network layer by Mobile IP when roaming among different wireless LANs. An example of such configuration is the wireless access of the IT University in Kista [2].

In this solution, the distribution system is implemented as an Ethernet network connecting all the access points and the router to other networks. Each access point behaves basically like an Ethernet switch between the radio interface and the wired Ethernet. In the uplink direction, when a frame is received through the radio interface, it is forwarded to the wired link. All the rest of the access points will receive the frame and use its bridging table to check if the destination is within the range of its own radio interface. If this is the case, it will forward the frame through its radio channel. Otherwise, the router will forward the frame to other networks. In the downlink direction, the router will forward the received frame to the shared Ethernet network, where all the access points will receive it. Again using the information in its bridging table, the access point that has the destination terminal in range will forward the frame to the air.

The usage of the Ethernet's broadcast facility eliminates the need to track the movement of the terminal among the access points. The information to update after a handover is the bridging table in the access points. This is not difficult since the terminal is forced to register with the new access point. This process is called association and produces two effects. First, the access point that receives the association message updates its bridging table. Second, this access point broadcasts a frame to the Ethernet so that the other access points can update their tables.

This process continues while the terminal roams among access points connected to the same Ethernet. When it roams to an access point in a different Ethernet network, the Mobile IP agents must act to maintain the connections. Typically the router of the wireless LAN behaves as home agent for local terminals and as foreign agent for visiting terminals. Thus, when the terminal reaches an access point of a different wireless LAN, it must register with the foreign agent in addition to the registration with the access point. Then the foreign agent establishes an IP tunnel to the terminal's home agent and assigns a temporary IP address to the terminal. The terminal uses this address while visiting the new network, so its traffic is encapsulated and forwarded to and from its home network. In this way, the terminal keeps working as if it was in the home network.

This solution has demonstrated that it works, although it has some drawbacks. Most probably efficiency is the main one. When the terminal is away from its home network, the traffic is forwarded first to the home network and then to the visited network. Moreover, the second hop implies additional overhead per packet, since an extra IP header will be added to each datagram as part of the encapsulation process. Another drawback is the need to install the Mobile IP agents in both routers and terminals. Our design tries to overcome these problems by not using IP-in-IP encapsulation for the traffic forwarding between access points and by limiting the modifications in the existing infrastructure to the access points; routers and terminals remain unchanged.

Other proposals avoid the introduction of Mobile IP in our scenario by modifying the architecture of the wireless network. In these proposals each terminal will obtain an IP address to use while roaming inside the organization between the different wireless LANs. Thus, in these schemes, the IP address is only used to identify the terminal and is never used for packet routing. Each access point will behave like an IP router with a wireless LAN interface, constituting an independent wireless LAN. Therefore there is no distribution system and no terminal mobility support needed in the link layer. The terminal can join any wireless LAN in range to send IP packets. In addition, an IP-based mobility protocol must be run among the routers connecting the access points to deliver datagrams to the terminal. This protocol will track the movement of the terminal and update the downlink routing information accordingly. Such protocols are called micro-mobility protocols and architectures based on them are under active research [3].

These micro-mobility protocols have been compared extensively to Mobile IP to show that they have better performance if the mobility is mainly a local process among neighbor routers [4]. However, in our scenario their drawbacks are similar to those of Mobile IP. All the proposals require the active participation in the micro-mobility protocol of the terminals and routers with the radio interface, thus they have to be modified. Moreover, some modifications in the other routers of the network are required. The number of intermediate routers affected depends on the forwarding strategy of the downlink traffic when the terminal roams. It varies from all intermediate routers being modified when per-host-forwarding schemes are used to only some particular routers altered when tunneling plus signaling strategies are used. There is little experience in the deployment of these schemes, but most probably the extensive modifications required to the existing infrastructure will limit their adoption.

3 Distribution system architecture

The solutions presented in the previous section cope with the limitations of the Ethernet-based distribution system in large wireless LANs by adding functionality to the network layer. Therefore, the modifications affect not only to the access points but also the routers and terminals. In contrast, we propose to use the original

IEEE 802.11 architecture presented in Figure 1 and only modify the distribution system, thus restricting the changes to the access points. In this section we describe our design of an IP-based distribution system for IEEE 802.11 wireless LAN networks. El-Hoiydi presents in [5] more details of the Ethernet-based option. Our design allows the installation of the access points belonging to the same wireless LAN in different IP networks, thus achieving the range extension while simplifying the installation of wireless access in campus networks. Our proposal maintains the architectural elements and functions of the IEEE 802.11 standard, modifying only the implementation of the distribution system in the access points. Our distribution system is an unnoticeable replacement to the Ethernet-based one seen from the terminals since they operate with the access point using the standardized 802.11 MAC protocol.

We will present our distribution system by describing the behavior of the access points. Each access point obtains a valid IP address through its wired interface from the IP network. This IP address is used in the communication with the access points in other IP networks. Using IP datagrams, the access points exchange the required information to accomplish the functionality of the IEEE 802.11 distribution system. This requires the design of an Inter-Access Point Protocol based on IP datagrams, which takes care of the terminal mobility and the forwarding of user data frames between access points. Figure 2 illustrates the protocol stack of the access point. In the Figure, two terminals are shown communicating through their respective access points. In the upper part the access points are connected using an Ethernet-based distribution system, while in the lower part our IP-based system is used.

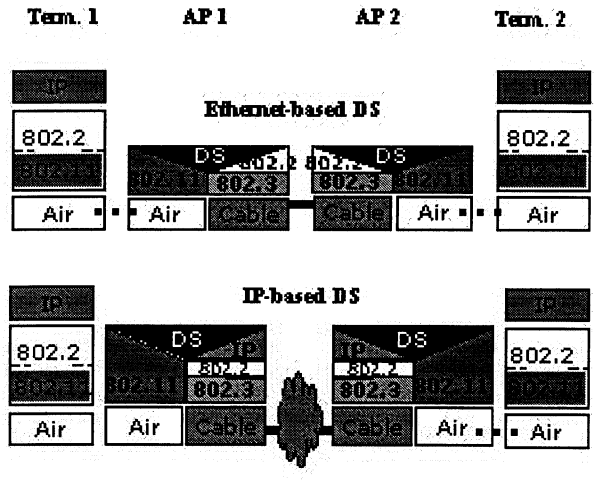


Figure 2: Protocol stack for different types of distribution system

From the terminal point of view, the set of access points and their interconnecting IP network will be perceived as a link-layer link with a gateway for the incoming traffic. We call it wireless gateway and it is the point where the incoming traffic enters to the distribution system. Then the distribution system forwards the traffic to the access point through which the destination terminal can be reached. Any router in the network can play this role. The only requirements are that it can forward IP packets to the access points and it implements the distribution system functionality. Every access point must know the IP address of this wireless gateway, since it will play a key role in the mobility management described later.

The terminal operates with the access points following the normal procedures established in the IEEE 802.11 MAC protocol. Before it can transmit any data, the terminal must authenticate and then associate with any access point in range. These processes allow the access point to be aware of the terminals in its coverage area. Upon the reception of the association message, the access point will register in the wireless gateway the current location of the terminal, so the incoming traffic will be forwarded by the wireless gateway to the access point, which in turn will forward it to the terminal. This is achieved through a two-step encapsulation of the incoming IP packets. Firstly, they are encapsulated in IP datagrams to cross the distribution system from the wireless gateway to the access point. There the original IP datagram is extracted and encapsulated again in a wireless LAN frame to cross the radio media to the terminal.

The packets generated from the terminal follow a similar process. Initially, they are encapsulated in wireless LAN frames to reach the access point through the radio link. Then the access point has to determine the location of the terminal whose MAC address appears as destination in the frame. It will first search in the list of its associated terminals. If it is not there, it will ask the wireless gateway whether the terminal is associated with any other access point. If the wireless gateway answers positively, the answer will contain the IP address of the destination access point. Thus the packet will be encapsulated in an IP datagram and sent to that destination. If the wireless gateway does not contain any record of the destination, the destination is considered outside this network and the access point will extract the original IP packet from the received wireless LAN frame. This IP packet will be routed through the wired interface using the standard routing table; it will not be forwarded to the wireless gateway through the distribution system. This contributes to reducing the traffic load in the wireless gateway. In any case, the access point will cache the answer from the wireless gateway for future use.

Since the terminal location functionality is based on querying to the wireless gateway, its information must be updated each time a terminal roams between access points. The update of the information is the responsibility of the access point receiving the terminal

and does it when the terminal associates with it. When the wireless gateway receives the update, it generates additional updates to all the access points that have entries for this terminal in their caches.

This procedure implies that the terminal will lose some packets after a handover because the update message will need some time to reach the wireless gateway. Before the update arrives, the wireless gateway will send incoming packets to the old location. To eliminate this problem, we use an effective movement prediction scheme based on the authentication information. Before the terminal can associate with an access point, it must authenticate itself with the access point. This process is made well in advance of the handover because it typically requires waiting while the access point contact external servers to authenticate the terminal. Pre-authentication helps to reduce the handover time.

The access points where a terminal is authenticated are the candidates for the next access point after a handover. The number of candidates will not be large because the relatively small number of available channels in wireless LAN networks limits the number of access points covering the same area without interfering with each other to three. In most of the situations, the terminal will find one access point in range and a second will be reachable only in the borders of the coverage area of the first one. This means that the wireless gateway can send copies of the incoming packets to all the access points where the terminal is authenticated, not only to the access point where the terminal is associated, without wasting excessive bandwidth. This guarantees that the packets will not be lost during handover.

4 Analysis

Our solution presents two main advantages compare to the Mobile IP and micro-mobility schemes. First, it is easier to deploy, since the new functionality is distributed between the access points and the wireless gateway. There is no need to modify existing routers, since they do not participate in the mobility support, or terminals, since they use the standardized messages of wireless LAN to roam among the access points. Second, the terminal mobility is handled at the link layer level with full access to its signaling during the handover between access points. In our design, we use the pre-authentication with the next access point before the actual handover happens to predict the movement of the terminal. With this knowledge, the distribution system delivers copies of the packets addressed to the terminal to the current and the future access points, thus minimizing the traffic flow disruption during handovers.

If we compared our IP-based distribution system with the more traditional Ethernet-based one, we can highlight as an advantage the higher flexibility of the first one, since the access points can be installed in different IP networks.

The main disadvantage of our design is the lack of broadcast in the IP network that connects the access points. This complicates the search of the access point to which the destination terminal is attached. Currently we are using the wireless gateway as a register for the location of each terminal. However this solution has several drawbacks. Firstly, the wireless gateway represents a single point of failure and it is a bottleneck. Secondly the position of the wireless gateway in the network topology is critical, since the delay between the access points and the wireless gateway must be maintained as low as possible to achieve fast answers when information is registered or queried. This is especially critical during handover, when the update message must travel first to the wireless gateway before reaching the old access point.

We are evaluating an alternative approach to the centralized register, in which the information is maintained distributed among the access points and wireless gateway. In this approach, we are using an IP multicast group which replaces the Ethernet's broadcast service. All access points join the multicast group to receive management messages, so every access point can maintain its own location information. In addition, the common multicast group can be used to search a terminal's access point. The problems are that multicast is not widely available in current networks and that the complexity of the access points is increased because they must support IP multicast transmission.

Another problem of our approach is the higher overhead per packet when the frames are transmitted between access points compared to the distribution system based on Ethernet. Using Ethernet frames as transport adds 19 bytes per frame, while using IP version 4 adds another 20 bytes of the IP header. For large packets the difference is not important, but for small packets like the typical 120 bytes packet of the Voice over IP traffic, this overhead cannot be neglected. Although it could be considered high for some applications, it is the same overhead that the traffic would suffer while crossing the Internet.

5 Conclusions

We have presented our IP-based distribution system for wireless LAN networks. Its main feature is the flexibility in the installation of the access points because there is no need to have a common Ethernet backbone among them. The access points will use standard IP datagrams to exchange signaling messages and packets. The drawback to this increased flexibility is the lower efficiency of the distribution system in terms of overhead per packet and handover time.

References

[1] C. E. Perkins, B. Woolf, S. R. Alpert, "Mobile IP Design Principles and Practices", Prentice Hall, 1st edition, January 1998.

[2] A. Escudero, B. Pehrson, E. Pelletta, J.O. Vatn, P. Wiatr, "Wireless access in the Flyinglinux.NET infrastructure: Mobile IPv4 integration in a IEEE 802.11b", 11th IEEE Workshop on Local and Metropolitan Area Networks, LANMAN2001, Co, USA, March 2001.

[3] A. Festag, H. Karl, G. Schäfer, Adam Wolisz, "Current Developments and Trends in Handover Design for ALL-IP Wireless Networks", Technical University Berlin, Telecommunication Networks Group, August 2000.

[4] J. Manner, A. López, A. Mihailovic, H. Velayos, E. Hepworth and Y. Khouaja, "Evaluation of Mobility and QoS Interaction", Computer Networks, Vol. 38, Issue 2, Feb. 2002, pp.137-163. Elsevier Science Publisher.

[5] Amre El-Hoiydi, "Implementation Options for the Distribution System in the 802.11 Wireless LAN Infrastructure Network", Proc. of 2000 IEEE Internacional Conference on Communications, vol 1, pp 164-169, New Orleans, USA, June 2000.

Privacy Extensions for Stateless Address Autoconfiguration in IPv6 "Requirements for unobservability"

Alberto Escudero Pascual <aep@kth.se>
Royal Institute of Technology (KTH)
Kista. Sweden

26th March 2002

Abstract

Stateless address autoconfiguration defines the mechanism for a IPv6 node to generate an address without the need of an external DHCP server based on the interface identifier. In the case of Ethernet the Interface Identifier is based on the EUI-64 identifier derived from the interface's built-in 48-bit IEEE 802 address (MAC address). The IPv6 address generated via Stateless Autoconfiguration contains the same interface identifier regardless of the location the mobile node is attached to the Internet. RFC3041 presents a privacy extension to Stateless Autoconfiguration based on the idea of generating random interface identifiers periodically.

The paper introduces the concept of "unobservability" of the privacy extension and studies in which scenarios a third party will be able to determine with high probability if a node is running RFC3041 or not. The paper shows the privacy implications of the universal/local bit of the current IPv6 addressing architecture and presents a set of suggested changes to enhance privacy.

1 Background

This paper is divided as follows: Section 1 contains a very brief overview of the Internet Protocol Version 6, Section 2 describes the different mechanisms to generate a global scope address including the privacy extension for stateless address configuration RFC3041 [6]. Section 3 defines unobservability of the privacy extension. Section 4 introduces the level of privacy extension observability for different scenarios and finally in Section 5 we present a set of suggested changes to the current IP Version 6 Addressing Architecture", RFC 2373 [3] to enhance privacy.

1.1 The Internet Protocol version 6

The Internet Protocol Version 6 (IP$v6$), also called "IPng" (IP Next Generation), is the latest version of the Internet Protocol (IP). Formally, IPv6 is a set of specifications from the Internet Engineering Task Force (IETF). IPv6 was designed as an evolutionary set of improvements to the current IP Version 4. The most obvious improvement in IPv6 over the IPv4 are that IP addresses are lengthened from 32 bits to 128 bits which anticipates the future growth of the Internet and provides relief for what was perceived as an impending shortage of network addresses. Besides, IPv6 offers technical advantages over IPv4, including self-configuration mechanisms, enhanced security, quality of service features and native mobility support [1].

IPv6 includes a security protocol in the network layer that provides cryptographic security services that supports combinations of authentication, integrity, access control, and confidentiality. The IP

Encapsulating Security Payload (ESP) and the IP Authentication Header (AH) are part of the IP Security architecture (IPSEC) described in RFC 2401 [2].

Both ESP and AH are mandatory parts of IPv6 and make sure that a third party eavesdropping on the channel can not read and/or modify the IP datagram. The IP Authentication Header seeks to provide security by adding authentication information to each IP datagram whilst confidentiality requires the use of ESP. Neither AH nor ESP hide the source and destination IP addresses of the communicating parties and hence their network location.

The protocol operation defined for mobility in IPv6 is known as MobileIPv6 and allows a mobile node to move from one link to another without changing the mobile node's IP address. A mobile node is always addressable by its "home address", an IP address assigned to the mobile node within its home subnet, i.e., with the prefix of its home link. Packets may be routed to the mobile node using this address regardless of the mobile node's current point of attachment to the Internet, and the mobile node may continue to communicate with other nodes while using this address, even after moving to a new link. With specific support for mobility in IPv6, packets destined to a mobile node would be able to reach it even while the mobile node is away from its home network.

In summary, IPv6 provides new security opportunities which include message integrity, authentication, and confidentiality (IPSEC) and the possibility for a mobile node to be always addressable by its "home address" (MobileIP). All these functionalities rely on treating the fixed IP address of the node as an indentifier. In the case of IPSEC end-to-end security uses the fixed IP address as part of the security association and mobility requires to the mobile node to send the fixed home address included in a destination option.

In the next section, as part of the description of the different mechanisms to obtain a global address, we present the possible threats for privacy when an IP identifier can be linked with personal identifiable information as a "personal device" and how the existing privacy extension has not taken into consideration that the user might be also interested in hiding the fact that is using the privacy extension itself.

2 Mechanisms to obtain a global address

We have clasiffied the mechanisms to obtain a global address as follows: stateless and stateful address configuration, manual, cryptographic generated and random addresses which include the privacy extension and IPv6 over PPP with no global identifier available.

2.1 Stateless Address Autoconfiguration in IPv6

Stateless address autoconfiguration defines the mechanism for a IP_{v6} node to generate an address without the need for an external DHCP server based on the interface identifier [8,4]. The stateless approach is used when a site is not particularly concerned with the exact addresses hosts use.

To insure that all configured addresses are likely to be unique on given link, nodes do "duplicate address detection" on addresses before assigning them to an interface. The Duplicate Address Detection (DAD) algorithm is used to check all addresses, *independent* of whether they are obtained via stateless or stateful autoconfiguration.

2.1.1 Interface Identifier

In the case of stateless address autoconfiguration for Ethernet the Interface Identifier is based on the EUI-64 identifier derived from the interface's built- in 48-bit IEEE 802 address (MAC address).

The EUI-64 is formed as follows:

1. The first three octets that corresponds to the Organizationally Unique Identifier (OUI) of the Ethernet address become the *company_id* of the EUI-64. The OUI blocks are assigned by IEEE.

2. The fourth and fifth octets of the EUI-64 are set to the fixed value FFFE hexadecimal (encapsulation of IEEE 802 address in EUI-64).

3. The last three octets of the Ethernet address (extension identifier values) become the last three octets of the EUI-64.

The Interface Identifier is then formed from the EUI-64 by complementing the "Universal/Local" (U/L) bit, which is the next-to-lowest order bit of the first octet of the EUI-64. Complementing this bit will generally change a 0 value to a 1, since an interface's built-in address is expected to be from a universally administered address space and hence have a **globally unique value.**

A universally administered IEEE 802 address or an EUI-64 is signified by a 0 in the U/L bit position, while a globally unique IPv6 Interface Identifier is signified by a 1 in the corresponding position. In brief, a globally unique IPv6 interface based on a hardware token that is unique will carry one bit indicating its uniqueness.

[3] states that motivation for inverting the "u" bit when forming the interface identifier is to make it easy for system administrators to hand configure local scope identifiers when hardware tokens are not available and to allow development of future technology that can take advantage of interface identifiers with global scope.

2.1.2 Link, site and global-scope addresses

The link-local address of an IP_{v6} node is the result of combining the global unique interface identifier with the reserved link-local prefix FE80. The site-local and global-scope addresses are created by combining prefixes advertised in Router Advertisements and Neighbor Discovery.

The IP_{v6} address generated via Stateless Autoconfiguration contains the same interface identifier *regardless* of the location the mobile node is attached to the Internet.

Even when higher communication layers encrypt their payloads (for example with ESP), there is not an easy mechanism to hide the addresses in packet headers as they appear in clear, this fact makes it very easy for an eavesdropper to track mobile nodes by analyzing the prefixes related to a certain interface identifier.

For example, let's consider a device with built-in 48 bit EUI-48 address 00:01:02:65:71:37 that moves from a network A with prefix 3ffe:200:15:1 to a network B with prefix 3ffe:200:17:2.

1. A unique EUI-64 value is generated by concatenating the company_id, an $FFFE_{16}$ valued label, and the extension identifier values, obtaining 00:01:02:ff:fe:65:71:37.

2. The global unique interface identifier is then formed by complementing the "Universal/Local" bit, resulting 02:01:02:ff:fe:65:71:37.

3. The link local address is fe80::201:2ff:fe65:7137 in both networks

4. When moving from network A to B the mobile device will change from 3ffe:200:15:1:201:2ff:fe65:7137 global IP_{v6} address to 3ffe:200:17:2:201:2ff:fe65:7137.

2.2 Privacy Extension for Stateless Address Configuration

Narten and Draves [6] developed a privacy extension for Stateless Address Configuration based on the idea of generating random interface identifiers periodically. They describe two approaches for the maintenance of the randomized interface identifier depending upon weather stable storage is present (history scheme) or not. The use of the history scheme tries to avoid the scenario where two nodes generates the same randomized interface identifier, both detect it via DAD, but then proceed to generate identical randomized interface identifier via the same flawed random number generation algorithm.

The first big difference between the EUI-64 based interface identifiers and RFC3041 is that the latest can not claim to be globaly unique and hence the universal bit must be set to zero.

2.3 Stateful address configuration

The dynamic host configuration protocol (DHCP) is the stateful counterpart to stateless autoconfiguration in which hosts obtain interface addresses and/or

configuration information and parameters from a server. Servers maintain a database that keeps track of which addresses have been assigned to which hosts. Stateless and stateful autoconfiguration complement each other. For example, a host can use stateless autoconfiguration to configure its own addresses, but use stateful autoconfiguration to obtain other information.

2.4 Manual configuration

If an IEEE global identifier is not available a different source of uniqueness should be used. Suggested sources of uniqueness include link-layer addresses, machine serial numbers, etc. In this case the "u" bit of the interface identifier must be set to zero.

If a good source of uniqueness cannot be found, it is recommended that a random number be generated. In this case the "u" bit of the interface identifier must also be set to zero.

2.5 Cryptographically Generated Addresses

The Cryptographically Generated Address where introduced to solve the problem of address ownership [14] in mobileip. In all the different proposals the CGA addresses have a strong cryptographic binding with a public key. CGA are obtained by means of a one-way hash functions.

For example in the case of SUCV IDs [7] the CGA addresses are created as follows:

$$CGA = 64bit \quad prefix + CGIID$$
$$CGIID = f(PublicKey, j)$$

where $f()$ is the least significant 64 bits of SHA-1 hash of Public Key concatenated with 16 bit counter j. The universal bit u is set to zero.

2.6 Classification of Interface Identifiers

The IPv6 interface identifiers can be clasified as follows:

1. Half of this space, with the universal bit set to 1, is given to IEEE EUI-64 identifiers.
2. The other half, with the universal bit set 0 is used for Interface Identifiers that are not globally unique including:

- manually assigned.
- local unique assigned by DHCPv6.
- random interface identifiers used in RFC 3041 [6] and RFC 2472 [5] section 4.1.
- cryptographically generated addresses.

3 Unobservability of the privacy extension

In [15] unobservability is defined as the property that ensures that a user may use a resource or service without others, especially third parties, being able to observe that the resource or service is being used.

Unobservability requires that a third party cannot determine whether an operation is being performed based on certain knowledge of the subject(s) of observation.

In our analysis we define as the item of interest for the attacker to be able to determine if the victim is using RFC3041. In order to try to describe different scenarios it is important to identify possible sources of knowledge for the attacker to make a decision. The basic source of information is the link local or global scope address of the victim but other sources could be available as:

- The victim's mac address.
- The presence of a DHCPv6 serving the prefix of the victim's network.
- If the victim has mobility support and if its address is a CGA.
- The victim's hardware type and/or operative system.

4 Scenarios

According to the amount of knowledge available to the attacker we can divide all scenarios in two.

4.1 Attacker is not present in the link

In this scenario the attacker is not present in the link and is in the path between the victim and a correspondent node. In our model we assume that the attacker knows about which OUIs has been assigned by IEEE but not the distribution of the "extension identifier values" for certain OUI.

In first place the attacker determines if the interface identifier is based on a EUI-64 identifier by checking the universal/local bit. If the universal/local bit is set to one, the attacker checks if the OUI has been assigned by IEEE. If the OUI has not been assigned by IEEE the attacker can presume that the victim has either modified its MAC address, configured manualy the interface identifier or generated a random interface identifier without setting the u-bit to zero.

If the universal/bit is set to zero, the attacker can choose among different possibilities : the interface has been configured by stateful address configuration, manualy or is random interface identifier (CGA, RFC3041, RFC2472).

The attacker can gather extra knowledge as follows:

- By observing the nature of the traffic between the victim and a correspondent node. Cryptographically Generated Addresses are included in authentified binding updates and are part of destination/routing header options.

- By discovering the presence of a DHCP server via the (FF05::1:3) site-scoped multicast address.

4.2 Attacker is present in the link

In the second main scenario the attacker is present in the link, most of the information required to observe if the victim is using the privacy extension can be gathered by victim's traffic analysis. The attacker will be able to observe not only the presence of a DHCP server but also the traffic exchange between the DHCP server and the victim. If the victim has a CGA, the attacker will be also able to observe the flow of authentified mobility bindings.

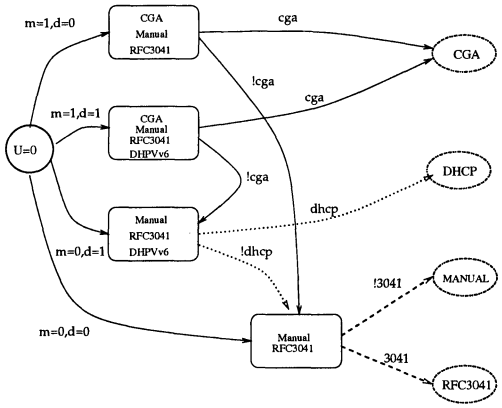


Figure 1: Possible scenarios

In the case that the victim is not using stateful address configuration or is a mobile node with a CGA, the attacker has to decide between two possible options [Fig. 1]: the victim is running RFC3041 privacy extension or has configured manualy its address. The attacker then, can try to match during a certain period of time the hardware addresses with the link or global scope addresses. Nodes running RFC3041 will change their addresses periodicaly while keeping the same hardware address.

4.3 Conclusions

[Fig. 1] shows all possible scenarios considering the amount of knowledge available to the attacker. In all the cases the attacker starts checking if the universal/local bit of the interface identifier is set to zero. If the node is not running MobileIP with CGA and there is not a DHCP server available in the victim's subnet, the attacker assumes that the victim is running RFC3041 or has configured the address manualy. Finally, the attacker can observe the addresses associated with a certain hardware address and determine if the victim is running RFC3041.

5 Recommendations

A better privacy protection can be achieved if the random interface identifier can not be distinguished from a *common* one. i.e. an eavesdropper can not determine if certain node is using or not the stateless address configuration privacy extension.

Unobservability can be garanteed as follows:

- All the hosts generate their interface identifier randomly by default. (suggested change in RFC2373).

- The universal/local bit is not reserved and hosts always rely in duplicate address detection (DaD).

- Alternatively, the host generates an interface identifier based on the addresses present in the link. The main idea is that the mobile node should keep statistical records of the presence of the different OUIs in the media and generate a random identifier based on that information. The host learns about the nodes in the media by sending a neigbor discovery message to the all hosts multicast address.

References

[1] **C. Huitema,** *IPv6, the new Internet Protocol.* 2nd Edition. Prentice Hall. 1997

[2] **S. Kent and R. Atkinson,** *"Security Architecture for the Internet Protocol".* RFC 2041.

[3] **R. Hinden and S. Deering,** *"IPv6 Addressing Architecture"*, RFC 2373, July 1998.

[4] **R. Hinden, M. O'Dell, S. Deering,** *"An IPv6 Aggregatable Global Unicast Address Format"*, RFC 2374, July 1998.

[5] **D. Haskin and H. Allen** *"IP Version 6 over PPP"*, RFC 2472, December 1998

[6] **T. Narten and R. Draves,** *"Privacy Extensions for Stateless Address Autoconfiguration in IPv6".* RFC3041, January 2001.

[7] **G. Montenegro and C. Castelluccia,** *"SUCV Identifiers and Addresses".* draft-montenegro-sucv-01.txt. July 2001.

[8] **S. Thomson and T. Narten,** *"IPv6 Address Autoconfiguration"*, RFC 2462, December 1998.

[9] **A. Escudero.** *"Location privacy in IP_{v6} internetworking - Pseudorandom interface identifiers".* IDMS2001. Lancaster. UK. August 2001.

[10] **D. Johnson and C. Perkins,** *"Mobility Support in IPv6"*, draft-ietf-mobileip-ipv6 v14, July 2000.

[11] **H. Soliman, C. Castelluccia, K. El-Malki and L. Bellier,** *"Hierarchical MIPv6 mobility management"*, draft-ietf-mobileip-hmipv6 v3. February 2001.

[12] **C. Perkins,** *"IP Mobility Support"*, RFC 2002, October 1996.

[13] **A. Escudero,** *"Anonymous and Untraceable Communications: Location Privacy in Mobile Internetworking".* Licentiate Thesis. July 2001.

[14] **Pekka Nikander,** *"An Address Ownership Problem in IPv6"*, draft-nikander-ipng-address-ownership-00.txt, February 2001.

[15] **ISO99** IS 15408, 1999, `http://www.csrc.nist.gov/cc`

Simultaneous Routing and Resource Allocation in Wireless Networks

Lin Xiao[1], Mikael Johansson[2] and Stephen Boyd[3]

[1]Packard 243, Stanford, CA 94305 E-mail: lxiao@stanford.edu

[2]Cory Hall 262M, U.C. Berkeley, CA 94703. E-mail: mikaelj@eecs.berkeley.edu.

[3]Packard 264, Stanford, CA 94305 E-mail: boyd@stanford.edu

Abstract

We consider a data network in which packets are transmitted over wireless communication links. We use a capacitated multicommodity network flow model to describe data flows within the network. The link capacities depend on the communication resource variables, such as transmit powers and bandwidths, allocated to the links. Adjusting the resource allocation changes the link capacities, influences the optimal routing of data flows and alters the network utility. We formulate the problem of simultaneous optimal routing and resource allocation for wireless data networks as a convex optimization problem, discuss efficient solution methods, and illustrate the approach on an example with several hundred links.

1 Introduction

Many relevant routing problems for data networks can be formulated as convex multicommodity network flow problems, for which efficient solution methods exist [1, 2]. The optimal routing depends on the link capacities which are usually assumed to be fixed. The emergence of wireless data networks gives new perspectives on optimal routing problems problems, and challenges the classical assumptions. Wireless networks are highly dynamic in channel characteristics and configuration. The capacities of wireless channels are determined by the channel characteristics and communication resources (transmit powers, bandwidths, time-slot fractions, etc.) allocated to the channels. Adjusting the resource allocation within the network changes the capacities of individual links, influences the optimal routing of data flows, and alters the total utility of the network.

In this paper, we consider the problem of simultaneously optimizing the routing and resource allocation in a wireless data network. We use a capacitated multicommodity flow model to describe the data flows in the network. Ignoring the detailed transmission protocols and mechanisms, the network flow variables should be interpreted as average flows in bits per second. We assume that the capacity of each wireless link is a concave increasing function of the communication resources allocated to the link, and that the communication resources for groups of links are limited. These assumptions hold for many common capacity formulas and networks. We formulate the simultaneous routing and resource allocation (SRRA) problem as a convex optimization problem over the network flow variables and communication variables, and illustrate the approach on a network with several hundred links.

2 Network flow model

2.1 Network topology

We model the topology of a data communication network by a directed graph. A collection of nodes $\mathcal{N} = \{1, \ldots, n\}$ represent physical radio stations which can send, receive, and relay data across the network. A communication link is an ordered pair (i, j) of distinct nodes. The presence of a link (i, j) means that the network supports data flow from the start node i to the end node j. We label all the links with integers, and denote the set of all links by $\mathcal{L} = \{1, \ldots, m\}$. We define $\mathcal{O}(i)$ as the set of links that are outgoing from node i, and $\mathcal{I}(i)$ as the set of links that are incoming to node i.

We can represent the network topology by its node-link incidence matrix. Assume that the network has n nodes and m links, then its incidence matrix is a matrix $A \in \mathbb{R}^{n \times m}$ whose entry a_{ik} is associated with node i and link k via

$$a_{ik} = \begin{cases} 1, & \text{if } k \in \mathcal{O}(i) \\ -1, & \text{if } k \in \mathcal{I}(i) \\ 0, & \text{otherwise.} \end{cases}$$

Hence, each column of A (representing a link) has exactly two nonzero entries, one equal to 1 and the other equal to -1, indicating the start and end nodes of the corresponding link. Each row of of A describes

link connections at a node: the $+1$ entries indicate outgoing links and the -1 entries indicate incoming links.

2.2 A capacitated multicommodity network flow formulation

We will use a multicommodity flow model to describe data flows across the network. In this model, each node can send data to many destinations and receive data from many sources. However, multicast is not considered in our model. We assume that data flows are lossless across links, and that they satisfy flow conservation laws at each node.

We identify the flows by their destinations. We let $\mathcal{D} \subseteq \mathcal{N}$ denote the set of all destination nodes. For each $d \in \mathcal{D}$, we define a source-sink vector $s^{(d)} \in \mathbb{R}^n$, whose ith entry ($i \neq d$) denotes the (nonnegative) amount of flow originating from node i (the source) and destined for node d (the sink). In light of the flow conservation law, we have

$$s_d^{(d)} = -\sum_{i \in \mathcal{N},\ i \neq d} s_i^{(d)}. \tag{1}$$

At each intermediate node, flows with the same destination node are aggregated and treated as a single commodity. The aggregated flow is transmitted according to the routing table, which possibly splits the flow between the outgoing links. On each link k, we let $x_k^{(d)}$ denote the amount of flow destined for node d. We call $x^{(d)} \in \mathbb{R}_+^m$ the flow vector with destination d. At each node i, the flow vector and the source-sink vector of the same destination d satisfy the following flow conservation law:

$$\sum_{k \in \mathcal{O}(i)} x_k^{(d)} - \sum_{k \in \mathcal{I}(i)} x_k^{(d)} = s_i^{(d)}.$$

The flow conservation law across the whole network can be expressed as

$$Ax^{(d)} = s^{(d)}, \quad \forall d \in \mathcal{D}.$$

The above expression also implies (1) since $\mathbf{1}^T A = 0$.

Finally, we will impose capacity constraints on the individual links. Let $t_k = \sum_{d \in \mathcal{D}} x_k^{(d)}$ be the total amount of traffic on link k and we impose the capacity constraints $t_k \leq c_k$.

In summary, our network flow model imposes the following group of constraints on the network flow variables $x^{(d)}$, $s^{(d)}$ and t:

$$\begin{array}{rcll} Ax^{(d)} & = & s^{(d)}, & \forall d \in \mathcal{D} \\ x^{(d)} & \succeq & 0, & \forall d \in \mathcal{D} \\ t_k & = & \sum_{d \in \mathcal{D}} x_k^{(d)}, & \forall k \in \mathcal{L} \\ t_k & \leq & c_k, & \forall k \in \mathcal{L} \end{array} \tag{2}$$

In capacitated multicommodity network flow problems, the capacities c_k are fixed and one is to optimize the performance of the network subject to the above constraints; see [3]. In a communication network, however, the link capacities c_k are typically functions of the communication resources allocated to the links. The nature of these capacity constraints will be described next.

3 Communications model

In this section we describe our model of the communication system. The capacities of the individual links depend on the media access scheme and the selection of certain critical parameters, such as transmission powers and bandwidths or time-slot fractions allocated to individual channels (or groups of channels). We refer to these critical communication parameters collectively as *communication variables*, and denote the vector of communication variables by θ. The communication variables are themselves limited by various resource constraints, such as limits on the total power at each node or total bandwidth available across the whole network. We will assume that the medium access methods and coding and modulation schemes are fixed, but that we can optimize over the underlying communication variables θ.

Let θ_k be a vector of communication variables associated with link k. In general, the capacity c_k depends not only on θ_k, but also on communication resources allocated to other communication links of the network (due to interferences). However, in this paper we will only consider the case $c_k = \phi_k(\theta_k)$.

We will use the following general model to relate the vector of total traffic t, and the vector of communication variables θ:

$$\begin{array}{ll} t_k \leq \phi_k(\theta_k), & \forall k \in \mathcal{L} \\ C\theta \preceq b & \\ \theta \succeq 0 & \end{array} \tag{3}$$

We make the following assumptions about the model.

- The functions ϕ_k are concave, monotone increasing functions of θ_k. The concavity of ϕ_k implies that the first set of constraints are jointly convex in t and θ. The monotonicity condition on ϕ_k means that, roughly speaking, the capacities of communication links increase with increasing resources.
- The second set of constraints describe resource limitations, such as a total available power for the links starting from the same node, or the total bandwidth for group of links or the whole network.
- The third constraint specifies that the communication variables (which represent powers, bandwidths, time-slot fractions) are nonnegative.

This generic model and the network flow model in section 2 will allow us to formulate the SRRA problem as a convex optimization problem in section 4. As shown in [4], capacity formulas of many important channel models satisfy the convexity and monotonicity assumptions of the generic model. We will only illustrate how the Gaussian broadcast channel with FDMA fits into this framework.

Example: Capacity of Gaussian broadcast channel with FDMA

In the Gaussian broadcast channel with frequency-division multiple access (FDMA), the transmitters at node i send information to receivers at the ending nodes of its outgoing links. The outgoing links $k \in \mathcal{O}(i)$ are assigned disjoint frequency bands with bandwidths $W_k \geq 0$ and powers $P_k \geq 0$. The communication variables $\theta_k = (P_k, W_k)$ are constrained by a total power limit and a total bandwidth limit at node i:

$$\sum_{k\in\mathcal{O}(i)} P_k \leq P_{\text{tot}}^{(i)} \qquad \text{and} \qquad \sum_{k\in\mathcal{O}(i)} W_k \leq W_{\text{tot}}^{(i)}.$$

These resource constraints conform to the generic constraint model $C\theta \preceq b$ in (3).

The receivers are subject to independent additive white Gaussian noises with power spectral densities N_k. The total traffic t_k on each outgoing link is constrained by the Shannon capacity:

$$t_k \leq W_k \log_2\left(1 + \frac{P_k}{N_k W_k}\right), \qquad \forall k \in \mathcal{O}(i). \quad (4)$$

In practice, only a fraction of this capacity can be achieved (it depends on the channel coding method). As shown in [4], the function

$$\phi_k(\theta_k) \triangleq W_k \log_2(1 + P_k/(N_k W_k)) \quad (5)$$

is concave and monotone increasing in θ_k. Hence the constraints relating t_k and θ_k have the generic form (3).

4 Problem formulation

Based on the network flow model (2) and communication model (3), the simultaneous routing and resource allocation problem can be formulated as the following convex optimization problem:

$$\begin{array}{ll} \text{minimize} & f_{\text{net}}(x,s,t) + f_{\text{com}}(\theta) \\ \text{subject to} & Ax^{(d)} = s^{(d)}, \quad \forall d \in \mathcal{D} \\ & x^{(d)} \succeq 0, \quad \forall d \in \mathcal{D} \\ & t_k = \sum_{d\in\mathcal{D}} x_k^{(d)}, \quad \forall k \in \mathcal{L} \\ & t_k \leq \phi_k(\theta_k), \quad \forall k \in \mathcal{L} \\ & C\theta \preceq b \\ & \theta \succeq 0 \end{array} \quad (6)$$

Here the optimization variables are the network flow variables x, s, t and the communication variables θ. We assume that the additive performance functions f_{net} and f_{com} are convex. In many cases they are separable in local variables, *i.e.*, they can be expressed as a sum of convex functions like $\sum_{d\in\mathcal{D}} f_d(s^{(d)})$ or $w^T\theta$ (w is a weighting vector). When the objective function is not strictly convex, the solution to SRRA problem may not be unique. Before discussing specific solution methods, we describe two variants of SRRA in more detail.

Example: Minimum power SRRA

Given a number of source and destination rates (*i.e.*, the source-sink vectors $s^{(d)}$ are given), it is natural to try to find the joint routing and resource allocation that minimizes the total power used by the network. This problem is readily put in the form (6) by letting

$$f_{\text{net}}(x,s,t) = 0, \qquad f_{\text{com}}(\theta) = w^T\theta$$

where $w_i = 1$ if θ_i is a power variable, and 0 otherwise. Many variations, such as minimizing the maximum transmit power used by any node in the network, or minimizing a weighted sum (hereby accounting for the relative "costs" of draining the different power sources) can be handled similarly.

In other cases, it can be of interest to minimize the total bandwidth needed to support the desired traffic. This problem can be described by letting $w_i = 1$ if θ_i is a bandwidth variable, and 0 otherwise.

Example: Maximum utility SRRA

Given the limits of communication resources at each node and across the network, it is natural to maximize some utility function (such as the generated revenue) of the data flows. As an example, consider the additive utility function

$$U(s) = \sum_{d\in\mathcal{D}} U_d(s^{(d)})$$

where each U_d is a concave increasing function of flows $s_i^{(d)}$ ($i \neq d$) from node i to destination d. Finding the optimal routing and resource allocation to maximize this utility function can be put in the form (6) by letting

$$f_{\text{net}}(x,s,t) = -\sum_{d\in\mathcal{D}} U_d(s^{(d)}), \qquad f_{\text{com}}(\theta) = 0.$$

4.1 Efficient Solution Methods

The SRRA problem is a convex optimization problem which can be solved efficiently by, for example, general interior point methods (see, *e.g.*, [5]). Moreover, in the above model, the matrices A and C are

sparse and highly structured, and far more efficient algorithms can be developed by exploiting the problem structure. In particular, if the resource allocation, *i.e.*, the vector θ, is fixed, then the problem becomes a convex multicommodity network flow problem for which very efficient large-scale algorithms exist (see, *e.g.*, [3, 6]).

An important property of the SRRA problem (6) is that the network flow variables and the communication variables are only coupled through the link capacity constraints $t_k \leq \phi_k(\theta_k)$. This almost separable structure can be exploited to develop efficient solution methods. We have developed a method based on dual decomposition, in which the SRRA problem is decomposed into an uncapacitated network flow problem and a communication resource allocation problem. The two subproblems are coordinated by a master problem, whose variables can be interpreted as prices for the available link capacities. The objective of the network subproblem is to minimize a cost function which includes the total cost of the link capacities used, while the resource allocation problem for each link amounts to maximizing the utility plus the total revenue collected from the capacities that the link supports. Both subproblems can be solved very efficiently, and the master problem can be solved effectively using a cutting-plane method (see [7] for full details). This is the approach that we have used to solve the examples in this paper. Since the necessary information for solving the master problem can be computed locally at each link, it might be possible to develop distributed solution algorithms where the master problem is solved using a distributed, possibly asynchronous, subgradient method.

5 Example

To illustrate the approach we consider the wireless network shown in Figure 1. This network has 50 nodes and 340 links, and is generated randomly in a way that guarantees that there exists at least on directed path between every pair of nodes.

We use a simple path loss model, in which the transmitted power p_k and received power P_k are related through

$$P_k = K(d_0/d_k)^\gamma p_k = p_k g_k$$

Here K is a unitless constant, γ is the path loss exponent, d_0 is a reference distance and d_k is the distance between the transmitter and receiver. We assume that frequency division multiplexing is used, and that there are no cross channel interferences. We use Shannon's capacity formula for additive white Gaussian noise (AWGN) channel

$$\phi_k = B_k \log\left(1 + \frac{g_k p_k}{N_k B_k}\right)$$

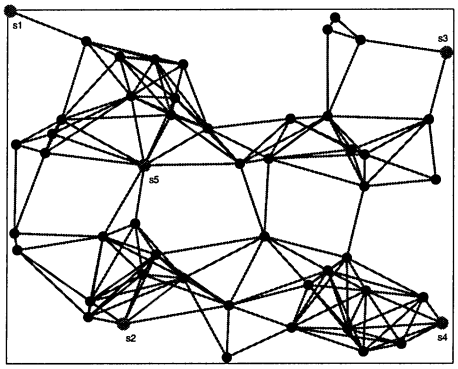


Figure 1: Network topology: 50 nodes and 340 links.

where B_k is the signal bandwidth, and N_k is the noise spectral density. For simplicity, we assume that all channels have equal bandwidth, and generate receiver noises by drawing from a uniform distribution such that $N_k B_k \in [0.01, 0.1]$.

The total transmit power at each node is limited by $p_{\text{tot}} = 100$, and we arbitrarily select five source nodes (marked $S_1, \ldots, S_5$ in Figure 1) that send data to each other. The objective of the SRRA problem is to determine the combined routing and power allocation that maximizes the utility

$$U(s) = \sum_{d=1}^{5} \sum_{i \neq d} \log(s_i^{(d)}) \tag{7}$$

The optimal routing of data flows and allocation of transmit powers are illustrated in Figures 2 and 3, while the optimal source-sink vectors are shown in Table 1. For comparison, we fix the power allocation by dividing p_{tot} equally between the transmitters at each node, and find the routing that optimizes the utility (7). The resulting source-sink vectors are shown in Table 2. We observe that the simultaneous routing and resource allocation achieves a utility of 17.27, a 35% increase over the utility 12.77 attained by the fixed power allocation strategy.

An example with more spectacular performance gains can be found in [7]. In that example, we allocated both powers and bandwidths across a small network, and found that the SRRA-procedure improved the throughput by more than 600%.

i	1	2	3	4	5
1	-3.8803	1.1148	0.9228	1.1148	1.1290
2	1.0275	-16.0537	2.9315	6.9823	6.9748
3	0.8401	2.6900	-9.4320	2.6900	2.7737
4	0.9632	4.8016	2.4620	-18.2341	4.7981
5	1.0495	7.4473	3.1158	7.4471	-15.6755

Table 1: Source-sink flows $s_j^{(i)}$ with SRRA.

i	1	2	3	4	5
1	-2.2591	1.0312	0.8795	1.0063	1.3727
2	0.5608	-13.9519	1.7316	9.5886	5.9157
3	0.5388	2.0710	-6.6064	1.9729	4.1389
4	0.5400	6.7046	1.5472	-16.3378	4.2041
5	0.6195	4.1451	2.4481	3.7700	-15.6314

Table 2: Source-sink flows $s_j^{(i)}$ with fixed capacity (uniform power allocation).

6 Conclusions and extensions

We have considered the problem of simultaneous optimal routing and resource allocation for wireless data networks. The network routing problem and the resource allocation problem interact through the capacity constraints on the total traffic supported on individual communication links. We formulated the SRRA problem as a convex optimization problem, using a multicommodity network flow model in which the link capacities are concave functions of the associated resource variables. We discussed efficient solution methods, and demonstrated the approach on a wireless network with several hundred links.

Acknowledgments

The authors are grateful to Haitham Hindi and Andrea Goldsmith for helpful discussions.

This research was sponsored in part by NSF Grants ECS-9707111 and ECS-9983874, by AFOSR Grants F49620-98-1-0147 and F49620-01-1-0365, by DARPA Contracts F33615-99-C-3014 and MDA972-02-1-0004, and by MARCO Contract 2001-CT-888.

References

[1] D. P. Bertsekas and R. G. Gallager. *Data Networks.* Prentice Hall, 1992.

[2] A. Ouorou, P. Mahey, and J.-Ph. Vial. A survey of algorithms for convex multicommodity flow problems. *Management Science*, 46(1), January 2000.

[3] D. P. Bertsekas. *Network Optimization: Continuous and Discrete Models.* Athena Scientific, 1998.

[4] L. Xiao, M. Johansson, H. Hindi, S. Boyd, and A. Goldsmith. Joint optimization of communication rates and linear systems. *IEEE Trans. Aut. Control*, 2001. Submitted. Also available at `http://www.stanford.edu/~boyd/`.

[5] S. P. Boyd and L. Vandenberghe. *Course reader for EE364: Introduction to Convex Optimization with Engineering Applications.* Stanford University, 1998.

[6] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation.* Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[7] L. Xiao, M. Johansson, and S. Boyd. Simultaneous routing and resource allocation for wireless networks. In *Proceedings of 39th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, October 2001.

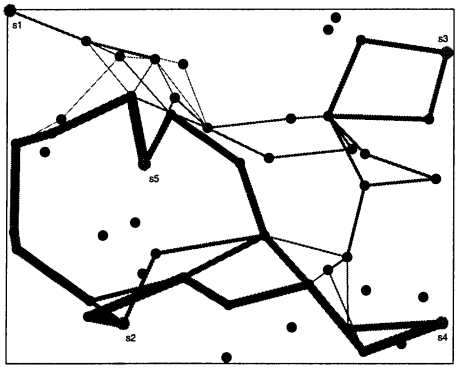


Figure 2: Optimal (aggregate) data flows.

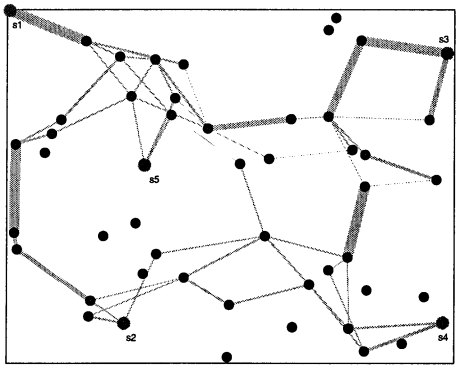


Figure 3: Optimal power allocations.

TCP over UMTS

Anders Dahlén[1] and Per Ernström[2]

[1]Telia Research AB, Vitsandsg. 9, B724, 123 86 Farsta. E-mail: Anders.A.Dahlen@telia.se.
[2]Telia Research AB, Vitsandsg. 9, B719, 123 86 Farsta. E-mail: Per.V.Ernstrom@telia.se.

Abstract

This study investigates the link deteriorating effects of TCP congestion control mechanisms when triggered by UTRAN RLC retransmissions. We analytically quantify the problem due to fast retransmissions for 384, 128 and 64 kbps radio bearers with 1% and 10% block error rate. Simulations are also included for comparison.

1 Introduction

This paper studies the well-known problem of low data throughput over an unreliable radio link due to the TCP. As is well known high performance real time radio links are very expensive. For traffic with less stringent demands on delays and delay variations the cost can be brought down considerably, by lowering the quality of the radio link, i.e., lowering the signal power. This allows for more simultaneous users per base station. The draw back is a higher radio block loss. The lost blocks are retransmitted by a radio retransmission protocol turning losses into delays. The end result is a much lower cost per transmitted bit, but larger delays and delay variations.

For UMTS it has been assumed that cheap radio links with 10% block error rate (BLER) should be used for best effort traffic (web browsing, e-mailing and ftp traffic). These applications are not in themselves delay sensitive, but the TCP protocol is. Delays and delay variations can trigger the TCP congestion control mechanisms and reduce the throughput on the radio link. The use of cheap, high BLER, radio links for end-to-end TCP connections have therefore been questioned. The TCP over UMTS problem has partially been studied in [1,2].

The Transmission Control Protocol (TCP) is a transport protocol providing reliable end-to-end transmission on the Internet. It handles retransmission of lost segments and prevents the Internet from being congested. TCP is designed for wired packet switched networks where segment packets are primarily lost or delayed due to congestions and only rarely due to unreliable transmission links. Hence, when a segment is delayed or lost TCP makes a retransmission of that segment and reduces the transmission rate to alleviate the congestion. TCP has no way of distinguishing whether a delay is due to congestion on the internet or a retransmission over the radio link, and thus TCP will reduce the transmission rate in both cases. Slowing down the transmission in mobile Internet access due to highly unreliable wireless transmission links causes unnecessary slow traffic and inefficient use of radio links.

Here we quantify the problem in terms of throughput reduction for a number of radio links when applying out of sequence delivery of packets to the IP layer. More precisely, a simplified analysis of the reduced throughput due to TCP fast retransmissions is given for 1% and 10% BLER of 384, 128 and 64 kbps wireless links.

2 Internet/TCP

TCP has the task to guarantee end-to-end delivery [3,4]. To achieve this, each TCP segment has a sequence number. When a segment arrives, the receiver replies with an acknowledgment (ACK) indicating which segment is expected next in sequence. Hence, for each segment received out of sequence the receiver sends a duplicate acknowledgment (DUPACK), which indicates the lowest sequence number not yet received.

Each segment is retransmitted if an acknowledgment (ACK) is not received in a certain period of time. This *timeout* value is a function of the observed round trip time (RTT), i.e., the time it takes from sending a packet to receiving the acknowledgement of that segment. If three DUPACKs are received the segment pointed out by the DUPACKs is retransmitted, without waiting for timeout. This is called *fast retransmission.*

TCP allows the sender to transmit as many segments as indicated by the *send window* before any ACK is received. At the beginning of a connection the send window is set to an initial window size, which is usually one SMSS (Sender Maximum Segment Size (TCP/IP header not included)). Then for each received ACK the window is increased by SMSS, which gives an exponential growth of the window. This is called *slow start.* To avoid congestions the send window is increased by a factor (SMSS*SMSS/congestion window) for each received ACK after the window has reached a certain threshold. This is called *congestion avoidance* and gives a linear growth of the window of one SMSS per RTT. If a timeout or fast retransmission occurs the threshold is updated to one-half of the current send window size (but not less than two SMSS). In case of a timeout the send

window is set to one SMSS and the window is increased according to slow start, whereas for fast retransmission the send window is first set to the updated threshold plus 3 times the SMSS, and each time another DUPACK arrives the window is increased by one SMSS. When the ACK of the retransmitted segment finally arrives then the send window is updated to the threshold, and the window increase state is congestion avoidance. This window procedure for fast retransmission is called *fast recovery*.

3 UMTS

The UMTS network handles transmission errors on the wireless links locally in the sending RLC (Radio Link Control). Hence, there are end-to-end retransmissions handled by TCP and radio link retransmissions handled by the RLC. Note that a lost packet over the radio link is not lost but only delayed. These delays might trigger TCP retransmissions. Then TCP assumes that packets are lost due to congestions and decreases transmission rate.

The packets can be delivered in sequence or out of sequence from the receiving RLC to the IP layer. In sequence delivery causes higher delays, which increase the rate of timeouts, whereas out of sequence delivery increases the rate of fast retransmissions performed by the TCP. As in [1], we will assume out of sequence delivery in our analysis, since a timeout in general decreases the throughput much more than a fast retransmission.

The IP packets that arrive at the (sending) RLC have the size of the TCP segments plus TCP and IP headers. Thus, the maximum size of the IP packets is SMSS plus 40 bytes of headers, and this is called the MTU (Maximum Transfer Unit). For Ethernet the MTU is 1500 bytes and if the MTU from end-to-end is not known the default MTU is 576 bytes. An assumption made when estimating the rate of the fast retransmissions is that all TCP segments have full size. Thus, all IP packets are assumed to be of size 1500 or 576 bytes. The IP packets are segmented into radio blocks at the RLC, and the transmission rate of the link and the transmission time interval (TTI), which can be 10, 20, 40 or 80 ms in UMTS, determines the radio block size. Here, TTI is 20 ms and the transfer rate R_{radio} is 384, 128 or 64 kbps.

The RTT across the Internet is assumed to be 150 ms (constant), and to be able to study the effect of the radio link delays on TCP in isolation, no packets are assumed to be lost over the Internet. The transmission delay of the core network (CN) is assumed to be 20 ms, and for the radio access network (RAN) 40 ms, accounting e.g. for channel coding, interleaving and radio frame synchronization. This gives a round trip time of 270 ms, (which is also the assumption in [1]) including Internet, CN and RAN delays. The total RTT also includes the expected transmission time of one MTU over radio (see Appendix), which depends on radio transfer rate and MTU size. The total RTT values are given in the table below.

R_{radio} [kbps]	384	128	64
RTT [ms], 1500 bytes MTU	321,25	383,75	477,5
RTT [ms], 576 bytes MTU	302	326	362

4 Analysis

Below we will evaluate the losses due to fast retransmission and fast recovery. In Section 4.1 we evaluate the losses due to one fast retransmission and fast recovery event, and in Section 4.2 we evaluate how often these events are triggered by retransmissions over radio. Finally, in Section 4.3 we use the results from Section 4.1 and Section 4.2 to evaluate the losses due to fast retransmissions triggered by radio retransmissions.

4.1 A fast retransmit transmission bit loss

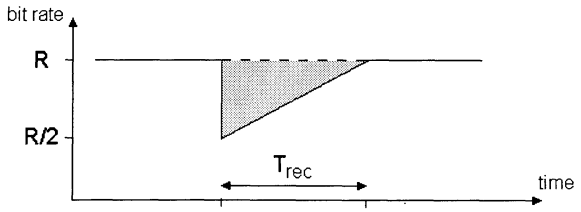


Figure 1. T_{rec} is the time to recover after a fast retransmission and the grey area is the bit loss during fast recovery.

Let R be the bit rate when no fast retransmits occur. In our case this bit rate is limited by the radio link. Assume that the send window has reached its maximum value, which is the buffer size at the receiver. Then, at a fast retransmit event the rate will be reduced to $R/2$, because the halving of the send window, see Figure 1. After the event (we ignore here the inflation phase of the send window) the send window will increase with $SMSS/RTT$, and the bit rate will increase linearly with MTU/RTT^2 [bits/s^2] (a sent packet also includes TCP and IP headers) and reach the normal bit rate R after a time

$$T_{rec} = \frac{R}{2} \cdot \frac{RTT^2}{MTU} [\text{s}].$$

During the recovery period T_{rec} the bit rate is lower than R, and thus fewer bits than normal are transmitted. The loss in transmitted bits (the grey area in Figure 1) is

$$L_{rec} = T_{rec} \cdot \frac{R}{4} = \frac{R^2 \cdot RTT^2}{8 \cdot MTU} [\text{bits}].$$

In addition to this, one IP packet of size MTU is unnecessarily retransmitted[1] by TCP:

$$L_{rtx} = MTU.$$

If fast retransmit occurs with an average rate R_{FR} then $R_{FR} * T_{rec} << 1$ assures that the probability of a new fast retransmit event during the recovery phase is low. Assuming this inequality holds, the effective radio transmissions rate after taking fast retransmits into account is

$$R_{eff} = R - R_{FR}(L_{rec} + L_{rtx}) = \\ = R - R_{FR}\left(\frac{R^2 \cdot RTT^2}{8 \cdot MTU} + MTU\right),$$

and the utilization of the radio bearer rate is reduced by

$$U = 1 - \frac{R_{eff}}{R} = R_{FR}\left(\frac{R \cdot RTT^2}{8 \cdot MTU} + \frac{MTU}{R}\right).$$

If the inequality $R_{FR} * T_{rec} << 1$ is not fulfilled the expression for R_{eff} above still gives an upper bound for the effective bit rate.

4.2 Fast retransmission rate

A radio block error followed by RLC retransmission causes a delay of the corresponding TCP segment. This delay disorders the TCP segments when out of sequence delivery is used. Recall that fast retransmission is triggered by the reception of four identical acknowledgements. Hence, if a TCP segment is delayed by RLC retransmission for such a long time that three new TCP segments are transmitted during the delay, then TCP will receive four identical acknowledgements and a fast retransmission will be triggered.

RLC retransmissions can cause delays and reordering both of initial TCP segments carrying application data and of TCP segments carrying TCP acknowledgements, but retransmissions of ACKs are not covered here.

The RLC retransmission delay for a lost radio block depends on the configuration of the RLC protocol as well as on the processing speed of the radio protocol implementation. An estimate of 80ms (20 ms for status report transmission (known), 20 ms for radio block retransmission (known), and 40 ms for protocol processing (estimate)) is used here, and should not be too far off. Out of these 80 ms only 60 ms can be used for transmission of new data, since one TTI (20 ms) is used for the actual retransmission of the lost radio block. We refer to the part of the delay during which new data can be transmitted as the *RLC retransmission reordering time.*

If the RLC retransmission reordering time is smaller than three times the TCP segment transmission time, no fast retransmission will be triggered. If the RLC retransmission reordering time is larger than four times the TCP segment transmission time a fast retransmission event will be triggered[2]. If the RLC retransmission reordering time is somewhere in between three and four times the TCP segment transmission time, fast retransmission will be triggered with some probability[3] ranging from zero for a reordering time of three times the TCP segment transmission time, to one for a delay of four times the TCP segment transmission time. Even though a fast retransmit will reduce the utilization of the radio link, we assume that the radio link always has packets to send after a lost radio block. This is realistic since the radio transmitter has a non-empty buffer of packets to transmit when the send window is at its maximum level. Moreover, we will assume that fast retransmission always occurs if the RLC retransmission reordering time is larger than three times the TCP segment transmission time. This is a worst-case scenario. Hence, each radio block that is retransmitted n times will trigger a fast retransmission or not, and the fast retransmission rate will be equal to the rate of an n times retransmission event.

Retransmission delays will occur with a rate of (to lowest order in BLER)

$$Rd = BLER / TTI.$$

A multiple reordering time can be caused by repeated loss of the same radio block as it is retransmitted over the radio link. The general formula of event rate for n loss delay events is (to lowest order in BLER)

$$Rdn = BLER^n / TTI.$$

Relevant values for the event rates are given in Table 1. Other types of double loss events that also can cause long delays are:

- Loss of radio block carrying initial TCP segment + loss of RLC status report.
- Loss of radio block carrying initial TCP segment + loss of radio block carrying TCP acknowledgement.

[1] Note that from a radio perspective the "retransmission loss" is worse than the "under utilisation loss", since "under utilisation" does not use up radio capacity, and thus with a good scheduler this capacity can to some extent be used by other users.

[2] Performing the analysis to lowest order in BLER, we neglect the possibility that one of the three segments following the delayed segment is also delayed.

[3] Fast retransmission triggering depends on the detailed segmentation/concatenation of TCP segments into radio blocks, and on the position of the retransmitted radio block in relation to the TCP segment border.

- Loss of radio block carrying TCP acknowledgement + loss of RLC status report.
- Loss of radio block carrying TCP acknowledgement + loss of retransmission of that radio block.

These events are not covered in the analysis here. However, it is clear that they give a higher rate of fast retransmissions than accounted for in this report.

Table 1. The rate of one (Rd), two (Rd2) and three (Rd3) radio retransmissions for some block error rates.

BLER	Rd [Hz]	Rd2 [Hz]	Rd3 [Hz]
1%	0,5	0,005	0,00005
10%	5,0	0,5	0,05

4.3 Evaluation of fast retransmits losses

Based on the two preceding sections we evaluate the fast retransmit losses for different radio bearers.

Studying first an IP MTU of 1500 bytes, we see in Table 2 that three times the TCP segment transmission time is larger than 60 ms for all three radio-bearers considered. A loss of a single radio block is thus not expected to cause a fast retransmission. A double radio block loss (120 ms delay) would however cause a fast retransmit for the 384 kbps radio bearer. The fast retransmission rate R_{FR} will then be 0,5 for 10% BLER, and the radio bearer utilization reduction U in Section 4.1 is given as Loss in Table 3. Of that total utilization loss, the re-tx loss in the table is the utilization reduction caused by unnecessarily retransmitted packets by the TCP. Note that $R_{FR}*T_{rec}= 0,83$ for 10% BLER.

For the 128 kbps and 64 kbps radio links the risk of fast retransmits will obviously be negligible for realistic values of the BLER.

For an IP MTU of 576 bytes we can see from Table 2 that a single radio block loss will definitely cause a fast retransmission over the 384 kbps radio bearer. For this case $R_{FR}*T_{rec}>1$ already at 1% BLER. Hence, TCP will be more or less constantly in fast recovery. The formula for fast retransmit loss is not valid in this case, but it is clear that the loss is very high. Deriving the utilization reduction, U, for 1% BLER anyway gives around 48% utilization loss of the radio link. For a BLER of 5% or 10% the real bit rate will be a small fraction of 384 kbps.

A double loss event will cause fast retransmission also over the 128 kbps radio bearer. In Table 4 the radio bearer utilization reduction is given for this radio link. Here $R_{FR}*T_{rec}= 0,74$ for 10% BLER.

For the 64 kbps bearer not even a triple delay will induce a fast retransmission and the fast retransmission losses are negligible. Obviously, it is important to remember that the RLC retransmission delay is only an estimate. A longer delay will cause more trouble.

Table 2. Radio transmission time of three MTUs.

R_{radio} [kbps]	384	128	64
tx time of 3*1500 bytes [ms]	93,75	281,25	562,5
3 * TCP segment tx time [ms]	36	108	216

Table 3. 384 kbps radio link utilization loss for 1500 bytes MTU.

BLER	Loss	re-tx loss
1%	0,22%	0,02%
10%	22,20%	1,56%

Table 4. 128 kbps radio link utilization loss for 576 bytes MTU.

BLER	Loss	re-tx loss
1%	0,20%	0,02%
10%	20,25%	1,80%

5 Simulations

Below simulation results of the TCP behaviour over UMTS will be presented for comparison with the analysis. The simulations are carried out in OPNET [5] and the assumptions are essentially the same as above. The TCP acknowledgements and the radio link status reports are not exposed to radio retransmissions, and a 64 kbps radio link is allocated for uplink. The delay (one direction) is assumed to be of normal distribution with expected value 0.135 seconds and variance 0.001 seconds. The buffer of the receiver is chosen as multiples of SMSS and as small as possible but with guaranteed full usage of the corresponding error free radio link (about 60% larger than the bandwidth delay product). TCP gives the possibility to delay acknowledgements up to a specified time (200 ms in Windows and 50 ms in Solaris). But here there is no delay.

The implementation of the radio link characteristics for the simulations is, however, somewhat simplified. The radio block segmentation is simplified to an IP fragmentation. Thus an extra overhead is introduced here in form of extra IP headers, and a radio block will only include data from one TCP segment. Hence, every last radio block of a TCP segment will not be a full radio block. The extra overhead has been taken in consideration when setting the size of the radio blocks so that the TCP segments are partitioned into the "correct" number of radio blocks. However, because of non-full radio blocks the radio links will have a lower throughput than the theoretical values (384, 128 and 64 kbps). To compensate for the lower throughput the retransmission reordering time is set to 80 ms instead of 60 ms.

A FTP file of size 16 Mbytes is transferred from a server to a mobile client. This large file transfer is chosen

to study the steady state behaviour of the send window and to obtain a measure of the average behaviour of TCP connections over UMTS.

The utilization loss obtained is displayed in Table 5 and 6. The number of timeouts observed was very few. The simulation result clearly verifies the result of the analysis in respect of when TCP causes a problem or not. The exact quantity of the utilization loss differs somewhat. Note however that the analysis accounts for a worst-case scenario when estimating the problem and in the simulations a simplified radio link model is used, which has lower bit rate than the theoretical values. Thus the simulation results probably under estimate the problem.

The send window for 384 kbps, 10% BLER and 576 bytes MTU is displayed in Figure 2. In this case the send window essentially never reaches its maximum of 14472 bytes.

Delaying the acknowledgements 200 ms as in Windows gave an even larger utilization loss in our simulations.

Table 5. Utilization loss for 1500 bytes MTU.

BLER	384 kbps	128 kbps	64 kbps
1%	0,2%	0,0%	0,0%
10%	11,8%	0,0%	0,0%

Table 6. Utilization loss for 576 bytes MTU.

BLER	384 kbps	128 kbps	64 kbps
1%	23,7%	0,1%	0,0%
10%	56,3%	11,7%	0,0%

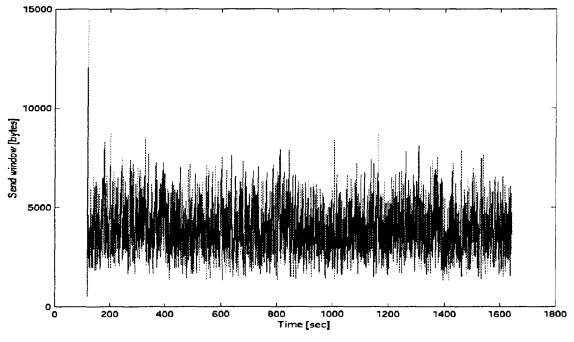


Figure 2. Send window for 384 kbps, 10% BLER and 576 bytes MTU.

6 Conclusions

In this study of TCP over UMTS we have analytically quantified the problem in terms of throughput reduction for a number of radio bearers.

The rate of fast retransmissions for 384, 128 and 64 kbps radio links have been estimated, and the under-utilization of the wireless links in a TCP connection due to these fast retransmits has analytically been derived. A wireless link of 384 kbps and small IP packets (576 bytes) gives a drastic reduction of throughput already at 1% BLER, whereas for 64 kbps radio links the rate of fast retransmissions are negligible for the packet sizes used here. Even 128 kbps wireless link and 1500 bytes IP packet traffic (Ethernet) experience essentially no problem with TCP. The utilization loss, due to fast retransmissions, is very distressing for 384 and 128 kbps radio link and 1500 and 576 bytes IP packets, respectively. Simulations verify the analytical results.

A lot of solutions have been suggested in the literature to mitigate the problem of TCP connections over unreliable wireless links. One of the most promising for UMTS is probably the Eifel algorithm [6]. Finding the best solution for UMTS is now work in progress.

7 References

[1] J. Sachs, M. Meyer. Mobile Internet – Performance Issues Beyond the Radio Interface. 10th Aachen Symposium on Signal Theory, Sep 2001.

[2] J. Peisa and M. Meyer. Analytical Model for TCP File Transfer over UMTS. Proceedings 3G Wireless 2001, San Francisco, May-June 2001.

[3] D. E. Comer. Internetworking with TCP/IP Volume 1: Principles, Protocols and Architectures. Prentice Hall, 2000.

[4] M. Allman, V. Paxson and W. Stevens. TCP Congestion Control. RFC 2581, April 1999.

[5] OPNET 8.0.C PL 15 http://www.opnet.com/

[6] R. Ludwig, R. H. Katz, The Eifel Algorithm: Making TCP Robust Against Spurious Retransmissions, ACM Computer Communication Review, Vol. 30, No. 1, January 2000.

8 Appendix

Define $[x]$ as the integer part of x and $\{x\}$ as the decimal remainder of x. One radio block size (RBS) is $RBS = R_{radio} * TTI$. The number of radio blocks per TCP segment is then $r = MTU / RBS$. If the relative position of radio blocks and TCP segments is random a TCP segment will be segmented into $[r]+1$ radio blocks with probability $1-\{r\}$ and into $r+2$ radio blocks with probability $\{r\}$. Note that the entire radio block has to be received before its content can be read and transferred further. Thus, the expected time to transfer a TCP segment (plus headers) is

$$T = TTI * (([r]+1) * (1-\{r\}) + ([r]+2) * \{r\}) = TTI * (r+1).$$

TCP with Forward Error Correction

Henrik Lundqvist[1] and Gunnar Karlsson[2]

[1]E-mail: henrik.lundqvist@it.kth.se.
[2]E-mail: gunnar.karlsson@it.kth.se.

Abstract

In this paper we investigate the effects of combining Forward Error Correction (FEC) with TCP. FEC reduces the number of retransmissions and results in different congestion control behaviour. Our initial analytical calculations suggest that FEC can yield a higher net throughput. We discuss implementation issues and conclude that FEC has to be adaptive and use information about the window size of TCP to work well. Through simulations we verify the analytical results and show that in scenarios with large bandwidth-delay products the throughput can be improved by adding FEC to TCP. Simulation results are presented for several TCP variants and the effect of FEC is studied in scenarios where either congestion-induced losses or noise-induced losses dominate. The benefits of FEC can be significant, especially if a major part of the losses occur due to noise.

1 Introduction

TCP is the most used reliable transport protocol on the Internet and much effort has been spent on optimising its performance. Recent research on TCP over wireless channels has suggested forward error correction (FEC) as a possible remedy for the performance problems. The idea has been to add FEC at the wireless link to reduce the losses and it has been shown that this can significantly improve the performance. But FEC can also have positive effects on a retransmission protocol since the number of retransmitted packets can be reduced, thereby reducing delay. In this paper we investigate the possibility to use FEC on an end-to-end basis in combination with TCP.

1.1 Forward Error Correction for Packet Loss Channels

Usually FEC is used for correction of bit errors, but the application we are considering is to reconstruct lost packets. The idea is to apply coding over a block of packets before transmission to generate one or more packets of redundant information per block, as illustrated in figure 1. Since the packets have sequence numbers the locations of the lost packets are known and only erasures have to be corrected. Therefore, the number of recoverable losses per block is the same as the number of added redundancy packets. With FEC less packets will be lost and have to be resent, thus reducing the number of timeouts and the delay.

Several coding schemes have been suggested for packet loss channels. For example, simplified versions of Reed Solomon coding have been suggested by McAuley and by Rizzo and shown to achieve high coding/decoding speeds with low complexity [1] [2]. In this paper we will only consider the simplest kind of coding which is simple parity bits resulting in a single added redundancy packet per block.

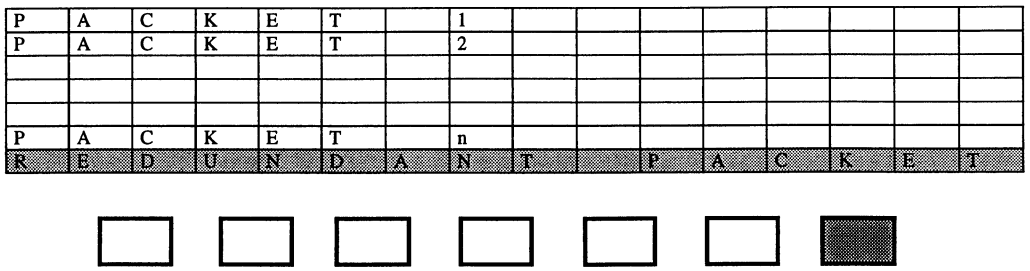


Figure 1. *Illustration of packet level coding. The coding is made vertically in the figure, that is the first bit of every packet is used to get the first bit of the redundancy packet. The packets are sent without delay and the redundancy packet is encoded using copies of the data packets.*

To achieve good performance with FEC it is necessary to adapt the parameters of the coding, i.e. the length of a coding block and the number of redundancy packets per block, to the characteristics of the channel. The most important characteristics of a channel that influence the performance are the loss process and the bandwidth-delay products, and since these parameters can change over time it is usually helpful to adjust the coding parameters dynamically.

Not only the loss probability but also the correlation of the losses is an important parameter of the loss process. If there are correlations in the loss process on timescales longer than the time of a block, this can be used to improve the performance of FEC by adapting to the variations of the loss process. On the other hand, if the losses come in bursts on time scales that are shorter than the block length the performance of FEC will be worse.

Losses occur either when buffers in the network overflow due to congestion or when bit errors caused by noise are detected and the packets are discarded as a consequence. The characteristics of these different loss processes and the ability of FEC to handle them are quite different. Noised-induced errors are usually independent of load and therefore FEC can be a good help to take care of losses when packets have to be discarded due to bit errors. For congestion-induced losses, on the other hand, FEC will increase the load and therefore increase the loss probability. The question is whether FEC is able to recover more losses than it causes by increasing the load.

The correlation of losses is also different between the different types of losses. For noise-induced errors the noise-process has to be correlated over time intervals longer than the time between two packets to make the packet loss process correlated. Congestion-related losses can also be correlated since the probability of arriving at a full buffer is usually higher if the previous packet of the same session arrived at a full buffer.

There are different reports about measurements of the characteristics of the loss processes in networks, and the conclusions are also different in different papers. For example, Paxson concluded that the losses experienced by TCP-flows in the internet are correlated [3]. Also Borella concluded that the losses are bursty and the length of loss bursts can be modelled by a Pareto distribution [4]. On the other hand, there were also many measurements where the losses were quite uncorrelated. In earlier measurements by Bolot the losses appeared almost independently [5].

1.2 TCP with FEC

Lately the focus of TCP-related research has been on possible ways to mitigate the poor performance of TCP over noisy links, which has become increasingly important with the more widespread use of wireless data networks. Bakin et al propose the use of packet-level FEC as one way to improve the throughput of TCP over wireless links [6]. The idea is to introduce special functionality for the noisy link. Another way of using FEC on a single link is to use it on the bit level, which is proposed by Allman et al in [7]. An important difference between applying FEC at bit level as opposed to packet level is that coding at bit level requires longer codes to cover the same time interval. A longer code is more powerful but also more complex to encode and decode.

This investigation, considers the effects of using FEC on an end-to-end basis and is not limited to wireless networks. It is not clear whether FEC will increase or decrease the throughput of TCP, but it can be expected that it is not possible to improve performance when the delay is small, since retransmission protocols are very efficient in that case. For longer delays, however, FEC may improve the performance of TCP since the time to recover the window size after a loss will be longer. In that case it could be better to handle this functionality end-to-end instead of adding it at special links within the network.

2 Analytical evaluation

We use an analytical model of the throughput of TCP to investigate whether FEC can improve the performance. The model deduced by Padhye et al showed that expression (1) is a good approximation of TCP throughput [8]:

$$T = \min\left(\frac{W_m}{RTT}, \frac{1}{RTT\sqrt{\frac{2bp}{3}} + T_0 \min\left(1, 3\sqrt{\frac{3bp}{8}}\right) p(1+32p^2)}\right) \quad (1)$$

where W_m is the maximum window size, b is the number of packets acknowledged by each acknowledgement, p is the loss probability, T_0 is the timeout period and RTT is the round trip time. This expression considers that window reductions occur both due to duplicate acknowledgements and timeouts. In this model it is assumed that losses occur independently.

To evaluate the reduction in loss probability that can be achieved by FEC we again assume that losses occur independently, then the resulting packet loss probability after loss recovery is given by the following expression:

$$p = \sum_{i=M+1}^{N+M} \binom{N+M}{i} (1-P_{pl})^{N+M-i} P_{pl}^{i} \frac{i}{N+M} \quad (2)$$

where M is the number of redundancy packets in a block, N is the number of data packets per block and P_{pl} is the loss probability before FEC is applied. We assume that some of the capacity is used to send redundant information thus reducing the throughput but keeping the loss probability P_{pl} constant. It should be pointed out that this does not exactly correspond to the scenario that is evaluated in the simulations but it provides some

information about whether FEC can improve the throughput of TCP.

In figure 2 the throughput according to this analytical model is plotted. We have chosen parameters corresponding to a channel with a bandwidth of 1 Mbit/s with one-way delays of 10 ms and 100 ms respectively. The graph reveals that the throughput can be substantially improved when the delay is large, whereas there is only a very small range of loss probabilities where the FEC will result in an improvement if the delay is 10 ms. For the longer delay the throughput is improved for loss probabilities ranging from 0.002 to 0.4. These results are not surprising since the lost throughput due to retransmissions is naturally higher for channels with large bandwidth-delay products. In fact, selective acknowledgement always reaches the channel capacity of the packet loss channel as long as the throughput is not limited by the window size.

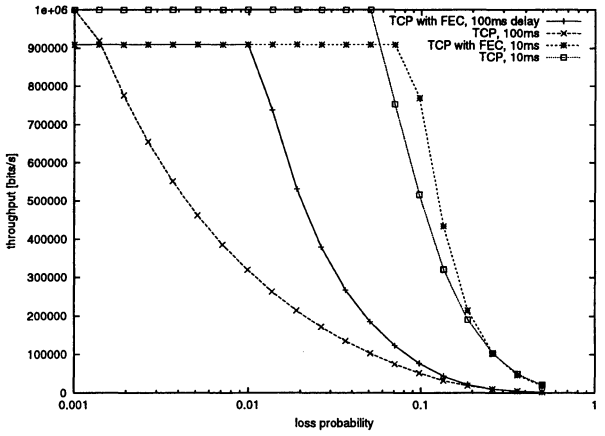


Figure 2. *The throughput of TCP as a function of loss probability. The loss probability on the X-axis does not include losses caused by congestion and the round-trip times are approximately twice the indicated delay.*

3 Implementation

We have not implemented TCP with FEC in a real system but we have made simulations with NS-2. In the simulation the FEC has been added to TCP in a way that actually increases the traffic load. The decoding is made before the TCP-segment is delivered to the acknowledgement function in the TCP-receiver. Thus, the congestion control will act only on the losses that are not correctable by FEC and the traffic load in the network will be higher since this is a more aggressive behaviour.

For the combination of FEC and TCP to work well some attention has to be paid to the length of the error correcting code. Since loss recovery cannot be performed until the last packet of the block has arrived a delay of one block will occur whenever it is necessary to recover a lost packet. Note that no extra delay is added as long as no packets are lost since packets are delivered immediately and a copy is saved by the encoding and decoding functions in order to perform encoding and decoding respectively. However, if a packet within a block is lost the decoding function at the receiver will delay the rest of the packets as long as it sees a possibility to recover the lost packet. It is important that the length of a block does not exceed the window size of TCP, since in that case the receiver would not acknowledge the packets that were delayed by the decoding function and the sender would not be able to deliver more packets. Thus, a timeout would occur and lead to a retransmission, even though it still might be possible to recover the lost packet.

Therefore, we adapt the block size to be half the window size but with an upper and a lower limit. As an upper limit we use 20 packets and as a lower limit we use 4 packets. When the size of the window is decreased below a lower limit the FEC is disabled. Thereby the amount of redundant information is kept between 5% and 25%.

4 Simulation results

The performance has been evaluated through simulations with NS-2 [9]. The throughput has been found in scenarios with different loss probabilities, delay and different TCP variants. The topology consists of two links, one error free and one where losses occur according to a specific probability distribution.

In figure 3 the simulation results from a scenario with a one-way delay of 100 ms and different loss probabilities is compared to the analytical results from the analytical model in the previous section. There are ten TCP-Tahoe sessions competing for a 10 Mb/s link where most of the losses occur independently with the probability indicated on the X-axis [10]. However, there are also some losses due to buffer overflows, these are not included in the probabilities on the X-axis, but the loss rate due to congestion is less than 0.01 in all cases. The packet size is chosen to 1000 bytes, but it is actually the bandwidth-delay product in terms of packets that determines the behaviour, thus the results can be directly translated to other packet-sizes and link-speeds. As can be seen the results are similar to the results of the model and the performance is improved significantly when FEC is included. For example, the throughput is increased by over 100% for loss probabilities of 0.01 and 0.02 in the simulations.

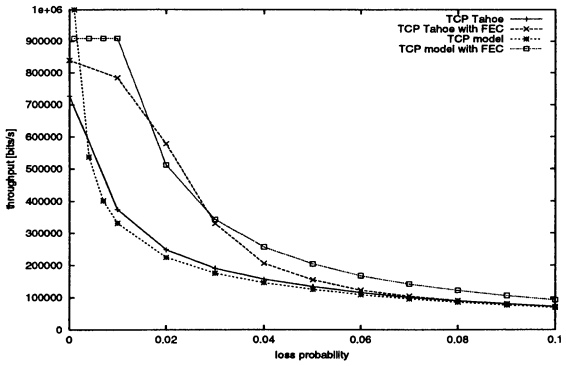


Figure 3. *Both analytical and simulation results for the throughput as a function of loss probability are shown. TCP with FEC performs better for most loss probabilities.*

We have also investigated the performance of different TCP-variants in combination with FEC. The implementations have been taken from NS-2 with the only difference that FEC has been added. In figure 4 the throughput for TCP Tahoe is compared to TCP SACK and TCP Vegas [11] [12]. The scenario is the same as in the first simulation, that is the delay is 100 ms. It can be seen that it is actually TCP Tahoe that performs best in combination with FEC, whereas for TCP Vegas the throughput only gets worse when FEC is added. This is a little surprising since Tahoe is the simplest TCP-variant which performs worst without FEC. It could be expected that TCP SACK would perform better than Tahoe also with FEC since the selective retransmissions should make the protocol more efficient, however, the simulation shows that the throughput is slightly lower for SACK than for Tahoe. TCP Vegas is quite different from the other variants, and the figure reveals that the throughput is reduced at all loss probabilities when FEC is added.

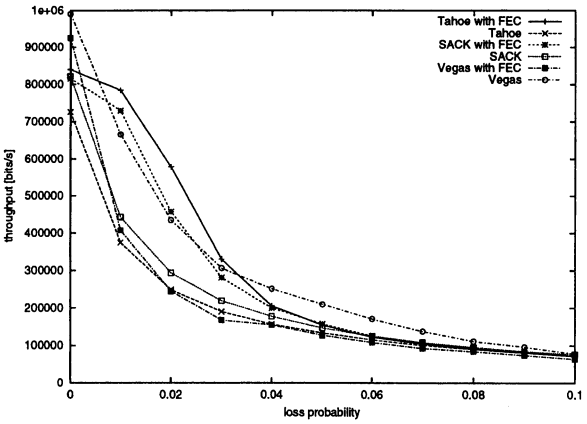


Figure 4. *Throughput for different TCP-variants. The throughput is improved by FEC for most TCP-variants, but Vegas works better without FEC, at least with this implementation of FEC.*

The effect of the delay has also been evaluated by simulations of links with different delay. The scenario is similar to the one above except that in this case there are no noise-induced losses. The one-way delay varies between 1 ms and 1000 ms and TCP Tahoe and TCP Vegas are simulated. Figure 5 shows that for Tahoe it can be beneficial to use FEC on paths with high delay even though all losses are due to congestion. The buffer-size in these simulations was 30 packets, in general the benefit of FEC is more significant for small buffers than for large buffers. For Vegas with FEC the graph shows that the performance is better without FEC. However, in this scenario where all losses are caused by buffer overflow we can see that Vegas works much better than Tahoe, also when FEC is used.

Note that the maximum window size we have used in the last simulations is 200 packets, since NS-2 uses packets rather than bytes as the unit for TCP window size. This large window size makes it possible to investigate the behaviour of the congestion control for large bandwidth-delay products, but at the largest delays in the simulations, the 64 kilobytes window size limit of TCP would actually limit the throughput to lower levels than our results indicate.

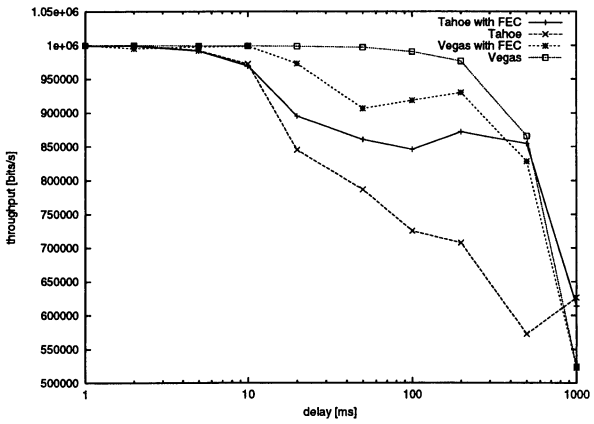


Figure 5. *Throughput as a function of delay. It can be seen that the performance is improved by FEC for long delays, even though all losses are caused by congestion.*

5 Conclusions

The purpose of this paper was to evaluate the performance of TCP with forward error correction in a few different scenarios. Both analytical methods and simulations were used and the results show that there are scenarios where the throughput can be substantially increased if FEC is used. For FEC to be beneficial the delay should be quite high and the losses should be fairly uncorrelated. Since we only use a very simple code with a single redundancy packet per block the performance is only improved at low loss probabilities. Probably it is possible to achieve better performance for a wider range

of parameters if a more advanced code is used. If both the number of redundancy packets and the block length can be varied it might be beneficial to use more feedback about the channel conditions to adapt the coding parameters. This remains to be investigated in future research.

References

[1] A. McAuley, "Reliable Broadband Communication using a Burst Erasure Correcting Code", Proc. SIGCOMM '90, Philadelphia, PA, Sept. 1990, pp. 287-306.

[2] L.Rizzo, "On the Feasibility of Software FEC", DEIT Technical Report LR-970131. URL: "http://www.iet.unipi.it/~luigi/softfec.ps".

[3] V. Paxson, "Measurements and Analysis of End-to-End Internet Dynamics", PhD thesis, University of California, Berkeley, April 1997.

[4] M. S. Borella, "Measurement and Interpretation of Internet Packet Loss", Journal of Communication and Networks, Vol. 2, No. 2, June 2000.

[5] J.-C. Bolot, "End-to-End Packet Delay and Loss Behaviour in the Internet", Proc. SIGCOMM'93, Pages 289 - 298, Sept. 1993.

[6] D.S.Bakin, M Joa-Ng, A.J. McAuley, "Quantifying TCP Performance Improvements in Noisy Environments Using Protocol Boosters", Computer and Communications, 2000. Proceedings. ISCC 2000. Fifth IEEE Symposium on, 2000. pp 92-97.

[7] M. Allman, D. Glover, L.Sanchez, "Enhancing TCP Over Satellite Channels using Standard Mechanisms", BCP 28, RFC 2488, Jan. 1999.

[8] J. Padhye V. Firoiu, D. Towsley, J. Kurose, "Modeling TCP Reno Performance: A Simple Model and Its Empirical Validation", IEEE/ACM Trans. On Networking, Vol. 8, no. 2, April 2000.

[9] The UCB/LBNL/VINT Network Simulator (NS). URL: "http://www-mash.cs.berkeley.edu/ns/".

[10] V. Jacobson, "Congestion Avoidance and Control", in Proc. SIGCOMM '88 Symp., Aug. 1988, pp. 314-329.

[11] M. Mathis, J. Mahdavi, S. Floyd and A. Romanow, "TCP Selective Acknowledgement Options", RFC 2018, Oct. 1996.

[12] L. Brakmo, L. Peterson, "TCP Vegas: End to End Congestion Avoidance on a Global Internet", IEEE JSAC, Vol. 13, No. 8, Oct 1995.

From Single to Dual Array Architectures A Systems Perspective

M.A.Beach[1], D.P. McNamara[2], P.N. Fletcher[3] and P. Karlsson[4]

[1]University of Bristol, Bristol, UK. E-mail: M.A.Beach@Bristol.ac.uk.
[2]University of Bristol, Bristol, UK. E-mail: Darren.McNamara@Bristol.ac.uk..
[3]University of Bristol, Bristol, UK. E-mail: Paul.Fletcher@Bristol.ac.uk..
[4]Telia AB, Malmo, SE. E-mail: Peter.C.Karlsson@telia.se.

Abstract

The application of Smart Antenna technology to wireless communication systems has attracted considerable attention over the last decade or so, as a means of capacity enhancement. Here, the application of space-time processing using a single antenna array at the base-station has been considered in order to obtain interference reduction (both co-channel and multi-path) as well as range extension. More recently, the application of dual antenna array architectures to create multiple-input multiple-output (MIMO) configurations, has attracted considerable interest both within academia and industry as a means of providing significant capacity gains over conventional array based enhancement methods. This paper provides an introduction to these MIMO communication systems and explains how such capacity gains can be achieved using this technology.

1 Introduction

In order to realise the full potential of, 'Any Where, Any Time Personal Communications', which includes high bit-rate real-time multimedia applications, space-time adaptive signal processing is now regarded as a core element of future system deployments [1]. Until recently, it was envisaged by those engaged in wireless system design that space-time processing in the form of adaptive or smart antennas employing a single antenna array [2][3], would provide the long-term means to offer sufficient user capacity within the confines of the limited radio spectrum allocated to these systems. Conventional 'smart' antenna systems yield a capacity enhancement or range extension over conventional fixed beam antenna installations by focussing their beam patterns towards an addressed volume associated with the signalling space of the desired traffic channel. This both reduces the interference [4] caused by, and also seen by, this channel, hence providing a capacity enhancement through Spatial Filtering Interference Reduction (SFIR), or Space Division Multiple Access (SDMA).

Numerous test-beds or validators have been constructed in order to provide further 'proofs-of-concept' as well as to inform the industry and network operators about this potential of this technology. This has been conducted both privately within numerous organisations [5][6], as well as publicly under initiatives such as the European RACE[1] and ACTS[2] programmes [7][8][9].

Despite these benefits offered by single array based solutions, 3rd Generation Wireless networks (e.g. UMTS and beyond) could still fall short of market expectations in terms of available capacity for 'e-commerce' and 'multi-media' based services. As recently reported in the literature [10], the smart antenna concept can be extended by employing multi-element arrays at both ends of the communications link. This MIMO architecture has the ability to exploit the multipath scattering present in the environment in order to realise substantial gains in both link and network capacities. If practically viable, this novel approach could offer a unique solution for supporting bandwidth hungry wireless applications in future generation networks.

2 Fundamentals of MIMO Systems

It is well understood that a multi-element antenna array can be employed to obtain independent fading signals from a multipath rich environment and a processing gain realised through the application of optimum combining [11]. This concept can be extended by employing multi-element arrays at both ends of the communications link [10], thereby exciting independent paths between each of the transmit and receive elements Figure 1.

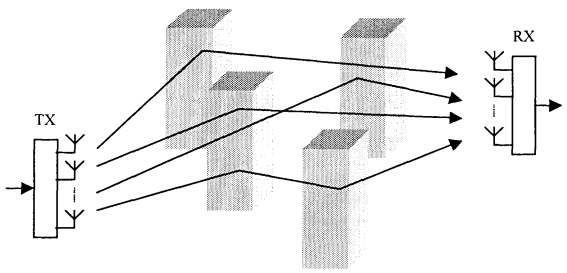


Figure 1: Creation of a MIMO channel

[1] "Research into Advanced Communications for Europe"
[2] "Advanced Communications Technologies Societies"

It has been shown [12] that when these dual arrays are deployed in a suitably rich scattering environment there is considerable potential for obtaining extremely high spectral efficiencies. Assuming a minimum of n antenna elements at each of the transmitter and receiver, then n parallel channels will effectively be created between the transmitter and receiver, thus increasing the spectral efficiency n times. (The number and quality of these parallel channels is dependent on the instantaneous propagation conditions). This implies a linear increase in the channel capacity with increasing numbers of antenna elements, as opposed to the logarithmic increase normally associated with multi-antenna diversity techniques.

For a narrowband system, the complex channel response between a single element at the transmitter and receiver can be represented as a single complex number. The full channel response of a system comprising of n_T transmit elements and n_R receive elements, can therefore be described by an n_R-by-n_T matrix, G, (where element G_{ij} is the response between receive element i and transmit element j).

The information theoretic capacity of this channel, for fixed numbers of antenna elements, is a function of both the average received signal to noise ratio at each receive element, ρ, and the normalised channel response matrix, H, as shown in (1) [12]. (This assumes that the transmitter does not have knowledge of the channel response, and hence distributes its power equally amongst the n_T elements).

$$C = \log_2\left[\det\left(I_{n_R} + \frac{\rho}{n_T} HH^*\right)\right] \quad \text{bits/s/Hz} \qquad (1)$$

The normalised channel matrix, H, is obtained from the absolute (measured) channel response, G, by removing the average path loss component [12], as follows

$$H = \left(\hat{P}/P\right)^{1/2} \cdot G \qquad (2)$$

where $\hat{P}$ is the total transmitted power and P is the average power over all receive elements.

As mentioned above, this matrix channel corresponds to the creation of multiple paths between the transmit and receive arrays. The relative (power) gains of each of these parallel channels are given by the eigenvalues, λ_i, of the channel covariance matrix, HH^* (with the absolute channel gains being given by the eigenvalues of GG^*). It can therefore be seen that in order to create n parallel channels, the channel matrix G must be of at least rank n.

In reality, this desire for the channel matrix to have high rank translates to the requirement for each receive antenna element to experience independent fading. Although an environment may exhibit sufficient multipath scattering to induce independent fading across the arrays, this independence still does not guarantee decorrelation of the channel matrix. There is therefore a finite probability that even in a highly scattering environment a MIMO channel matrix can become ill-conditioned. This therefore causes the lowest of the gains of the spatial channels to significantly drop, as shown by the absolute channel gains in Figure 2 at about 2.4 seconds (if the correlation between elements was worse then more of the gains would also be seen to drop).

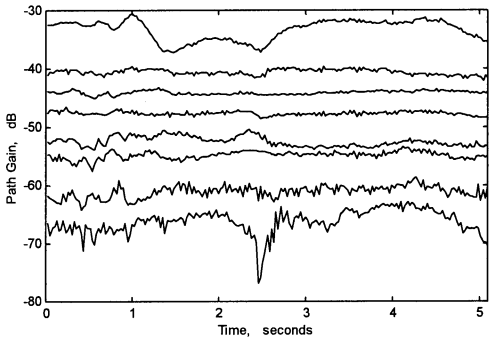


Figure 2: Measured gains of all 'spatial channels' created by an 8x8 MIMO system in an indoor environment [13].

It is the creation of these parallel channels that gives rise to the high capacities of MIMO systems. Since each of these 'spatial channels' is orthogonal to each of the others, they are all capable of supporting independent data streams. The capacity of the overall MIMO channel can therefore also be calculated as the sum of the classical Shannon capacities ($\log_2(1+SNR)$) of each of these spatial channels (modified by their individual channel gains), as in (3), which can be shown to be equal to (1).

$$C = \sum_i \log_2\left(1 + \frac{\rho}{n_T}\lambda_i\right) \qquad (3)$$

In order to quantify some of these claims of huge capacities, Figure 3 shows an example of the complementary cumulative distribution functions (CCDFs) of MIMO channel capacity for both measured and statistically generated channels. The statistically generated channels represent an ideal Rayleigh fading environment and hence the best possible conditions for generating high MIMO capacities. All elements of these statistically generated channel matrices are therefore modelled as independent identically distributed (i.i.d.), complex, zero mean, unit variance random variables [12], such that:

$$H_{ij} = \text{Normal}(0,1/\sqrt{2}) + \sqrt{-1}\,\text{Normal}(0,1/\sqrt{2}) \qquad (4)$$

In comparison, it can be seen that although the capacities of the measured channels (from an indoor environment [14]) are indeed lower than those of an ideal Rayleigh environment, significant spectral efficiencies are still achieved with high probability.

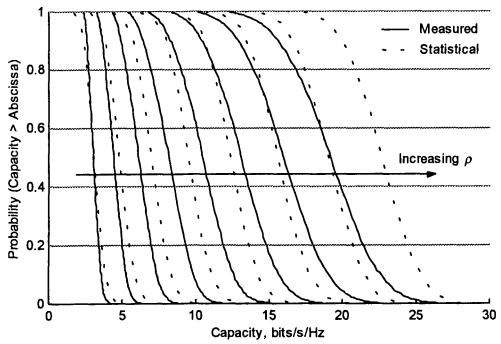


Figure 3: Capacity CCDFs of measured and statistically generated channels for ρ increasing in 3dB steps between 0 and 21dB. (4-element 0.5λ-spaced arrays).

Figure 3 also illustrates how, even for relatively low numbers of transmit and receive elements, very large spectral efficiencies can be obtained. For example, with a signal to noise ratio of 21dB and an outage probability of 10%, the measured channels achieved a capacity of just over 15bits/s/Hz. To achieve the same capacity using a single transmitter would require a modulation scheme with $2^{15} = 32{,}768$ constellation points!

3 MIMO Architeures

a) Bell Labs Layered Space-Time (BLAST) [15]

Bell Labs were the first to propose an architecture and present a reconstruction algorithm to exploit MIMO channels. This was given the acronym BLAST (Bell Labs Layered Space Time). Two variations of BLAST exist; Diagonal (D-BLAST) [10] and Vertical (V-BLAST) [16] (the prefix denoting the direction in which the received signals are processed in space and time).

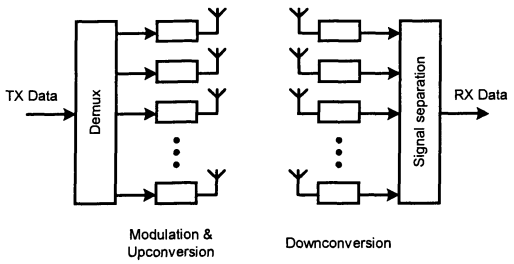


Figure 4: BLAST system architecture

The same general system architecture is common to both BLAST schemes (Figure 4), although the demultiplexing and signal separation operations are different.

For both schemes, the demultiplexer divides the incoming data into n_T individual bit streams (labelled *a* to *d* for the 4 tx-element example in Figure 5). For V-BLAST, each demultiplexed data stream is always assigned (uncoded) to the same transmit antenna element as in Figure 5a. (Although the original V-BLAST scheme itself does not incorporate any coding, it can be concatenated with an outer code for improved performance [17]). However, in D-BLAST each data stream is encoded individually and then cyclically assigned to a different antenna in each symbol period (Figure 5b). The transmission of each coded stream is therefore distributed in both space and time.

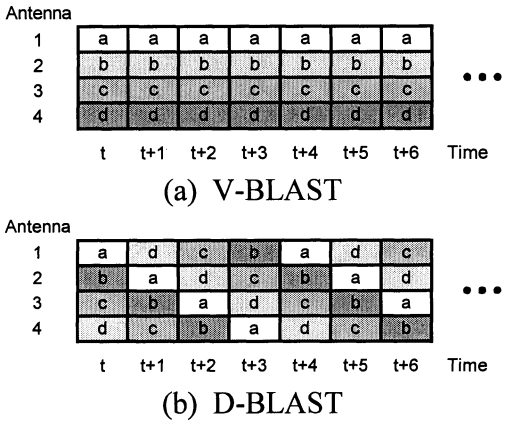


Figure 5: Assignment of data streams to transmit antenna elements

The signal separation and detection of the V-BLAST algorithm works on a symbol by symbol basis, using a combination of nulling and symbol cancellation [16][18]. Conversely, D-BLAST processes the received signals corresponding to each coded stream individually [10] (the diagonal layers in space-time). For example, having detected one diagonal containing stream *a* in Figure 5b this would be cancelled out and processing would continue on diagonals containing streams *d*, *c*, *b* and *a* again, etc.

For both these schemes to function there must be at least as many receive antenna elements as there are transmit elements, although any further increase in the number of receive elements above this can be used to help improve detection performance [10]. Secondly, both detection processes also require accurate knowledge of the MIMO channel response matrix and hence any implementation must also include channel estimation [19][20].

As an example of the performance achievable by these architectures, a demonstrator implementing the V-BLAST scheme has been reported [16] by Bell Labs to achieve spectral efficiencies of between 20 and 40bps/Hz in an indoor environment.

b) Space-Time Coding

In many environments it may not be practical to exploit the very high spectral efficiency properties of the MIMO channel. This could be true in situations where the high variability of the channel characteristics gives rise to a significant variation in the number of parallel spatial paths capable of carrying data. In such a case, architectures trying to extract maximum spectral efficiency from the channel, as described previously, could suffer from catastrophic bit error failure.

Space-time coding techniques provide an alternative method for extracting an advantage from MIMO channels

[21]. Rather than using the parallel channels for increased spectral efficiency, they are employed so as to give a greatly increased diversity order and, in some cases, coding gain. In this way, a reduction in the number of parallel channels will only result in decreased diversity and not necessarily in system failure.

These techniques were initially derived from investigations into transmit diversity [22]. This can be seen from inspection of the delay diversity scheme in Figure 6, where the channel encoder is a rate ½ repetition code. Although this gives a diversity advantage without bandwidth expansion (since two symbols are transmitted simultaneously), it was questioned whether a more efficient channel encoder could be found [23] that would also realise some coding gain.

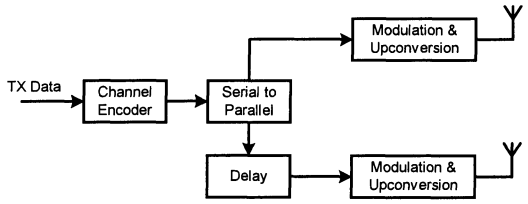


Figure 6: Delay diversity transmission scheme

The removal of the delay block shown in Figure 6, and an extension to allow any number of outputs from the serial to parallel converter gives a generalised architecture. This has led to the development of space-time channel encoders of which, as in conventional coding theory, both block and trellis forms exist.

Space-time trellis coding is a combination of coding and modulation for multiple transmit antennas that achieves both a diversity and coding gain. These codes can be designed according to criteria derived in [23] in order to trade off the diversity and coding gains against each other. The encoder is based on a finite state-machine with the state transitions being determined by the input data bits. Each of these state transitions identifies the symbols that are transmitted from each of the antennas as shown in Figure 7. Maximum likelihood sequence estimation is then employed at the receiver to reconstruct the transmitted data stream.

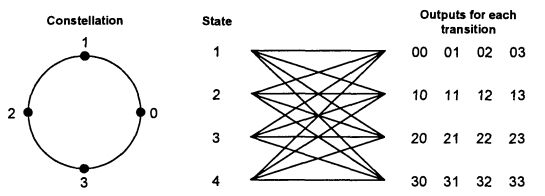


Figure 7: Example of a 4-PSK 4-state space-time trellis code for 2 Transmit Antennas [23]

These space-time trellis codes are able to offer significant coding gains, although this is at the expense of larger numbers of states and hence higher complexity which also increases exponentially with the number of antenna elements. Space-time block codes are therefore attractive since schemes have been found [24][25] which allow the use of only linear processing at the receiver. These codes are able to achieve the full diversity order possible for a given number of antennas and without bandwidth expansion, although they do not offer the coding gains possible with space-time trellis codes. It is this provision of transmit diversity whilst maintaining relatively simple processing that has made the scheme in [24] the first space-time code to be included in a standard for future systems (3GPP proposal for UMTS [26]). This scheme uses a simple block coding algorithm to encode pairs of symbols for transmission by two antennas. Blocks of two consecutive symbols, s_0 and s_1, are taken from the input data stream and used to generate the transmission sequence for each antenna as described in Table 1 (where T is the symbol period and * is complex conjugate).

Time	Antenna 0	Antenna 1
t	s_0	s_1
t + T	$-s_1^*$	s_0^*

Table 1: Space-time block encoder for 2 transmit antenna elements [24]

Although this is only specified for two transmit and one receive element, the scheme can also accommodate additional antennas at the receiver, increasing the overall diversity order to $2n_R$, where n_R is the number of antenna elements at the receiver.

4 Application of MIMO Techniques

Even given the current extent of research into MIMO and space-time techniques, there is still much work to be undertaken before a complete understanding of all aspects of this these systems is approached. Due to the obvious commercial value of this technology in the light of the cost and scarcity of radio spectrum, this research is progressing rapidly, with some products and standards already being proposed (e.g. [27][28] amongst others). The majority of these are currently being developed for the broadband fixed wireless access (FWA) market, for two main reasons. Firstly, there is an undoubted demand for this service [29] as the requirement for delivery of high bandwidth services direct to homes and offices grows. Secondly though, from a technical perspective, the fixed nature of such a system provides a more forgiving propagation environment than high mobility cellular systems. This allows some of the constraints on aspects such as channel estimation and/or equalisation to be relaxed, therefore easing initial practical implementation.

Future wireless LANs are also receiving attention as systems that could benefit from the high data rates MIMO architectures could support [30], especially since indoor environments inherently provide high levels of multipath scattering [14].

Finally though, with further development of MIMO technology and increasing integration, there will be a

point at which even cellular networks will be able to implement these techniques [31].

5 Conclusion

Given the predicted capacity gains of MIMO techniques and the current volume research and development activities within both industry and academia, considerable technological advances within this field are now likely.

Acknowledgements

The authors gratefully acknowledge the financial support from the UK EPSRC, DERA and the CEC IST SATURN project in conjunction with the preparation of this paper.

REFERENCES

[1] Y.J. Guo, S.Vadgama & Y. Tanaka, "Advanced Base Station Technologies for UTRAN", *IEE E.C.E.J.*, Vol.12, No.3, June 2000, p123-132.

[2] S.C. Swales, M.A. Beach, D. Edwards & J.P. McGeehan, "The performance enhancement of multi-beam adaptive base station antennas for cellular land mobile radio systems", *IEEE Trans. Vehicular Tech.*, Vol. 39, 1990, pp.56-67.

[3] A.F. Naguib, A Paulraj & T. Kailath, "Capacity improvement with base-station antenna arrays in cellular CDMA", *IEEE Trans. Vehicular. Tech.*, Vol. 43, pp691-698.

[4] J Litva & T Lo, *Digital Beamforming in Wireless Communications*, Artech House, 1996.

[5] C Ward, D Adams, F Wilson, A Bush, "The Live-Air Trials of a Multibeam Cellular Base Station Antenna System", *Proc. IEE ICAP'99*, University of York, 1999, pp169-172.

[6] S Anderson, B Hagerman, D Henrik, U Forssin, J Karlsson, F Kronestedt, S Mazur, "Adaptive Antennas for GSM and TDMA systems", *IEEE Personal Comm. Magazine*, June 1999, pp74-86.

[7] G Tsoulos, J McGeehan & M.A. Beach, "Space Division Multiple Access (SDMA) Field Trials. Part 1: Tracking and BER Performance", *IEE Proc. Radar, Sonar Navig.*, Vol.145, No. 1, Feb 1998, p73-78.

[8] C. Evci. et al, "AWACS: ATM Wireless Access System – potential candidate for European HIPERLINK standard", *Int. J. of Comp. & Telecom. Net.*, Elsevier, Vol 31, 1999.

[9] C.M. Simmonds, P.B. Darwood, M.A. Beach & P.Howard, "TSUNAMI (II) Macrocellular Field Trial System Performance in the Presence of Deliberate Interference", *IEE ICAP 1999*, p73-76.

[10] G.J. Foschini, "Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multiple Antennas", *Bell Labs. Tech. Journal*, Vol 1, No 2, Autumn 1996, pp. 41-59.

[11] J.D. Parsons, *Mobile Radio Propagation Channel*, 2nd Edition, Wiley, 2000.

[12] G.J. Foschini and M. J. Gans, "On Limits of Wireless Communications in a Fading Environment When Using Multiple Antennas", *Wireless Personal Communications*, Vol. 6, No. 3, March 1998, p. 311.

[13] D.P. McNamara, M.A. Beach, & P.N. Fletcher, "Temporal Variation of Multiple-Input Multiple-Output (MIMO) Channels in Indoor Environments", *IEE Int. Conf. On Antennas & Propagation* (ICAP'01), April 2001.

[14] D.P. McNamara, M.A. Beach, P.N. Fletcher & P. Karlsson, "Initial Investigation of Multiple-Input Multiple-Output (MIMO) Channels in Indoor Environments", *IEEE Benelux Chapter Syp. On Veh. Tech. & Comms*, Leuven, Oct. 2000.

[15] http://www.bell-labs.com/project/blast/

[16] P.W. Wolniansky, G.J. Foschini, G.D. Golden, R.A. Valenzuela, "V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel", *Proc. ISSSE'98*, Pisa, Italy, Sept. 29, 1998.

[17] X. Li, H. Huang, G.J. Foschini & R.A. Valenzuela, "Effects of Iterative Detection and Decoding on the Performance of BLAST", *Proc. IEEE Globecom*, San Francisco, Nov. 2000.

[18] G.J. Foschini, G.D. Golden, R.A. Valenzuela & P.W. Wolniansky, "Simplified Processing for High Spectral Efficiency Wireless Communication Employing Multi-Element Arrays", *IEEE J.S.A.C.*, Vol. 17, No. 11, Nov. 1999.

[19] T.L. Marzetta, "Blast Training: Estimating Channel Characteristics for High Capacity Space-Time Wireless", *Proc. 37th Allerton Conf.*, Monticello, Sept. 22-24th, 1999.

[20] B. Hassibi & B.M. Hochwald, "How Much Training is Needed in Multiple-Antenna Wireless Links?", *Submitted to IEEE Trans. Info. Theory.*

[21] A.F. Naguib, N. Seshadri & A.R. Calderbank, "Increasing Data Rate Over Wireless Channels", *IEEE Signal Processing Magazine*, May 2000.

[22] A. Wittneben, "A New Bandwidth Efficient Transmit Antenna Modulation Diversity Scheme for Linear Digital Modulation", *Proc. IEEE ICC'93*, Geneva, 1993.

[23] V. Tarokh, N. Seshadri, A.R. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction", *IEEE Trans. on Info. Theory*, Vol 44, No 2, March 1998, pp. 744-765.

[24] S.M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications", *IEEE J.S.A.C.*, Vol 16, No 8, October 1998, pp. 1451-1458.

[25] V. Tarokh, H. Jafarkhani, A.R. Calderbank, "Space-Time Block Codes from Orthogonal Designs", *IEEE Trans. on Info. Theory*, Vol 45, No 5, July 1999, pp. 1456-1467.

[26] 3GPP Technical Specification, TS 25.211, V3.1.1, Dec. 1999.

[27] http://www.iospanwireless.com

[28] http://www.bwif.com

[29] R. Dettmer, "broadband@home.com", *IEE Review Magazine*, Vol 46, No 1, January 2000, pp. 13-18.

[30] http://www.ist-saturn.org

[31] http://www.ist-metra.org

Advanced Antenna Systems in WCDMA Networks - Downlink Capacity and Coverage Trade-offs

Bo Hagerman[1] and József Barta[2] and Sven Petersson[3] and David Astély[4]

[1]Ericsson Research, 164 80 Stockholm, Sweden E-mail: Bo.Hagerman@era.ericsson.se

[2]Ericsson Research, H-1037 Budapest, Laborc u. 1. Hungary, E-mail: Jozsef.Barta@eth.ericsson.se

[3]Ericsson Research, 431 84 Mölndal, Sweden, E-mail: Sven.Petersson@emw.ericsson.se

[4]Ericsson Research, 164 80 Stockholm, Sweden E-mail: David.Astely@era.ericsson.se

Abstract

Advanced antenna systems are identified as one of the key techniques to improve the capacity and coverage of third generation WCDMA FDD systems. Advanced antenna systems can be used to solve capacity limitations in certain urban areas and to extend the coverage of large rural cells.

This paper studies the downlink capacity and coverage of a WCDMA system with uniformly distributed speech users. This is done for a so called fixed beam system, which employs a number of beams with fixed pointing directions. Soft handover and the overhead from common channels are included. Since the number of orthogonal channelization codes is limited, the use of multiple scrambling codes is also considered.

1 Introduction

Advanced antenna systems can be used to improve the system performance of third generation WCDMA systems. Compared to systems with three-sector sites, significant gains in capacity and coverage can be obtained, both in the uplink and in the downlink. Advanced antenna systems can therefore be used to extend the coverage of large rural cells, and as the subscriber population grows, the introduction of such systems can be used to meet needs for increased capacity in certain urban areas.

Although the downlink and uplink must be considered jointly when the coverage and capacity of a deployment are studied, the focus of this work is on the downlink, and we note that it in some cases can be the limiting link. This may be due to the expected asymmetry of the data traffic. Another reason for downlink capacity limitation can be that the available base station transmit power is limited. This latter reason may have a large impact on the available downlink capacity in large cells.

Advanced antenna systems provide higher antenna gains and spatial selectivity [1]. For small cells, in which the interference is the limiting factor, the downlink capacity can be increased. For each link, less power is required and less interference is spread. In power limited cases, the antenna gain offered by an advanced antenna system can be used to extend the coverage so that larger cells can be supported. In intermediate cases the capacity and coverage gain can be traded depending on the actual system deployment.

In the WCDMA downlink, so called orthogonal variable spreading factor (OVSF) codes are used as channelization codes to spread different physical channels [2]. The signal transmitted from the base station is then scrambled using a common scrambling code. Since the number of orthogonal channelization codes is limited, the WCDMA standard allows for the use of multiple scrambling codes. The set of channelization codes may thus be reused. The drawback of this is that the orthogonality between different downlink channels is degraded.

An earlier study for conventional WCDMA systems estimated the downlink capacity [3]. Furthermore, in [4],[5] and [6], the capacity of different advanced antenna technologies was investigated. In [7], the capacity and coverage as well as potential trade-offs for a WCDMA downlink with fixed multi beam antennas were studied. In the present work, which is an extension of [7], the code limitation and the use of multiple scrambling codes are considered.

2 Advanced antenna systems

In this paper, antenna arrays with four or eight elements covering a 120° sector are considered. There are two main groups of advanced antenna systems [1], concepts using fixed beams and concepts using so called steered beams. A fixed beam antenna concept uses beams with fixed pointing directions and predefined beam patterns. In uplink all beams are used to collect the incoming energy and to provide diversity gain. In downlink the beam receiving the

Parameter	Value	Unit
Chip rate	3.84	Mchip/s
Sector antenna gain	18	dBi
Speech data rate	12.2	kbps
Eb/No	7	dB
CIR target - speech	-18	dB
Thermal noise power	-100	dBm
Receiver noise figure	7	dB
Users in Soft Handover	25%	-
Soft handover combining gain	0.7	-

Table 1: Summary of system simulation parameters.

largest uplink power from the mobile station will be used for transmission. A steered beam concept steers beams with pointing directions towards the different terminals, and the beams may be shaped according to the actual conditions. Consequently, the steered beam solution may eliminate any pointing errors of the fixed beam solution.

The system architecture requirements are different for the concepts, usually the requirements on hardware are more strict for a steered beam system due to coherency requirements of the radio frequency chains. In this paper we focus on the fixed beam solution with simplified modeling of the antenna diagrams.

3 System modeling

In this section, the modeling of the downlink for a single 5 MHz carrier is described. Although only a single cell is modeled explicitly, several system level characteristics such as intercell interference from other cells and soft handover are included.

Only speech users are considered in all investigated cases. The angular distribution of the users within a cell is an important parameter that affects the system performance of an advanced antenna system. Throughout this paper we assume the users to be distributed uniformly within the cell.

The cells must be able to provide the connectivity also in uplink. Therefore, a similar investigation of the uplink coverage and capacity should also be done. The focus of this paper is however on the downlink.

3.1 Model for the three-sector case

The model used in the case with three-sector sites is based on [3]. The C/I target (γ_k) of each mobile connection is calculated based on (1) and (2).

$$\gamma_k = \frac{P_k \; g_k \; (1+\Delta_k)}{(\alpha_k + F_k) P_{TOT} \; g_k + (1-\alpha) \; P_{SCH} \; g_k + N_f \; N} \quad (1)$$

$$P_{TOT} = \sum_{j=1}^{(1-SHO)\; n} P_j + \sum_{j=(1-SHO)n+1}^{(1+SHO)\; n} P_j + P_{CCH} + P_{SCH} \quad (2)$$

where:

- γ_k: carrier-to-interference ratio target of user k;
- n: number of users in the cell;
- P_k: base station transmit power towards user k;
- g_k: path gain between base station and user k at distance d_k; including path loss, sector antenna gain;
- α_k: orthogonality factor of user k;
- F_k: inter-to-intracell interference ratio of user k;
- P_{SCH}: average synchronization channel power;
- P_{CCH}: average common channel power;
- Δ_k: soft handover combining gain $\Delta_k = 0$ for users not in soft handover;
- N_f: mobile receiver noise figure;
- N: thermal noise in the mobile receiver;
- SHO: fraction of mobile users in soft handover.

The equations above model the synchronization channel (SCH) as non-orthogonal, while only the α_k fraction of common channel (CCH) power is experienced as intracell interference. For simplicity, only the case with a single scrambling code is presented in the equations above.

In Table 1 some of the main parameters are summarized.

3.2 Antenna diagrams

Simplified, idealized antenna diagrams were used in this study. The conventional three-sector case was simulated with an antenna diagram that had a constant gain of 18 dBi in the 120 degree sector. Each beam of a fixed beam system with M beams had an antenna diagram in which the mainbeam had a constant gain of M times the gain of the sector antenna and a beamwidth of $120/M$ degrees. In the first neighboring beams, the sidelobe level was set to -10 dB relative to the gain in the mainbeam, while the gain in all the other beams was set to -20 dB. This kind of idealized antenna diagrams are intended to approximate a fixed multi beam antenna system under close to ideal circumstances. The approximation also neglects pointing errors.

For the case with a fixed multi beam antenna, equations (1) and (2) where modified to take the approximate antenna diagrams into account. This is outlined in [7]. Note that the sector antenna diagram is used for the common channels (CCH and SCH). This since they have to be broadcast over the entire cell.

3.3 Propagation model

A simple path loss model for suburban environment was used. The path loss in dB was modeled as

$$a + b\ log_{10}(d) + l\,,$$

where d is the distance measured in kilometers, $a = 25.54$ and $b = 34.4$. The term l represents extra losses, such as e.g. cable loss. In this study, $l = 4$ dB was used.

3.4 Intercell interference

The intercell interference is one of the key characteristics of WCDMA systems and therefore very important to model. It is approximated by the inter-to-intracell interference ratio F_k. The model is based on the assumption that the F_k ratio in a three-sector cell follows an exponential increase as a function of the cell range.

The ratio F_k is assumed to be unchanged when using advanced antenna systems at the base stations. Such a modeling assumption may be valid in a system where all base stations are employed with advanced antenna systems. In such a full deployment, the capacity increase in each cell is then assumed to be the same, so that the ratio of intra and intercell interference remains approximately the same.

3.5 Downlink common channels

The total downlink power of a WCDMA base station consists of two components. First, the power allocated to the dedicated channels (including radio links in soft handover), then the power required for common channels to maintain the cell.

The power allocated on common channels is determined by two factors. Naturally, the first parameter is the cell range (or path loss). The second factor is the interference generated by the users *i.e.* the cell load. Since the activity factor of some common channels depend on the load of the cell, the power allocated to these control channels also depends on the cell load. A detailed description and a model of the downlink common channel power allocation is presented in [8].

In this study, the common channel power is set according to the path loss at the cell border assuming a fully loaded cell. The synchronization channel (SCH) is assumed to be completely non-orthogonal to all other channels (common and dedicated channels) in the cell. All other common channels (referred as CCH) are transmitted on the first scrambling code.

3.6 Soft handover

Soft handover is modeled in the way that the 25% of the users closest to the cell border are connected to the neighboring base station as well, see (2). These connections experience a soft handover combining gain of $\Delta = 0.7$, see (1), which corresponds to a soft handover margin of approximately 3 dB in the investigated environment [3]. Only two-leg soft handovers are considered, and soft handover with a neighboring sector of the same site (softer handover) is not included in the model.

3.7 Downlink orthogonality

The orthogonality factor for a given user, α_k, depends on the power delay profile, and the power delay profile does in general depend on the environment and the position in the cell. Despite this, in the numerical examples presented, an average orthogonality factor will be used for all links. In fact, two average orthogonality factors, $\alpha_k = 0.15$ and $\alpha_k = 0.4$, will be considered.

The orthogonality factor represents the fraction of the interference that a terminal despreading a certain channelization code will experience from signals transmitted on other channelization codes with the same scrambling code. For interference transmitted with different scrambling codes, an orthogonality factor of one is used.

Since multiple scrambling codes are considered, it should be noted that a scrambling code has to be associated with each link, and that equation (1) has to be modified to take this into account. However, these modifications are straight forward.

The following cases were considered.

1. No code limitation. In this case, the limitation of the number of available orthogonal channelization codes is neglected.

2. Beam based code allocation. The spatial dimension is taken into account in the sense that the number of scrambling codes used in each beam is minimized.

3. Power based code allocation. The code allocation of multiple scrambling codes is based on downlink power towards the users in the way that a new scrambling code is allocated to terminals closest to the base station. In this case all scrambling codes of the cell may be used in all beams.

Note that in all cases, the common channels are transmitted over the entire cell using the first scrambling code. 1/16 of the channelization codes on the first scrambling codes are reserved for this and cannot be used to serve dedicated links. The channelization codes required to establish the (two-leg) soft handover connections are also considered.

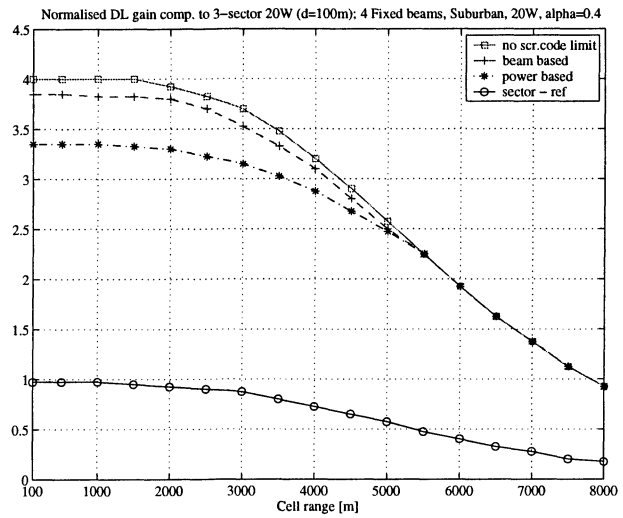


Figure 1: Downlink capacity gain of the 4 fixed beam concept normalized to the reference case at 100m cell range ($\alpha_k = 0.4$). The total transmit power used by all systems is 20W per cell.

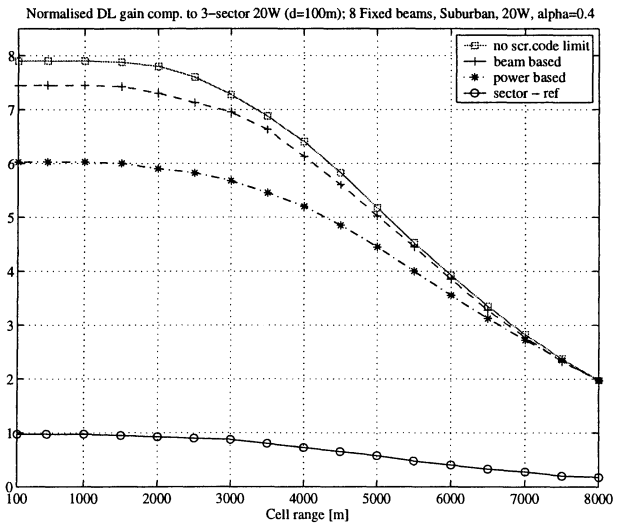


Figure 2: Downlink capacity gain of the 8 fixed beam concept normalized to the reference case at 100m cell range ($\alpha_k = 0.4$). The total transmit power used by all systems is 20W per cell.

It should also be stressed, that the code allocation schemes are not necessarily optimal in the sense that system performance is optimized.

4 Simulation results

We have selected the three-sector (single beam) scenario with 20W total downlink power per cell as the reference case. Figures 1 - 4 present downlink capacity gains which are normalized with this reference case at a cell range of 100m. The total transmit power used by the fixed beam systems was in the first three Figures the same as in the three-sector case, 20W.

Figure 1 shows normalized capacity gains of the four fixed beam concept for the three different scrambling code allocation methods described above. Figures 2 and 3 present similar results of the eight fixed beam concept for $\alpha_k = 0.4$ and 0.15, respectively. In Figure 4 results are presented for one (sector), four and eight beams examining 20, 40 and 80W of total downlink power available for the cell with beam based multiple scrambling code allocation.

As can be seen from the results, the use of advanced antenna systems may provide substantial gains as compared to conventional three-sector systems. At a given cell range, an advanced antenna system can be used to increase the capacity, and for a given capacity, the cell range can be extended significantly. It is also possible to trade the coverage and capacity gains depending on the requirements. For a given deployment, both the coverage and the capacity can be increased at the same time.

The results also show that there is difference between different scrambling code allocation strategies. The beam based scrambling code allocation method, which considers the spatial dimension, substantially outperforms the power based method. The power based method can be viewed as a non-coordinated scrambling code allocation method from a spatial point of view. The difference between the beam based and the power based scrambling code allocation is even larger for the case with a low average orthogonality factor, see Figure 3 which shows the case with $\alpha_k = 0.15$. Note here that the absolute capacity numbers is higher for the case with lower orthogonality factor.

In the presented results the effects of different downlink total power levels were also investigated (see Figure 4). They show that any additional downlink power which can be used for dedicated channels increase the WCDMA downlink capacity and coverage.

Similar results of detailed system level investigations for mixed service WCDMA systems are reported in [5] and [6]. The results of four fixed beam investigations show capacity gains of more than three times in cell ranges where mostly the intracell interference is the limiting factor. Even larger gains are found at larger cell ranges where the increased antenna directivity contribute to the capacity/coverage increase.

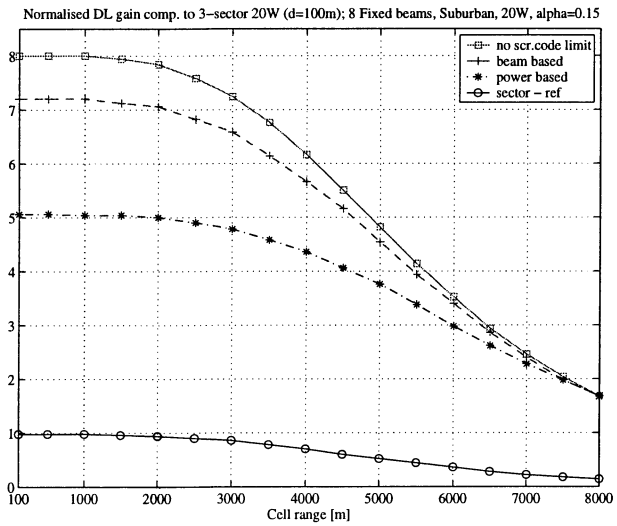


Figure 3: Downlink capacity gain of the 8 fixed beam concept normalized to the reference case at 100m cell range in suburban environment ($\alpha_k = 0.15$). The total transmit power used by all systems is 20W per cell.

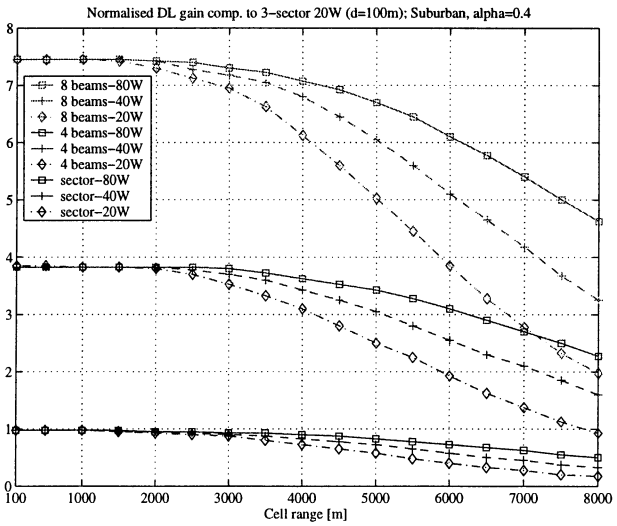


Figure 4: Downlink capacity gain of the single beam, 4 and 8 fixed beam concept with beam based multiple scrambling code allocation normalized to the reference case at 100m cell range ($\alpha_k = 0.4$).

5 Concluding remarks

In this paper, the coverage and capacity of a WCDMA system with fixed multi beam antennas were studied. In small interference limited cells, the gain offered can be used to increase the capacity. In power limited cases, it can be used to extend the coverage. Depending on the requirements, the gains, as compared to a conventional three-sector system, can be traded as to obtain both a capacity gain and coverage gain.

The use of multiple scrambling codes were considered. A degradation as compared to the case where the code limitation is neglected was observed. The degradation depends on the scrambling code allocation strategy. The results show that it is important to consider the spatial dimension in order to limit the degradation.

The paper studied only the downlink performance. However, it should be remembered that the uplink coverage and capacity also must be considered in order to balance the links of the system. Such an investigation then needs to consider the impact of the advanced antenna system on the uplink as well.

References

[1] J.C. Liberti, T.S. Rappaport, "Smart Antennas for Wireless Communications: IS-95 and Third Generation CDMA Applications", Prentice Hall, 1999

[2] 3GPP TSG RAN 25.213 v3.6.0, "Spreading and modulation (FDD), Release 1999".

[3] K. Hiltunen, R. Bernardi, "WCDMA Downlink Capacity Estimation", IEEE VTC 2000 Spring, May 2000, Tokyo

[4] B. Göransson, B. Hagerman, S. Petersson, J. Sorelius, "Advanced Antenna Systems for WCDMA: Link and System Level Results", PIMRC 2000, 2000, London

[5] M. Ericson, A. Osseiran, J. Barta, B. Göransson, B. Hagerman, "Capacity Study for Fixed Multi Beam Antenna Systems in a Mixed Service WCDMA System", PIMRC 2001, 2001. San Diego

[6] A. Osseiran, M. Ericson, J. Barta, B. Göransson, B. Hagerman, "Downlink Capacity Comparison between Different Smart Antenna Concepts in a Mixed Service WCDMA System", IEEE VTC2001 Fall, 2001 Atlantic City

[7] J. Barta, S. Petersson, B. Hagerman, "Downlink Capacity and Coverage Trade-offs in WCDMA with Advanced Antenna Systems", appearing in IEEE VTC 2002 Spring, 2002, Birmingham, Alabama

[8] K. Hiltunen, N. Binucci, "WCDMA Downlink Coverage: Interference Margin for Users Located at the Cell Coverage Border", appearing in IEEE VTC 2002 Spring, 2002, Birmingham, Alabama

Direction-of-Arrival Estimation in the Presence of Correlated Signals with Non-Ideal Uniform Circular Arrays

Buon Kiong Lau[1], Greg Cook[2], and Yee Hong Leung[3]

[1]ATCRC, Curtin University of Technology, Australia. Email: vincent@atri.curtin.edu.au
[2]ATCRC, Curtin University of Technology, Australia. Email: gcook@atri.curtin.edu.au
[3]ATCRC, Curtin University of Technology, Australia. Email: leung@atri.curtin.edu.au

Abstract

The Davies transformation is a method to transform the steering vector of a uniform circular array (UCA) to a vector with Vandermonde form. As this form is similar to that of the steering vector of a uniform linear array (ULA), we can apply to UCAs the many tools that have been developed for ULAs. However, the Davies transformation can be highly sensitive to perturbations of the underlying ideal array model. In this paper, we present two different methods for deriving a more robust transformation. The main underlying idea is to trade off robustness against accuracy of the Vandermonde form. In particular, we consider its application to direction-of-arrival estimation in the presence of correlated signals using root Weighted Subspace Fitting (root-WSF). The effectiveness of these methods are illustrated through numerical examples.

1 Introduction

Uniform circular arrays (UCAs), as a result of their two-dimensional array structure, are able to provide all azimuth coverage. Furthermore, when called for, they are able to provide 180° coverage in elevation. In comparison to uniform linear arrays (ULAs), the special features of UCAs come, however, at the cost of a less friendly steering vector: it is non-Vandermonde. It is well known that, as a direct consequence of the Vandermonde property of a ULA's steering vector, many powerful array-processing techniques have been devised. Some examples include root-MUSIC [1], Dolph-Chebyshev pattern synthesis [2] and spatial smoothing/averaging [3], [4]. Significant efforts have been directed to bringing these techniques over to UCAs via preprocessing on the array outputs. The philosophy behind these preprocessing techniques is to transform the steering vectors of UCAs to Vandermonde form.

There are currently two main approaches to the preprocessing step. Bronez et al [5], [6], [7] map the array outputs of a general two-dimensional array (e.g. a UCA) to a virtual uniform linear array, commonly called an interpolated array. However, the large difference in form between the steering vector of a UCA and a ULA means the mapping can be achieved only for a sector of angles to the minimize approximation error. The preprocessing step is then repeated to obtain full azimuth coverage.

A second approach, first devised by Davies [8], makes use of a spatial-DFT based transformation to map the UCA to a virtual array. This virtual array has no physical interpretation except that its steering vector is Vandermonde. Unlike the Bronez approach, the Davies approach uses only a single closed form transformation matrix and the resulting virtual array can provide all azimuth coverage. Different studies have demonstrated the successful applications of this approach in Dolph-Chebyshev pattern synthesis [9], root-MUSIC DOA estimation with spatial smoothing [10], [11] and optimum beamforming [12] in the presence of correlated signals.

In this paper, we apply the Davies transformation to enable DOA estimation with UCAs in the presence of correlated signals using the root Weighted Subspace Fitting (root-WSF) technique, also known as MODE [7], [13]. However, it has been shown that virtual arrays can be highly sensitive to array imperfections such as element position errors, mutual coupling, and gain and phase mismatches [14]. We outline two different techniques to improve the robustness of virtual arrays to array imperfections – a novel semi-infinite optimization technique [14]-[16] and a simple least squares technique. We demonstrate the effectiveness of our approach via numerical examples. In the examples, we also compare the performance of root-WSF against that of root-MUSIC with spatial smoothing. It is well known that the performance of root-WSF approaches that of maximum likelihood (ML) for sufficient number of snapshots. However, root-MUSIC can outperform root-WSF in the presence of array imperfections [17]. Here we investigate the impact of the robustness techniques on both methods.

2 Problem Formulation

2.1 The Davies Transformation

Consider a UCA with N elements, adjacent-element

spacing d and radius r. The nth component of the array response (or *steering*) vector $\mathbf{a}(\theta)$, $n=1,\ldots,N$, to a narrowband signal of wavelength λ arriving from angle θ, $\theta\in[-\pi,\pi]$, is given by

$$a_n(\theta)=G_n(\theta)\exp\left[j\frac{2\pi r}{\lambda}\cos\left(\theta-\frac{2\pi(n-1)}{N}\right)\right] \quad (1)$$

where $G_n(\theta)$ is the complex gain pattern of the nth element.

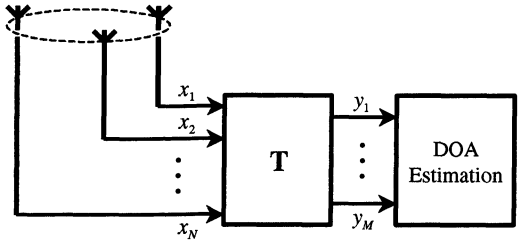


Fig. 1. Transformation for UCAs

Suppose the array elements are all identical and isotropic, i.e., $G_n(\theta)=1$, $n=1,\ldots,N$. Suppose further the antenna element outputs are processed as shown in Fig. 1 where $x_1,\ldots,x_N$ represent the baseband complex output signals of the "real" array and $y_1,\ldots,y_M$, $M<N$, represent the baseband complex output signals of the virtual array. Define the transformation matrix $\mathbf{T}=\mathbf{JF}$ where the matrices $\mathbf{J}\in\mathbb{C}^{M\times M}$ and $\mathbf{F}\in\mathbb{C}^{M\times N}$ are given by

$$\mathbf{J}=\mathrm{diag}\left\{\left[j^{m-1-H}\sqrt{N}J_{m-1-H}\left(\frac{2\pi r}{\lambda}\right)\right]^{-1}\right\} \quad (2)$$

and

$$[\mathbf{F}]_{mn}=\frac{1}{\sqrt{N}}e^{j2\pi(m-1-H)(n-1)/N}, \quad (3)$$

and where $m=1,\ldots,M$, $n=1,\ldots,N$, $J_k(\cdot)$ denotes a kth order Bessel function of the first kind, and

$$H=(M-1)/2\in\mathbb{Z}. \quad (4)$$

In [10], it is shown that the M-dimensional steering vector of the virtual array, $\tilde{\mathbf{b}}(\theta)$, will take on, approximately, the Vandermonde form

$$\tilde{\mathbf{b}}(\theta)=\mathbf{T}\mathbf{a}(\theta)\cong\left[e^{-jH\theta}\ \cdots\ 1\ \cdots\ e^{+jH\theta}\right]\triangleq\mathbf{b}(\theta). \quad (5)$$

Note, in view of (4), M is odd. Note also that $\mathbf{T}$ is a fixed matrix, i.e., it can be computed off-line.

2.2 Problem Statement

The lack of robustness of the Davies transformation can be traced to the construction of $\mathbf{J}$. As can be seen from (2), for some choices of m, H, and r/λ, the magnitude of one or more of the diagonal elements of $\mathbf{J}$ can approach infinity as the corresponding value of $J_{m-1-H}(2\pi r/\lambda)$ approaches zero. Accordingly, the norm of $\mathbf{T}$ can become very large. But the square of the norm of $\mathbf{T}$ gives a measure of the noise amplification of the transformation matrix. Therefore, for a $\mathbf{T}$ with large norm, small perturbations in $\mathbf{a}(\theta)$ will translate to large perturbations in $\tilde{\mathbf{b}}(\theta)$. Fig. 2 plots $|J_{m-1-H}(2\pi r/\lambda)|$ with zero-crossings in the range of interest for a 15-element UCA and a corresponding 13-element virtual array.

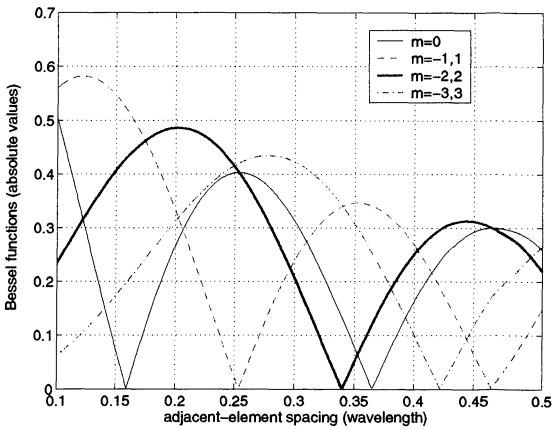


Fig. 2. Bessel functions in $\mathbf{J}$ vs. adjacent-element spacing

Based on the above observation, we formulate the following semi-infinite optimization (SIP) problem to find a more robust transformation matrix. The basic idea is to trade-off the approximation error in the transformation of $\mathbf{a}(\theta)$ for robustness.

Denote the robust transformation matrix by $\mathbf{U}\in\mathbb{C}^{M\times N}$. We find $\mathbf{U}$ as follows:

$$\min_{\mathbf{U}}\|\mathbf{U}\|_F^2 \quad (\mathcal{P}1)$$

subject to $|\mathbf{U}\mathbf{a}(\theta)-\mathbf{b}(\theta)|\le\mathbf{e}$, $\forall\theta\in[-\pi,\pi]$

where $\mathbf{e}=[\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_M]^T$, $\varepsilon_m\in\mathbb{R}_+$, $m=1,\ldots,M$, $\|\cdot\|_F$ denotes Frobenius norm, and $|\cdot|$ is the absolute value norm

$$|\mathbf{x}|=\begin{bmatrix}\max\{|\mathrm{Re}\{x_1\}|,|\mathrm{Im}\{x_1\}|\}\\ \vdots\\ \max\{|\mathrm{Re}\{x_M\}|,|\mathrm{Im}\{x_M\}|\}\end{bmatrix},$$

and where $\mathbf{x}\in\mathbb{C}^M$ and x_m is the mth element of $\mathbf{x}$.

Now, since the rows of $\mathbf{U}$ are not related in the above formulation, $(\mathcal{P}1)$ can be solved, row-by-row, as follows:

For $m=1,\ldots,M$, $\min_{\mathbf{u}_m}\|\mathbf{u}_m\|_F^2$ $(\mathcal{P}2)$

subject to $\left|\mathrm{Re}\left\{\mathbf{u}_m^T\mathbf{a}(\theta)-b_m(\theta)\right\}\right|\le\varepsilon_m$

and $\left|\mathrm{Im}\left\{\mathbf{u}_m^T\mathbf{a}(\theta)-b_m(\theta)\right\}\right|\le\varepsilon_m$, $\forall\theta\in[-\pi,\pi]$,

where $\mathbf{u}_m^T$ is the mth row of $\mathbf{U}$. The advantage of $(\mathcal{P}2)$ is that it allows the original problem $(\mathcal{P}1)$ to be solved efficiently by the dual parameterization method given in [14]-[16].

Remarks

1. The robustness of $\mathbf{U}$ depends on the choice of ε_m, $m=1,\ldots,M$. The larger we choose ε_m, the greater is the approximation error but correspondingly, the smaller is $\|\mathbf{u}_m\|_F$.
2. If $\varepsilon_m\ge 1$, then for that m, $(\mathcal{P}2)$ has the trivial solution $\mathbf{u}_m=\mathbf{0}$. This follows since $|b_m(\theta)|=1$.
3. As a rough guide to robustness, the square of the norm of each row of $\mathbf{U}$ should not greatly exceed N/M.

The reasoning is as follows. Suppose array imperfection can be modeled as a (complex) noise output from the antenna elements. Suppose the real and imaginary parts of these "imperfection noise" terms are independent with identical variance σ_x^2, and the imperfection noise terms of all the antenna elements are mutually independent. The total imperfection noise from the array of N elements is then given by $2N\sigma_x^2$. Suppose the transformation matrix has Frobenius norm $\|\mathbf{U}\|_F$. The total imperfection noise at the output of the virtual array is then given by $2\|\mathbf{U}\|_F^2\,\sigma_x^2$. If the transformation is required to not increase noise, then we require $2N\sigma_x^2 \geq 2\|\mathbf{U}\|_F^2\,\sigma_x^2$, or $\|\mathbf{U}\|_F^2 \leq N$. Finally, suppose the noise gain is distributed uniformly over the elements of the virtual array. This yields $\|\mathbf{u}_m\|_F^2 = \|\mathbf{U}\|_F^2/M \leq N/M$.

Alternatively, the idea of trading-off approximation error for a more robust transformation can be captured in the following least squares (LS) formulation:

$$\min_{\mathbf{u}_m}\left\{\alpha\int_{-\pi}^{\pi}\left\|\mathbf{u}_m^H\mathbf{a}(\theta)-b_m(\theta)\right\|_F^2 d\theta+(1-\alpha)\|\mathbf{u}_m\|_F^2\right\} \quad (\mathcal{P}3)$$

where $\alpha \in (0,1)$ is the weighting factor determined by the desired level of trade-off. $(\mathcal{P}3)$ can be solved easily by differentiating the objective function w.r.t $\mathbf{u}_m$ and setting the derivative to $\mathbf{0}$.

3 Numerical Examples

3.1 The UCA

We considered a scenario of highly correlated signals (correlation of $0.99\exp(j\pi/4)$), with wavelength λ and a 15-element UCA with $d = 0.46\lambda$. A 13-element virtual array was chosen for root-WSF [7], [13]. A whitening procedure similar to [7] was incorporated into root-WSF since noise in the virtual array is not spatially white [10]. For root-MUSIC, forward-backward spatial smoothing [4] was applied to the 13-element virtual array to restore the rank of the covariance matrix, followed by a whitening procedure [10].

Table 1 summarizes the squared-norm and maximum real and imaginary errors of each row of the Davies matrix for this UCA. As can be seen, the squared-norms of rows 6 and 8 greatly exceed $N/M = 15/13 = 1.1538$. Indeed, it is these very rows that render the Davies matrix non-robust. For the robust transformation matrix, our strategy is to retain as many rows of the Davies matrix as possible except for rows with large squared-norms. Accordingly, for the SIP technique, we replaced rows 6 and 8 with rows found by solving $(\mathcal{P}2)$ with ε_m set to 0.55[1]. Note that the resulting ε_m in Table 1 is slightly different from 0.55 due to the finite number of iterations in the optimization procedure. The characteristics of the robust transformation matrix are summarized also in Table 1. Note the increase in approximation error in rows 6 and 8. Note also that the resulting transformation vectors for rows 6 and 8 are related by a phase rotation. This also holds true when the optimization was carried out for any other similar pairs of rows in Table 1. In other words, we only need to compute $(\mathcal{P}2)$ once for each such pairs of rows.

	Davies Matrix		Robust Matrix	
Row #	**Squared-norm**	**Max Error**	**Squared-norm**	**Max Error**
1, 13	0.5896	0.1677		
2, 12	0.5375	0.0634		
3, 11	2.2956	0.0460		
4, 10	2.7382	0.0159		
5, 9	0.7161	0.0023		
6, 8	174.437	0.0102	18.7656	0.5534
7	0.7433	0.0003		

Table 1. Characteristics of the Davies and robust transformation matrices for $N = 15$, $M = 13$, $r = 1.106\lambda$

3.2 Case Study 1

First we look at the case of 2 signals arriving from directions 0° and 30°. The signal-to-noise-ratios (SNRs) of both signals are varied between 0 and 15 dB. We used 2 subarrays of 12 elements each for spatial smoothing. The performance is measured in terms of the average root mean squared error (RMSE) of the DOAs over all signals present. Fig. 3 summarizes the average RMSE of the DOA estimates for the ideal, non-ideal and robust UCAs. The non-ideal and robust UCAs were subject to perturbations drawn from standard normal distributions with zero means. The standard deviation for either real or imaginary part of gain perturbations is 0.05 (relative to 1); for either x- or y-axis element position perturbation, 0.01λ; and for either real or imaginary part of mutual coupling, 0.01 (relative to 1). We assumed that mutual coupling is significant only for adjacent UCA elements. Here we used 200 snapshots and 1000 Monte-Carlo trials, where each Monte-Carlo trial corresponds to a realization of array perturbations. From Fig. 3, it is clear that in this case root-WSF with robust UCA has significantly better performance than that with non-ideal UCA. At low SNR, root-WSF with robust UCA appears to outperform root-WSF with ideal UCA. This could be due to the sensitivity of root-WSF to amplification of "noise" (for rows in Davies matrix with large norm) resulting from the number of finite snapshots at low SNR.

As for root-MUSIC which is inherently more robust towards array imperfections [17], the performance of the robust UCA is roughly the same as the non-ideal UCA for $\varepsilon_m = 0.55$. A very marginal improvement may be obtained by optimizing ε_m for the best robust UCA performance. We also note that the approximation involved in the transformation introduces a small bias in the DOA estimates of both methods [11] which has been taken into account in the RMSE measure.

[1] A simple search procedure is used to find the value of ε_m that gives the optimum robust performance w.r.t root-WSF.

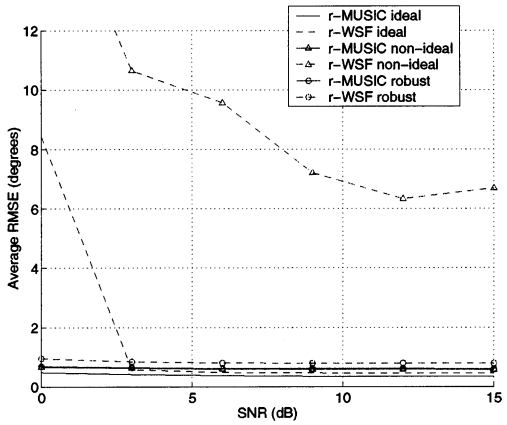


Fig. 3. RMSE DOA performance of UCAs: 2-signal case

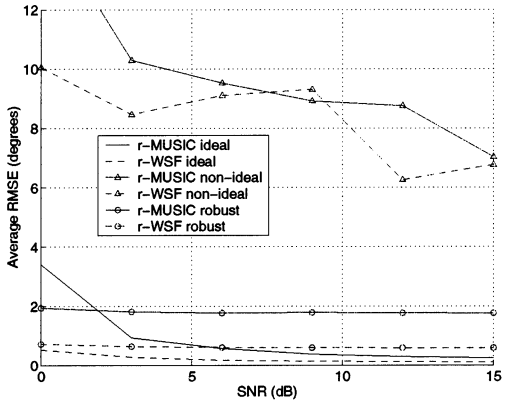


Fig. 4. RMSE DOA performance of UCAs: 5 signal case

3.3 Case Study 2

Next, we had 5 signals arriving from the directions $-150°$, $-90°$, $0°$, $37°$, $85°$ with SNRs varied between 0 and 15 dB. For this case, we used 5 subarrays of 9 elements each for the spatial smoothing. For this case, setting $\varepsilon_m = 0.55$ again gives the best performance for root-WSF. In this case, we observe significant improvement in the robust UCAs for both root-WSF and root-MUSIC. The presence of more highly correlated signals presents a bigger challenge for root-MUSIC as more subarrays of a smaller subarray size are used. As a result, root-MUSIC is less robust than the 2-signal case. In fact, if the correlation between signals is real or close to real, then the performance of root-MUSIC (with spatial smoothing) deteriorates further as the forward-backward smoothing effectively reduces to forward only smoothing under this condition [18].

3.4 Discussions

With the LS technique to find the optimum $\mathbf{u}_m$ and α, we found in our simulations that the procedure, though computationally simple and also effective, does not perform as well as the SIP technique[2]. This could be due to the constrained formulation of ($\mathcal{P}2$) which gives a more uniform approximation error for the virtual array.

It is observed from the case studies that the relative performance of root-WSF and root-MUSIC with spatial smoothing depends on several parameters which include the number of highly correlated signals present (which also influences the required size of subarray), the nature of signal correlation (real or complex) [18] and the sensitivity to array imperfections [17]. While root-WSF has higher computational complexity than root-MUSIC, root-WSF demonstrates a more consistent RMSE performance than root-MUSIC for the robust UCAs where ε_m is optimized separately for each method.

It is worthwhile to note that while the robustness procedure is able to improve the performance of non-ideal UCAs with critical array parameters, the best robust UCA performance is still poorer than that of UCAs with non-critical parameters. Here, we give an example of the performance of a UCA with non-critical parameters, viz. $N = 15$, $M = 13$, $d/\lambda = 0.4$. There are 5 signals and 5 subarrays of 9-element each as described in case study 2. In Fig. 2, we see that $d/\lambda = 0.4$ is far from any zero-crossings. Indeed, Table 2 further confirms that the norm values are acceptable. In Fig. 5, we show the RMSE performance for the ideal, non-ideal and robust UCAs. The robust UCA operates on rows 4 and 10 and uses $\varepsilon_m = 0.1$ for the best performance. The degradation in performance in non-ideal UCA is smaller than case studies 1 and 2, and the robust procedure, as expected, gives only very marginal improvements, if at all. Further we note that in this case the performance of root-WSF is better than that of root-MUSIC in the ideal, non-ideal and robust UCAs.

	Davies Matrix		Robust Matrix	
Row #	**Squared-norm**	**Max Error**	**Squared-norm**	**Max Error**
1, 13	1.0587	0.0888		
2, 12	0.5020	0.0203		
3, 11	0.5380	0.0062		
4, 10	6.4713	0.0058	5.2178	0.101
5, 9	1.0560	0.0006		
6, 8	0.9297	0.0001		
7	2.5200	0.0001		

Table 2. Characteristics of the Davies transformation matrices for $N = 15$, $M = 13$ and $r = 0.962\lambda$

The above observation also indicates that while robustness may be introduced in the vicinity of critical parameters, it is prudent to avoid them altogether if possible. Nevertheless, for large arrays with more frequent zero crossings than that of Fig. 2, and where these parameters are not easily avoided, our proposed method can be effective.

[2] A simple search procedure is used to find the value of α that gives the optimum robust performance w.r.t root-WSF.

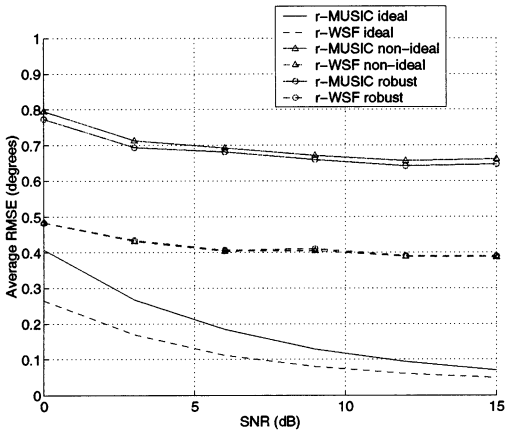


Fig. 5. RMSE DOA performance of UCAs: 5 signal case for $r = 0.962\lambda$

5 Conclusions

In this paper, we addressed the important problem of direction finding with imperfect UCAs in correlated signal environments. We showed that a crucial step in the solution is to find a robust transformation matrix to transform the steering vector of the UCA to one with Vandermonde form. The robust matrix is found by posing and solving a quadratic semi-infinite optimization problem which trades-off the Vandermonde approximation error with a matrix of lower norm. We showed that, by an appropriate formulation, we can decompose the problem into a set of much simpler optimization problems which can then be solved efficiently using the dual parameterization method of [14]-[16]. Each sub-problem yields a row of the robust transformation matrix. Alternatively, we also proposed a least squares formulation that operates on similar principles. The robustness of the new transformation matrix is demonstrated by numerical examples where we investigated the performance of the robust UCA with respect to root-WSF and root-MUSIC with spatial smoothing. We note that the robustness procedure appears to be particularly effective for root-WSF. Finally, we remark that it is trivial to extend the presented work to elevation angles for hemispherical coverage [11] where a different **J** is used for each elevation angle of interest.

References

[1] A. J. Barabell. Improving the resolution performance of eigenstructure-based direction-finding algorithms. *Proc. IEEE ICASSP'83*, pp. 336-339, Boston, MA, 1983.

[2] C. L. Dolph. A current distribution for broadside arrays which optimizes the relationship between beam width and side-lobe level. *Proc. IRE*, 34(6), pp. 335-348, Jun. 1946.

[3] T. J. Shan and T. Kailath. Adaptive beamforming for coherent signals and interference. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(3), pp. 527-536, Mar. 1985.

[4] S. U. Pillai and B. H. Kwon. Forward-backward spatial smoothing techniques for coherent signal identification. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(1), pp. 8-15, Jan. 1989.

[5] T. P. Bronez. Sector interpolation of nonuniform arrays for efficient high resolution bearing estimation. *Proc. IEEEICASSP'88*, pp. 2885-2888, NY, USA, Apr. 1988.

[6] B. Friedlander and A. J. Weiss. Direction finding using spatial smoothing with interpolated arrays. *IEEE Trans. Aerosp. Electron. Syst.*, vol.28, pp.574-587, Apr. 1992.

[7] A. J. Weiss, B. Friedlander, and P. Stoica. Direction-of-arrival estimation using MODE with interpolated arrays. *IEEE Trans. Signal Processing*, vol.43, pp.296-300, Jan. 1995.

[8] D. E. N. Davies. A transformation between the phasing techniques required for linear and circular aerial arrays. *Proc. IEE*, 112(11), pp. 2041-2045, 1965

[9] B. K. Lau and Y. H. Leung. A Dolph-Chebyshev approach to the synthesis of array patterns for uniform circular arrays. *Proc. IEEE ISCAS*, vol. 1, pp. 124-127, Geneva, Switzerland, May 2000.

[10] M. Wax and J. Sheinvald. Direction finding of coherent signals via spatial smoothing for uniform circular arrays. *IEEE Trans. Antennas Propagat.*, 42(5), pp. 613-620, May 1994.

[11] C. P. Mathews and M. D. Zoltowski. Eigenstructure techniques for 2-D angle estimation with uniform circular arrays. *IEEE Trans. Signal Processing*, 42(9), pp. 2395-2407, Sep. 1994.

[12] B. K. Lau and Y. H. Leung. Optimum beamformers for uniform circular arrays in a correlated signal environment. *Proc. IEEE ICASSP*, vol. 5, pp. 3093-3096, Istanbul, Turkey, Jun. 2000.

[13] P. Stoica and K. Sharman. Novel eigenanalysis method for direction estimation. *IEE Proc.*, vol. 137, pt. F, pp. 19-26, Feb. 1990.

[14] B. K. Lau. Y. H. Leung, Y. Liu, and K. L. Teo. Direction of Arrival Estimation in the Presence of Correlated Signals and Array Imperfections with Uniform Circular Arrays. *IEEE ICASSP'2002*, accepted.

[15] Y. Liu, K. L. Teo and S. Ito. Global optimization in quadratic semi-infinite programming. *Computing*, Supplement 15, pp. 119-132, 2001.

[16] Y. Liu and K. L. Teo. Adaptive dual parameterization algorithm for quadratic semi-infinite programming problems. *J. of Global Optimization*, (accepted).

[17] A. L. Swindlehurst and T. Kailath. A performance analysis of subspace-based methods in the presence of model errors: part II – multidimensional algorithms. *IEEE Trans. Signal Processing*, 41(9), pp. 2882-2890, Sep. 1993.

[18] K. M. Reddy and V. U. Reddy. Analysis of spatial smoothing with uniform circular arrays. *IEEE Trans. Signal Processing*, 47(6), pp. 1726-1730, Jun. 1999.

Model of an array antenna consisting of circular apertures on a truncated ground plane

Ola Forslund[1,2] and Lennart Mattsson[2]

[1] Div. of Electromagnetic Theory, Alfvén Lab., Royal Institute of Technology, SE-100 44, Sweden.

[2] Saab Bofors Dynamics AB, Järfälla, SE-175 88, Sweden.

Abstract

When analysing systems consisting of array antennas "simple" models are often requested concerning coupling and radiation. A rigorous numerical model based on e.g. FEM- or integral equation solvers are often too complicated and time consuming to be used in system analysis models. This contribution describes a simple model of an array antenna consisting of circular waveguide apertures on a truncated ground plane. A few modes in each waveguide are assumed to excite the structure considered. Based on equivalence principles, the waveguide modes are assumed to excite some characteristic electric, and magnetic surface modes in an exterior problem. Based on made assumptions, a model for the coupling matrix and element patterns is derived. The simple model is compared with calculations made using the FEM-program HFSS. Good agreements are obtained for the copolar radiation patterns. The cross-polarised diagrams show no agreement in detail.

1 Introduction

An array antenna consisting of circular waveguide apertures on a finite groundplane has been designed. The antenna consists of 19 elements arranged in a hexagonal grid. (See Figure 5) The element spacing is $0.75\lambda_0$ and the size of the apertures $0.47\lambda_0$ (in terms of a reference wavelength λ_0). Each aperture is filled with a dielectric plug. The elements are surrounded by circular quarter wavelength grooves in order to match the antenna and reduce the interelement coupling. The grooves considerably reduce the current flown in the finite groundplane so that the interaction with the edges of the groundplane is lowered. As a consequence, the "short-circuited" radiation pattern of an individual aperture becomes rather insensitive to the location on the groundplane. (By "short-circuited" element pattern we mean the element pattern obtained for a particular aperture when the surrounding apertures are short-circuited.) This implicates that the radiation pattern of the array can be approximated by a pattern factorised in element- and array parts. The antenna can thus be modeled by an array of the typical "short-circuited" element pattern. The typical element pattern is here calculated using the FEM-program HFSS. Comparisons are made using the simplified model consisting of an array of elements having the typical element pattern and the antenna as a whole rigorously analysed using HFSS.

2 Waveguide apertures in a PEC

Consider N waveguide apertures in a PEC (Perfect Electric Conductor). See Figure 1 . Suppose that we have a plane wave incident from the outside as well as waves incident upon the aperture from the waveguides. Denote by $\mathbf{E}_0^{in}$ and $\mathbf{H}_0^{in}$ the fields generated by the incident plane waves and by $\mathbf{E}_n^{in}$ and $\mathbf{H}_n^{in}$ the waves incident from the waveguides. Using equivalence principles [1] this problem can be divided into $N+1$ subproblems, one exterior problem and N interior problems. The different subproblems are coupled together by applying the boundary conditions for the original problem. The equivalent problem for the exterior region is shown in Figure 2. In this problem we have sources in the form of the fields incident in the original problem $\mathbf{E}_0^{in}$, $\mathbf{H}_0^{in}$ and in the form of magnetic surface currents in $\mathbf{M}_n$ where $\mathbf{M}_n(\mathbf{r}') = -\hat{n}\times\ \mathbf{E}^{tot}(\mathbf{r}')$ and finally electrical surface currents $\mathbf{J}_s$ where $\mathbf{J}_s = \hat{n}\times\ \mathbf{H}^{tot}(\mathbf{r}')$. The total electric field in the exterior region can then be written

$$\mathbf{E}^{tot} = \mathbf{E}^{in}(\mathbf{M}_1)+\ldots+\mathbf{E}^{in}(\mathbf{M}_N)+\mathbf{E}^{s}(\mathbf{J}_s)+\mathbf{E}_0^{in} \quad (1)$$

where $\mathbf{E}^s(\mathbf{J}_s)$ is the scattered field whose sources $\mathbf{J}_s$ are induced at the surface of the inner region. The sources $\mathbf{M}_n(\mathbf{r}')$ are located at the positions of the apertures in the original problem. The terms in (1) can in terms of the free space Green's function be written as

$$\mathbf{E}^{in}(\mathbf{M}_n) = -\nabla \times \int_{S_n} \mathbf{M}_n(\mathbf{r}')\, G(\mathbf{r},\mathbf{r}')\, dS$$

$$\mathbf{E}^{s}(\mathbf{J}_s) =$$

$$-i\omega\mu_0 \int_S \left[\mathbf{J}(\mathbf{r}')G(\mathbf{r},\mathbf{r}') + \tfrac{1}{k_0^2}\nabla\left(\nabla\cdot(\mathbf{J}(\mathbf{r}')G(\mathbf{r},\mathbf{r}')\right)\right] dS$$

where

$$G(\mathbf{r},\mathbf{r}') = \tfrac{e^{-ik_0|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|}$$

The Interior problems can be dealt with in a way similar to the exterior though we do not use the free space Green's functions for the interior problems. We have in interior problem No. n the sources $\mathbf{E}_n^{in}$ and $\mathbf{E}(-\mathbf{M}_n) = -\mathbf{E}(\mathbf{M}_n)$ where the $(-)$ sign appears because of the continuity of the electric field in the aperture No. n. The magnetic sources $\mathbf{M}_n$ and the induced electric surface current $\mathbf{J}_s$ can be rigorously determined by formulating a system of integral equations for the magnetic field which couples the exterior problem to the $N-1$ interior. The magnetic sources $\mathbf{M}_n$ could then preferably be expanded in waveguide modes. In the real problem with circular apertures, viewing the antenna as a transmitter and assuming the waveguides terminated, we would have one (or two) propagating modes incident upon the aperture and an infinite set of backscattered modes of which two (degenerate fundamental modes) would be propagating and the rest evanescent.

We will not solve any integral equations here, however, the exercise above is included as a fundament for the argumentation regarding the approximations later made when deriving the coupling matrix and radiation patterns. In the sections to follow, an approximation approach, similar to what sometimes is denoted as the "single mode approximation" is described. (See e.g. [2].)

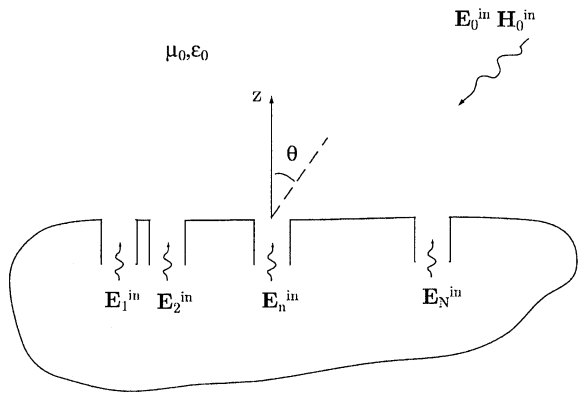


Figure 1: Original problem

3 Problem considered

3.1 Apertures short-circuited

In the following we consider, without loss of generality, the antenna as a transmitter and assume that no incident fields $\mathbf{E}_0^{in}$, $\mathbf{H}_0^{in}$ from the exterior occur. (The antenna has the same transmit and receive patterns which in general differ from the scattering pattern.) Assume that we short circuit all apertures except No. n. In the equivalent problem for the exterior we then only have the source $\mathbf{M}_n$ present. We

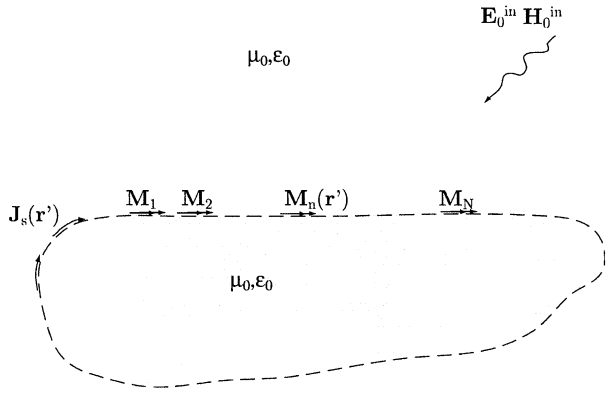


Figure 2: Equivalent problem for the exterior region

write $\mathbf{M}_n = b_n^1\mathbf{m}_n^1 + b_n^2\mathbf{m}_n^2$ (where the superscripts are to be considered as indices). These magnetic current modes should be thought of as characteristic current configurations obtained when the fundamental waveguide modes are incident upon aperture n. The magnetic surface current modes will excite and coexist with a particular electric surface current configuration $\mathbf{J}_s = b_n^1\mathbf{j}_n^1 + b_n^2\mathbf{j}_n^2$. The electric surface current modes $\mathbf{j}_n^1$ and $\mathbf{j}_n^2$ have support over the entire boundary of the exterior domain in the equivalent problem as opposed to the magnetic currents $\mathbf{m}_n^1$ and $\mathbf{m}_n^2$ which only exist in the original location of aperture n. The surface currents are excited by incident fundamental waveguide modes of type 1 and 2 respectively. These modes are propagating, orthogonal and have degenerate eigenvalues. We obtain

$$\begin{bmatrix} b_n^1 \\ b_n^2 \end{bmatrix} = \begin{bmatrix} 1+s_{11}^{s;n} & s_{12}^{s;n} \\ s_{21}^{s;n} & 1+s_{22}^{s;n} \end{bmatrix} \begin{bmatrix} b_n^{1\prime} \\ b_n^{2\prime} \end{bmatrix} \quad (2)$$

where $b_n^{1\prime}$ and $b_n^{2\prime}$ are coefficients for the incident orthogonal modes that we denote by $\mathbf{E}_1^+$ and $\mathbf{E}_2^+$. The inward propagating fields can be written $(s_{11}^s b_n^{1\prime} + s_{12}^s b_n^{2\prime})\mathbf{E}_1^-$ and $(s_{21}^s b_n^{1\prime} + s_{22}^s b_n^{2\prime})\mathbf{E}_2^-$ where $\mathbf{E}_1^-$ and $\mathbf{E}_2^-$ denote the inward propagating modes. One could argue that this would violate the boundary condition (of the continuity of the fields in the aperture) but we will only interpret (2) as $b_n^{1\prime}\mathbf{E}_1^+$ and $b_n^{2\prime}\mathbf{E}_2^+$ exciting $b_n^1\mathbf{m}_n^1$ and $b_n^2\mathbf{m}_n^2$ in the aperture.

3.2 Apertures terminated

Here we still consider the antenna as a transmitter and assume that no incident fields $\mathbf{E}_0^{in}$, $\mathbf{H}_0^{in}$ from the exterior exist. Furthermore we assume that all apertures are present and that the waveguides are terminated looking inwards. Feeding waveguide n we then more or less excite all apertures. At aperture m we obtain due to excitation of aperture n that

$\mathbf{M}_m = a^1_{mn}\mathbf{m}^1_m + a^2_{mn}\mathbf{m}^2_m$ where

$$\begin{bmatrix} a^1_{mn} \\ a^2_{mn} \end{bmatrix} = \begin{bmatrix} \delta_{mn} + s^{11}_{mn} & s^{12}_{mn} \\ s^{21}_{mn} & \delta_{mn} + s^{22}_{mn} \end{bmatrix} \begin{bmatrix} a^{1\prime}_n \\ a^{2\prime}_n \end{bmatrix} \tag{3}$$

and where $a^{1\prime}_n$ and $a^{2\prime}_n$ are coefficients for the incident orthogonal modes in waveguide n and where $\delta_{mn} = 1$ when $m = n$ and 0 otherwise.

3.3 Transformation between real and hypothetical antenna

Now assume that we have the same excitation for the (hypothetical) array antenna consisting of a sum of elements with short-circuited surroundings (which we imagine is appropriate for signal processing) and the real array consisting of elements with terminated waveguides. In the "short-circuited" case we have

$$\mathbf{M}_n = b^1_n\mathbf{m}^1_n + b^2_n\mathbf{m}^2_n, \qquad n = 1, 2, \ldots, N$$

In the real case we have

$$\mathbf{M}_n = \sum_{m=1}^{N}(a^1_{nm}\mathbf{m}^1_n + a^2_{nm}\mathbf{m}^2_n), \qquad n = 1, 2, \ldots, N$$

If we let these two be equal we obtain

$$\begin{bmatrix} b^1_n \\ b^2_n \end{bmatrix} = \sum_{m=1}^{N} \begin{bmatrix} a^1_{nm} \\ a^2_{nm} \end{bmatrix} = \sum_{m=1}^{N} \begin{bmatrix} \delta_{nm} + s^{11}_{nm} & s^{12}_{nm} \\ s^{21}_{nm} & \delta_{nm} + s^{22}_{nm} \end{bmatrix} \begin{bmatrix} a^{1\prime}_m \\ a^{2\prime}_m \end{bmatrix} \tag{4}$$

We then obtain

$$\begin{bmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_N \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \vdots \\ c_{N1} & \cdots & c_{NN} \end{bmatrix} \begin{bmatrix} \tilde{a}'_1 \\ \vdots \\ \tilde{a}'_N \end{bmatrix}$$

where $\tilde{b}_n = \begin{bmatrix} b^1_n \\ b^2_n \end{bmatrix}$, $\tilde{a}_m = \begin{bmatrix} a^{1\prime}_m \\ a^{2\prime}_m \end{bmatrix}$ and $c_{nm} = \begin{bmatrix} \delta_{nm} + s^{11}_{nm} & s^{12}_{nm} \\ s^{21}_{nm} & \delta_{nm} + s^{22}_{nm} \end{bmatrix}$. The matrix c_{nm} is the coupling matrix for the problem. We thus have a relationship between the excitation coefficients in the hypothetical array and the real one. Say that we want to design an array with a certain aperture distribution (i.e. sampled Taylor). We then decide the coefficients $\{\tilde{b}_n\}$ in the hypothetical array according to this and then we obtain $\{\tilde{a}'_n\}$ by

$$\begin{bmatrix} \tilde{a}'_1 \\ \vdots \\ \tilde{a}'_N \end{bmatrix} = \mathcal{C}^{-1} \begin{bmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_N \end{bmatrix}$$

The radiation patterns can be decided either by considering the "short circuited" element patterns or the ones with a terminated surrounding. In the 1:st case we write

$$\begin{bmatrix} \mathbf{F}_{co}(\theta,\varphi) \\ \mathbf{F}_{xo}(\theta,\varphi) \end{bmatrix} = \sum_{n=1}^{N} \begin{bmatrix} \mathbf{d}^{cc}_n(\theta,\varphi) & \mathbf{d}^{cx}_n(\theta,\varphi) \\ \mathbf{f}^{xc}_n(\theta,\varphi) & \mathbf{f}^{xx}_n(\theta,\varphi) \end{bmatrix} \begin{bmatrix} b^1_n \\ b^2_n \end{bmatrix} \mathrm{e}^{ik_0\hat{r}\cdot\mathbf{r}'_{t;n}} \tag{5}$$

where $\mathbf{d}^{cc}_n(\theta,\varphi)$ and $\mathbf{f}^{xx}_n(\theta,\varphi)$ are the copolar radiation patterns for modes of type 1 and type 2 respectively. Furthermore $\mathbf{d}^{cx}_n(\theta,\varphi)$ is the cross polar pattern given by excitation of mode 2 while $\mathbf{f}^{xc}_n(\theta,\varphi)$ is the copolar pattern given by excitation of mode 1. Co- and cross-polarisation is preferably defined according to some of Ludwigs principles [3]. The factor $\mathrm{e}^{ik_0\hat{r}\cdot\mathbf{r}'_{t;n}}$ can be excluded if a common reference point is chosen for the element pattern calculations. As a comparison we can in the case of a terminated surrounding write

$$\begin{bmatrix} \mathbf{F}_{co}(\theta,\varphi) \\ \mathbf{F}_{xo}(\theta,\varphi) \end{bmatrix} = \sum_{m=1}^{N} \begin{bmatrix} \mathbf{g}^{cc}_m(\theta,\varphi) & \mathbf{g}^{cx}_m(\theta,\varphi) \\ \mathbf{h}^{xc}_m(\theta,\varphi) & \mathbf{h}^{xx}_m(\theta,\varphi) \end{bmatrix} \begin{bmatrix} a^1_{mm} \\ a^2_{mm} \end{bmatrix} \mathrm{e}^{ik_0\hat{r}\cdot\mathbf{r}'_{t;n}} \tag{6}$$

If (4) is inserted into (5) we obtain

$$\begin{bmatrix} \mathbf{F}_{co}(\theta,\varphi) \\ \mathbf{F}_{xo}(\theta,\varphi) \end{bmatrix} = \sum_{n,m} \begin{bmatrix} \mathbf{d}^{cc}_n(\theta,\varphi) & \mathbf{d}^{cx}_n(\theta,\varphi) \\ \mathbf{f}^{xc}_n(\theta,\varphi) & \mathbf{f}^{xx}_n(\theta,\varphi) \end{bmatrix} \begin{bmatrix} \delta_{nm} + s^{11}_{nm} & s^{12}_{nm} \\ s^{21}_{nm} & \delta_{nm} + s^{22}_{nm} \end{bmatrix} \begin{bmatrix} a^{1\prime}_m \\ a^{2\prime}_m \end{bmatrix} \mathrm{e}^{ik_0\hat{r}\cdot\mathbf{r}'_{t;n}} \tag{7}$$

3.4 Further approximations

Until now it has been assumed that all element patterns are individual. However, due to the quarter-wavelength grooves it is found that the "short-circuited" element patterns are fairly independent of the location of the individual aperture. This is because the groove significantly reduces the current on the truncated ground plane and thus the interaction with the edge is considerably smaller than in the case of no grooves. Assuming that the "short-circuited" element patterns are equal for all elements we write

$$\begin{bmatrix} \mathbf{F}_{co}(\theta,\varphi) \\ \mathbf{F}_{xo}(\theta,\varphi) \end{bmatrix} = \begin{bmatrix} \mathbf{d}^{cc}(\theta,\varphi) & \mathbf{d}^{cx}(\theta,\varphi) \\ \mathbf{f}^{xc}(\theta,\varphi) & \mathbf{f}^{xx}(\theta,\varphi) \end{bmatrix} \sum_{n=1}^{N} \left\{ \mathrm{e}^{ik_0\hat{r}\cdot\mathbf{r}'_{t;n}} \sum_{m=1}^{N} \begin{bmatrix} \delta_{nm} + s^{11}_{nm} & s^{12}_{nm} \\ s^{21}_{nm} & \delta_{nm} + s^{22}_{nm} \end{bmatrix} \begin{bmatrix} a^{1\prime}_m \\ a^{2\prime}_m \end{bmatrix} \right\} \tag{8}$$

Now let $\mathbf{d}^{cc}(\theta,\varphi) = \Theta^{cc}(\theta,\varphi)\hat{\theta}$, $\mathbf{d}^{cx}(\theta,\varphi) = \Theta^{cx}(\theta,\varphi)\hat{\theta}$, $\mathbf{f}^{xx}(\theta,\varphi) = \Phi^{xx}(\theta,\varphi)\hat{\varphi}$ and $\mathbf{f}^{xc}(\theta,\varphi) = \Phi^{xc}(\theta,\varphi)\hat{\varphi}$. The intuitively most appropriate polarisation definition is Ludwig's 3:rd [3]. Letting $\hat{x}$ -polarised modes in the waveguide denote copolarisation for the sources, according to Ludwig's 3:rd we then obtain

$$\begin{aligned} \hat{e}_{co} &= \cos(\varphi)\hat{\theta} - \sin(\varphi)\hat{\varphi} \\ \hat{e}_{xo} &= \sin(\varphi)\hat{\theta} + \cos(\varphi)\hat{\varphi} \end{aligned}$$

Writing $\mathbf{F}_{co}(\theta,\varphi) = F_{co}^{(3)}(\theta,\varphi)\hat{e}_{co}$ and $\mathbf{F}_{xo}(\theta,\varphi) = F_{xo}^{(3)}(\theta,\varphi)\hat{e}_{xo}$ we then finally obtain

$$\begin{bmatrix} F_{co}^{(3)}(\theta,\varphi) \\ F_{xo}^{(3)}(\theta,\varphi) \end{bmatrix} = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} \cdot \begin{bmatrix} \Theta^{cc}(\theta,\varphi) & \Theta^{cx}(\theta,\varphi) \\ \Phi^{xc}(\theta,\varphi) & \Phi^{xx}(\theta,\varphi) \end{bmatrix} \cdot \sum_{n=1}^{N} \left\{ e^{ik_0\hat{r}\cdot\mathbf{r}'_{t;n}} \sum_{m=1}^{N} \begin{bmatrix} \delta_{nm} + s_{nm}^{11} & s_{nm}^{12} \\ s_{nm}^{21} & \delta_{nm} + s_{nm}^{22} \end{bmatrix} \begin{bmatrix} a_m^{1\prime} \\ a_m^{2\prime} \end{bmatrix} \right\} \quad (9)$$

Note that the set of modes $\{\mathbf{j}_n^1(\mathbf{r}')\}$ and $\{\mathbf{j}_n^2(\mathbf{r}')\}$ do not constitute a complete set [4] on the domain of the exterior region, neither do the modes $\{\mathbf{m}_n^1(\mathbf{r}')\}$ and $\{\mathbf{m}_n^2(\mathbf{r}')\}$ in the apertures. If the currents on the real antenna for a specific excitation cannot be represented by the the given modes, then we naturally don't have a good representation and approximation of the antenna.

4 Numerical comparisons and comments

The radiation patterns of the antenna have been calculated for two cases of uniform excitation. In Figure 3 we have $\theta_0 = 30°, \varphi = 0°$ and in Figure 4 $\theta_0 = 30°, \varphi = 90°$. The dash dotted curves have been calculated using the simplified model (9) with one typical "short-circuited" element pattern used; the typical element pattern as well as the coupling coefficients have been calculated using HFSS. The other curves have been calculated using a rigorous HFSS model of the antenna. It can be seen that the agreement between the different cases is very good for the copolar patterns. The cross-polar patterns show no agreement in detail.

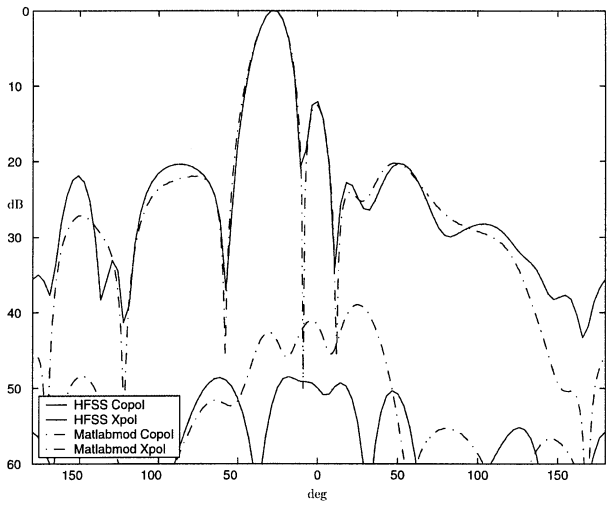


Figure 3: E-plane scan

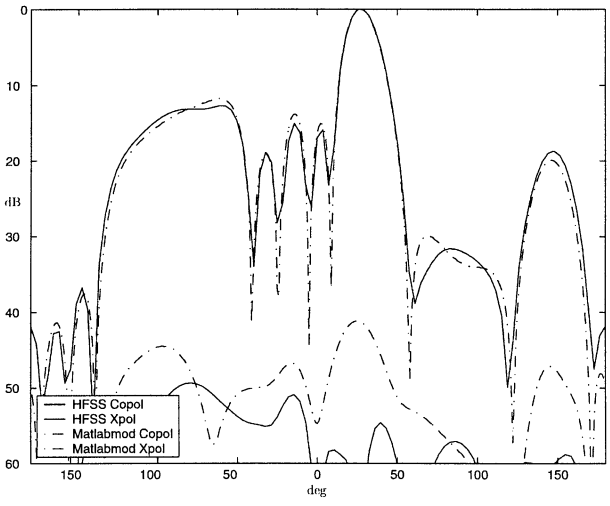


Figure 4: H-plane scan

References

[1] C. A. Balanis. *Advanced Engineering Electromagnetics.* Wiley, 1989.

[2] H. Steyskal, J. S. Herd. "Mutual Coupling Compensation in Small Array Antennas". *IEEE Transactions on Antennas and Propagation,* **38**(12), 1971–1975, Dec. 1990.

[3] A. C. Ludwig. "The Definition of Cross Polarisation". *IEEE Transactions on Antennas and Propagation,* **21**(1), 116 – 119, Jan. 1973.

[4] R. F. Harrington, J. R. Mautz. "Theory of Characteristic Modes for Conducting Bodies". *IEEE Transactions on Antennas and Propagation,* **19**(5), 622 – 628, Sept. 1971.

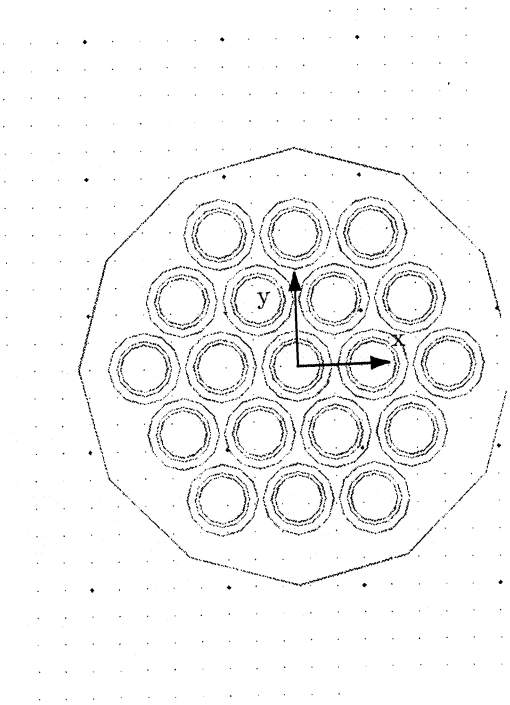


Figure 5: Antenna layout

Mutal coupling between dipoles on faceted conformal antenna arrays

Carina Marcus[1], Lars Pettersson[2],

[2] IFM, Linköpings Universitet, SE-581 83 Linköping, Sweden. E-mail: carma@ifm.liu.se

[2] IFM, Linköpings Universitet, SE-581 83 Linköping, Sweden. E-mail: larpe@ifm.liu.se.

Abstract

As a first part of an ongoing work to control the properties of faceted conformal antenna arrays, analysis of the coupling between small elements within the array and the element patterns have been made.

The magnetic infinitesimal dipole is given special attention, since it will be a part of building blocks for future realistic antenna designs including waveguide apertures.

The modeling of the antenna structure has so far been done using Uniform Theory of Diffraction, UTD. This metod is suitable because of its ability to give insight into the physics of the problem as well as fast results even for structures with many facets that otherwise would require extensive calculations.

1 Introduction

Most efforts on conformal antenna arrays have so far been focused on smoothly curved surfaces, e.g. [1]. Our aim here is to be able to analyze and control the properties of a non-planar array that consists of polygonal facets. The array will thus consist of facets which are used as building blocks to create a structure that approximates a given surface. This is a very practical approach when building an array from several flat subarray surfaces. In order to control the properties such as the antenna pattern of a faceted conformal antenna important parameters are the coupling between elements and the antenna patterns of the separate elements within the array.

The mutual coupling and perturbation of magnetic dipoles may be presented as admittances normalized to the conductance of a small magnetic dipole on an infinite groundplane,

$$G_{11} = \frac{4\pi}{3Z_0}\left(\frac{l}{\lambda}\right)^2 \tag{1}$$

The normalized admittance matrix of a two element magnetic dipole array can then be written as:

$$\mathrm{y} = \begin{bmatrix} 1+\Delta y_{11} & y_{12} \\ y_{21} & 1+\Delta y_{22} \end{bmatrix} \tag{2}$$

where Δy_{11} and Δy_{22} represent the change in self impedance due to the reflections introduced by the edges of the facets while y_{12} and y_{21} are the normalized mutual admittances. For a larger array the admittance matrix will increase in size but the elements above remain the same for small (single mode) dipoles.

For electric monopoles, an impedance matrix is more appropriate, using the resistance of an electric monpole over an infinite groundplane,

$$R_{11} = \frac{4\pi}{3}Z_0\left(\frac{h}{\lambda}\right)^2 \tag{3}$$

as normalization factor.

The element pattern is best presented as isolated patterns, when the other magnetic dipoles (electric monpoles) are short- (open-) circuited.

2 2D structures and fields

2.1 Reference case: FDTD in two dimensions

To test the accuracy of coupling calculations performed using UTD [2], FDTD (Finite Difference Time Domain) [3] caluclations in two dimensions have been used as a reference. Both source and target are infinite electric line sources that are perpendicular to the cross-section shown in fig 1. The structure is a right angle of perfect electric conductor for which FDTD is expected to give very accurate results. The parameters are

$$s_{in} = 0.31m \quad s_{out} = 0.36m$$
$$a = 18^o \qquad b = 37^o$$

The distance between the source and the edge is at the highest and lowest frequencies approximately 3λ and $\frac{\lambda}{4}$, respectively. As can be seen in fig 2, the agreement is very good, even when the distance between the source and edge is well below one wavelength. Clearly, UTD is a reliable method for calculations of coupling between small elements.

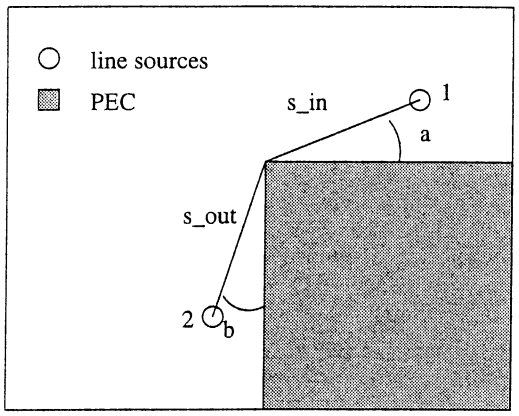


Figure 1: Configuration of the 2D simulation.

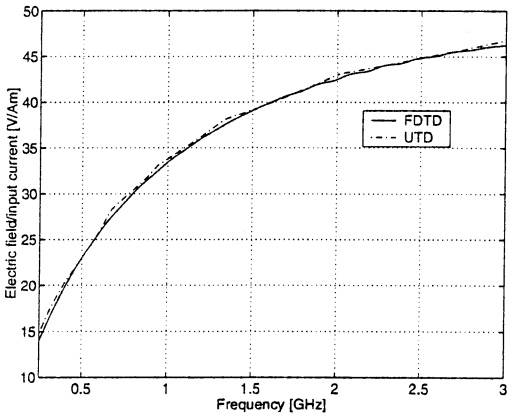


Figure 2: Results from 2D calculations.

3 2D structures with 3D fields

3.1 Coupling

Two small, identical magnetic dipoles are placed paralell to the edge of a 2D infinite wedge according to fig 3. One dipole is placed 3λ from the edge and the other one at the distance 2λ on the other face of the wedge. The magnitude of change in the self-admittance due to the presence of the edge decreases as the wedgeangle increases, fig 4. When the dipole is placed on an infinite groundplane, corresponding to a wedgeangle of 180^o, there is obviously no influence. On the other hand, the coupling between the elements is highest in that case and the UTD solution converges nicely to the theoretical value for two short dipoles on an infinite groundplane. The coupling between the element decreases rapidly as the angle decreases from 180^o. The angles between the facets in a realistic design are likely to deviate only moderately from 180^o, which indicates that the coupling between the elements will be an important factor during synthesis of an array antenna.

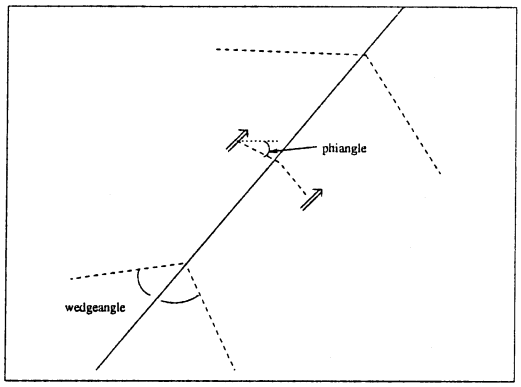


Figure 3: 2D wedge for 3D fields.

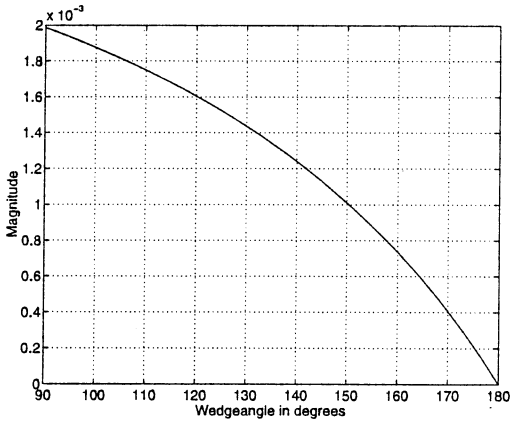


Figure 4: Magnitude of the normalized difference in self-impedance, $|\Delta y_{11}|$

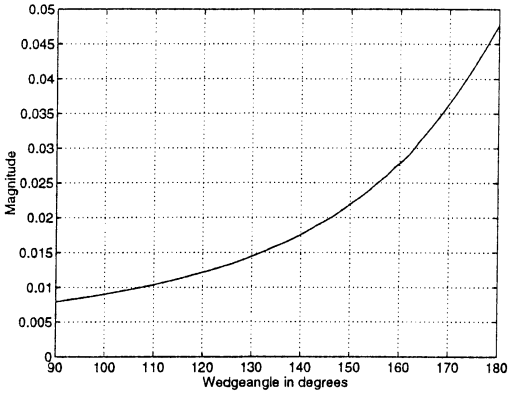


Figure 5: Magnitude of the normalized mutual admittance, $|y_{21}|$

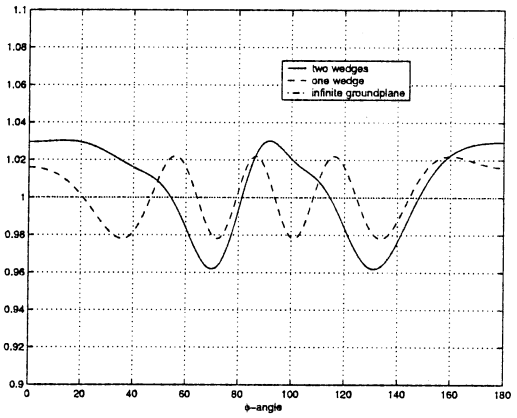


Figure 7: Antenna patterns: one and two edges.

3.2 Antenna pattern

The element pattern is important since it has direct bearing on the pattern of the entire array. Here, a single dipole is placed parallel to, and 3λ from the edge. The antenna patterns in the plane perpendicular to the dipole and to the wedge are shown in fig 6 for a few discrete wedgeangles. An undisturbed dipole produces the field with unit amplitude, and the smaller the wedgeangle the more influence on the behavior. The deviations from the pattern of the undisturbed dipole are here less than 0.5dB.

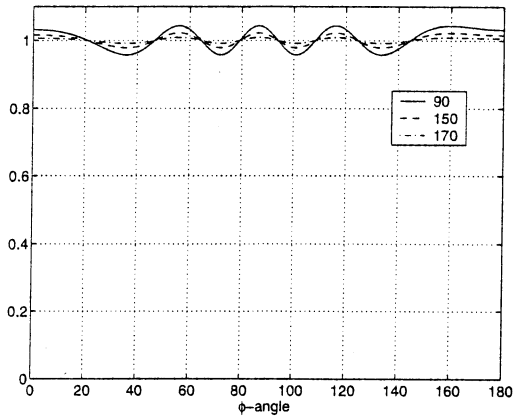


Figure 6: Antenna pattern: one edge.

In fig 7, the dipole is placed between two edges at distances 1λ and 2λ. The wedgeangles are 150^o. Here the increase in disturbance compared to the case with only one wedge is evident. This in an interesting result, since it is shows the result of a situation that will occur in a complete array, where the element pattern will be affected by all the edges on the facet where it is situated. Its importance will of course depend on the requirements put on the array pattern.

3.3 FDTD as a reference

The only way to get 3D field on an 2D structure in FDTD is to extend one of the the objects dimensions. In this case, a box with a quadratic cross-sectionwith the side of 0.4 m and length of 2.85 m is used. Two monopoles, each 0.01m long are placed according to fig 8 which shows a plane at half the length of the box. The distance between monpole and edge is 2λ at the higest and 0.5λ at the lowest frequencies shown.

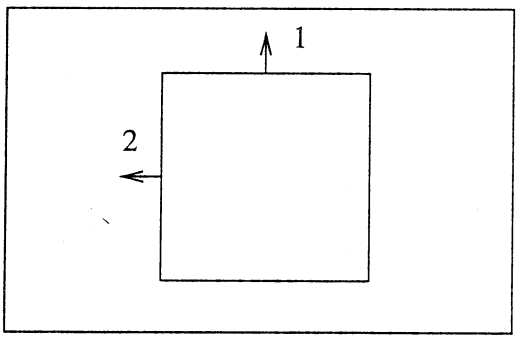


Figure 8: Cross-section of the 3D structure.

Since the result from a FDTD simulation is in time-domain, there is a possibility to distinguish between the contributions from the different parts of the geometry, provided that the distances between them are large enough. In this case, the length of the box allows filtering of the signal so that the portions transformed into frequency domain only contains contribution from the long edges which in some sense makes them infinite.

The contributions to the field at monopole 2 caused by monopole 1 are, apart from the obvious shortest

path and the one around the back of the structure, composed by a number of reflections between and diffractions via the edges. Each diffraction reduces the amplitude, and obviously, propagation over the surfaces does the very same thing. To obtain the result UTD shown in fig 9, where the magnitude of the mutual impedance is normalized to R_{11} in equation 3, diffractions via three edges were included. Except for the larger deviation around 1GHz, the results are very good. Since diffractions via four edges and more is omitted from the UTD calculations, while the signal from the FDTD simulation contains componets that have been diffracted more than three times. This might be the reason why the results differ.

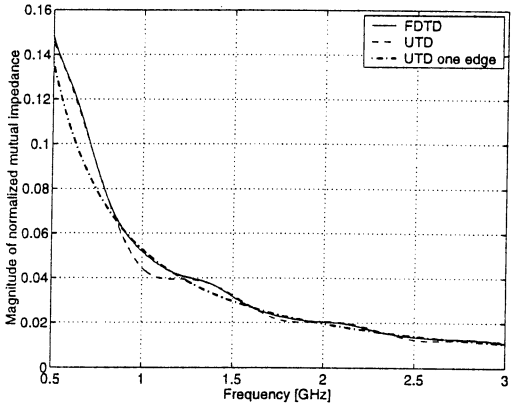


Figure 9: Results from the FDTD simulation compared to a UTD calculations including one single and up to three diffractions.

4 Future

The next step towards a realistic antenna is implementation of coupling calculations between apertures. The small magnetic dipole can be used to model the aperture distribution. The coupling is then calculated via integration in two or four dimensions, as for smooth surfaces [1]. When dealing with apertures, not only the coupling between the apertures, but also the coupling between the different modes that can be found in the aperture.

5 Conclusion

Tests with coupling calculations have shown that UTD is a suitable technique for modeling the properties of faceted conformal antenna arrays.

References

[1] Patrik Persson. *Analysis and Design of Conformal Array Antennas*, PhD thesis, Royal Institute of Technology, Division of Electromagnetic Theory, Stockholm, Sweden, 2001.

[2] McNamara, Pistorius, Malherbe. *Introduction to The Uniform Geometrical Theory of Diffration* Artech House, 1990.

[3] Torleif Martin. *Broadband Electromagnetic Scattering and Sheilding Analysis using the Finite Differece Time Domain Method*, Department of Physics and Measurement Technology, Linköpings universitet, Linköping, Sweden, 2001

On the Capacity of Polarization Diversity Antennas in Satellite Communications

Sven Nordebo[1], Abbas Mohammed[2] and Jonas Lundbäck[3]

[1]Växjö University, 351 95 Växjö, Sweden. E-mail: sven.nordebo@msi.vxu.se

[2]Blekinge Institute of Technology, 372 25 Ronneby, Sweden. E-mail: abbas.mohammed@its.bth.se

[3]Växjö University, 351 95 Växjö, Sweden. E-mail: jonas.lundback@msi.vxu.se

Abstract

This paper addresses the potential gain of using multiple antennas and MIMO–OFDM techniques in order to increase the bandwidth efficiency in satellite communication systems. In particular, we consider the increase in channel capacity that is possible by exploiting satellite and polarization diversity. A fundamental case is studied with three satellite branches and where each transmit/receive antenna unit consists of six elemental electric and magnetic dipoles yielding six distinguishable parallel polarization channels.

1 Introduction

It has been widely recognized that the capacity in wireless communication systems can be greatly increased by exploiting environments with rich scattering such as urban areas or indoors [1]-[4]. Independent spatial or polarization channels can be accessed by means of multiple antennas at both the transmitter and the receiver and the technique is thus referred to as Multiple-Input Multiple-Output (MIMO) communications. For a fixed total power and bandwidth, and with a matrix transfer function of independent complex Gaussian random variables, the MIMO wireless communication channel has an information theoretic capacity that (initially) grows linearly with the number of antenna elements [1].

Most often the increase in capacity is attributed to spatial channels and scalar properties of the propagating waves. Classically, it is well understood that the electromagnetic polarization of plane waves possesses two independent channels, or polarization states. However, in [2] it has been shown that the existence of multipath propagation can increase the number of polarization channels to six distinguishable electric and magnetic states of polarization and thus effectively increase the channel capacity of a multiple antenna system by a factor of three.

In this paper we consider the synthetic multiscattering environment created by employing satellite diversity [5]-[6]. The aim is to study the theoretical potential of exploiting polarization diversity to achieve MIMO capacity in satellite communication systems. A fundamental case is studied with three satellite branches and where each transmit/receive antenna unit consists of six elemental electric and magnetic dipoles yielding six distinguishable parallel polarization channels.

We consider the MIMO-OFDM technique [3], [7], [8] as a means of achieving independent channels over polarization space as well as over frequency. In essence, the MIMO technique achieves orthogonality over space by employing the Singular Value Decomposition (SVD), see [3], and the OFDM technique achieves orthogonality over frequency by employing the Discrete Fourier Transform (DFT) in conjunction with a Cyclic Prefix (CP), see [7].

Satellite communication systems today operate in a variety of frequency bands, ranging from 100 MHz to 100 GHz [6]. Low Earth Orbit (LEO) satellite systems like ORBCOMM, E-SAT and LEO ONE [6] operate in the frequency range 100 to 400 MHz and are examples of low bit–rate systems that can benefit significantly from the potential capacity of a MIMO satellite diversity system.

2 The Polarization Channel

Consider the satellite communication system depicted in Fig. 1. The system consists of a transmitter, a receiver, and three satellite branches (a, b and c). Each transmit/receive antenna unit consists of six elemental electric and magnetic dipoles. All vectors below are defined in the same cartesian coordinate system as depicted in Fig. 1.

The electromagnetic far–fields generated by elemental electric and magnetic dipoles are given by

$$\begin{pmatrix} \mathbf{E} \\ \eta\mathbf{H} \end{pmatrix} = \frac{\eta(j\omega)^2}{4\pi c^2}\frac{e^{-j\omega\frac{r}{c}}}{r}\begin{pmatrix} \hat{\mathbf{r}}\times\hat{\mathbf{r}}\times(c\mathbf{p})+\hat{\mathbf{r}}\times\mathbf{m} \\ -\hat{\mathbf{r}}\times(c\mathbf{p})+\hat{\mathbf{r}}\times\hat{\mathbf{r}}\times\mathbf{m} \end{pmatrix} \quad (1)$$

in standard vector notation. Here $\mathbf{E}$ and $\mathbf{H}$ are the

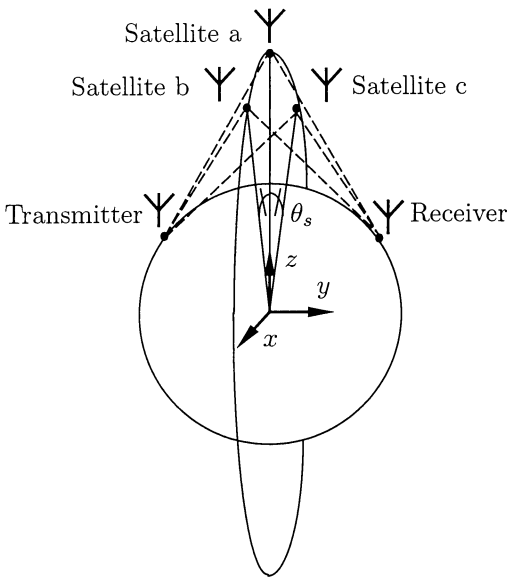


Figure 1: The fundamental satellite communication system with three satellites (a, b and c).

electric and magnetic vector fields in $[V/m]$ and $[A/m]$, respectively, and $\eta = 377$ $[\Omega]$ is the wave impedance in free space, $c = 3 \cdot 10^8$ $[m/s]$ is the speed of light, $\omega = 2\pi f$ is the angular frequency where f is the frequency in $[Hz]$, r is the radial distance in $[m]$, $\hat{\mathbf{r}}$ is the directional unit vector, $\mathbf{p}$ is the electric moment vector in $[Ams]$, and $\mathbf{m}$ is the magnetic moment vector in $[Am^2]$.

Let

$$j\omega\mathbf{p} = \mathbf{I}_e dl \quad (2)$$

$$\mathbf{m} = \mathbf{I}_m dS \quad (3)$$

where $\mathbf{I}_e$ and $\mathbf{I}_m$ are the three–dimensional current vectors driving the electric and magnetic dipoles with lengths dl and surfaces dS, respectively.

By adopting matrix notation, and writing the vector cross product operator $\hat{\mathbf{r}}\times = (r_x\hat{\mathbf{x}} + r_y\hat{\mathbf{y}} + r_z\hat{\mathbf{z}})\times$ in Cartesian coordinates as

$$\mathbf{J}(\hat{\mathbf{r}}) = \begin{pmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{pmatrix} \quad (4)$$

we may write the fields $\mathbf{E}$ and $\eta\mathbf{H}$ in terms of the driving currents $\mathbf{I}_e$ and $\mathbf{I}_m$ as

$$\begin{pmatrix} \mathbf{E} \\ \eta\mathbf{H} \end{pmatrix} = \mathbf{A}(\mathbf{r},\omega) \begin{pmatrix} \mathbf{I}_e \\ \mathbf{I}_m \end{pmatrix} \quad (5)$$

where

$$\mathbf{A}(\mathbf{r},\omega) = \frac{\eta \cdot j\omega}{4\pi c} \frac{e^{-j\omega\frac{r}{c}}}{r} \begin{pmatrix} \mathbf{J}^2(\hat{\mathbf{r}})dl & \frac{j\omega}{c}\mathbf{J}(\hat{\mathbf{r}})dS \\ -\mathbf{J}(\hat{\mathbf{r}})dl & \frac{j\omega}{c}\mathbf{J}^2(\hat{\mathbf{r}})dS \end{pmatrix} \quad (6)$$

and the position vector is $\mathbf{r} = r\hat{\mathbf{r}}$.

In order to balance the first and second block–columns above, we choose $dl = \frac{\omega_c}{c} dS$ where ω_c is the angular frequency of the carrier signal.

The total channel consists of three paths, $\mathbf{r}_1$–$\mathbf{r}_4$ (Tx→Sat. a→Rx), $\mathbf{r}_2$–$\mathbf{r}_5$ (Tx→Sat. b→Rx) and $\mathbf{r}_3$–$\mathbf{r}_6$ (Tx→Sat. c→Rx), as in Fig. 1. The total channel is therefore given by

$$\begin{pmatrix} \mathbf{E} \\ \eta\mathbf{H} \end{pmatrix} = \mathbf{H}(\theta_s,\omega) \begin{pmatrix} \mathbf{I}_e \\ \mathbf{I}_m \end{pmatrix} \quad (7)$$

where

$$\mathbf{H}(\theta_s,\omega) = \mathbf{A}(\mathbf{r}_4,\omega) \cdot \mathbf{A}(\mathbf{r}_1,\omega) + \\ + \mathbf{A}(\mathbf{r}_5,\omega) \cdot \mathbf{A}(\mathbf{r}_2,\omega) + \mathbf{A}(\mathbf{r}_6,\omega) \cdot \mathbf{A}(\mathbf{r}_3,\omega) \quad (8)$$

In our channel model, we assume that Satellite a is fixed in the direction (0,0,1) and Satellites b and c lie on a circle in the $x - z$ plane with an angle θ_s towards Satellite a, as depicted in Fig. 1.

We note that the matrix $\mathbf{A}(\mathbf{r},\omega)$ above is of rank 2, see [2], whereas the matrix $\mathbf{H}(\theta_s,\omega)$ is of rank 6 as soon as $\theta_s \neq 0$. Hence, one satellite branch corresponds to an ordinary plane wave and can carry only two independent polarization channels whereas the total channel can carry six distinguishable and independent polarization channels.

As a comparison case for the numerical evaluation we choose the path $\mathbf{r}_1$–$\mathbf{r}_4$ (Tx→Sat. a→Rx) with

$$\mathbf{H}(\omega) = \mathbf{A}(\mathbf{r}_4,\omega) \cdot \mathbf{A}(\mathbf{r}_1,\omega) \quad (9)$$

which is of rank 2.

3 The MIMO-OFDM System

The basic discrete–time, complex, baseband MIMO–OFDM system is depicted in Fig. 2. We assume a multiple antenna system with N transmitting and M receiving antennas. Here $s_n(t)$ and $r_m(t)$ denote the transmitted and received antenna signals, respectively, and $h_{mn}(t)$ denotes the impulse response connecting the transmitting antenna n to the receiving antenna m. Let the total delay spread be denoted τ_d.

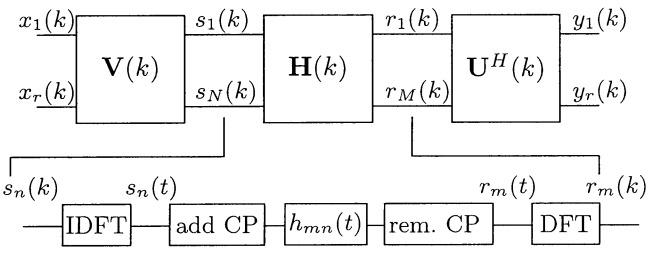


Figure 2: The baseband MIMO–OFDM system.

Given that the cyclic prefix $CP > \tau_d \cdot f_s$ where $f_s = 1/T_s$ and T_s is the symbol time, linear convolution becomes identical with circular convolution over the entire DFT–frame, and we have

$$r_m(t) = \sum_{n=1}^{N} h_{mn}(t) \otimes s_n(t) + v_m(t), \qquad (10)$$

for $t \in [0, \frac{N_{DFT}-1}{f_s}]$ where N_{DFT} is the size of the DFT and $m = 1, \ldots, M$. Here $v_m(t)$ is assumed to be uncorrelated Gaussian noise.

Taking the DFT of (10), we obtain

$$r_m(k) = \sum_{n=1}^{N} h_{mn}(k) s_n(k) + v_m(k), \qquad (11)$$

for $k \in [0, N_{DFT} - 1]$, which can be written in matrix form as

$$\mathbf{r}(k) = \mathbf{H}(k)\mathbf{s}(k) + \mathbf{v}(k), \qquad (12)$$

where $\mathbf{r}(k) = (r_m(k))$ is $M \times 1$, $\mathbf{s}(k) = (s_n(k))$ is $N \times 1$, $\mathbf{v}(k) = (v_m(k))$ is $M \times 1$, and $\mathbf{H}(k) = (h_{mn}(k))$ is $M \times N$. The correlation matrix of the noise vector $\mathbf{v}(k)$ is given by $E\{\mathbf{v}(k)\mathbf{v}^H(k)\} = \sigma_n^2 \cdot \mathbf{I}_M$ where σ_n^2 is the variance of the noise and $\mathbf{I}_M$ is the $M \times M$ identity matrix.

The Singular Value Decomposition (SVD) of the matrix $\mathbf{H}(k)$ is given by

$$\mathbf{H}(k) = \mathbf{U}(k)\mathbf{\Sigma}(k)\mathbf{V}^H(k) \qquad (13)$$

where $\mathbf{U}(k)$ is $M \times M$, $\mathbf{V}(k)$ is $N \times N$, $\mathbf{\Sigma}(k)$ is $M \times N$ containing r singular values greater than zero $\sigma_1(k) \geq \cdots \geq \sigma_r(k) > 0$ and r is the rank of the matrix $\mathbf{H}(k)$.

By defining

$$\mathbf{y}(k) = \mathbf{U}^H(k)\mathbf{r}(k) \qquad (14)$$
$$\mathbf{s}(k) = \mathbf{V}(k)\mathbf{x}(k) \qquad (15)$$
$$\mathbf{n}(k) = \mathbf{U}^H(k)\mathbf{v}(k) \qquad (16)$$

we obtain the diagonalized system

$$\mathbf{y}(k) = \mathbf{\Sigma}(k)\mathbf{x}(k) + \mathbf{n}(k) \qquad (17)$$

or

$$\begin{cases} y_1(k) = \sigma_1(k)x_1(k) + n_1(k) \\ \quad\vdots \\ y_r(k) = \sigma_r(k)x_r(k) + n_r(k). \end{cases} \qquad (18)$$

The correlation matrix of the noise vector $\mathbf{n}(k)$ is given by $E\{\mathbf{n}(k)\mathbf{n}^H(k)\} = \sigma_n^2 \cdot \mathbf{I}_M$ since $\mathbf{U}(k)$ is orthonormal.

Since the communication channels in (18) are completely decoupled and the noise terms $n_m(k)$ are uncorrelated over space as well as over frequency, the maximum information capacity can be calculated as

$$C = \sum_{k=0}^{N_{DFT}-1} \sum_{m=1}^{r} \log_2(1 + \sigma_{x_m}^2(k) \cdot \frac{\sigma_m^2(k)}{\sigma_n^2}) \qquad (19)$$

where $\sigma_{x_m}^2(k)$ denotes the variance of the uncorrelated input signals $x_m(k)$. The capacity is constrained by the total transmitted power

$$E = \sum_{k=0}^{N_{DFT}-1} \sum_{m=1}^{r} \sigma_{x_m}^2(k). \qquad (20)$$

The optimum water–filling solution is easily obtained as follows: Given a testing level B. For $k = 0, \ldots, N_{DFT} - 1$ and $m = 1, \ldots, r$, choose

- $\sigma_{x_m}^2(k) = B - \sigma_n^2/\sigma_m^2(k)$ if $B - \sigma_n^2/\sigma_m^2(k) > 0$
- $\sigma_{x_m}^2(k) = 0$ if $B - \sigma_n^2/\sigma_m^2(k) \leq 0$.

Then calculate the corresponding transmitted power level E according to (20). The capacity given by (19) is now the optimum water–filling solution for this particular power level E.

In our numerical examples we plot the capacity versus the average signal–to–noise ratio at the receiver, SNR_{av}, defined as

$$SNR_{av} = \frac{1}{N_{act}} \sum_{k=0}^{N_{DFT}-1} \sum_{m=1}^{r} \sigma_{x_m}^2(k) \cdot \frac{\sigma_m^2(k)}{\sigma_n^2} \qquad (21)$$

where N_{act} is the number of active channels for which $\sigma_{x_m}^2(k) \neq 0$.

For large signal–to–noise ratios we see that the capacity in (19) can be approximated by

$$C \sim N_{act} \log_2(SNR_{av}). \qquad (22)$$

4 Numerical Examples

The channel matrix $\mathbf{H}(k)$ in (12) is given by

$$\mathbf{H}(k) = \mathbf{H}(\theta_s, \omega_k) \qquad (23)$$

where $\mathbf{H}(\theta_s, \omega)$ is given in (8) and $\omega_k = 2\pi f_k$, $f_c - f_s/2 \leq f_k \leq f_c + f_s/2$ and $k = 0, \ldots, N_{DFT} - 1$. Here f_c is the carrier frequency, $f_s = 1/T_s$ is the baseband sampling frequency and T_s is the symbol time. In the examples below we have chosen $f_c = 150$ MHz, $f_s = 1$ MHz and $N_{DFT} = 128$.

Fig. 3 shows the singular value distribution $1/\sigma_m^2(k)$ for the channel $\mathbf{H}(\theta_s, \omega)$ in (8) with $\theta_s = 10$ degrees, as well as for the comparison case $\mathbf{H}(\omega)$ in (9). Fig. 4 shows the number of active channels N_{act} versus the average signal–to–noise ratio at the receiver SNR_{av} for the same cases. Fig. 5 shows the corresponding channel capacity in bits/transmission. The quantities in Figs. 4–5 are calculated according to the optimum water–filling principle and plotted versus SNR_{av} using the total transmitted power E as parameter.

Figs. 6–11 show similarly the cases with $\theta_s = 20$ and $\theta_s = 30$ degrees. The dotted lines show the comparison case.

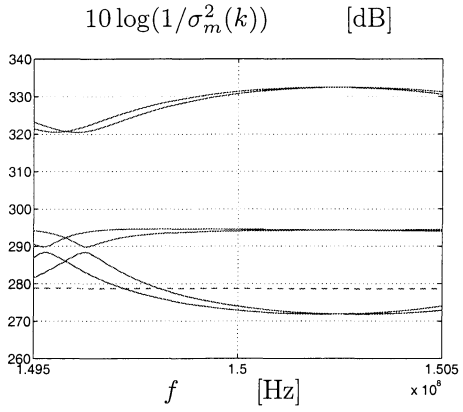


Figure 3: Singular value distribution $1/\sigma_m^2(k)$ versus frequency. Solid lines: Channel case $\theta_s = 10$ degrees.

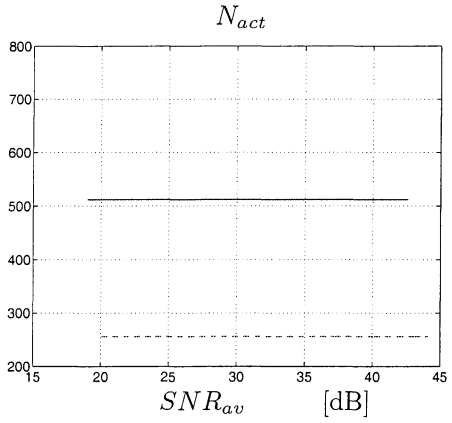


Figure 4: Number of active channels N_{act} versus SNR_{av}. Solid line: Channel case $\theta_s = 10$ degrees.

5 Summary and Conclusion

We have studied the potential gain of using multiple antennas and MIMO–OFDM techniques in order to exploit polarization diversity to increase the bandwidth efficiency in satellite communication systems. A fundamental case is studied with three satellite branches and where each transmit/receive antenna unit consists of six elemental electric and magnetic dipoles yielding six distinguishable parallel polarization channels. We have demonstrated by numerical examples that capacity increases linearly on a logarithmic signal–to–noise ratio scale where the constant of proportionality is the number of parallel channels.

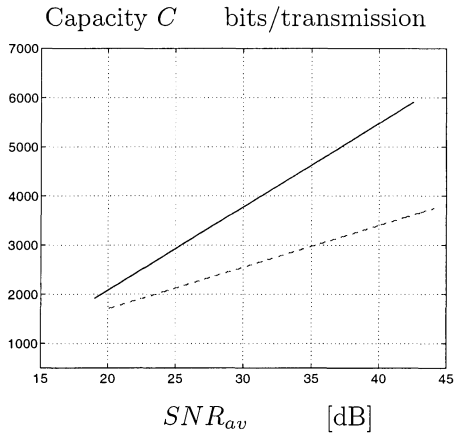


Figure 5: Capacity C versus SNR_{av}. Solid line: Channel case $\theta_s = 10$ degrees.

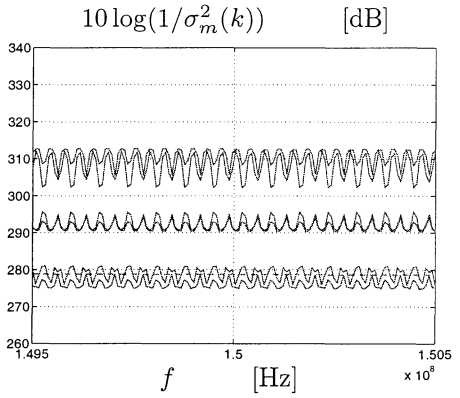


Figure 6: Singular value distribution $1/\sigma_m^2(k)$ versus frequency. Solid lines: Channel case $\theta_s = 20$ degrees.

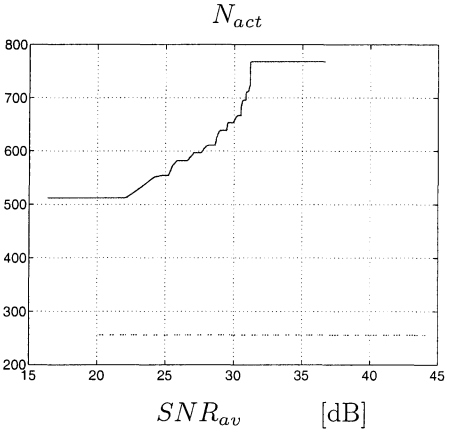


Figure 7: Number of active channels N_{act} versus SNR_{av}. Solid line: Channel case $\theta_s = 20$ degrees.

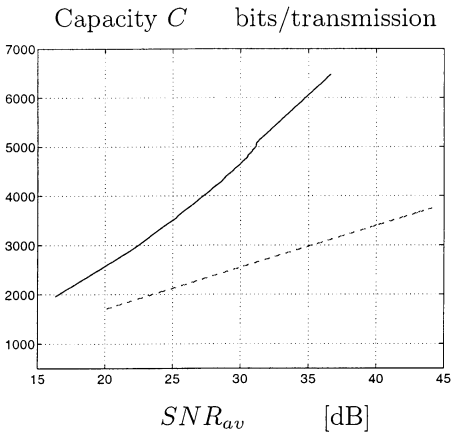


Figure 8: Capacity C versus SNR_{av}. Solid line: Channel case $\theta_s = 20$ degrees.

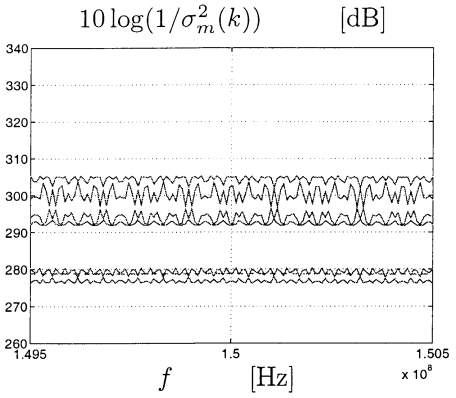


Figure 9: Singular value distribution $1/\sigma_m^2(k)$ versus frequency. Solid lines: Channel case $\theta_s = 30$ degrees.

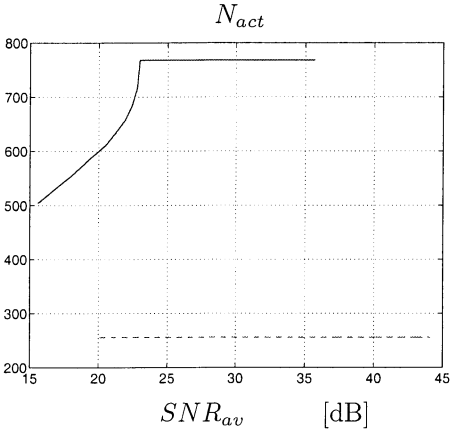


Figure 10: Number of active channels N_{act} versus SNR_{av}. Solid line: Channel case $\theta_s = 30$ degrees.

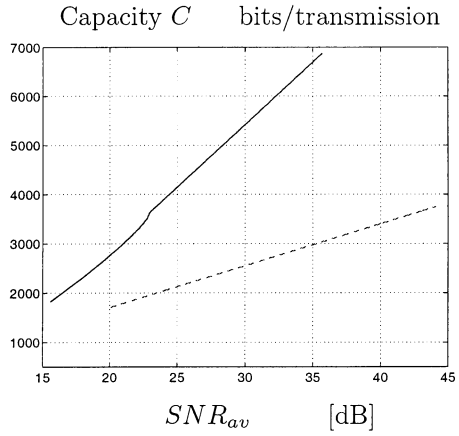


Figure 11: Capacity C versus SNR_{av}. Solid line: Channel case $\theta_s = 30$ degrees.

References

[1] G. J. Foschini, M. J. Gans,"On Limits of Wireless Communications in a Fading Environment when using Multiple Antennas", Wireless Personal Communications 6: 311-335, 1998.

[2] M. R. Andrews, P. P. Mitra, R. deCarvalho, "Tripling the capacity of wireless communications using electromagnetic polarization", Nature, vol. 409, pp. 316-318, January 2001.

[3] G. G. Raleigh, J. M. Cioffi, "Spatio-Temporal Coding for Wireless Communication", IEEE Transactions on Communications, vol. 46, no. 3, pp. 357-366, March 1998.

[4] A. F. Naguib, N. Seshadri, A. R. Calderbank, "Increasing Data Rate over Wireless Channels", IEEE Signal Processing Magazine, pp. 76-92, May 2000.

[5] R. Tesi, L. Mucchi, D. Tujkovic, E. Kunnari, "Transmit Diversity for Multi-Satellite UMTS", International Symposium on 3rd Generation Infrastructure and Services 3GIS '01, Athens, Greece, July 2-3, 2001.

[6] R. E. Sheriff, Y. F. Hu, Mobile Satellite Communications Networks, John Wiley & Sons, 2001.

[7] J. A. C. Bingham, "Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come", IEEE Communications Magazine, vol. 28, pp. 5-14, May 1990.

[8] R. S. Blum, Y. G. Li, J. H. Winters, Q. Yan, "Improved Space-Time Coding for MIMO-OFDM Wireless Communications", IEEE Transactions on Communications, vol. 49, no. 11, pp. 1873-1878, November 2001.

Patch antennas on inhomogeneous substrates

Thomas Rylander[1], Peter Waller[2], Gerhard Kristensson[3] and Anders Derneryd[4]

[1] Dept. of Electromagnetics, Chalmers, SE-41296 Göteborg. E-mail: rylander@elmagn.chalmers.se

[2] Dept. of Electroscience, Lund Univ., P.O. Box 118, SE-22100 Lund. E-mail: peter.waller@es.lth.se

[3] Dept. of Electroscience, Lund Univ., P.O. Box 118, SE-22100 Lund. E-mail: gerhard.kristensson@es.lth.se

[4] Ericsson AB, SE-43184 Mölndal. E-mail: anders.derneryd@ericsson.com

Abstract

Patch antennas on inhomogeneous substrates are analyzed with a stable FEM-FDTD hybrid method and an analytic propagator technique. For inhomogeneous substrates with a *fixed* homogenized value of the permittivity, our results indicate that a decrease of the permittivity in the vicinity of the patch increases the bandwidth and the resonance frequency. Furthermore, we find that inhomogeneous substrates with the permittivity only varying with the height influence the field pattern within the substrate, but not the radiation efficiency.

1 Introduction

Patch antennas are attractive in many applications since they are low-profile, cheap, mechanically robust and simple to manufacture. However, the traditional patch antenna suffers from, *e.g.* low efficiency and narrow frequency bandwidth. Thus, improved patch antenna designs have been pursued for many years. One parameter that can be optimized is the substrate [1]. This can be done by designing different regions with constant permittivity [2]. Recently, Vegni *et. al* [3, 4] have proposed an inhomogeneous substrate with the permittivity continuously varying with the height coordinate. Improved bandwidth and directivity were reported. This problem has also been investigated in [5] under the assumption of a fixed current distribution on the patch. It was shown that the permittivity profile has much influence on the field pattern inside the substrate but only a slight influence on the radiation pattern and efficiency.

In the present work, we optimize such antennas for large bandwidth operation using numerical computations and compare with recent analytical results. For the numerical computations, we use the stable hybrid method [6] which combines the Finite-Difference Time-Domain (FDTD) scheme and the Finite Element Method (FEM). It can handle complex geometry without resorting to the staircase approximation and the small structures of the coaxial cable feed are discretized by FEM. The rest of the computational domain, including the major part of the inhomogeneous substrate, is treated by the FDTD scheme. The time-step is set by the Courant condition of the FDTD and it is not limited by the small cells needed to resolve the geometry of the feed. We also use an analytical technique [5] that assumes a given current distribution on the patch and an infinite substrate with a permittivity depending only on the height from the ground plane.

2 Methods for analysis

2.1 Analytical model

The analytical model, which treats the time harmonic case, is carefully described in [5]. The model is based on the assumption that we have an infinite substrate and an infinite ground plane. Thus, it is possible to Fourier transform the fields in the lateral directions (the x- and y-directions). The computation of the far field leads to a system of ordinary differential equations in the height coordinate (z-coordinate), see Fig. 1. This approach can treat bianisotropic materials [7] and can also be extended to a MoM approach, *e.g.*, [3, 8].

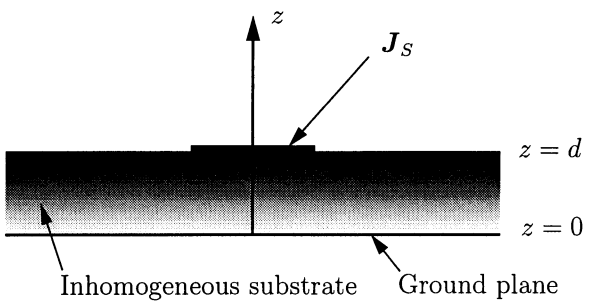


Figure 1: Geometry of the simulated structure, $\boldsymbol{J}_S$ is the given surface current density.

A circular patch is modeled using a given current placed above the substrate, corresponding to the

dominant mode of the cavity model. The resonance frequency can be calculated according to the cavity model (without taking the fringing fields into account [9]), *i.e.*

$$f_r = \frac{1.841 c_0}{2\pi a \sqrt{\epsilon^{hom}}}$$

where a is the radius of the circular patch. The relative permittivity ϵ^{hom} is the homogenized value of the relative permittivity function $\epsilon(z)$, and it is given by

$$\frac{d}{\epsilon^{hom}} = \int_0^d \frac{1}{\epsilon(z)} \mathrm{d}z, \tag{1}$$

where d is the thickness of the substrate. Here, $z = 0$ at the ground plane and $z = d$ at the air-dielectric interface. The electromagnetic fields inside and outside the substrate are solved using propagators [5] with the boundary conditions given by the Perfect Electric Conductor (PEC) ground plane and the radiation condition. We use a wave splitting technique to impose the radiation boundary condition.

2.2 Stable FEM-FDTD hybrid method

A patch antenna on an infinite substrate and an infinite ground plane can also be analyzed by the stable FEM-FDTD hybrid method [6]. Furthermore, the hybrid method can treat finite substrate and/or finite ground plane. Here, we feed the antenna by a coaxial cable with the relative permittivity $\epsilon = 1.86$, inner radius 0.48 mm and outer radius 1.5 mm. The FEM is used in the vicinity of the feed and parts of a typical grid are shown in Fig. 2.

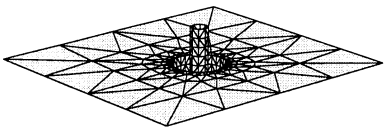


Figure 2: Finite element discretization of the coaxial cable feed.

The geometry for our test case involving finite substrate and finite ground plane is shown Fig. 3, where the dimensions of the patch are kept fixed when the substrate and the ground plane are extended to infinity.

For computations, the antenna is enclosed in an appropriate metal box. The walls of the box are covered by a so called sponge layer [10] which absorbs outward propagating waves. The antenna is driven by the TEM-mode launched from a port terminating

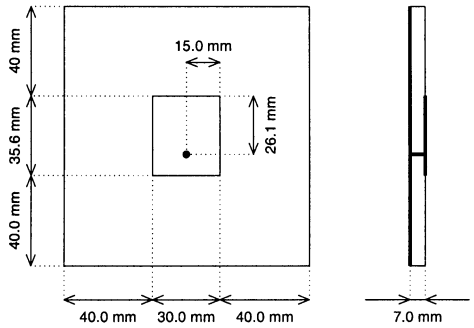


Figure 3: The patch antenna on a finite rectangular ground plane. The view from the above and the side are shown to the left and right, respectively, and the position of the feed is shown by the dot in the view from above.

the coaxial cable. Here, the launched wave has the time dependence

$$E_i(t) = E_0 \exp\left[-\left(\frac{t - t_0}{d_0}\right)^2 \right] \sin(2\pi f_0 (t - t_0)), \tag{2}$$

where $f_0 = 2$ GHz, $d_0 = 2/(3f_0)$ and $t_0 = 3/f_0$. The reflected wave $E_r(t)$ is recorded at the port and the reflection coefficient is $S_{11}(f) = E_r(f)/E_i(f)$. The radiation pattern can be calculated by means of the near-to-far-field transformation [11].

In order to treat inhomogeneous materials, the hybrid code has been extended with numerical integration [12–14] for the finite elements. We use the PETSc [15] for the solution of the linear system of equations associated with the FEM part of the computations.

3 Results

In this section, we present results on the bandwidth and radiation efficiency for patch antennas on infinite substrates with different spatial variations of the permittivity. The dispersion curves of surface waves for such substrates are also shown. Furthermore, we investigate the influence of finite inhomogeneous substrates on the bandwidth.

Here, we refer to bandwidth as the half-power bandwidth $\Delta f/f_0$ for the total radiated power $P_t = 1 - |S_{11}|^2$, where f_0 is the resonant frequency. Furthermore, $\Delta f = f_2 - f_1$ corresponds to the frequency interval $[f_1, f_2]$ where P_t is at least half of the maximum power at f_0. It should be emphasized that the results presented for the hybrid method are computed on one finite grid size and, consequently, the value of *e.g.* S_{11}

is subject to some discretization errors. However, the main objective at this point is to find the parameters which have major influence on the bandwidth and it should be safe to compare the FEM-FDTD computations with each other to identify such trends.

3.1 Infinite substrate

In the case of an infinite substrate and an infinite ground plane, we let the permittivity be a function of the height above the ground plane using the profiles shown in Tab. 1. Here, the constant relative permittivity $\epsilon = 8/3$ corresponds to the homogenized value of the increasing and decreasing hyperbolic profiles according to Eq. (1).

Permittivity profile	$\epsilon(z)$
Constant	$8/3$
Increasing hyperbolic	$8/(4-2z/d)$
Decreasing hyperbolic	$8/(2+2z/d)$

Table 1: Permittivity profiles.

The hybrid method was used to compute the reflection coefficient S_{11} for the rectangular patch shown in Fig. 3. The Padé approximation (of the computed S_{11}) is shown in Fig. 4 by the solid, dashed and dash-dotted curves for the constant, increasing and decreasing permittivity profile, respectively. The computed values are shown by circles. We found that the half-power-bandwidth for P_t was 14.4% for the constant profile, 13.8% for the increasing hyperbolic profile and 15.2% for the decreasing hyperbolic profile.

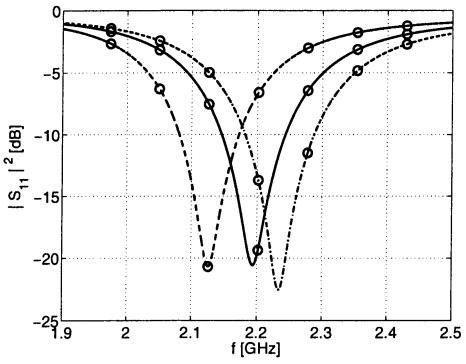


Figure 4: The reflection coefficient S_{11} for the three profiles shown in Tab. 1 computed by the hybrid method: constant profile – solid curve, increasing hyperbolic profile – dashed curve, and decreasing hyperbolic profile – dash-dotted curve.

However, some of the power P_t will be bound to surface waves, i.e. $P_t = P_r + P_{sw}$ where P_r is the radiated power associated with the space wave and P_{sw} is the corresponding power for the surface wave. We have used the analytical model outlined in Sec. 2.1 to investigate such effects for circular patch antennas on a number of different inhomogeneous substrates [5], including those in Tab.1. We found that the ratio P_r/P_{sw} increases with lower values of the homogenized permittivity and thinner substrates. Furthermore, profiles with an increasing permittivity profile gave somewhat higher P_r/P_{sw} compared to the corresponding decreasing profiles.

Several dispersion curves of the surface waves for inhomogeneous substrates were calculated using the analytical model and they are shown in Fig. 5 for the constant (lower figure) and increasing hyperbolic (upper figure) permittivity profile. We note that the dispersion curves cross each other for an increasing profile. For the constant and decreasing permittivity profiles the dispersion curves do not cross one another.

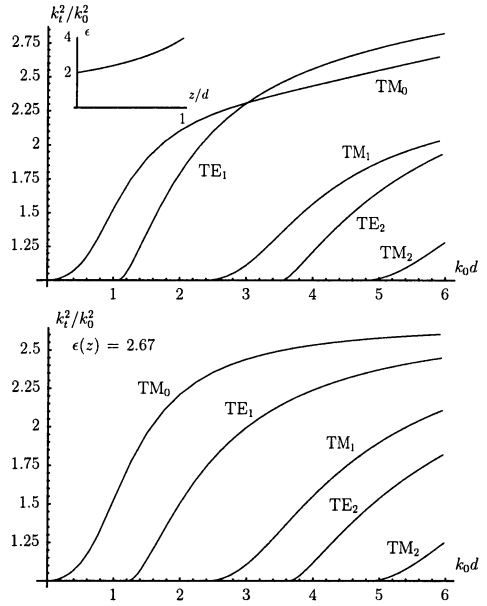


Figure 5: The dispersion curves for the increasing hyperbolic and constant permittivity profiles are shown in the upper and lower plot, respectively. The transverse wave number is denoted k_t, d is the thickness of the substrate, and k_0 is the free space wave number.

Figure 6 shows the radiation efficiency, defined as $P_r/(P_r + P_{sw})$, by the solid and dashed curves for a circular patch on a substrate with decreasing hyperbolic and constant permittivity profiles, respectively.

3.2 Finite substrate

For the finite geometry shown in Fig. 3, we allow the relative permittivity ϵ of the slab to be a function of

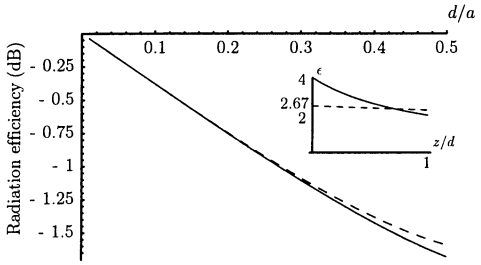


Figure 6: The solid and dashed line shows the radiation efficiency for the decreasing hyperbolic profile and the homogenized profile, respectively.

x, y and z given by

$$\epsilon = \epsilon_c + \epsilon_\Delta \left[\left(\frac{x - x_0}{\Delta x} \right)^2 + \left(\frac{y - y_0}{\Delta y} \right)^2 + \left(\frac{z - z_0}{\Delta z} \right)^2 \right] \tag{3}$$

where we fix $\epsilon_c = \epsilon_\Delta = 2$. Furthermore, we let x_0 and y_0 correspond to the midpoint of the patch. We chose Δx to take the value $l_x/2, l_x$ or ∞ and Δy to take the value $l_y/2$, l_y or ∞. Here, $l_x = 30.0$ mm is the width of the patch, *i.e.* perpendicular to the current on the patch, and $l_y = 35.6$ mm is the length of the patch. The parameters z_0 and Δz are given in Tab. 2. Note that the substrates with the height variation labeled C in Tab. 2 are identical to turning the corresponding substrates labeled A upside down.

Permittivity profile	z_0	Δz
A	0	d
B	-	∞
C	d	d

Table 2: Permittivity profiles.

In Fig. 7, the half-power bandwidth for P_t is shown in percent as a function of Δx and Δy for the height variation labeled A in Tab. 2. The corresponding bandwidth is shown in Fig. 8 and 9 for profile B and C, respectively. For $l_x/\Delta x = l_y/\Delta y = 0$, the constant relative permittivity $\epsilon = 2$ for profile B does not correspond to the homogenized relative permittivity $\epsilon^{hom} = 8/\pi$ of profile A and C. An additional computation for a homogeneous substrate with $\epsilon^{hom} = 8/\pi$ gave a bandwidth of 14.8%. This confirms the result that a permittivity decreasing with height increases the bandwidth. Furthermore, Figs. 7, 8 and 9 do not indicate any clear advantage of an inhomogeneous substrate. The voltage standing wave ratio (VSWR) is less than 1.6 and the resonant frequency ranges from 1.5 GHz to 2.5 GHz for the profiles presented in Fig. 7, 8 and 9.

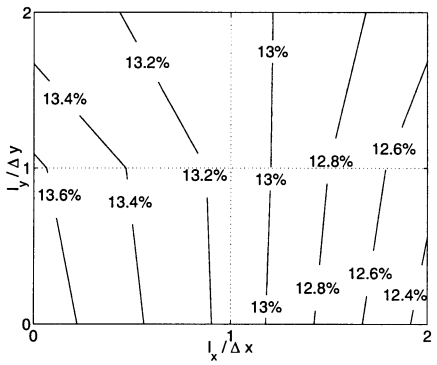


Figure 7: Half-power bandwidth for P_t for profile A in Tab. 2.

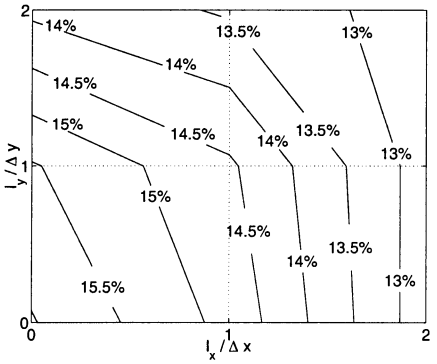


Figure 8: Half-power bandwidth for P_t for profile B in Tab. 2.

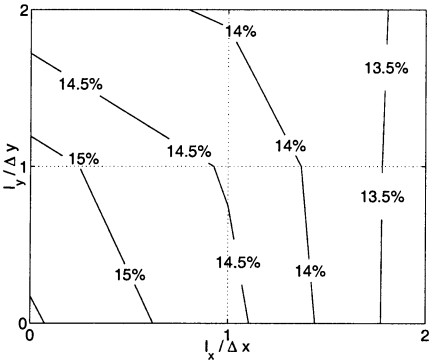


Figure 9: Half-power bandwidth for P_t for profile C in Tab. 2

4 Conclusion

We have presented a study on how inhomogeneous substrates for patch antennas influence the performance in terms of surface waves, radiation efficiency and bandwidth. It is well known [9] that, for a homo-

geneous substrate, a higher permittivity gives a more concentrated field below the patch, a lower resonance frequency and a smaller bandwidth. For inhomogeneous substrates with a *fixed* homogenized value of the permittivity, our results indicate that an increase of the permittivity in the vicinity of the patch lowers the bandwidth and the resonance frequency. To be specific, the FEM-FDTD computations indicate that a permittivity profile which decreases (increases) as a function of height from the ground plane somewhat improves (lowers) the bandwidth compared to a substrate with the corresponding homogenized permittivity. We also see that an increasing (decreasing) permittivity profile yields a lower (higher) resonance frequency than the corresponding homogenized permittivity. Furthermore, the analytical technique showed that inhomogeneous substrates with the permittivity only varying with the height influence the field pattern within the substrate, but not the radiation efficiency.

References

[1] N. G. Alexopoulos, P. B. Katehi, and D. Rutledge, "Substrate optimization for integrated circuit antennas," *IEEE Trans. Microwave Theory Tech.*, vol. 31, no. 7, pp. 550–557, 1983.

[2] C. S. Lee, V. Nalbandian, and F. Schwering, "Dual-frequency microstrip antenna with inhomogeneously filled dielectric substrate," *Microwave Opt. Techn. Lett.*, vol. 6, no. 11, pp. 629–632, 1993.

[3] L. Vegni, F. Bilotti, and A. Toscano, "Scattering properties of patch antennas loaded with inhomogeneous substrates via a combined spectral domain-moment method," *Journal of Modern Optics*, vol. 48, no. 3, pp. 425–438, 2001.

[4] L. Vegni, F. Bilotti, and A. Toscano, "Microstrip disk antennas with inhomogeneous artificial dielectrics," *J. Electro. Waves Applic.*, vol. 14, no. 9, pp. 1203–1227, 2000.

[5] G. Kristensson, P. Waller, and A. Derneryd, "Radiation effiency and surface waves for patch antennas on inhomogeneous substrates," Tech. Rep. LUTEDX/(TEAT-7100)/1–49/(2001), Lund Institute of Technology, Department of Electroscience, P.O. Box 118, S-221 00 Lund, Sweden, 2001.

[6] T. Rylander and A. Bondeson, "Stable FEM-FDTD hybrid method for Maxwell's equations," *Comput. Phys. Comm.*, vol. 125, pp. 75–82, March 2000.

[7] I. V. Lindell, A. H. Sihvola, S. A. Tretyakov, and A. J. Viitanen, *Electromagnetic Waves in Chiral and Bi-isotropic Media.* Boston, London: Artech House, 1994.

[8] G. Kristensson, S. Poulsen, and S. Rikte, "Propagators and scattering of electromagnetic waves in planar bianisotropic slabs — an application to frequency selective structures," Tech. Rep. LUTEDX/(TEAT-7099)/1–32/(2001), Lund Institute of Technology, Department of Electroscience, P.O. Box 118, S-221 00 Lund, Sweden, 2001.

[9] C. A. Balanis, *Antenna Theory.* New York: John Wiley & Sons, second ed., 1997.

[10] P. G. Petropoulos, L. Zhao, and A. C. Cangellaris, "A Reflectionless Sponge Layer Absorbing Boundary Condition for the Solution of Maxwell's Equations with High-Order Staggered Finite Difference Schemes," *J. Comput. Phys.*, vol. 139, pp. 184–208, 1998.

[11] A. Taflove, *Computational Electrodynamics: The Finite-Difference Time-Domain Method.* Norwood, MA: Artech House, 1995.

[12] R. Cools and P. Rabinowitz, "Monomial cubature rules since "Stroud": a compilation," *J. Comput. Appl. Math.*, vol. 48, pp. 309–326, 1993.

[13] R. Cools, "Monomial cubature rules since "Stroud": a compilation – part 2," *J. Comput. Appl. Math.*, vol. 112, no. 1-2, pp. 21–27, 1999.

[14] X. F. Zgainski, J. L. Coulomb, and Y. Maréchal, "A New Family of Finite Elements: The Pyramidal Elements," *IEEE Trans. Magnetics*, vol. 32, pp. 1393–1396, May 1996.

[15] S. Balay, W. Gropp, L. C. McInnes, and B. Smith, "The portable, extensible toolkit for scientific computation." `http://www-fp.mcs.anl.gov/petsc/index.html`.

Test of Smart Antenna Properties in a Simulated Scattered Field Environment

Bo Olsson

[1]Telia Research AB, 123 86 Farsta, Sweden. E-mail: bo.g.olsson@telia.se.

Abstract

This paper investigates some basic techniques to direct the transmission from a Base Station (BS) to a Mobile Station (MS) in simulated scattered field environments.
Simulation results show that, although an adaptive BS adaptive array can give high down link C/I median gain, up to 14 dB with 8 elements, in low angular spread (6°) environments, this is not possible in high angular spread environments. However, in combination with MS two (2) antennas Optimum Combining (OC) diversity reception, still a considerable C/I median gain is possible. I.e. 6 dB total C/I median gain with a 4 elements BS array and 9 dB with an 8 elements BS array in 35° average angular spread environments, compared with 1 dB and 2 dB for 4 and 8 element BS arrays alone. A minor improvement of 1 – 3 dB is achieved increasing MS complexity to 4 element OC.

1 Introduction

For 3G mobile communication systems, i.e. UMTS, the down link C/(I+N) or actually the C/I, for a heavy loaded base station, will determine the down link capacity (at a given coverage). Multi path propagation causes the orthogonal codes to interfere. Given that the C/I distribution remain the same, every doubling (3dB) of the median down link C/I thus doubles the Base Station (BS) capacity. For broadband services, up to 2 Mbits/s per mobile terminal, it is thus important to find techniques to improve the C/I. One way is to utilize "smart antennas". I.e. to use antenna diversity reception and adaptive antenna beam forming for transmission.
This report compares the down link performance of some basic methods to control antenna arrays in simulated scattered field environments. This means how the standard antenna concepts like Adaptive Narrow Beam (phased array), Equal Gain Combining, Maximal Ratio Combining and Optimum Combining performed in up link can be used to direct the transmission in the down link duplex frequency band in a scattered field environment. The additional gain of MS diversity reception is also included.

2 Model of the scattered field environment

The Non-Line-Of-Sight (NLOS) scenario can be modeled by a set of single (or multiple, clustered) scatters, normally surrounding the mobile terminal and with random weights for the signal components that are reflected or diffracted at these points, [1]. The used model is a narrow band scattered field model. I.e. it does not include the time domain properties, like the impulse response and the associated gain in an equalizer. It only combines the narrow band signals from multiple antennas. In up link this could simulate a combining after a number of RAKE receivers, one from each antenna element. The model consists of a number of signal paths from the mobile terminal antenna to a base station antenna via a chosen number of single scatters. There is no direct Line Of Sight (LOS) component, so the received signal envelope will be close to Rayleigh distributed. The base station antenna is a linear array with a number of elements that will receive or transmit signals. Mutual coupling between elements is neglected. The mobile terminal is moving on a circle with chosen radius. For the purpose of simulating interference, there is also a second mobile terminal (same cell), having a linear movement, which have a second set of scatters and thus separate signal paths to the base station. There is no lognormal shadow fading or distance dependence included in the model. Polarization effects are not considered. I.e. all antennas are considered vertically polarized.
Three different scenarios have been compared, first a small cell scenario with a 120° sector cell where the average angular spread is 6°. The second scenario is considered to be a micro cell with an average angular spread of 25°. The third scenario is considered to be a micro cell or an indoor cell with an average angular spread of 35°. In all three cases the scatters are limited to a half plane. I.e. they are situated at only one side of the array.

Figure 1, below, shows one random micro cell scenario with 15 scatters for both the wanted signal and the interference.

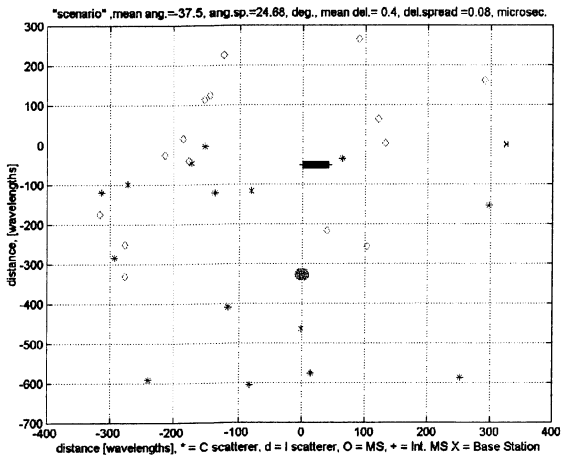


Figure 1. A micro cell scenario, where the average distance from the mobile station, MS, circle at (0,-325), to the base station, BS, X at (325,0) is 50-100 m at the chosen frequency 1950 MHz. The interfering MS2 (+) starts at (0, -50). Wanted signal have 15 scatters (*) and interference have 15 scatters (◊) randomly placed in a square with 100 m side. Angular spread for the wanted signal is 24.7 ° and delay spread is 0. 080 μs. BS antenna, X, is in this case an 8 element array with 0.45 λ separation at RX frequency and 0.49 λ at TX frequency.

3 Diversity reception

For 3G mobile communication systems, i.e. UMTS, the Four different beam forming and diversity reception techniques have been used, adaptive Narrow Beam (NB), Equal Gain Combining (EGC), Maximal Ratio Combining (MRC) and Optimal Combining (OC).

3.1 Adaptive Narrow Beam

The NB antenna is the simple linear phase controlled array where the complex weights for the signal voltages at elements, i = 1 – N, are:

$$W_i = e^{-j\varphi_i} \quad (1)$$

$$\varphi_i = 2\pi(i-1)D\sin(\phi) \quad (2)$$

Where ϕ is the NB horizontal lobe angle, $-\pi/2 < \phi < \pi/2$. I.e. the signal voltages at the elements will be added in phase for a signal coming from direction φ. Figure 2 below show two lobes directed at different angles. Antenna array element separation D = 0.45 λ (0.49 at transmit frequency).

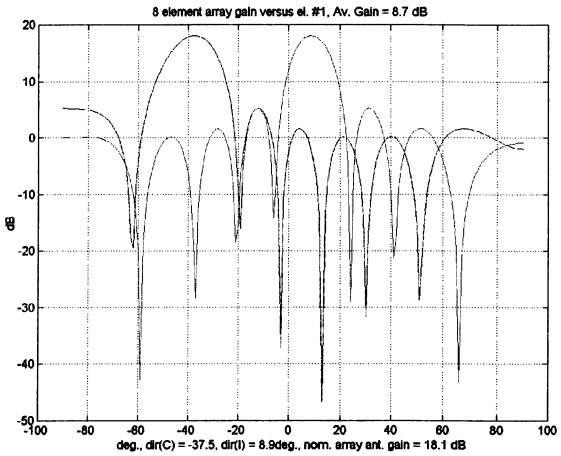


Figure 2. Two Narrow Beams for an 8 element antenna array illustrating the interference rejection that will be possible with an adaptive antenna array. ϕ = -37.5 ° and 8.9 °. Note that the gain in noise is less (1/2) since there will be 8 times the receiver noise. I.e. SNRgain = $20*^{10}\log(8)-10*^{10}\log(8)$ = 9.0 dB.

3.2 Equal Gain Combining, EGC

In this case the voltage at each antenna element from the scattered field is added in phase. I.e. the instantaneous phase of the voltage from each antenna element is compensated for and the voltages can thus be added in phase.

If the signal voltage at each antenna element is:

$$V_i = A_i e^{j\varphi_i} \qquad i = 1 - N \quad (3)$$

$$V_{EGC} = \sum_i V_i e^{-j\varphi_i} = \sum_i A_i \quad (4)$$

Note that in the case of Line Of Sight (LOS), that is equivalent to one scatter, EGC and NB will give the same result.

3.3 Maximal Ratio Combining, MRC

Maximal Ratio Combining is similar to EGC. The voltage at each antenna element is phase compensated and weighted by the Signal to Noise Ratio (SNR) and normalised. I.e. the sum of the weights, that is the same as for the NB and EGC antennas, is equal to the sum of the elements.

$$V_i = A_i e^{j\varphi_i} \qquad i = 1 - N \quad (5)$$

$$V_{MRC} = \sum_i W_i V_i e^{-j\varphi_i} = \frac{1}{k}\sum_i (A_i)^2 \quad (6)$$

$$Wi = Ai/k, \ \sum_i W_i^2 = N \quad (7)$$

Note: k~s^2, s^2 = variance of noise (noise power), [2]

3.4 Optimum Combining, OC

Optimum Combining is similar to EGC and MRC. The voltage at each antenna element is phase compensated and weighted by the Signal to Interference (voltage) Ratio (SIR) at each element and normalised. I.e. the sum of the weights, that is the same as for the NB EGC and MRC antennas, is equal to the sum of the elements.

$$V_i = A_i e^{j\varphi_i} \qquad i = 1 - N \qquad (5)$$

$$V_{OC} = \sum_i W_i V_i e^{-j\varphi_i} = \sum_i \frac{1}{k_i} (A_i)^2 \qquad (6)$$

$$W_i = A_i / k_i \;,\; \sum_i W_i^2 = N \qquad (7)$$

Note: $k_i \sim s_i^2$, s_i^2 = variance of interference (power)

3.5 Reception (RX) noise performance

With an N element adaptive antenna array the noise power will be higher than in the single element receiver. This is obvious in the NB and EGC array with unity gain weights where the noise power is N times higher. In a LOS situation this would give a maximum SNRgain = $20*^{10}\log(N)-10*^{10}\log(N)$, (see fig. 2). With the sum of the squared weights normalised to N, this will also be true for MRC and Optimum Combining.

4 Results

4.1 Direction of main lobe for maximum received signal

Figure 3, below, shows how the direction of the maximum lobe varies when the Mobile Station (MS) moves along the circle (fig. 1) for the NB, EGC, MRC and OC (called OPT in the diagrams) array antennas (lower curves). I.e. the NB antenna is directed towards maximum received power in each of the 1024 samples, when the mobile (MS) antenna moves along the circle, and will hereafter be called NBa (adaptive). The average directions are -39.4, -38.5, -39.4 and -38.2° for NB, EGC, MRC and OC. Note that for NB and MRC not only the average direction is the same, but also the maximum lobe direction is the same for every sample along the circle. The direction of the NB antenna lobe for maximum reception by MS2 (NB2(r)) is also included with average direction 6.0°.

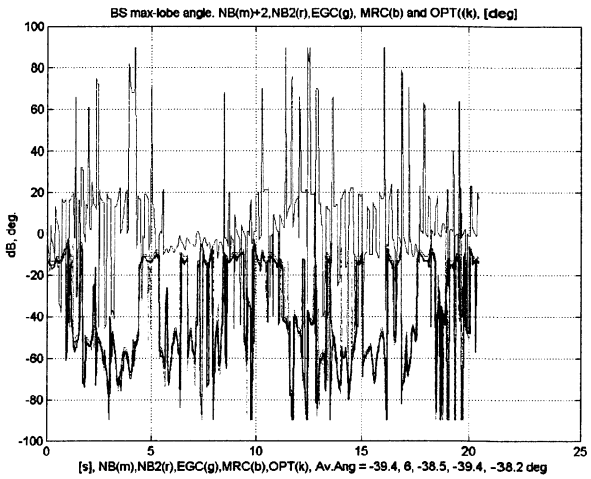


Figure 3. The directions of the main lobes for reception with NB, EGC, MRC and OC diversity reception antennas (lower curves). Note that the main lobes are pointing in almost the same directions. NB2(R) is the direction of the Narrow Beam adaptive antenna receiving maximum interference (upper curve).

4.2 Received signal to noise gain (SNR), up link

Figure 4 shows the cumulative SNR distributions of 100 simulations. Also a slow or average direction NB antenna is used directed in the average (or mean) direction of NBa. This antenna is called NBm (with main lobe in -39.4° in fig. 3;).

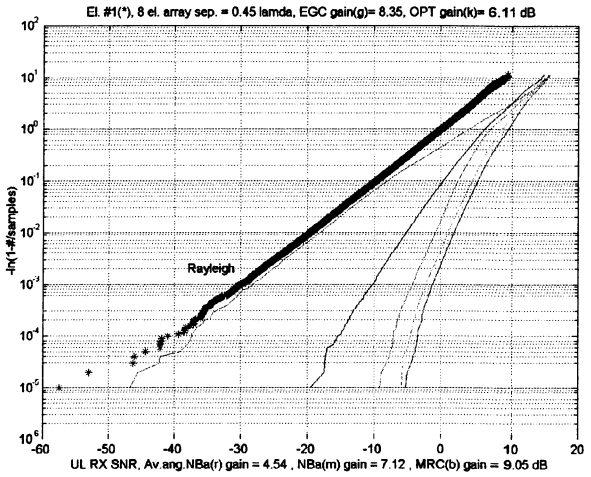


Figure 4. Received signal SNR CPDF (curves from left to right) for one element (*), NBm in average direction, OC, adaptive NBa, EGC and MRC. Note: 100 random simulated environments (fig.1) with average angular spread 25°.

4.3 Adaptive BS antenna transmission with MS diversity reception

In the following these BS array antennas have been used also to beam form in the frequency shifted duplex band

(from 1950 MHz to 2140 MHz; UMTS). I.e. for the adaptive NB antennas the phase is compensated for the shift in frequency, so the transmitting main lobe is directed in the same angle as the receiver antenna. For EGC, MRC and OC the complex weights used for reception are used also for transmission in the downlink. All combinations of BS transmit adaptive arrays (NBm, NBa, EGC, MRC and OC) have been simulated with 100 sets of random environments with 6°, 25°and 35° angular spread. The median (50%) C/I gain of the Cumulative Probability Density Function (CPDF) relative to one (1) BS antenna to one (1) MS antenna has been used as performance measure.

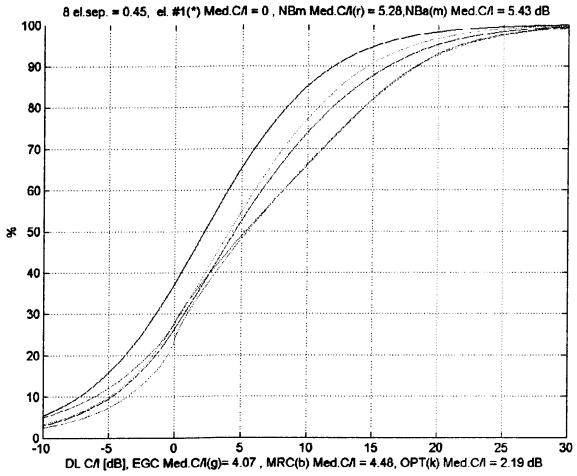


Figure 5. Received SIR CPDF in the mobile (MS) (interference to MS2) for the different beam forming algorithms; from left to right: OC, EGC, MRC, NBm and NBa. Note: 100 random environments with average angular spread = 25° in a 120° cell.

The SIR gain is quite low for the adaptive arrays used to beam form in the transmit frequency band compared to the SIR gain with the array used for reception. Median SIR gain is maximum 5.4 dB for the NBa array whereas the uplink reception gain is maximum 15.0 dB for OC. See figure 6.

4.4 Adaptive BS antenna transmission with MS diversity reception

The above down link simulations, 4.3, have also been performed with MS 2 and 4 antenna (square configuration) reception space diversity with four different algorithms, EGC, MRC, OC and C/I selection combining. The spacing of the antenna elements are 0.45/0.49 λ at the two frequency bands (UL/DL).

Here it is worth noting that MS diversity reception will only give down link intra cell SIR gain if the BS transmit antenna is an adaptive array that can decorrellate the desired signal from the interfering signals (codes) transmitted to other MSs in the cell.

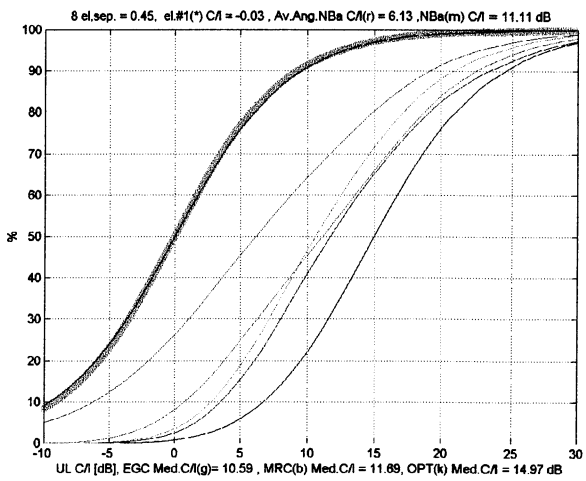


Figure 6. Up link received signal to interference (SIR) CPDF with (from left at 50% level) one element reference (*), NBm in average direction, EGC, NBa, MRC and OC BS reception with intra cell interference from MS2. Note: 100 random environments with average angular spread = 25° in a 120° cell.

The results from the simulations have been condensated into to a table with the highest median C/I gains in the three different scenarios. I.e. NBa and MRC BS transmit arrays and MS 2 and 4 elements optimum combining, although MS 4 elements C/I selection combining seems to give slightly higher median SIR gain than MS 4 element OC for some scenarios. See figure 7.

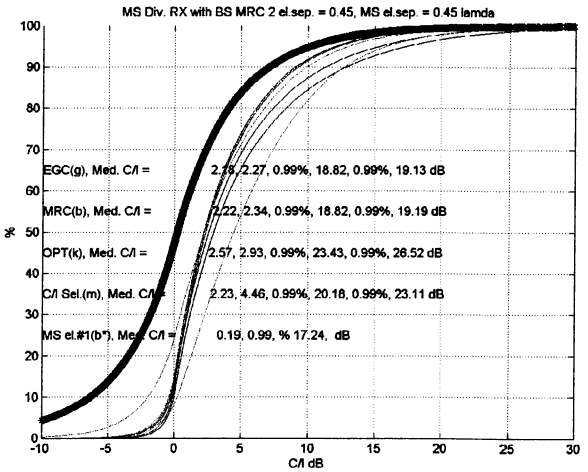


Figure 7. MS 2 and 4 element EGC, MRC, OC and maximum. C/I selection diversity reception (2 el. BS MRC array in 35° angular spread environment). MS Optimum combining has the highest 99% level C/I gain, but MS max C/I selection diversity with four antennas has higher gain in the 10 - 95% level. Median MS 4 elements C/I selection C/I gain is 4.5 dB while 4 el. OC gain is 2.9 dB.

5 Summary of results

MS ant. elements	BS Angular spread°	1 BS ant element gain, dB	2 BS ant. elements		4 BS ant. elements		8 BS ant. elements	
			NBm	MRC	NBm	MRC	NBm	MRC
1	6	0	2.3	2.4	9.8	8.6	14.4	12.6
	25	0	1.4	0.8	3.1	2.5	5.3	4.5
	35	0	0.1	0.2	0.3	0.8	0.7	2.1
2	6	-	2.6	3.1	11.9	10.7	18.0	16.3
	25	-	2.9	2.5	7.1	6.9	10.4	10.8
	35	-	0.4	2.6	2.9	6.1	5.0	9.5
4	6	-	2.7	3.1	12.7	11.2	20.3	17.2
	25	-	4.1	2.8	10.5	7.6	14.1	12.0
	35	-	0.7	2.9	5.0	6.8	8.0	10.3

Table 1. MS 2 and 4 antenna elements Optimum Combining with BS NBm (average direction) and MRC downlink beam forming array total median intracell C/I gain. Note that the MRC BS antenna give higher total median C/I gain for 2 and 4 antenna MS Optimum Combining in the high angular spread ($\sigma = 35°$) environment.

MS ant. elements	BS Angular spread°	1 BS ant element gain, dB	2 BS ant. elements		4 BS ant. elements		8 BS ant. elements	
			NBm	MRC	NBm	MRC	NBm	MRC
1	6	0	2.3	2.4	9.8	8.6	14.4	12.6
	25	0	1.4	0.8	3.1	2.5	5.3	4.5
	35	0	0.1	0.2	0.3	0.8	0.7	2.1
2	6	-	2.9	3.0	12.0	11.0	17.5	15.4
	25	-	3.1	2.4	6.1	5.1	8.8	7.7
	35	-	0.7	2.2	2.3	3.7	3.2	5.3
4	6	-	3.4	3.6	13.4	12.6	19.9	17.9
	25	-	4.8	4.1	8.9	7.6	11.7	10.6
	35	-	1.6	4.5	4.3	6.4	5.3	8.2

Table 2. MS 2 and 4 antenna element C/I Selection Combining with BS NBm (average direction) and MRC downlink beam forming array total median intracell C/I gain.

6 Conclusions

For low angular spread environments, cells with over roof top BS antennas, a phased array that can find an average MS direction, NBm, is a good solution with high median intracell down link C/I gain (SIR), 14.4 dB for the 8 elements array in the 6° average angular spread simulated model. In a high angular spread environment, 35°, this NBm BS array median gain is reduced to only 0.7 dB. Using the 8 elements BS MRC array increases the median gain to 2.1 dB in the 35° angular spread environment and MS 2 element OC boosts the median gain to 9.5 dB (table 1).

Interesting is to note that MS C/I selection with 4 antennas seems to give 3.6 – 4.5 dB median DL intracell C/I gain with only 2 elements BS MRC array (table 2 and fig.7).

[1] Overview of Spatial Channel Models for Antenna Array Communication Systems, R. Ertel, P. Cardieri, K. Sowerby and T. Rappaport, IEEE Personal Com. Feb. 1998, p. 10-22.

[2] W.C Jakes, Microwave Mobile Communications, Wiley, 1974

The AMC-9 Ku-band Receive antenna

Maria Holm[1]

[1]Saab Ericsson Space AB, SE-405-15 Göteborg, Email: maria.holm@space.se.

Abstract

This paper describes the AMC-9 Ku-band Receive antenna operating at 14-14.5GHz, and using both vertical and horisontal polarisation. The antenna concept chosen is a shaped offset fed Gregorian dual reflector system. The feed chain consists of a circular corrugated horn, an OrthoModeTransducer and two waveguide runs. The system geometry parameters are presented as well as the predicted and measured performance Design and manufacture was performed at Saab Ericsson Space. RF-tests were performed at Intespace, France in a compact range with a quiet zone of 5.5x5x6 m^3 (WxHxD). The customer is Alcatel, France.

1 Introduction

The AMC-9 satellite is aimed at TV-broadcasting directly to customer for North-America and Mexico. A transmit antenna for the same frequencies and polarisation is placed on an East/West side panel of the satellite.

The AMC-9 satellite will be launched December 2002 from Baikonur, Kazakhstahn in a Proton launcher. It will have a lifetime of 15 years and have four control station, with the main control station in Vernon Valley, New Jersey.

2 Design Requirements

The AMC-9 Ku-band Receive antenna has very strict requirements on gain, isolated cross-polarisation, sidelobes over a large area and gain variation. Some of the requirements are summarised in the table below. The antenna has two mainlobe coverages, CONUS (CONtinental US) and Mexico, and strict sidelobe suppression requirement over a coverage over South America. The coverages were given as longitude/latitude polygons.

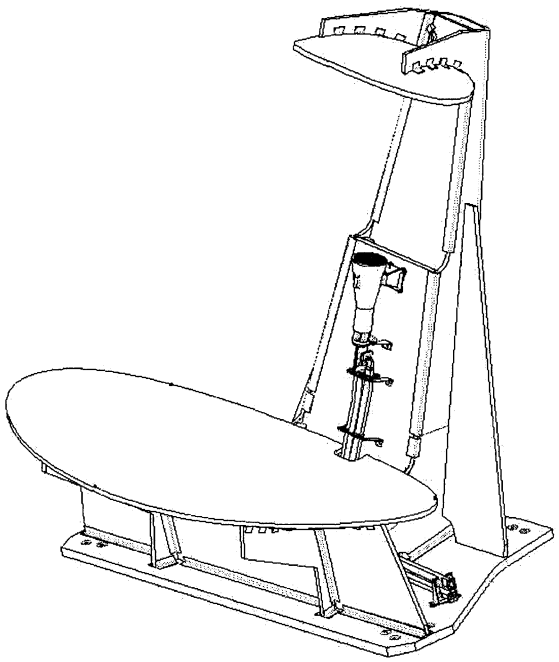


Figure 1-1: *The AMC-9 Ku-band Receive antenna.*

Requirement	Value	Unit
Minimum gain in CONUS +Mexico city	28.8	dBi
Minimum gain in Mexico	26.1	dBi
Maximum gain in South America	0	dBi
Minimum isolated crosspolarisation in CONUS + Mexico city (Analysis)	34	dB
Minimum isolated crosspolarisation in CONUS + Mexico city (Measured)	30	dB

Table 2-1 Summary of some requirements.

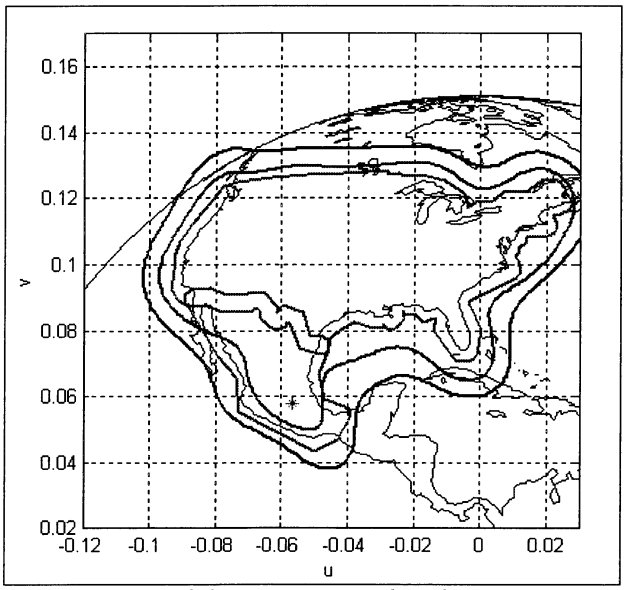


Figure 2-1 *Main lobe coverages and performance.*

3 Design Description

The specification defined a limited allocated volume on earth-deck for the antenna. From this volume it was decided that the best concept to chose was a Gregorian reflector system. The reflector system contains the following parts.

- Main reflector
- Subreflector
- Ku-band Feed chain
- Support structure

To obtain low thermal distortion to the RF-performance, the antenna was manufactured in CFRP (Carbon Fibre Reinforced Plastic). Due to a hot thermal case in the launch phase, the reflectors had to be painted white. The feed chain was manufactured in aluminium, except from a flexible section in the waveguides which was made out of beryllium cupper. Both waveguide runs were silverplated to reduce losses.

1 Minimising the cross-polarisation by introducing an angle of a certain magnitude between the sub and main reflector axes

3.1 Reflector system design

The offset plane of the reflector system was put in a east/west direction. The main reflector rim was chosen to be elliptical to suite the coverage. The optimum taper of the subreflector was found to be -16dB, however, the rim of the subreflector was increased to reduce back scattering from the subreflector. To obtain the strict crosspolarisation requirement, a careful design of the reflector system in terms of fulfilling Mitzugutchs conditions[1] and an advanced design of the circular corrugated horn was performed.

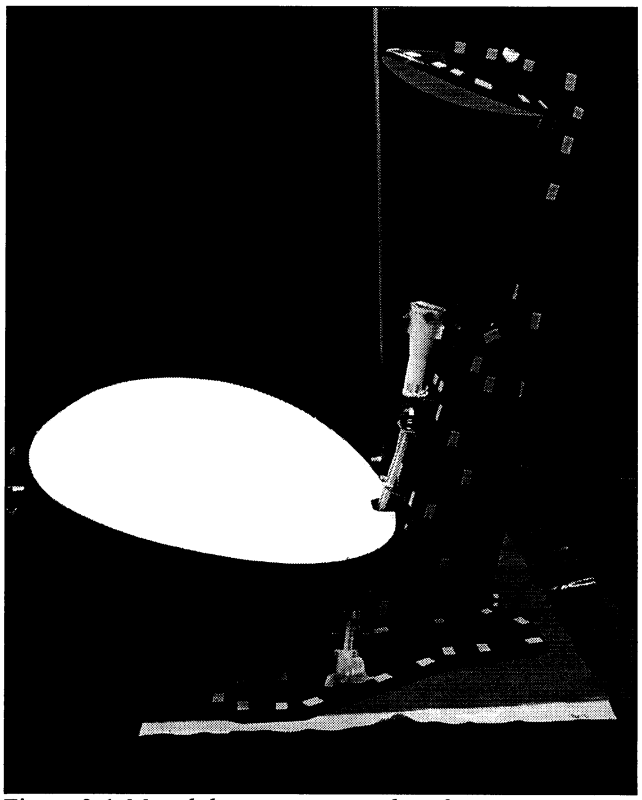


Figure 3-1 *Main lobe coverages and performance.*

Geometry parameters of the reflector design is presented in the table below.

	Value	Unit
Main reflector		
Focal distance	800	mm
Aperture radius (x,y)	475/675	mm
Offset	500	mm
Subreflector		
Feed/focal axis tilt	17.02	deg
Half distance between focal points	150	mm
Half major axis	525	mm
Project rim radius (x,y)	226/272	mm

Table 3-1 *Main lobe coverages and performance.*

The main and sub reflector surfaces were shaped using an optimising software. However, this software did only take into account the field generated in the main reflector. This proved neither to be sufficient for the low levels in which we had sidelobe and cross-polarisation requirements nor for the requirement of maximum allowed gain variation in the mainlobe.

To get the exact radiation pattern we therefore had to analyse the performance in more detail with another software. Using this software we could include direct field from the feed, back scattering from the subreflector, and reflections in the structure. This software was also used to perform in-orbit analysis by adding deformations on the reflectors and displacement of feed for different thermal cases.

In Figure 3-2 the deviation from a parabolic surface for the main reflector of the AMC-9 Ku-band receive antennas main reflector is presented.

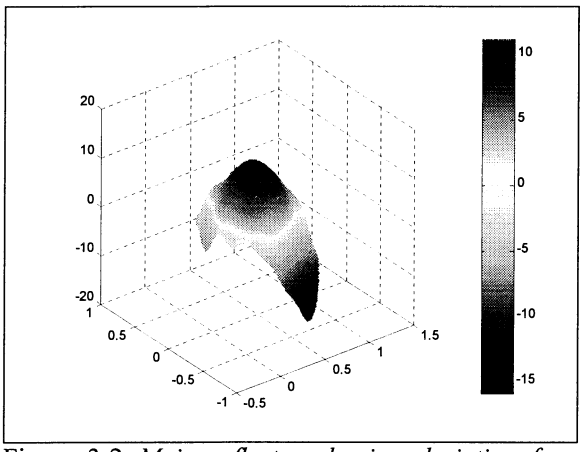


Figure 3-2 *Main reflector shaping deviation from parabolic shape [mm].*

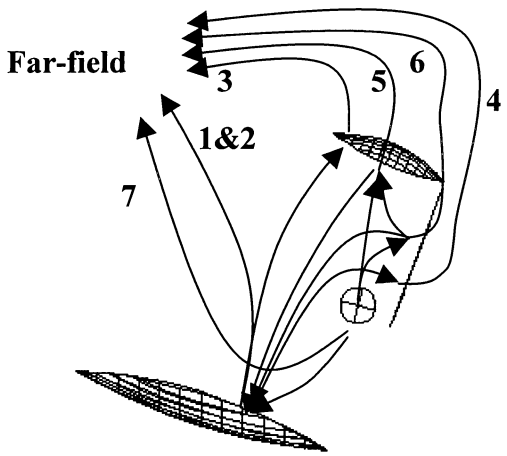


Figure 3-3 Field contributors to detailed analysis.

The analysis included the following contributors each corresponding to a number in the figure above.

1. Direct radiation
   Feed → sub → main → far field

2. Reflection in support tower
   Feed → tower → sub → main → far field

3. Scattering from sub-reflector
   Feed → sub → main → sub → far field
   Feed → tower → sub
   Feed → tower → main

4. Scattering from support tower
   Feed → sub → main → tower → far field
   Feed → tower → sub
   Feed → tower → main

5. Back radiation from sub reflector
   Feed → sub → far field

6. Back radiation from support tower
   Feed → tower → far field

7. Direct radiation from feed
   Feed → far field

3.2 Feed chain design

The feed chain consists of 4 components.

- Circular corrugated horn
- OMT (Ortho Mode Transducer)
- 2 Waveguide runs

The circular corrugated horn was designed in a modematching software. An analysed pattern is presented in figure 3-4 where the low cross-polarisation due to corrugations is evident.

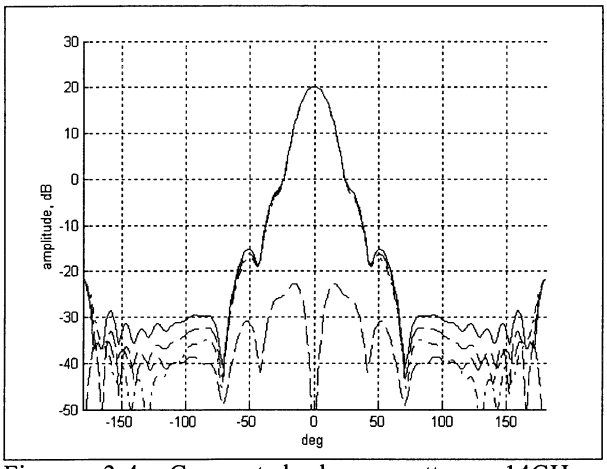


Figure 3-4 Corrugated horn pattern 14GHz Horisontal polarisation, - E-plane --45 degrees .- H-plane.

An OMT was designed in an FDTD software to split the signal into a horisontal and a vertical polarisation component.

A standard WR75 waveguide was designed to route the signals down to a defined waveguide interface. The waveguides were silverplated to reduce losses and had a flexible section to account for thermal stresses.

4 Design Verification

During manufacturing, the surface shapes of the reflectors were measured. This measured data was used in an analysis which thus includes manufacturing errors. An exact measurement of the position of the reflectors after mounting was also included.

In the test phase, two pattern measurements were performed. The initial measurement was made prior to environmental tests (thermal vacuum, vibration and acoustic tests) and the final after the environmental tests.

Figure 4-1 indicates the impact of the manufacturing tolerances we have. Errors origins both from the manufacturing of reflectors and the assemblage of them. The grid is in u-v co-ordinates (equivalent of radians for small angles).

An indication of the measurement tolerances is seen in Figure 4-2, where measurements from the initial and final measurements are shown. The data presented is suppressed cross-polarisation (co-polarisation over cross-polarisation on a point by point basis).

Figure 4-3 presents analysed data of measured manufactured surfaces in their actual position vs measured performance.

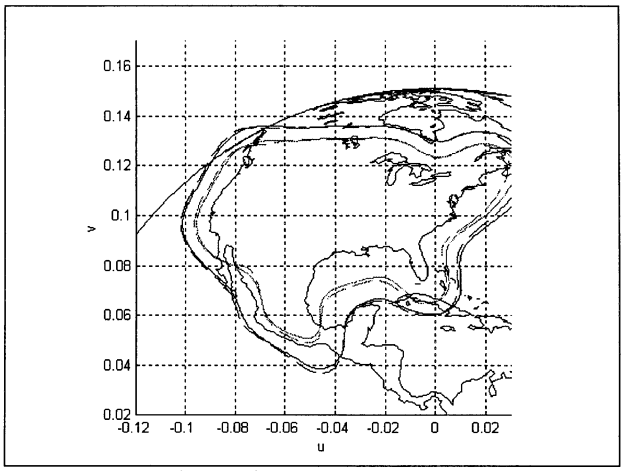


Figure 4-1 *Analysis of mainlobe performance 14.5GHz Vertical polarisation [26.1 28.8dBi].*
- Analysis using measurement on manufactured devices and positions.
-- Analysis of ideal structure

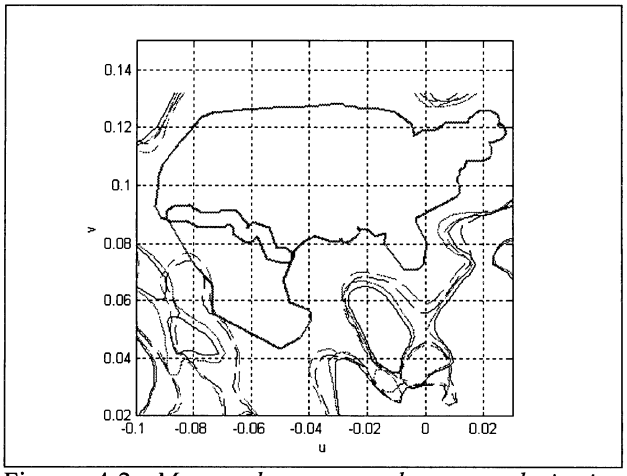


Figure 4-2 *Measured suppressed cross-polarisation performance 14.26GHz, horisontal polarisation [29 30 dB]. - Initial measurement -- Final measurement*

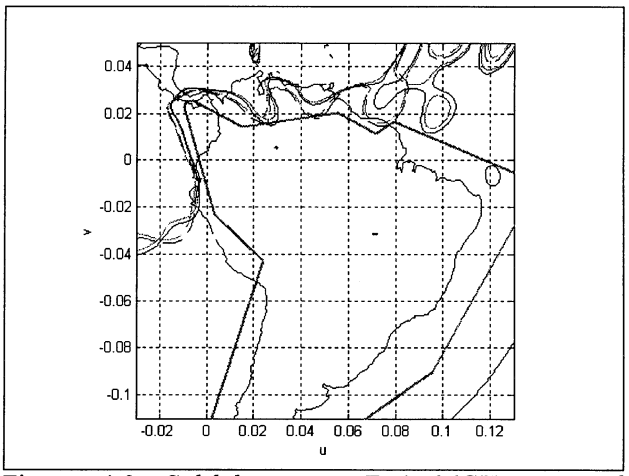


Figure 4-3 *Sidelobe area F=14.36GHz, vertical polarisation [1, 0 dBi]. - Analysed with measured data*
-- Initial measurement

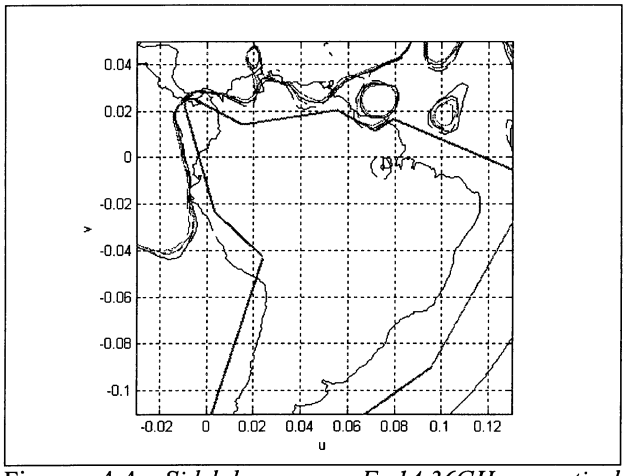


Figure 4-4 *Sidelobe area F=14.36GHz, vertical polarisation [1, 0 dBi]. - Initial measurement -- Final measurement*

To compare the difference between analysis and measurement we have also included a plot of sidelobes at both initial and final measurement as seen in Figure 4-4. It can be noticed that the accuracy of the detailed analysis is in the same region as the measurement accuracy.

5 Conclusion

The AMC-9 antenna fulfils the mainlobe and the strict cross-polarisation requirements. It was predicted in the design phase that it would not be possible to achieve the sidelobe requirements in combination with the mainlobe requirements as the way to reduce the sidelobes were to increase the subreflector taper. What was important to us was that the predictions proved to be very accurate.

Our conclusion from the analysis and measured performances of the AMC-9 antenna is that both software uncertainties and measurement inaccuracies at the test range are small. Thus, we can predict the performance from analysis with high confidence.

Advanced Antenna Systems for Cellular Networks

Bo Hagerman and David Astély

Ericsson Research, Ericsson AB, SE-164 80 Stockholm, Sweden.
Bo.Hagerman@era.ericsson.se, David.Astely@era.ericsson.se

Abstract

The deployment of second generation cellular networks, such as GSM, has been a success. The growth of the subscriber population and the introduction of data services in cellular network motivate the study of how capacity and coverage of current and future systems can be increased to meet demands. Advanced antenna systems, employing multiple antennas for reception and/or transmission, is a promising technology for this purpose. In this talk, advanced antenna systems and their application to cellular networks are discussed.

1 Background

Cellular systems provide area coverage and supports mobility. First generation cellular systems, such as NMT and AMPS, were analog systems using frequency division multiple access (FDMA). In combination with FDMA, second and third generation systems, such as GSM, EDGE, PDC, D-AMPS, CDMAOne, CDMA2000 and WCDMA are digital and use code division multiple access (CDMA) or time division multiple access (TDMA). To provide coverage and to improve the spectral efficiency per area unit, the coverage area is divided into cells. In each cell the limited resource, such as spectrum, time-slots, and/or code-sets, can be re-used and the requirements on the output power can be reduced.

The deployment of second generation cellular systems, such as GSM, has been very successful with a steady growth of the subscriber population during the past decade. At the same time, the number of computers connected to the Internet has increased exponentially. Today evolved second-generation systems and third generation systems offering not only voice and low data rate services, but also packet switched data services, are being deployed. Previous experience on how the use of cellular systems has increased with time, and the introduction of wireless internet services motivate studies of how coverage and capacity can be developed to meet increasing demands.

Coverage and capacity can be improved by using more resources such as power and spectrum. However, so far, a limited spectrum has been allocated to cellular systems, and the cost to obtain more spectrum, if it even has been possible due to regulatory reasons, has in many cases been very high. An alternative, which in general requires standard changes, is to use more efficient access methods. Therefore, other techniques such as using the spatial dimension efficiently are interesting, given that they are cost-effective.

A natural first step can be to tune the networks by optimizing planning parameters and by improving resource management. Further, the cellular concept itself is a method to improve coverage and capacity, and the introduction of micro and even smaller cells in addition to sectorisation, is a way to increase the capacity. Indoor coverage can be offered with distributed antenna systems. Another technique with the potential to utilize the spatial dimension even more efficiently is the use of multiple antennas for reception and/or transmission. Such antenna arrangements in combination with signal processing functionality integrated in cellular networks are referred to as advanced antenna system and treated herein.

2 Advanced Antennas

We define an advanced antenna as "an antenna arrangement in combination with signal processing functionality". Signal processing is a central part of an advanced antenna. By processing the signals to or from a number of static antennas at different locations, the spatial dimension of the radio channel may be exploited and the effective radiation or reception pattern can be adapted to changes in the environment.

Several approaches can be appealing for utilizing the additional degrees of freedom supported by operating with multiple antennas. Examples are adaptive antennas, space-time coding and MIMO (Multiple Input Multiple Output) concepts which employs multiple antennas at both the transmitter and the receiver.

A typical adaptive antenna uses a uniform linear array with elements placed along a line and separated half a wavelength. For each antenna, an appropriate complex weight, such as a phase shift, is multiplied with the signal to be transmitted, and a beam is formed. In this way, the radiated power may be directed in a narrow beam towards an individual user. The target, besides the antenna gain improvement, is to minimize the interference spread at transmission. In the same way, the signals may be linearly combined at reception to suppress interference. Several approaches can be used to create the narrow beam. A phase front on the antenna elements corresponding to a beam can be generated at baseband using digital beamforming or at radio frequency (RF)

using a passive network or phase shifters. The basic implementation difference between the two approaches is schematically shown in Figure 1. Baseband transmit beamforming techniques always require phase coherency all the way to the antenna elements, but enables the possibility of beam shaping and nulling. One main advantage in terms of implementation complexity of using one beam from a passive beamforming network is that it does not require phase coherency between the radio transmitter and the beamformer. However, if multiple beams are transmitting simultaneously using the same carrier frequency, phase coherency between the beams is essential. An example of a passive network is the Butler matrix, which generates a set of orthogonal beams. Orthogonal beams, however, have a gain loss at the beam crossover points.

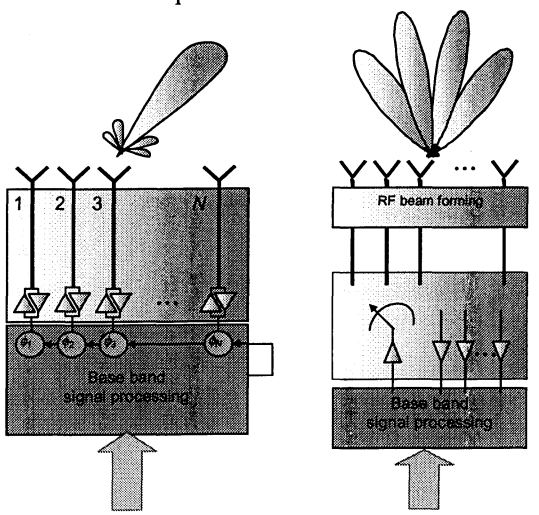


Figure 1: Examples of adaptive antennas. Beams can be generated with digital beamforming in the baseband or with a passive network at radio frequency.

In a cellular system with both reception and transmission, the digital beamforming weights for transmission can be determined by processing the received signals or by using feedback signaling. Digital beamforming can of course also be used with rather arbitrary antenna arrangements with elements of different polarization and separated several wavelengths. This is sometimes referred to as receive and transmit diversity. It should be noted that the filtering applied could be done not only in the spatial domain as described above, but also in the joint spatial and temporal domain. For each antenna, a frequency selective filter is then applied to the signal to be transmitted.

Recently there has been a lot of interest in the use of space-time coding and so called layered schemes to obtain transmit diversity for fading channels and to exploit the potential of MIMO channels with rich scattering. Data to be transmitted is then encoded in the spatial and temporal domain so that a unique stream of modulated symbols are transmitted from each antenna. This is depicted in Figure 2. Block codes, trellis codes and turbo codes have been proposed for this purpose. Typical for such space-time encoding schemes are that they are blind in the sense that no channel knowledge is required at the transmitter, and that uncorrelated fading is desirable. Adaptation to the characteristics of the environment can be done by means of measurements and feedback to select appropriate encoding.

For space-time encoded transmission, the main signal processing burden is in general moved from the transmitter side to the receiver side, and in addition to rate and diversity gain, receiver complexity is an important design parameter. This is thus a difference as compared to transmission with adaptive antennas where no modifications to the receiver are required. However, for simple space-time encoded transmission schemes, such as delay diversity, which transmits the same signal but with different delays from each antenna, and antenna hopping, the same receiver can in general be re-used.

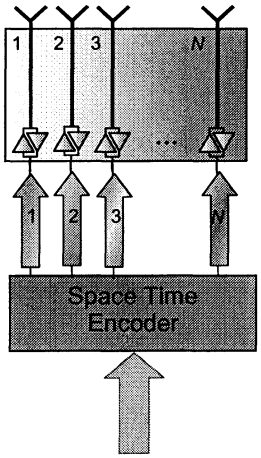


Figure 2: Encoding in the spatial and temporal domain.

The use of MIMO concepts with multiple antennas at both the receive and transmit side has lately attracted significant interest. This, since information-theoretic studies have shown that the capacity of a single MIMO link is significantly higher than the capacity of a link with single antennas at both ends. However, this requires that the radio channel offers enough scattering to guarantee sufficient decorrelation of the different channels. If knowledge of the channel is available at both ends, then techniques such as spatial pre-distortion and water filling may be employed at the transmitter. In combination with linear processing at the receiver, multiple parallel channels, over which different data streams are sent, can then be realized. If channel knowledge is not available at the transmitter, then various space-time encoding schemes in combination with signal processing functionality in the receiver can be used. In both cases, signal processing is an important component. It should also be noted that a MIMO link

may be viewed as a link with advanced antennas at both the transmitter and receiver, and that such an arrangement in a cellular network can be used in many different ways.

3 Advanced Antenna Systems in Cellular Networks

As mentioned above, cellular networks provide area coverage and supports mobility. The coverage area of a cellular network is divided into cells, see Figure 3. In each cell the limited resources, which are determined by the access method, can be re-used. The available amount of physical layer resources in a given area is decided via static cellular network planning and by dynamic network radio resource management.

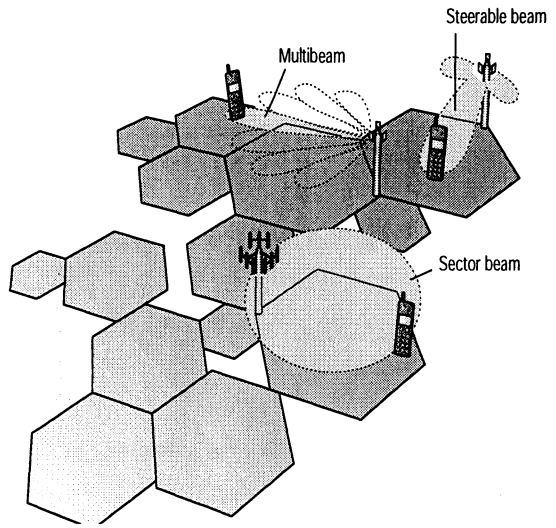


Figure 3: A cellular network.

In order to utilize the scarce physical layer resources efficiently, the cellular network planning parameters need to be optimized. Important as well is adequate interaction between base stations and radio network controllers so that the cellular network is observable and controllable. For example, efficient radio resource management requires relevant measurements of the physical layer conditions. To summarize, the air-interface and the radio resource management functionality shall be integrated to enable efficient utilization of available spectrum.

Advanced antennas offer additional means to reduce the multiple access interference and the total transmit power, while keeping the link quality at the desired level in an ever-changing wireless environment. The improvement from the deployment of advanced antenna system experienced by a single link can in general be used in several different ways on a system level. Examples include combinations of reduced transmit power of the terminals in the uplink and of the base station in the downlink, improved system capacity, extended coverage and support of more high data rate users in a mixed service system.

The deployment of advanced antennas must be considered in both the static and dynamic interference management. The integration and co-optimization with the air-interface, the radio network functionality and the protocols are in general both beneficial from a performance point of view and required to realize the improvement on a system level. Some examples of this are as follows.

— Admission and congestion control should preferably take the spatial domain into account so that the gain of the advanced antenna is considered. It is possible that some parts of a cell or the network are overloaded, while user can be admitted into the network in other areas, where the load is lower.

— To maximize throughput, MAC (Medium Access Control) operation with multiplexing and scheduling require more advanced algorithms, taking into account e.g. the users position and mobility in addition to the expected performance of the advanced antenna. Also power allocation to different channels may be affected by the introduction of advanced antennas.

— Cell selection/re-selection and handover will be affected, since conventional network design commonly is based on the assumption that both actual traffic channels and broadcast channels are transmitted in the same way with identical sector covering patterns. The situation can be different when advanced antennas are deployed. Using strategies integrating this channel inequality and utilizing the spatial domain is important to efficiently support mobility.

— Protocols require that common channels, such as the random access channels in the uplink and the broadcast channels in the downlink, must be received and transmitted over the entire cell. In fact, the area covered by the common channels defines the cell. Minor problems are expected for the uplink, but the downlink may require special considerations.

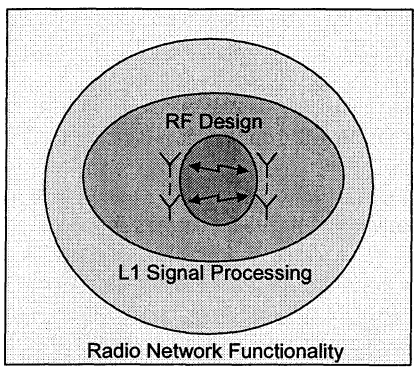


Figure 4: An advanced antenna system in a cellular network

As illustrated in Figure 4, an advanced antenna system consists of a plurality of antennas or antenna elements with RF front ends, and signal processing functionality for modulation, demodulation and measurements. The advanced antenna system interacts with the radio propagation, the radio network functionality, protocols and network architecture. In order to efficiently exploit the potential gains of an advanced antenna in a cellular network, it is important to understand the interaction with the other parts.

4 Advanced Antenna Systems Design

So far we have described some advanced antennas and the potential benefits of their use in cellular networks. From a design point of view, the problem may be formulated with the following question in mind:
"How shall an advanced antenna system be designed in order to efficiently improve the performance of a cellular network?"
This means that relevant advanced antenna concepts must be identified and related to each other. In addition to understanding how much various performance metrics can be improved, one also needs to understand how the performance improvements of different concepts are related. Gains can be additive or complementary, in the sense that the concepts are applicable in different environments. It is also of great interest to evaluate the trade-off between complexity and performance in order to find a cost-efficient solution. Compliance with the network architecture, the standard and the protocols of the given cellular network must also be ensured.
For the design and evaluation, the spatial dimension must be taken into account. When it comes to modeling of the radio channel, not only temporal characteristics are of interest, but also the spatial properties. The use of antennas with different polarization puts even more requirements on the measurements, modeling and understanding. Different deployments, such as macro and micro cells in addition to indoor environments are also to be considered.
To ensure that relevant interference operating points are considered, traffic models that captures the users' distribution in space are used. The interaction with radio network functionality, such as radio resource management and mobility, is also addressed.
Different concepts also put different requirements on the radio network functionality if their full benefit is to be exploited. A part of these requirements includes the definition of measurements of the physical layer conditions. Also, in addition to ensuring that relevant information is measured, the protocols should be such that this information is available in the appropriate node when needed. From a signal processing point of view, transmission of known pilot symbols for synchronization, channel estimation and for measurements is another issue that needs to be addressed.

In the standardization of WCDMA, some effort to support advanced antennas has been made. A simple space-time block code and low rate feedback for transmit diversity have been standardized for the case with two antennas at the base station. Further, certain channels may be transmitted to only a part of the cell and so-called dedicated pilot bits are transmitted. This in order to support adaptive antennas. Space-time encoding and low rate feedback schemes transmit diversity with more than two antennas are currently being discussed in the standardization forum. MIMO concepts with multiple terminal antennas for high-speed downlink packet access and measurements to improve the support for adaptive antennas are also considered.

5 Concluding Remarks

Advanced antenna system is a key tool to evolve the system performance in terms of capacity and coverage of current cellular networks. An advanced antenna system in this context consists of antenna arrangement in combination with signal processing functionality integrated in the radio network. It should here be stressed that signal processing is a central part of an advanced antenna system.
Although, the focus here has been on current cellular systems, the use of advanced antenna systems to exploit the spatial dimension is believed to be an important part also of future wireless systems. Examples of interesting topics include the use of multiple terminal antennas in wireless systems, the use of feedback for downlink beamforming, and the interaction with radio resource management such as fast packet data scheduling and mobility. A relevant question to address is then how a wireless system should be designed to efficiently exploit the spatial dimension with advanced antenna systems.

6 References

[1] S.Andersson et.al, *"Adaptive Antennas for Global System for Mobile Communications and Time Division Multiple Access (interim standard-136)"*, Chapter 24 of Handbook of Antennas in Wireless Communication, CRC Press, New York, 2002, Edited by Lal C. Godara.

[2] B.Hagerman et.al, *"Smart Antenna Systems for WCDMA – Downlink performance in Mixed Service Traffic Environments"*, in Proceedings of Future Telecommunications Conference 2001, November 2001, Beijing, China.

[3] U. Wachsmann, J. Thielecke, and H. Schotten, *"Exploiting the Data-Rate Potential of MIMO Channel: Multi-Stratum Space-Time Coding*, in Proceedings of IEEE Vehicular Technology Conference (VTC2001 Spring), May 2001, Rhodes, Greece.

Array Mapping: Reduced Bias Transformation Matrix Design

*Per Hyberg, †Magnus Jansson , †Björn Ottersten

Swedish Defense Research Agency FOI
Stockholm, Sweden

†Royal Institute of Technology KTH,
Stockholm, Sweden

Abstract

Mapping of the data output vector from an existing antenna array onto the data vector of an imaginary array of more suitable configuration is well known in array signal processing. By mapping onto an array manifold of lower dimension or uniform structure f. ex, processing speed can be improved. Conditions for the retaining of DOA accuracy under such mapping have been formulated by several authors but the equally important systematic mapping errors, the bias, has been less treated to date.

This paper uses a geometrical interpretation of a Taylor expansion of the DOA estimator cost function to derive an alternative design of the mapping matrix that almost completely removes the bias. The key feature of the proposed design is that it takes the orthogonality between the manifold mapping errors and certain gradients of the estimator cost function into account. Verifying simulations show a bias reduction factor of up to 100.

1. Introduction

The concept of array mapping (interpolation) was introduced around 1990 by Friedlander [1], and Weiss & Friedlander [2], who in several papers suggested and analyzed different applications such as interpolation between calibrated directions, a. s. o. An application studied also by many other authors, Eriksson [4], is mapping from non-uniform wide spaced linear arrays, NULA:s, onto $\lambda/2$ spaced ULA:s in order to allow the faster rooting estimators and still keep the superior resolution or bandwidth of the larger non-uniform array.

In most of this earlier work, performance under mapping has been assessed as the variance of the estimates in relation to the Cramér-Rao bound. General conditions for attaining this bound under mapping were derived by Andersson [5].

In many scenarios however, especially when the bandwidth of a given array is stressed, *mapping bias* can dominate over variance and become the main source of error. One such example is decade bandwidth signal reconnaissance in the 3-30 MHz band using one single array with a limited number of antenna elements. To retain resolution these arrays often get spaced in excess of $\lambda/2$ at the upper band edge.

Usually the transformation matrix $\mathbf{T}$ is designed to provide a best least square fit between the spaces spanned by the two sets of response vectors, the real set $\mathbf{A}_r(\boldsymbol{\theta})$ and the virtual set $\mathbf{A}_v(\boldsymbol{\theta})$. This is done by taking $\mathbf{T}=\mathbf{A}_v(\boldsymbol{\theta})\mathbf{A}_r(\boldsymbol{\theta})^{\dagger}$, where † denotes the Moore-Penrose pseudo-inverse.

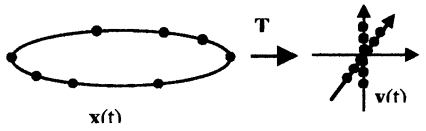


Figure 1: A sparse non-uniform circular array, NUCA, whose output vector $\mathbf{x}(t)$ is mapped by the matrix $\mathbf{T}$ onto the output vectors $\mathbf{y}(t)$ of a crossed ULA. Both stochastic and systematic errors may arise if $\mathbf{T}$ is not designed properly.

In many mapping situations where the least square fit is less than perfect this design of $\mathbf{T}$ is not optimal, especially when the mapping is between arrays of dissimilar shapes and has to work over large sectors. In such situations systematic measurement errors (bias) can occur. This bias can sometimes exceed standard deviation and become the main source of error.

2. Contributions and paper outline

The present paper proposes an alternative design of $\mathbf{T}$ that greatly can improve the performance of a mapped array by reducing the bias caused by the above mentioned mismatch.

Instead of matching the spaces spanned by the two sets of response vectors the alternative design puts the highlight on rotating the mapping errors so that they get orthogonal to the conjugate of the gradient of the estimator cost function derivative along the relevant signal eigenvectors. The design is derived using a Taylor expansion of said cost function around the true directions of arrival and the corresponding true signal eigenvectors.

3. Data models and assumptions

Consider a general planar real array of m_r isotropic antenna elements. Then introduce a spherical coordinate system with azimuth ϕ and elevation θ. Assume a Gaussian distributed signal model $\mathbf{s}(t)$ with covariance matrix $\mathbf{S}$, and the usual properties of being both temporally and spatially white.

The array output is modeled as an m_r x N vector function $\mathbf{x}(t)$

$$\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\theta}) \cdot \mathbf{s}(t) + \mathbf{n}(t) \quad \ldots (1)$$

where the columns of the m_r x d matrix $\mathbf{A}(\boldsymbol{\theta})$ transform from the impinging signals $\mathbf{s}(t)$ to the array output $\mathbf{x}(t)$, and $\mathbf{n}(t)$ is the noise contribution from the m_r receiver channels. The p signal parameters, here azimuth and elevation, are collected in the parameter matrix $\boldsymbol{\theta}$ of dimension p x d.

To distinguish the real array from the virtual, we will denote the former $\mathbf{A}_r$ and the latter $\mathbf{A}_v$. The symbols $\hat{\boldsymbol{\theta}}$, $\hat{\phi}$ and $\hat{\theta}$ will be used in the sequel for estimates of the parameter matrix $\boldsymbol{\theta}$, azimuth ϕ and elevation θ respectively.

The array output covariance matrix $\mathbf{R}$ is:

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(t) \cdot \mathbf{x}^*(t)\right\} = \mathbf{A}(\boldsymbol{\theta}) \cdot \mathbf{S} \cdot \mathbf{A}^*(\boldsymbol{\theta}) + \sigma^2 \mathbf{I} \quad \ldots (2)$$

The estimated signal- and noise subspaces $\hat{\mathbf{E}}_s$ and $\hat{\mathbf{E}}_n$, are formed by eigenvalue decomposition of an estimate of $\mathbf{R}$ obtained as

$$\hat{\mathbf{R}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}(t) \cdot \mathbf{x}^*(t) \quad \ldots (3)$$

where N is the number of snapshots.

4. Array transformations

In a practical application it is the output vector $\mathbf{x}_r(t)$ of the real array that is mapped onto the corresponding output vector $\mathbf{x}_v(t)$ of the virtual array, but from (2) this is equivalent to doing the same mapping between the two linear spaces spanned by the vectors in $\mathbf{A}_r$ and $\mathbf{A}_v$.

Usually the transformation matrix $\mathbf{T}$ is designed for a best least square fit between these two spaces over a certain sector

$$\mathbf{T}^* = \arg\min_{\mathbf{T}^*} \left\| \mathbf{T}^* \cdot \mathbf{A}_r(\theta^{(c)}) - \mathbf{A}_v(\theta^{(c)}) \right\|_F^2 \quad \ldots (4)$$

where F denotes the Frobenius norm. The set of directions $\boldsymbol{\theta}^{(c)}$ is chosen denser than the beamwidth of the real array and if $\mathbf{A}_r$ is not known analytically, $\boldsymbol{\theta}^{(c)}$ can preferably be a set of calibrated directions, hence the superscript $^{(c)}$.

Using the pseudo-inverse $\mathbf{A}^\dagger = \mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^{-1}$ the solution to (4) becomes $\mathbf{T}^*_{opt} = \mathbf{A}_v(\boldsymbol{\theta}^{(c)})\mathbf{A}_r^\dagger(\boldsymbol{\theta}^{(c)})$.

The virtual array can preferably be a $\lambda/2$ spaced ULA at all frequencies and always be oriented perpendicular to the bisector of the chosen calibration sector.

If the norm (4) can be put to zero, i. e.

$$\Delta\mathbf{A} = \mathbf{T}^* \cdot \mathbf{A}_r(\boldsymbol{\theta}^{(c)}) - \mathbf{A}_v(\boldsymbol{\theta}^{(c)}) = \mathbf{0} \quad \ldots (5)$$

the mapping will be perfect at the calibrated points $\boldsymbol{\theta}^{(c)}$ and will leave no bias or increased variance there. However, for mapping between dissimilar or wide spaced arrays over large sectors, perfect match will seldom be the case. In general the mapping will then cause both extra variance and bias. The latter may dominate, especially if array geometries are different and bandwidth is stressed.

5. A criterion for the removal of mapping bias

In this section we will derive general principles that govern the magnitude of DOA mapping bias by giving a geometrical interpretation to a Taylor expansion of the DOA estimator cost function derivative. The condition for zero bias is condensed into the criterion (8).

5.1 An analytical description of mapping bias

For all DOA estimators that use a scalar cost function $\mathrm{V}(\boldsymbol{\theta})$, the bias $\Delta\boldsymbol{\theta}$ can be described by the off-set in the position of the zeros of the derivative w.r.t. the searched parameter $\boldsymbol{\theta}$. This derivative will henceforth be denoted $\dot{\mathrm{V}}(\boldsymbol{\theta})$ and the corresponding second derivatives $\ddot{\mathrm{V}}(\boldsymbol{\theta})$.

In the sequel we will analyse one DOA error at a time. Then $\dot{\mathrm{V}}(\boldsymbol{\theta})$ and $\ddot{\mathrm{V}}(\boldsymbol{\theta})$ are scalars and since they also depend on the (complex) vectors in the signal subspace we will use the notation $\dot{\mathrm{V}}(\theta, \mathbf{E}_s)$ and $\ddot{\mathrm{V}}(\theta, \mathbf{E}_s)$ respectively.

For (scalar) DOA errors $\Delta\theta$ a first order Taylor expansion of $\dot{\mathrm{V}}(\hat{\theta}, \hat{\mathbf{E}}_s)$ around the true DOA θ and the vectors in the true signal subspace $\mathbf{E}_s$ takes the form, Brandwood [7]

$$\dot{\mathrm{V}}(\hat{\theta}, \hat{\mathbf{E}}_s) = \dot{\mathrm{V}}(\theta, \mathbf{E}_s) + \ddot{\mathrm{V}}(\theta, \mathbf{E}_s) \cdot \Delta\theta + \ldots$$
$$\ldots + 2 \sum_{i=1}^{d} \mathrm{Re}\left\{ (\nabla_{\mathbf{e}_i} \dot{\mathrm{V}}(\theta, \mathbf{E}_s))^{\mathrm{T}} \Delta\mathbf{e}_i \right\} + \mathrm{rem} \quad \ldots (6)$$

where $\nabla_{\mathbf{e}_i}$ is the complex gradient along vector $\mathbf{e}_i$ in the signal subspace $\mathbf{E}_s$. The remainder term 'rem' is at least quadratic in $\Delta\theta$ and $\Delta\mathbf{E}_s$, and can be neglected if $\dot{\mathrm{V}}(\boldsymbol{\theta})$

behaves "well" and the amount of bias $\Delta\theta$ is limited.

We now observe that $\hat{\theta}$ and $\hat{\mathbf{E}}_s$ are estimated at the extremes of the cost function, i. e. where the scalar entities $\dot{V}(\hat{\theta},\hat{\mathbf{E}}_s)$ and $\dot{V}(\theta,\mathbf{E}_s)$ both equal zero. Hence, using (6) we can express the modulus of the bias $\Delta\theta$ as

$$|\Delta\theta| = \left|\left[\ddot{V}(\theta,\mathbf{E}_s)\right]^{-1}\right| \cdot \left|2\sum_{i=1}^{d} \mathrm{Re}\left\{(\nabla_{\mathbf{e}_i}\dot{V}(\theta,\mathbf{E}_s))^T \Delta\mathbf{e}_i\right\}\right| \quad (7)$$

an entity that we want to minimise.

The leading Hessian inverse is difficult to manipulate but the second term with the gradients and the signal eigenvector mapping errors $\Delta\mathbf{e}_i$ offers a possibility that now will be exploited.

Inside the Re operator each term can be interpreted as the inner product between the conjugate of the gradient of $\dot{V}$ along the i:th signal eigenvector and the mapping error in that particular signal eigenvector. This geometrical interpretation will now be used to form a criterion that errors of all kind must fulfil in order not to result in systematic DOA errors, i. e. bias.

5.2 A criterion for zero mapping bias

In view of the structure of (7) we conclude that a *sufficient condition for zero bias* is

$$\bar{\mathbf{g}}^{(i)} \stackrel{\Delta}{=} \overline{\nabla_{\mathbf{e}_i}\dot{V}_{MU}(\theta,\mathbf{E}_s)} \perp \Delta\mathbf{e}_i, \quad i \in (1,d) \qquad \ldots (8)$$

This means that in a given mapping scenario, for each signal eigenvector of $\mathbf{R}$ the mapping matrix $\mathbf{T}$ should have the property to leave signal eigenvector mapping errors that are orthogonal to the conjugate of the gradient of the cost function $\dot{V}$ in that particular signal eigenvector direction. Where this is the case, the mapping errors do not effect the cost function and hence do not cause bias. This holds for all errors $\Delta\mathbf{e}_i$ that are small enough for the Taylor expansion (6) to be valid.

If we design $\mathbf{T}$ so that the spaces $\Re\left\{\mathbf{T}^*\mathbf{A}_r^{(i)}\right\} \approx \Re\left\{\mathbf{A}_v^{(i)}\right\}$, the mapping error $\Delta\mathbf{A}(\theta^{(i)})$ can be used in stead of $\Delta\mathbf{e}_i$. We can then reformulate the condition (8) into the following more useful requirement:

$$\Delta\mathbf{A}(\theta^{(i)}) \perp \bar{\mathbf{g}}^{(i)}, \quad \forall i = 1,\ldots,N_{cal} \qquad \ldots (9)$$

Generally if $\Delta\mathbf{A}\neq\mathbf{0}$, by maintaining the orthogonalities (9) in some directions, bias can be made arbitrarily small in those directions.

Condition (9) holds for all types of errors, not only mapping errors. It will now be used across large sectors where $N_{cal}>m_r$ and a least square compromise for $\mathbf{T}$ has to be found. This case is of interest in signal surveillance where the DOAs initially are completely unknown.

5.3 A design of T for minimum bias

Since bias is created by the real part of the projection of $\Delta\mathbf{A}(\boldsymbol{\theta}^{(i)})$ onto $\Re\left\{\bar{\mathbf{g}}(\theta)\right\}$ an optimal transformation matrix $\mathbf{T}$ should minimize the real part of all N_{cal} scalar products $\bar{\mathbf{g}}(\theta^{(i)})\cdot\Delta\mathbf{A}(\boldsymbol{\theta}^{(i)})$ across the calibration sector. Taking this into account and introducing the weighting constant k, $0\leq k\leq 1$, the following cost function for the design of $\mathbf{T}$ is proposed

$$\mathbf{T}_{opt}^* = \arg\min_{k,\mathbf{T}^*}\left\{(1-k)\left\|\mathbf{T}^*\cdot\mathbf{A}_r(\theta)-\mathbf{A}_v(\theta)\right\|_F^2 + \ldots\right.$$
$$\left. + k\cdot\sum_{i=1}^{N_{cal}}\left\|\mathrm{Re}\left\{\mathbf{g}^{(i)T}\cdot\mathbf{T}^*\mathbf{A}_r(\theta^{(i)})-\mathbf{g}^{(i)T}\mathbf{A}_v(\theta^{(i)})\right\}\right\|_F^2\right\} \quad (10)$$

It consists of the original manifold matching expression in (4) plus one penalty term for each calibration direction. The latter terms penalize non-orthogonalities in (9) to an amount proportional to k. We expect the optimum k to be close to 1 if bias is to be kept down.

The minimization problem (10) can be solved in many ways. One approach is to apply the vec operator to both terms inside each Frobenius norm and solve for a vectorized version $\mathbf{t}$ of $\mathbf{T}$. In doing so we get a least square solution of the type

$$\mathbf{t} = \begin{bmatrix}\mathrm{vec}(\mathrm{Re}\{\mathbf{T}\})\\ \mathrm{vec}(\mathrm{Im}\{\mathbf{T}\})\end{bmatrix} = \mathbf{M}^{\dagger}\cdot\mathbf{m} \qquad \ldots (11)$$

It involves calculating the pseudo-inverse of the large (full rank) matrix $\mathbf{M}$ of size $(2m_v+1)N_{cal}$ x $2m_vm_r$. If, as a typical example, $m_v=m_r=8$ and $N_{cal}=15$, the size of the matrix is 255 x128. This will not slow down real time operation of a radio DF system however, since all mapping matrices, one for each sector, can be calculated in advance in conjunction with a calibration process.

Note that in the above derivation of the expression (7) no restrictions on the cost function V were made other than the existence of the necessary derivatives. Hence we conclude the bias minimization criterion (9) to be applicable to all DOA estimators that are based on cost functions of this class.

Also note that if the virtual array in combination with the used DOA estimator is exactly bias free, then $\mathbf{A}_v^*(\theta^{(i)})\bar{\mathbf{g}}^{(i)} = 0, \forall i$ and (10) can be somewhat simplified. Since this is the usual case we see that we cannot allow k=1 for then $\mathbf{T}=\mathbf{0}$ would be an obvious, but not wanted, solution to (10).

The real part of the scalar product between the manifold mapping error $\Delta\mathbf{A}$ and the gradient $\bar{\mathbf{g}}$, a product which we want to minimize, can be made small via two mechanisms:

(i) Orthogonality between the two involved vectors
(ii) Small length for at least one involved vector

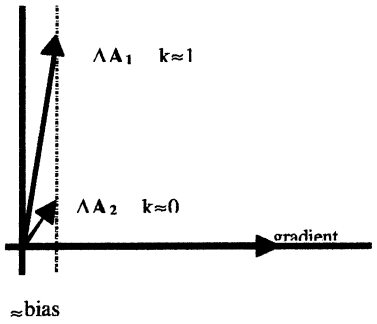


Figure 2: For a given acceptable bias the larger the k, the less control over the size of the mapping error ΔA and the more perfect an orthogonality we have to enforce.

The sum term in (10) secures (i) and the manifold matching term (ii). The weighting constant k thus provides a possibility to find a suitable mix between (i) and (ii).

To illustrate the problem of picking the optimal k we consider the calibration process (carried out against one emitter at a time). Then both the gradient $\bar{\mathbf{g}}$ and the mapping error $\Delta\mathbf{A}$ are m_r x 1 vectors, see figure 2 for a 2D illustration.

From figure 2 we conclude that using a k-value near 1 requires good accuracy in the orthogonality if bias is to be kept down. As for the required SNR, using noise pre-whitening to compensate for the generally non-unitary **T** that results from (10), simulations show 6-10 dB SNR to be sufficient for good bias reduction.

6. MUSIC verifying simulations

In the usual MUSIC power spectrum

$$P_{MU} = \frac{\mathbf{a}^*(\boldsymbol{\theta})\mathbf{a}(\boldsymbol{\theta})}{\mathbf{a}^*(\boldsymbol{\theta})\mathbf{E}_n\mathbf{E}_n^*\mathbf{a}(\boldsymbol{\theta})} \qquad \ldots (11)$$

the null spectrum in the denominator and its derivative can be used as the cost function. Then, for MUSIC the orthogonality condition (8) is the real part of the relation

$$\bar{\mathbf{g}}^{(i)} = \ldots$$
$$\overline{\dot{\mathbf{a}}(\theta^{(i)})\mathbf{e}_i^*\mathbf{a}(\theta^{(i)}) + \bar{\mathbf{a}}(\theta^{(i)})\mathbf{e}_i^*\dot{\mathbf{a}}(\theta^{(i)})} \perp \Delta\mathbf{e}_i, \quad i \in (1, N_{cal}) \qquad \ldots (12)$$

In the simulations to follow a single emitter that sweeps a 30^0 wide calibration sector was used. The sector has 15 equi-spaced calibrated directions and the corresponding 15 calibration response vectors are used in the calculation of **T** according to (10), and, for reference, also according to (4).

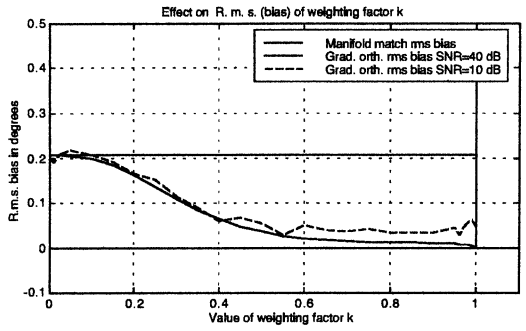


Figure 3: R.m.s. of the all the mapped DOA bias values inside a 30^0 wide sector as function of the weighting constant k with SNR as parameter. The reference line corresponds to pure manifold match. As seen the k-value is not critical.

The number of snapshots N was 100 and SNR per array element rather large, 40 and 10 dB respectively, to highlight mapping effects. Bias values are generated as the difference between the true directions and estimates from the mapped data. Each such estimate uses 400 standard root-MUSIC Monte Carlo runs.

Figure 3 shows the influence of different values for the weighting parameter k on the overall bias across a mapped sector. As seen, in the simulated scenario the particular k-value is not critical.

In figures 3-6 the real array is a circular 8 element UCA spaced at 4λ i. e. well in excess of the usual $\lambda/2$ limit which tends to enlarge the bias. The virtual array is an 8 element ULA spaced at $\lambda/2$ and oriented perpendicular to the bisector of the calibration sector.

The weighting constant k is kept at a rather large value, 0.99, to achieve maximum bias reduction.

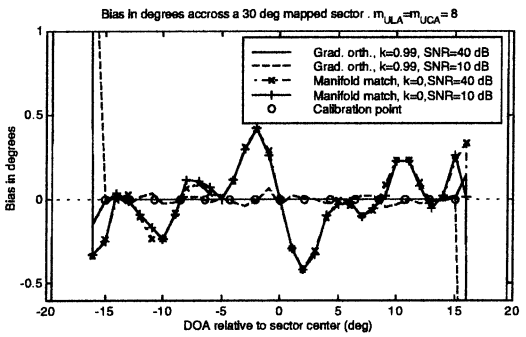


Figure 4: At k=0.99 maximum bias is reduced 50-100 times by the proposed design of the transformation matrix as compared to the pure manifold matching design. See fig. 5 for a magnified version.

From figures 4 and 5 we observe a bias reduction factor of up to 100 at SNR=40 dB. For lower SNRs the reduction becomes smaller but is still useful at 10 dB.

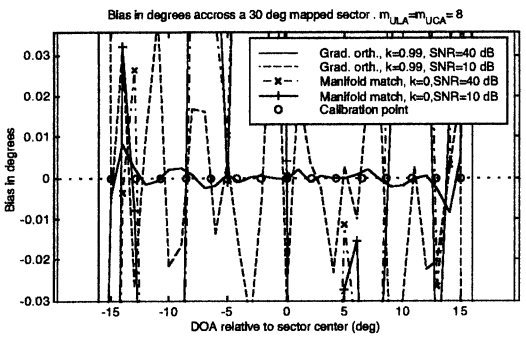


Figure 5. The scenario as in figure 4 but magnified to highlight the small bias that results from the proposed algorithm (10). This algorithm is quite useful also at SNR = 10 dB.

Mapping between less dissimilar arrays generate less bias. In such cases the reduction will also be smaller.

7. Conclusions

When mapping array output signals over sectors we have found that bias can dominate over variance, especially if the mapping is between arrays of dissimilar geometry and bandwidth is stressed.

We then derived (i) a general description of DOA bias, and (ii), a general criterion for its removal. The key feature of the new criterion is that, rather than minimizing the mapping errors, we rotate them into orthogonality with certain gradients of the estimator cost function.

Using the presented theory we have through simulations shown that circular arrays can be operated much in excess of the usual $\lambda/2$ element spacing, and still be mapped over a 30^0 wide sector with negligible bias. The large bias that otherwise would result was reduced by a factor of up to 100 by the proposed transformation matrix design.

8. References

[1] B. Friedlander, "The Root-MUSIC Algorithm for Direction Finding with Interpolated Arrays", *Signal Processing 30*, 1993, pp. 15-29

[2] A. J. Weiss, B. Friedlander, "Preprocessing for Direction Finding with Minimal Variance Degradation", *IEEE Transactions on Signal Processing*, June 1994

[3] A. Weiss, B. Friedlander, P. Stoica, "Direction of Arrival Estimation Using MODE with Interpolated Arrays", *IEEE Transactions of Signal Processing*, January 1995

[4] J. Eriksson, "On Detection and Estimation of Multiple Sources in Radar Array Processing", Thesis, Chalmers University of Technology, Göteborg 2002. ISBN 91-7291-111-5

[5] S. Andersson: , "On Dimension Reduction in Sensor Array Processing", *PhD Thesis*, Linköping University, 1992.

[6] Per Hyberg, " Circular to Linear Array Mapping using Calibration Data". *IEE Eighth International Conference on HF Radio Systems and Techniques*, July 2000, pp. 71-75

[7] Per Hyberg, "Circular to Linear Array Mapping and Bias Reduction", Lic. thesis, Royal Institute of Technology, Stockholm 2001. ISSN 1103-8039

[8] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory", *Proc. Inst. Electr. Eng.*, vol 130, parts F and H, no. 1, pp. 11-16, Feb. 1983

Complex Spreading Sequence Design

Hai Huyen Dam[1], Hans-Jürgen Zepernick[2], Jörgen Nordberg[2] and Sven Nordholm[1]

[1]Australian Telecommunications Research Institute
[2]Australian Telecommunications Cooperative Research Centre
GPO Box U1987, Perth, WA 6845, Australia
E-mail: dam@atri.curtin.edu.au

Abstract

The performance of a code division multiple access system depends on the correlation properties of the employed spreading code. Low cross-correlation values between spreading sequences are desired to suppress multiple access interference. An auto-correlation function with a distinct peak enables proper synchronization and suppresses intersymbol interference. However, these requirements contradict each other and a trade-off needs to be established. In this paper, a global two dimensional optimization method is proposed to minimize auto-correlation with cross-correlation being allowed to lie within a fixed region. This approach is applied to design sets of complex valued spreading sequences.

1 Introduction

In direct sequence code division multiple access (DS-CDMA) systems, each user has its own spreading sequence which distinguishes it from other users. Since mobile stations are spatially separated, the signals arrive at the base station with different delays. Thus, it is not possible to maintain spreading sequences to be synchronized among each other. As a result, multiple access interference (MAI) will be experienced. In addition, multipath propagation causes multiple copies of a transmitted signal and generates intersymbol interference (ISI). Therefore, sequences in a spreading code should have low cross-correlation (CC) values to suppress MAI. The auto-correlation (AC) function of a spreading sequence on the other hand, should have a narrow peak to avoid ISI and to enable proper synchronization. Unfortunately, both correlation requirements contradict each other. A trade-off can be obtained by properly optimizing respective correlation measures of the spreading code in a joint manner.

This paper focuses on the design of spreading sequences with optimized correlation properties for asynchronous CDMA systems. Major focus is given to a family of complex valued polyphase spreading sequences, which offers a wide range of correlation properties. An optimization problem is formulated to minimize average mean square AC, R_{ac}, for a certain range of average mean square CC, R_{cc}. This problem can be formulated as a two dimensional optimization problem with two parameters m and n. A modified bridging method is proposed to minimize R_{ac} for a certain range of R_{cc}.

2 Correlation Measures

Let N denote the length of the spreading sequences. For a pair of sequences $u_k = [u_k(0), \ldots, u_k(N-1)]$ and $u_l = [u_l(0), \ldots, u_l(N-1)]$, the aperiodic correlation between u_k and u_l is defined as

$$C_{k,l}(i) = \begin{cases} \frac{1}{N} \sum_{v=0}^{N-1-i} u_k(v) u_l^*(v+i), & 0 \le i \le N-1 \\ \frac{1}{N} \sum_{v=0}^{N-1+i} u_k(v-i) u_l^*(v), & 1-N \le i < 0 \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

where '*' denotes the complex conjugate of a complex variable.

As far as performance analysis for wireless communication systems is concerned, the focus has been changed from periodic to aperiodic correlation measures, e.g. [1, 2]. Instead of worst case scenarios extracted from maximum side lobes of periodic correlations, it seems to be more reasonable to consider mean square values of AC and CC of spreading sequences. Especially, the latter measure is much more appropriate for asynchronous CDMA systems in which the average bit error rate (BER) and eventually the capacity of the system depends on the average interference caused by aperiodic CC properties of the employed set of spreading sequences. Thus, this work is focused on optimizing average mean square aperiodic AC and average mean square aperiodic CC characteristics of spreading sequences.

Denote S as the number of sequences taken from a spreading code. The average mean square AC for S sequences $u_1, \cdots, u_S$ is defined as

$$R_{ac} = \frac{1}{S}\sum_{k=1}^{S}\sum_{i=1-N, i\neq 0}^{N-1} |C_{k,k}(i)|^2, \quad (2)$$

and the average mean square CC is given by

$$R_{cc} = \frac{1}{S(S-1)}\sum_{k=1}^{S}\sum_{l=1, l\neq k}^{S}\sum_{i=1-N}^{N-1} |C_{k,l}(i)|^2. \quad (3)$$

3 Complex Spreading Sequences

A family of complex spreading sequences with a wide range of correlation properties has been proposed in [1]. For $1 \leq k \leq N-1$, the v^{th} element $u_k(v)$ of a sequence u_k is defined by:

$$u_k(v) = (-1)^{k(v+1)} \exp\left[\frac{j\pi(k^m(v+1)^p + (v+1)^n)}{N}\right], \quad (4)$$

where $0 \leq v \leq N-1$ and parameters m, n and p take real values. For different combinations of m, n and p, the sequences have different correlation properties.

The mean square AC and the mean square CC are defined for a fixed combination of m, n and p. Thus, they are functions of the parameters m, n and p. The following theorems give some properties for the correlations of the sequences given in (4).

Theorem 1 *The aperiodic correlation is a continuous function of m, n and p for all k, l and i.*

Proof: This follows from (1) since $u_k(v)$, $u_l(v+i)$, $u_k(v-i)$ and $u_l(v)$ are continuous functions of m, n and p.

Theorem 2 *Average mean square AC and average mean square CC are continuous functions of m, n and p.*

Proof: This follows from (2) and (3) since $C_{k,l}(i)$ is a continuous function of m, n and p for all k, l and i.

Theorem 3 *For a fixed combination of (m,n,p), all the sequences have the same AC magnitude. This magnitude depends only on n if $p=1$, and is given $\forall\, k,i$ by [1]*

$$|C_{k,k}(i)| = \left|\frac{1}{N}\sum_{v=0}^{N-1-i} \exp\left[\frac{j\pi}{N}((v+1)^n - (v+i+1)^n)\right]\right|. \quad (5)$$

Note that $i=0$ gives

$$|C_{k,k}(0)| = 1.$$

Proof: For $1 \leq k \leq N-1$, the auto-correlation $C_{k,k}(i)$ for a sequence u_k is derived by substituting $l=k$ and $p=1$ to (1) and (4). We have

$$\begin{aligned} C_{k,k}(i) &= \tfrac{1}{N}\sum_{v=0}^{N-1-i} u_k(v)u_k^*(v+i) \\ &= \tfrac{1}{N}\sum_{v=0}^{N-1-i} (-1)^{2kv+2k+ki} \\ &\times \exp\left[\tfrac{j\pi}{N}(k^m(v+1) + (v+1)^n)\right] \\ &\times \exp\left[\tfrac{-j\pi}{N}(k^m(v+i+1) + (v+i+1)^n)\right] \end{aligned} \quad (6)$$

for $0 \leq i \leq N-1$ and

$$\begin{aligned} C_{k,k}(i) &= \tfrac{1}{N}\sum_{v=0}^{N-1+i} u_k(v-i)u_k^*(v) \\ &= \tfrac{1}{N}\sum_{v=0}^{N-1+i} (-1)^{2kv+2k-ki} \\ &\times \exp\left[\tfrac{j\pi}{N}(k^m(v-i+1) + (v-i+1)^n)\right] \\ &\times \exp\left[\tfrac{-j\pi}{N}(k^m(v+1) + (v+1)^n)\right] \end{aligned} \quad (7)$$

for $1-N \leq i < 0$.

It follows from (6) and (7) that

$$C_{k,k}(i) = \begin{cases} \frac{(-1)^{ki}}{N}\exp\left[\frac{-j\pi}{N}k^m i\right] \\ \times \sum_{v=0}^{N-1-i}\exp\left[\frac{j\pi}{N}((v+1)^n - (v+i+1)^n)\right] \\ \text{for } 0 \leq i \leq N-1, \\ \frac{(-1)^{ki}}{N}\exp\left[\frac{-j\pi}{N}k^m i\right] \\ \times \sum_{v=0}^{N-1+i}\exp\left[\frac{j\pi}{N}((v-i+1)^n - (v+1)^n)\right] \\ \text{for } 1-N \leq i < 0. \end{cases} \quad (8)$$

Thus, the auto-correlation magnitude depends only on the parameter n and is given by

$$|C_{k,k}(i)| = \left|\frac{1}{N}\sum_{v=0}^{N-1-i}\exp\left[\frac{j\pi}{N}((v+1)^n - (v+i+1)^n)\right]\right| \quad (9)$$

with $0 \leq i \leq N-1$ and

$$|C_{k,k}(i)| = |C_{k,k}(-i)|, \quad 1-N \leq i < 0. \quad (10)$$

For $i=0$, we have $|C_{k,k}(0)| = 1$.

In view of Theorem 3, we will concentrate only on the case $p=1$.

4 Problem Formulation

Sequences with low mean square AC approximate a delta function in time domain. Thus, the frequency spectrum of these sequences is wide and flat as desired for CDMA systems. This characteristic comes at the

expense of the mean square CC values, which are large between different sequences in the set when the mean square AC is low. In most situations, however, an acceptable region for R_{cc} can be defined for sufficient cross-correlation. Then, the problem becomes:

For a fixed region of R_{cc}, obtain a set of parameters m and n so that R_{ac} is minimized with the corresponding R_{cc} lying within a bound α.

The respective optimization problem is formulated as

$$\begin{cases} \min\limits_{m,n} R_{ac} \\ \text{subject to } R_{cc} \leq \alpha. \end{cases} \tag{11}$$

Note that (11) is a two dimensional optimization problem with two continuous variables m and n. The problem has a non-linear continuous cost function and a non-linear continuous constraint. A modified approach based on the idea of using the bridge function for one dimensional problem [3, 4] is proposed to solve this two dimensional optimization problem with a non-convex constraint.

5 Global Two Dimensional Optimization Method

The main idea of the bridging method is to use the bridged functions in the search for the global minimum so that the method does not get stuck at a local minimum point. The bridging method finds a local minimum point of the bridged function starting from one end of the feasible region. The bridged function is then updated every time a local solution is obtained until the global minimum of the original cost function is found.

By incorporating a local search method such as the golden section search or the parabolic interpolation [5, 6], the modified bridging method can be improved so that no smoothing procedure is needed. Thus, it is applicable to cost functions which are only continuous instead of continuously differentiable. The modified algorithm is developed to solve the two dimensional global optimization problem (11) with a general form

$$\begin{cases} \min\limits_{m,n} f(m,n) \\ \text{subject to } g(m,n) \leq \alpha, \end{cases} \tag{12}$$

where $m \in [m_1, m_2]$, $n \in [n_1, n_2]$ and $f(m,n)$, $g(m,n)$ are continuous functions of m and n. For a fixed value of m, (12) becomes a one dimensional global optimization problem, which can be solved by extending the bridging method [3, 4] to include a non-linear constraint.

The algorithm starts by searching for the first feasible point (m_0, n_0) of the problem (12), i.e. $m_0 \in [m_1, m_2]$, $n_0 \in [n_1, n_2]$ and $g(m_0, n_0) \leq \alpha$. This is done as follows:

Procedure 1 *Search for feasible point from (m_0, n_0)*

1. *Specify step size h.*
2. *Fix the value of m_0 and search for the left most feasible solution of (12). This is done by finding the smallest value of ν such that*

$$g(m_0, n_0 + \nu h) \leq \alpha. \tag{13}$$

3. *If ν exists, set $n_0 = n_0 + \nu h$ and go to Step 4.*

   If ν does not exist, then

   • *If $m_0 + h \leq m_2$, set $m_0 = m_0 + h$, $n_0 = n_1$ and go back to Step 2.*

   • *If $m_0 + h > m_2$, the optimization is too restrictive and no feasible solution exists from (m_0, n_0). Go to Step 4.*
4. *Stop the procedure.*

The value of (m_0, n_0) will be used as the initial search point for problem (12). For each (m_0, n_0), define the following minimization problem:

$$\begin{cases} \min\limits_{n \in [n_0, n_2]} b(m_0, n) \\ \text{subject to } g(m_0, n) \leq \alpha, \end{cases} \tag{14}$$

where $b(m_0, n) = \min\{f(m_0, n), f(\widehat{m}, \widehat{n})\}$. The function $b(m_0, n)$ is called the bridged function and $f(\widehat{m}, \widehat{n})$ gives the value of the cost function $f(m, n)$ at the local minimum $(\widehat{m}, \widehat{n})$. The modified bridging algorithm incorporates Procedure 1 and is given as follows:

Procedure 2 *Modified bridging method*

1. *Initialize $(m_0, n_0) = (m_1, n_1)$.*
2. *Obtain the first feasible point from (m_0, n_0) by using Procedure 1. If this feasible point does not exist, go to Step 5. Otherwise, update (m_0, n_0) as this feasible point. If the optimum solution has not been initialized, then set $(\widehat{m}, \widehat{n}) = (m_0, n_0)$.*
3. *Find the left most local optimum solution $(\widehat{m}, \widehat{n})$ of problem (14). This is done by searching for the smallest value of ν such that*

$$\begin{aligned} b(m_0, n_0 + \nu h) &< b(m_0, n_0 + (\nu+1)h), \\ n_0 + (\nu+1)h &\leq n_2 \end{aligned} \tag{15}$$

   or

$$g(m_0, n_0 + \nu h) > \alpha. \tag{16}$$

4. *If ν exists, then*

   • *If (15) occurs, then $b(m_0, n_0 + \nu h) \leq b(m_0, n_0 + (\nu - 1)h)$ and a local minimum with $m = m_0$ and $n \in [n_0 + (\nu - 1)h, n_0 + (\nu + 1)h]$ exists. The two dimensional local minimization method [5] can be used to locate the local minimum solution*

around this point. Update the local minimum $(\widehat{m},\widehat{n})$. *Set* $n_0 = n_0 + (\nu+1)h$. *Go back to Step 3.*

- *If (16) occurs and* $f(m_0, n_0 + (\nu-1)h) < f(\widehat{m},\widehat{n})$, *then update* $(\widehat{m},\widehat{n}) = (m_0, n_0 + (\nu-1)h)$, *set* $n_0 = n_0 + \nu h$ *and return to Step 2.*

If ν *does not exist, then*

- *If* $m_0 + h \leq m_2$, *assign* $m_0 = m_0 + h$ *and* $n_0 = n_1$ *and return to Step 2.*
- *If* $m_0 + h > m_2$, $(\widehat{m},\widehat{n})$ *is the global solution and go to Step 5.*

5. *Stop the procedure.*

The solution obtained from Procedure 2 is the global minimum solution of the function $f(m,n)$ over the interval $[m_1, m_2] \times [n_1, n_2]$ for a small step size h. By introducing the bridged function at (m_0, n), we can eliminate local minima of the original cost function $f(m,n)$ with values larger than $f(\widehat{m},\widehat{n})$.

The main advantage of this algorithm is that it only finds a small number of useful local minima. Once a global solution is obtained, the procedure moves quickly through the rest of the interval.

6 Design Example

Let us consider the design of a set of $S = 30$ complex spreading sequences with length $N = 31$ based on the construction rule given in (4) with parameter $p = 1$.

Results for this design example are shown in Table 1. The minimum value for R_{ac} and its corresponding parameters m and n are given for different bounds on R_{cc}. The first and the last rows are for the cases of optimum R_{cc} and R_{ac}, respectively. Depending on the requirements of a particular application, the table can be used to find a trade-off between R_{ac} and R_{cc}.

Figs. 1–4 plot the auto-correlation functions and the frequency spectra for the two extreme cases given in Table 1. Since we are only interested in the interaction of the spectrum for different sequences, all the spectrum plots are normalized. The spectrum for three sequences out of a complex spreading code are plotted, where (m,n) specifies the code.

Figs. 1 and 2 plot the auto-correlation function and the frequency spectra for three sequences taken from the spreading code with minimum R_{cc}. It can be seen that the auto-correlation function is broad, covering several lags while the overlap between the frequency spectra is low.

Figs. 3 and 4 plot the auto-correlation function and the frequency spectra for the extreme case with lowest possible $R_{ac} = 0.1107$ but very high $R_{cc} = 0.9984$. In practice, this characteristic would ease the synchronization task. The overlapping spectrum on the other hand, would cause some interference among users in an asynchronous CDMA system. This may result in poor system performance in terms of average bit error rate. Thus, given the system specification, the complex spreading sequence can be designed accordingly to obtain the best correlation properties for the particular application.

Table 1: Optimized R_{ac} for different upper bounds α on R_{cc} for $S = 30$ sequences of length $N = 31$.

R_{cc}	m	n	R_{ac}	R_{cc}
Optimum R_{cc}	1.0038	1.0000	19.6774	0.3419
$R_{cc} \leq 0.4$	1.0038	1.3000	17.9373	0.4000
$R_{cc} \leq 0.5$	1.0038	1.3837	15.2413	0.4900
$R_{cc} \leq 0.6$	1.0038	1.4468	11.9485	0.6000
$R_{cc} \leq 0.7$	1.0037	1.4966	8.9562	0.7000
$R_{cc} \leq 0.8$	1.0036	1.5573	5.9707	0.8000
$R_{cc} \leq 0.9$	1.0032	1.6807	2.9888	0.9000
Optimum R_{ac}	2.9200	2.0072	0.1107	0.9984

7 Conclusion

In this paper, we investigated the design of spreading sequences with minimum average mean square auto-correlation for an upper bound on the average mean square cross-correlation. This problem was formulated as a two dimensional global optimization problem with a non-linear cost function and a non-convex constraint. The modified bridging method was proposed for solving this optimization problem and was applied to the design of complex spreading sequences. The method may be used for sequence design in asynchronous CDMA systems.

References

[1] I. Oppermann and B.S. Vucetic, "Complex Spreading Sequences with a Wide Range of Correlation Properties," *IEEE Trans. Commun.*, vol. 45, no. 3, pp. 365–375, Mar. 1997.

[2] I. Oppermann, "Orthogonal Complex Valued Spreading Sequences with a Wide Range of Correlation Properties," *IEEE Trans. on Commun.*, vol. 45, no. 11, pp. 1379–1380, Nov. 1997.

[3] Y. Liu and K. L. Teo, "A Bridging Method for Global Optimization," in *Optmization Techniques and Applications*, vol. 2, pp. 558–565, 1998.

[4] H. H. Dam, K. L. Teo, Y. Liu, and S. Nordebo, "Optimum Pole Position for Digital Laguerre Network with Least Square Error Criterion," *Optimization Methods and Applications*, pp. 321–330, Boston: Kluwer Academic Publishers, 2000.

[5] Matlab: "Reference Guide," The Math Works, 1996.

[6] M. Minoux, "Mathematical Programming Theory and Applications," New York: Wiley, 1986.

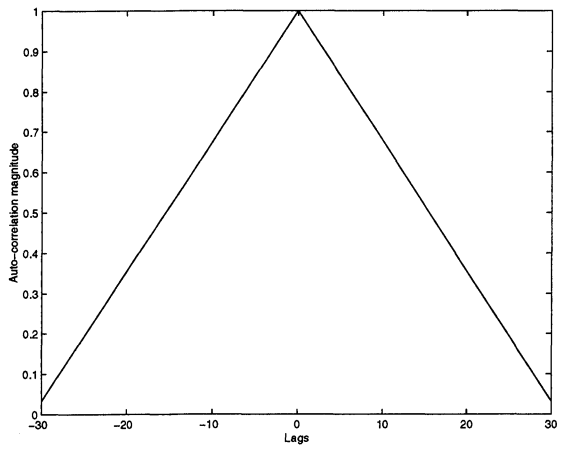


Fig. 1. Magnitude of AC for $m = 1.0038$, $n = 1.0000$ with optimum $R_{cc} = 0.3419$.

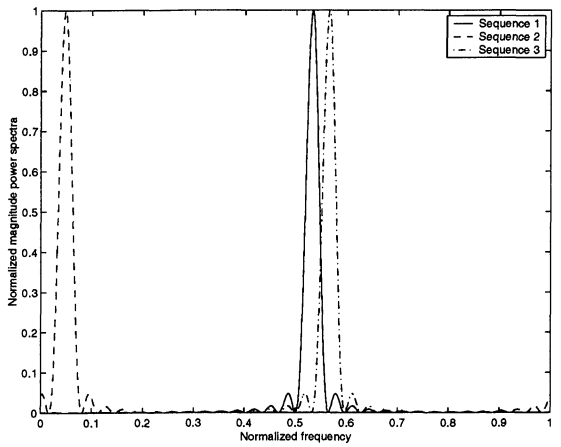


Fig. 2. Frequency spectra for $m = 1.0038$, $n = 1.0000$ with optimum $R_{cc} = 0.3419$.

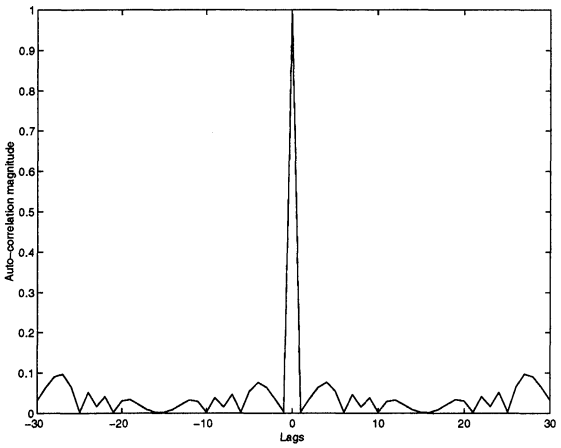


Fig. 3. Magnitude of AC for $m = 2.9200$, $n = 2.0072$ with optimum $R_{ac} = 0.1107$.

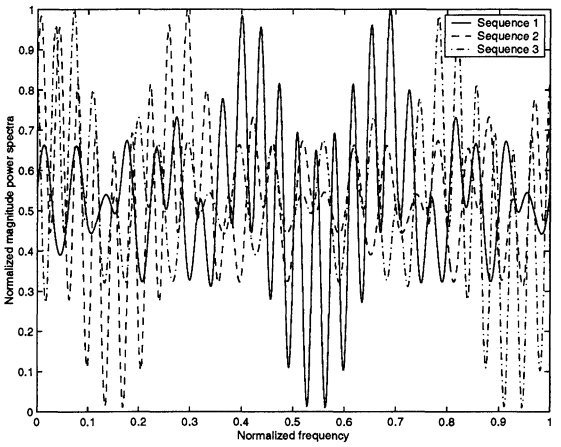


Fig. 4. Frequency spectra for $m = 2.9200$, $n = 2.0072$ with optimum $R_{ac} = 0.1107$.

Lower Symbol Error Probability Bounds for CPM Signaling over AWGN Channels

Zihuai Lin and Tor Aulin

Telecommunication Theory Group, Dept. of Computer Engineering, Chalmers Univ. of Technology, 412 96 Göteborg.
E-mail: {zihuai,tor}@ce.chalmers.se

Abstract

Bounds on the symbol error probability for Viterbi detected Continuous Phase Modulation (CPM) signals are considered. The detector is coherent and the channel is an Additive White Gaussian Noise (AWGN) channel. A new method for constructing a lower bound for CPM systems is suggested. It is shown that the lower bound can be substantially improved compared to previous approaches by considering the entire set of transmitted data sequences and not only the worst case. This generalized algorithm might be applied to any system which can be described as a £nite state machine.

1 Introduction

The class of Continuous Phase Modulation (CPM) schemes has gained attention in recent years due to its attractive properties such as spectral ef£ciency and a constant envelope [1]. The constant envelope property makes it possible to use non-linear ampli£ers, which have lower cost and signi£cantly higher power ef£ciency. Due to the memory in the CPM signal, a trellis diagram can be used to characterize it. Thus, the Viterbi Algorithm (VA) can be exploited for Maximum Likelihood Sequence Detection (MLSD) in AWGN.

In this paper, upper and lower bounds on the symbol error probability for coherently Viterbi detected CPM signals have been investigated. The upper bound is based on the analysis method in [2], using the well-known transfer function technique, previously used for other modulation/coding schemes [3]. The result is the union bound.

The lower bound is based on the genie's argument that gives the receiver side information [4, 5]. This was used originally for constructing a lower bound for Intersymbol Interference (ISI) channels. In [4], an asymptotically tight lower bound was obtained by considering the two transmitted sequences that give the overall minimum squared Euclidean distance d^2_{min}, and weighting the resulting error probabilities with the *a priori* probability of those sequences. As pointed out in [6], the Forney's genie argument used to develop the lower bound in [4] is ¤awed if the side information scheme introduces *a priori* bias between the two transmitted sequences before the channel output is observed. Instead, a Uniform Side Information (USI) scheme was proposed in [5] to construct the lower bound for ISI-Like channels. This lower bound can be improved if the entire set of transmitted sequences is considered [7]. The new method for constructing the lower bound can also be used in our system, because a CPM signal on an AWGN channel can be described using a trellis diagram as well.

In section II, CPM systems are brie¤y described. Upper and lower bounds on the symbol error probability for coherently Viterbi detected CPM signals are analyzed in section III and numerical and simulation results are presented in section IV. Finally, section V gives some conclusions.

2 CPM System Description

For an M-ary CPM system, the transmitted signal can be described [1] by

$$s(t,\boldsymbol{\alpha}) = \mathrm{Re}\left\{\sqrt{\frac{2E}{T}}e^{j(2\pi f_c t+\phi(t,\boldsymbol{\alpha}))}\right\} \qquad (1)$$

where the information carrying phase is

$$\phi(t,\boldsymbol{\alpha}) = 2\pi h\sum_i \alpha_i q(t-iT) \qquad (2)$$

and where h is the modulation index. The data symbols in the sequence $\boldsymbol{\alpha} = \alpha_0, \alpha_1, \cdots$ are assumed to be statistically independent and identically distributed (i.i.d) and $\alpha_i \in \{\pm 1, \pm 3, \cdots, \pm(M-1)\}$ with equal probability, where $M \in \{2, 4, 8, \cdots\}$. E is the symbol energy, and $E = \log_2 M \cdot E_b$, where E_b is the bit energy. T and f_c is the symbol interval duration and the carrier frequency, respectively. The phase response $q(t)$ is de£ned by

$$q(t) = \int_{-\infty}^{t} g(\tau)d\tau; -\infty < t < \infty \qquad (3)$$

where $g(t)$ is the frequency response of duration L symbol intervals. The raised cosine (RC) subclass of CPM systems

$$g(t) = \begin{cases} \frac{1}{2LT}\left(1-\cos\frac{2\pi t}{LT}\right); & 0 \le t \le LT \\ 0; & \text{otherwise} \end{cases}$$

is used for this investigation. The modulation index is assumed to be rational, $h = 2k/p$, $(k, p = 1, 2, \cdots)$. As pointed out in [2], for any rational modulation indexes, the CPM signal can be modelled as a £nite state time-invariant discrete time Markov process.

The detector used in our simulation is a coherent Viterbi detector which is described in [1, 2]. The receiver computes the metric correlation increment $Z_n(\tilde{\boldsymbol{\alpha}}_n, \tilde{\theta}_n)$ for all M^L possible sequences $\tilde{\boldsymbol{\alpha}}_n = \tilde{\alpha}_n, \tilde{\alpha}_{n-1}, \ldots, \tilde{\alpha}_{n-L+1}$ and all p possible phase states $\tilde{\theta}_n$. This makes pM^L different values of

$$Z_n(\tilde{\boldsymbol{\alpha}}_n, \tilde{\theta}_n) = \int_{nT}^{(n+1)T} \{r(t, \boldsymbol{\alpha}) \cos(2\pi f_c t + \theta(t, \tilde{\boldsymbol{\alpha}}_n) + \tilde{\theta}_n)\} dt \quad (4)$$

caused by the transmitter, where

$$\theta(t, \tilde{\boldsymbol{\alpha}}_n) = 2\pi h \sum_{i=n-L+1}^{n} \tilde{\alpha}_i q(t - iT)$$

and the phase state

$$\tilde{\theta}_n = \pi h \sum_{i \le n-L} \tilde{\alpha}_i$$

see [2] for details, and where $r(t, \boldsymbol{\alpha})$ is the received CPM signal,

$$r(t, \boldsymbol{\alpha}) = s(t, \boldsymbol{\alpha}) + n(t) \quad (5)$$

where $n(t)$ is a Gaussian random process having zero mean and the autocorrelation function $E[n(t)n(t+\tau)] = \frac{N_0}{2}\delta(\tau)$ and $\boldsymbol{\alpha}$ is the transmitted information sequence.

3 Symbol Error Probability Bounds

The error performance of a Viterbi detected CPM system is more dif£cult to analyze than e.g. systems using the VA to detect convolutional codes with memoryless modulation such as PSK. The reason is that CPM signals are non-linearly modulated with memory and it is not possible in the analysis to use Hamming distance [8]; moreover, the error probability for VA detected CPM systems depends on the transmitted data sequence. Regardless of these dif£culties, our analysis method still results in a tight upper and lower bound. Thus, the true symbol error probability is bracketed with high accuracy.

Ideally, the MLSD receiver must defer its decision until the entire sequence has been processed. This results in an unbounded decoding delay. In practice, it is usually desirable to reduce the delay as far as possible without seriously degrading performance. With the VA, this is done by truncating the path memory to N_T symbols, where N_T is typically a few constraint lengths. The truncation effect for CPM systems has been studied in [2]. In this paper, for the analysis of the upper bound on symbol error probability, we assume $N_T = \infty$.

3.1 Upper Bounds

The upper bound for CPM systems is investigated in [2]. This is based on the well-known transfer function technique, brie¤y summarized below.

The upper bound on error event probability is given by

$$P(\epsilon) \le \sum_{i=1}^{\infty} r_i Q(\sqrt{d_i^2 \frac{E_b}{N_0}}) \quad (6)$$

and the upper bound on the symbol error probability is given by

$$P_s \le \sum_{i=1}^{\infty} r_i v_i Q(\sqrt{d_i^2 \frac{E_b}{N_0}}) \quad (7)$$

where the factor r_i accounts for the dependence upon the transmitted data symbol, v_i is the number of symbol errors for ith path from V_s to V_t, (V_s and V_t are the start and terminal state of the error event state graph, respectively), and where E_b/N_0 is the signal-to-noise ratio per transmitted channel bit, and the Q function is de£ned as $Q(x) = (\sqrt{2\pi})^{-1} \int_x^\infty e^{-z^2/2} dz$. The quantities d_i^2 are the normalized squared Euclidean distances for the respective error events [2],

$$d_i^2 = \frac{1}{2E_b} \int [s(t, \boldsymbol{\alpha_i}) - s(t, \hat{\boldsymbol{\alpha}}_i)]^2 dt. \quad (8)$$

For CPM, this simpli£es to

$$\begin{aligned} d_i^2 &= \frac{1}{T} \int [1 - \cos(\phi(t, \boldsymbol{\alpha_i}) - \phi(t, \hat{\boldsymbol{\alpha}}_i))] dt \cdot \log_2 M \\ &= \frac{1}{T} \int [1 - \cos \phi(t, \boldsymbol{\alpha_i} - \hat{\boldsymbol{\alpha}}_i)] dt \cdot \log_2 M \end{aligned} \quad (9)$$

where $\boldsymbol{\alpha_i}$ and $\hat{\boldsymbol{\alpha}}_i$ are two possible transmitted information sequences.

3.2 Lower Bounds

In this subsection, we present a tight lower bound on the symbol error probability for CPM systems on the AWGN channel. Usually, lower bounds on the error probability for systems with memory (e.g. ISI like channel) are based on Forney's "friendly genie" that gives the receiver side information [4]. Since the receiver is optimum and has side information (from a genie), no actual detection scheme can do any better. As pointed out in [5, 6], the above reasoning is invalid if the side information introduces *apriori* bias between a pair of the transmitted sequences before observing the channel output. Instead, a lower bound based on USI scheme is suggested in [5].

Consider a transmitted data sequence $\boldsymbol{\alpha}$ and a receiver candidate sequence $\hat{\boldsymbol{\alpha}}$, with difference sequence $\boldsymbol{\gamma}$ de£ned by $\boldsymbol{\gamma} = \boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}$. Let ϑ denotes the side information. The USI is that either $\boldsymbol{\alpha}$ or $\hat{\boldsymbol{\alpha}}$ is the correct sequence, i.e., $P(\vartheta = \{\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}\}|\boldsymbol{\alpha}) = P(\vartheta = \{\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}\}|\hat{\boldsymbol{\alpha}}) = 1$. Under the

Maximum Likelihood (ML) sequence decision rule, the error probability is,

$$P(\epsilon) = Q(\sqrt{d^2 \cdot E_b/N_0}) \qquad (10)$$

where d^2 is the normalized squared Euclidean distance for the difference sequence γ.

Suppose now that the receiver was only presented with two sequences α and $\hat{\alpha}$. We know from the friendly genie that one of them is correct (transmitted). The detection problem will then reduce to a simple two hypothesis testing problem in white Gaussian noise. The error probability of this is given by equation (10). This is now a lower bound to the true error probability. By selecting α and $\hat{\alpha}$ such that d^2 is minimized, this lower bound is as large as possible. This is the error event which results in the minimum squared Euclidean distance d^2_{min}. By weighing the *a priori* probability $P(\alpha)$ of the transmitted sequence which corresponds to the error event, we get a lower bound on the average error probability, i.e., $P_s(\epsilon) \geq P(\alpha)P(\epsilon)$.

We now de£ne the set of transmitted sequences for distance d, as all sequences α that have a neighbor $\hat{\alpha} \in \zeta(\alpha)$ at squared Euclidean distance d^2 and denote this set by $\kappa(d^2)$, where $\zeta(\alpha)$ denotes the set of all valid different data sequences. We partitioning the set $\zeta(\alpha)$ into many subsets $\kappa(d^2)$ as shown in Figure 1, where $\kappa(d_1^2) \cap \kappa(d_2^2) \cap \kappa(d_3^2) \cap \kappa(\cdots) = \emptyset$ and $\kappa(d_1^2) \cup \kappa(d_2^2) \cup \kappa(d_3^2) \cup \kappa(\cdots) = \zeta(\alpha)$. Here $d_1^2, d_2^2, d_3^2, \cdots$ denote the normalized squared Euclidean distance for the respective error events in increasing order, i.e., $d_1^2 < d_2^2 < d_3^2 < \cdots$. Then, we can see that Forney's lower bound based on USI was actually constructed by only using the data sequences within the set $\kappa(d^2_{min})$. Now assume the friendly genie does this for every possible set $\kappa(d^2)$ of transmitted sequences for distance d, since these sets are mutually exclusive, by weighing the *a priori* probabilities of the sequences in these sets, an average lower bound (among all of the transmitted sequences) on the symbol error probability is obtained

$$P_s \geq \sum_{d^2 \geq d^2_{min}} B_d^2 \cdot Q(\sqrt{d^2 \cdot E_b/N_0}) \qquad (11)$$

where B_d^2 denotes the $P(\alpha)$ for transmitted sequence α in the set $\kappa(d^2)$ whose most probable error events ϵ have squared Euclidean distance d^2. Since there is no actual detection scheme can do any better than the optimum and USI-aided, equation (11) must be the lower bound for any scheme.

A trellis search algorithm for constructing the improved lower bound on the symbol error probability for ISI channels is given in [7]. The method can also be applied to CPM. The algorithm in [7] starts with £nding all error events, and then registers the transmitted sequence from these events. Care must be taken in that algorithm, because one error event might correspond to several transmitted sequences. If the same transmitted sequence is counted more than once, double counting [1] results and the lower bound

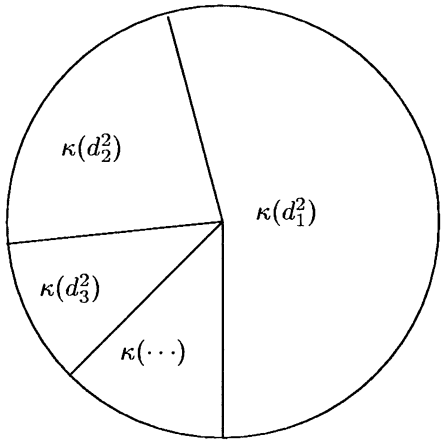


Figure 1: The partitioning of the set $\zeta(\alpha)$ of transmitted data sequences.

becomes useless. In this paper, we give a new search procedure for £nding the error events from the transmitted sequences. By doing this, many problems [7] are avoided. The procedure is listed below,

- Find the set $\kappa(d^2_{min})$ of the transmitted data sequences which results in an error event with minimum Euclidean distance d^2_{min}.
- Take away the set $\kappa(d^2_{min})$ from the entire set $\zeta(\alpha)$, start with the sequence which there is only one symbol difference from the one in the set $\kappa(d^2_{min})$, £nd the set $\kappa(d_1^2)$ of the transmitted sequence which results in a shortest error event with Euclidean distance d_1^2.
- Find the set $\kappa(d^2)$ for new distance d^2 by taking away the sets of the transmitted sequences which already found.
- For each set $\kappa(d^2)$, update the bound with the d^2, B_{d^2} of those sequences.
- When all transmitted sequences are registered, the bound is complete and can be calculated with equation (11).

4 Numerical and Simulation Results

Computer simulations have been performed for three CPM schemes. The £rst one is the binary 2RC system with modulation index $h = 1/2$ when $N_T = 50$ (where N_T is the path memory truncation length for the Viterbi detector). The simulation continued until 10000 symbol errors occurred for every considered SNR. For this system, the normalized minimum squared Euclidean distance is 1.9666 when the observation interval $N \geq N_B = 3$, where N_B is the observation symbol intervals which needed to reach the upper bound on the minimum Euclidean distance [1].

The components used for constructing the improved lower bound are listed in the £rst row of Table I. Figure 2 shows the lower and upper bounds on the symbol error probability when $N_T = \infty$. The simulation result is shown by stars, and the true symbol error probability for binary PSK is shown for comparison. It can be noted that our lower bound is signi£cantly improved when it is compared with the Forney's lower bound based on USI. The simulation result is close to our improved lower bounds. The reason is that the next to the minimum Euclidean distance is quite close to the minimum distance (see Table I), and at the range of considered SNRs, the error events corresponding to these two Euclidean distances occur with a very high probability.

By increasing the modulation index to $h = 3/4$, the minimum distance for the Viterbi detector increases to 2.6559 and N_B increases to 4. The early part of the distance spectrum for this scheme is listed in the second row of Table I. Figure 3 shows the symbol error probability bounds when the path memory is in£nite. The simulation result is shown by stars, it is obtained by running the simulation until 15000 symbol errors occurred for 0, 1, 2 dB and 5000 symbol errors for the rest of the considered SNRs. As before, the true symbol error probability for binary PSK is shown for comparison. There is a gain of about 1.23 dB in SNR compared to BPSK, and this is because the minimum Euclidean distance for BPSK is 2.

The last scheme is quaternary 2RC partial response system with $h = 3/4$. In this case, $N_B = 4$, since one symbol contains two bits, the bit observation interval is 8. Figure 4 shows the upper and lower bounds on the symbol error probability when a Viterbi detector having an in£nitely long path memory is used. The lower bound on the symbol error probability for the quaternary PSK modulation scheme is also shown for comparison. The normalized minimum squared Euclidean distance for this scheme with $N \geq N_B$ is 3.7094, and the minimum distance for 4PSK is 2. The components used for constructing the new lower bounds for this scheme are listed in the last row of table I.

Table I. The initial part of the distance spectrum for the considered CPM schemes

CPM Schemes	Components of lower bound B_{d^2}
$2RC, M = 2, h = 1/2$	$B_{1.9666} = 1/2$ $B_{2.0345} = 1/2$
$2RC, M = 2, h = 3/4$	$B_{2.6559} = 1/2$ $B_{3.6947} = 1/4$ $B_{4.0732} = 1/8$ $B_{4.6947} = 1/8$
$2RC, M = 4, h = 3/4$	$B_{3.7094} = 1/2$ $B_{4.4336} = 1/2$

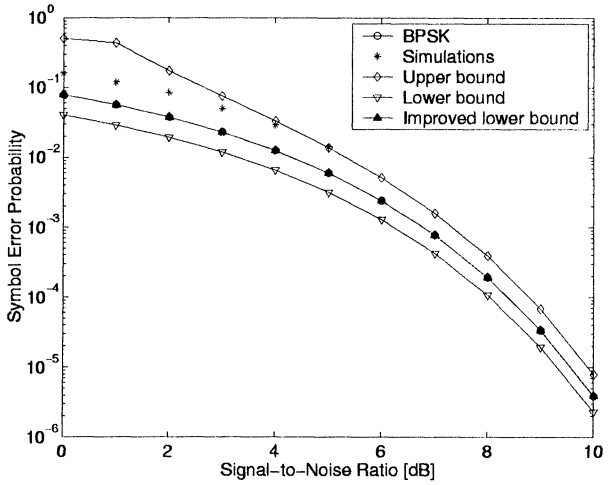


Figure 2: Symbol Error probability bounds and simulation results for binary 2RC system with h=1/2.

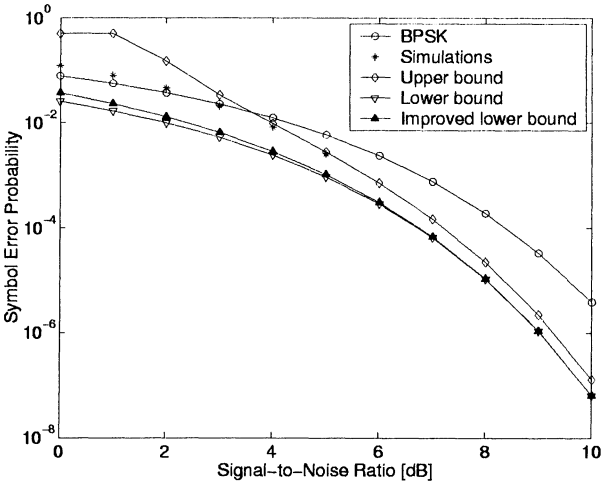


Figure 3: Symbol Error probability bounds and simulation results for binary 2RC system with h=3/4.

5 Conclusions

Symbol error probability bounds for coherently Viterbi detected CPM signals have been studied. The channel was assumed to be the AWGN channel.

The analysis method for constructing a lower bound on the symbol error probability for ISI channels [4, 5, 7] was generalized to be applicable to CPM systems. The reason why we can do this is because the signal can be modelled as a £nite state time-invariant discrete time Markov process.

The lower bounds have been obtained by considering a detector aided by USI provided by a "friendly genie". This lower bound can be improved by considering the entire set of transmitted sequences and not only the ones generating the minimum squared Euclidean distance. The algorithm presented here is general. Actually, any system that can be described as a £nite-state machine may be treated in the

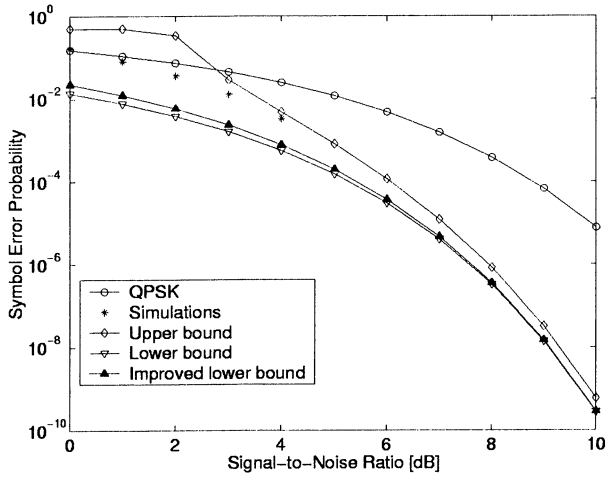


Figure 4: Symbol Error probability bounds and simulation results for Quaternary 2RC system with h=3/4.

same manner.

The improved lower bounds are shown to be useful even for fairly low SNR's for some certain CPM systems, such as the binary 2RC CPM system with $h = 1/2$. It is easy to see from equation (11) that the improvement depends on the distance spectrum and the spectrum coef£cient B_{d^2}. The largest improvements occur when the components in the distance spectrum are close together. Also, the B_d^2 should be fairly large.

References

[1] J. B. Anderson, T. Aulin, and C. E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, 1986.

[2] T. Aulin, "Symbol Error Probability Bounds for Coherently Viterbi Detected Continuous Phase Modulated Signals," *IEEE Transactions on Communications*, vol. COM-29, no. 11, pp. 1707–1715, Nov. 1981.

[3] G. D. Forney, Jr., "The Viterbi Algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, March 1973.

[4] G. D. Forney, Jr., "Lower Bounds on Error Probability in the Presence of Large Intersymbol Interference," *IEEE Transactions on Communications*, pp. 76–77, Feb. 1972.

[5] K. M. Chugg and A. Anastasopoulos, "On Symbol Error Probability Bounds for ISI-Like Channels," *IEEE Transactions on Communications*, vol. 49, no. 10, pp. 1704–1709, Oct. 2001.

[6] S. Verdú, *Multiuser Detection*, U. K. : Cambridge University Press, Cambridge, 1988.

[7] P. Moqvist and T. Aulin, "Improved lower bounds on the symbol error probability for ISI channels," *IEE Proceedings on Communication*, vol. 147, no. 6, pp. 351–354, Dec 2000.

[8] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice Hall, New York, 1983.

Robust Cumulant-Based Blind Beamforming

Otto Tronarp[1], Jouni Rantakokko[2]

[1] Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping, Sweden. E-mail: ottot@foi.se.
[2] Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping, Sweden. E-mail: jounir@foi.se.

Abstract

In this paper we examine the performance of a cumulant-based blind beamforming algorithm. The cumulant-based beamformer uses the well known Minimum Variance Distorsionless Response (MVDR) beamformer; however, it replaces the calibrated array response vector with estimates calculated from the 4:th order cumulants of the received data. Hence, the cumulant-based beamformer requires no knowledge about the array response vector. Furthermore, it does not require knowledge about the direction-of-arrival (DOA) of the signal-of-interest.

The cumulant-based beamformer is shown to outperform the standard MVDR beamformer, especially in scenarios with few samples and strong signals-of-interest. Furthermore, it is robust in scenarios with unknown antenna element failure. However, in the presence of strong interference signals the performance of the cumulant-based blind beamformer can degrade catastrophically. Finally, it is more sensitive to Gaussian interference signals compared to PSK-modulated signals.

1 Introduction

A secure and robust communication system with high capacity and low latency is of paramount importance in the quest to achieve information superiority on the future digitized battlefield. The concept of network-centric warfare depends heavily on the ability to transport large amounts of information between various platforms. Thus, the demand for robust high capacity radio networks is expected to rise in an unprecedented way.

Currently, there is a large interest in examining if civilian technology (*e.g.* UMTS) can be used in future tactical communication systems. The civilian systems will not fulfill all the requirements that are now put on tactical radio systems. However, by employing add-on antenna arrays these systems can be provided with robustness against jamming and other disturbances, as well as improving other aspects of the system (capacity, coverage, etc). Add-on antenna arrays can be a cost effective way to make the use of civilian communications technology feasible for military use.

Furthermore, in a software defined radio equipped with a digital beamforming antenna array it is possible to incorporate electronic warfare capabilities, such as direction finding and directional jamming. Every software defined radio node equipped with an antenna array has the potential to provide communication, radar and/or electronic warfare services.

Antenna array systems combined with more or less "intelligent" signal processing schemes, such as diversity methods, space-time coding, non-adaptive and adaptive beamforming algorithms, are envisaged to play an important role in future tactical radio systems. For example, antenna array systems have the potential to increase the capacity, availability, coverage, robustness (against jamming, intersystem interference, co-location disturbances and multipath propagation effects), power efficiency, and the covert communication capabilities.

An adaptive antenna array can adjust its array pattern automatically in response to changes in the signal environment. Numerous adaptive beamforming algorithms exist, and they can be divided into classes depending on what *a priori* information they use,

- information about the direction-of-arrival (DOA) of the incoming signals [1],
- reference signal [2] which is highly correlated with the signal-of-interest and has a low correlation with noise, interference and jammers,
- known characteristics of the communication signal, *e.g.*, cyclo-stationarity, finite alphabet or the constant modulus properties that arise for certain modulations [3]. A reference signal is created *implicitly* and the algorithms try to restore the known properties of the signal.

The latter type of algorithms are often named blind beamforming algorithms. They do not require information about the DOA of signals or any information about the array response vector (which includes the effects of array geometry and receiver channel responses). No reference signal (*e.g.* a known training sequence) is needed. Blind beamforming methods are of interest for use in many military applications and scenarios.

In this paper we examine the performance of the fourth-order cumulant-based blind beamforming method which was proposed in [4]. We will demonstrate and discuss its most important strengths and vulnerabilities when used in a military setting.

The paper is organized as follows. In the next section the MVDR beamformer is outlined. Thereafter, the

fourth-order cumulant-based blind beamformer is briefly described in section 3. Thereafter, the results from extensive simulations are described in section 4. Finally, the main conclusions drawn from this work are presented.

2 MVDR beamforming

If the direction to the signal-of-interest and the array response vectors (steering vectors) are known or can be estimated, then a constrained minimum variance beamformer can be used. It attempts to minimize the expected power (variance) out from the beamformer stemming from interference, jamming, and noise in other directions than for the signal-of-interest, θ, while keeping a constant gain in that direction. If a unity gain is chosen in the constraint direction (*i.e.*, $\boldsymbol{w}^*\boldsymbol{a}(\theta)=1$) then the beamformer is known as the minimum variance distorsionless response (MVDR) beamformer.

Let L be the number of antenna elements and M the number of (narrowband) source signals. N samples of data are available. The received signal at the antenna array at time t is then

$$\boldsymbol{x}(t)=[x_1(t) \quad \ldots \quad x_L(t)]^T = \boldsymbol{A}\boldsymbol{s}(t)+\boldsymbol{n}(t), \quad (1)$$

where $\boldsymbol{s}(t)$ is a $M\times 1$ vector containing the signals at time t, $\boldsymbol{A}$ is a $M\times L$ matrix containing the corresponding array response vectors and $\boldsymbol{n}(t)$ is an additive noise term.

The MVDR beamforming weight vector is calculated as [1]

$$\boldsymbol{w}_{MVDR} = \boldsymbol{R}^{-1}\boldsymbol{a}(\theta)\left(\boldsymbol{a}^H(\theta)\boldsymbol{R}^{-1}\boldsymbol{a}(\theta)\right)^{-1}, \quad (2)$$

where $\boldsymbol{a}(\theta)$ is the array response vector. $\boldsymbol{R}$ is the theoretic covariance matrix, $\boldsymbol{R} = E\{\boldsymbol{x}\boldsymbol{x}^H\}$, where $\boldsymbol{x}$ is the received data at the different antenna elements and $E\{\cdot\}$ denotes the statistical expectation operation. The array response vectors can be obtained by calibrating the array. The covariance matrix is usually estimated as

$$\hat{\boldsymbol{R}} = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^H, \quad (3)$$

where $\boldsymbol{X}$ is a $M\times N$ data matrix containing the received data vectors $\boldsymbol{x}(t)$ at different times as its columns.

Constrained minimum variance beamformers are known to be sensitive against various kinds of array response errors, caused for instance by calibration errors or discrepancies between the constraint direction and the true direction of the signal-of-interest. They also perform badly in scenarios with few samples, due to the mismatch between the theoretic covariance matrix and the estimated sample covariance matrix. The performance deterioration generally increases when the strength of the signal-of-interest increases. Several modifications can be applied in order to make the MVDR beamformer more robust, *e.g.*, derivative or eigenvector constraints, diagonal loading, noise eigenvalue equalization, or steering vector projection [5, 6].

3 Cumulant-based beamforming

The examined cumulant-based beamformer does not require any knowledge about the array response vector. Hence, no calibration is needed. Furthermore, it does not require knowledge about the direction-of-arrival of the signal-of-interest. It can be used for arbitrary (unambiguous) antenna array configurations. The cumulant-based beamformer maximizes the signal-to-interference-and-noise ratio (SINR) after beamforming. Also, it performs well in scenarios with coherent (perfectly correlated) multipath components [4]. In [7], fourth-order cumulants were used to suppress Gaussian interference and noise. However, that beamformer assumes that there only exists one non-Gaussian signal. For the cumulant-based beamformer examined in this paper the number of non-Gaussian signals is only restricted by the number of antenna elements.

In [4] it was shown that the fourth-order cumulants can be used to obtain estimates of the array response vectors, to within a complex constant. The cumulant-based blind beamformer then uses the well known Minimum Variance Distorsionless Response (MVDR) algorithm but replaces the array response vector with estimates derived from the fourth-order cumulants of the received data.

The following assumptions enable the estimation of the array response vectors for each non-Gaussian source to within a complex constant.

A.1 Desired source signals must have nonzero fourth-order cumulants.
A.2 The sources are statistically independent.
A.3 Narrowband source signals.
A.4 The array is unambiguous.
A.5 More antenna elements than non-Gaussian sources.
A.6 The additive noise can be Gaussian, non-Gaussian with a symmetrical distribution, or a mixture of both.
A.7 The reference elements are chosen so that they have nonzero responses to each source.

Further, in this work it is assumed that no multipath propagation occurs. However, simulations indicate that the cumulant-based beamformer is robust also in scenarios with highly correlated multipath components, but these results are not presented here.

The estimation of the array response vectors is based on the same principle as the ESPRIT algorithm [8]. In the ESPRIT algorithm, the rotational invariance of the signal subspace is induced by the translational invariance of the array. The signal subspace is extracted from the eigendecomposition of the covariance matrix of the concatenated measurements from the main array and an identical copy which is displaced in space. The cumulant-based algorithm achieves the same invariance without the use of an identical array. Instead, the signal subspace is

extracted from the SVD of concatenated fourth-order cumulant matrices [4].

The fourth-order cumulant matrices are calculated as

$$\boldsymbol{C}_1 = \mathrm{cum}\left(x_1(t), x_1^*(t), \boldsymbol{x}(t), \boldsymbol{x}^H(t)\right), \quad (4)$$

$$\boldsymbol{C}_2 = \mathrm{cum}\left(x_2(t), x_1^*(t), \boldsymbol{x}(t), \boldsymbol{x}^H(t)\right). \quad (5)$$

The TLS ESPRIT algorithm can be used to estimate the array response vectors from the above cumulant matrices. The interested reader is referred to the original paper [4] for a comprehensive description of the cumulant-based blind beamformer. The cumulant-based blind beamformer can be summarized as follows:

1. The cumulant-based beamformer estimates a separate array response vector for each non-Gaussian signal.
2. Beamforming weights are calculated for each non-Gaussian signal by substituting the array response vectors in the MVDR algorithm with the estimates, $\boldsymbol{w}_{CBBF} = \boldsymbol{R}^{-1}\hat{\boldsymbol{a}}_{cum}\left(\hat{\boldsymbol{a}}_{cum}^H \boldsymbol{R}^{-1}\hat{\boldsymbol{a}}_{cum}\right)^{-1}$.
3. Beamforming is performed for each signal.
4. Finally, the beamformer(s) that enhances the signal(s)-of-interest is identified, for instance by use of known data bits in the transmitted packets.

The fourth-order cumulants are in theory insensitive to Gaussian noise and interference signals with zero fourth-order cumulants [4]. However, since the cumulants are estimated from a finite set of data, a Gaussian interference can severely corrupt the cumulant estimates.

4 Results

We have examined the performance of the cumulant-based blind adaptive beamforming algorithm through extensive simulations. A 10 element uniform linear array (ULA), with a half wavelength spacing between the (isotropic) antenna elements, was used in the simulations. DOA's are calculated from the broadside of the ULA. All results are based on 100 Monte Carlo simulations.

In Figure 1 the interference suppression capability of the cumulant-based beamformer is shown. The signal-to-interference ratio out from the beamformer ($SINR_{out}$) was simulated for different sample sizes and interference-to-noise ratios (INR). The scenario is described in Table 1. The cumulant-based beamformer succeeds in canceling the BPSK-modulated interference as long as the difference in power between the signal-of-interest and the interference is sufficiently low. For a processing block size of 10000 samples, the performance of the cumulant-based beamformer deteriorates when the interference is about 20 dB stronger than the signal-of-interest. For even stronger interference signals, the cumulant-based blind beamformer begins to suppress the signal-of-interest. The reason for this degradation is that an increased interference power leads to a substantially increased variance of the fourth-order cumulants. Hence, the cumulant estimates for weaker signals are corrupted by strong interference.

As can be seen in Figure 2, strong Gaussian interference degrades the performance of the cumulant-based beamformer. The variance of the cumulant estimates increase rapidly when the strength of the interference increases. Thus, for limited sample sizes, a strong Gaussian signal corrupts the cumulant estimates, thereby degrading the beamformers performance.

Table 1: *Scenario 1. DOA, power, and signal type for the signal-of-interest (SOI) and interference.*

	DOA	Power	Type
SOI	10°	0/10 dB	BPSK
Interference	-30°		BPSK/random

Table 2: *Scenario 2. 2000 samples.*

	DOA	Power	Type
SOI	0°	0 dB	BPSK
Interference		10 dB	BPSK

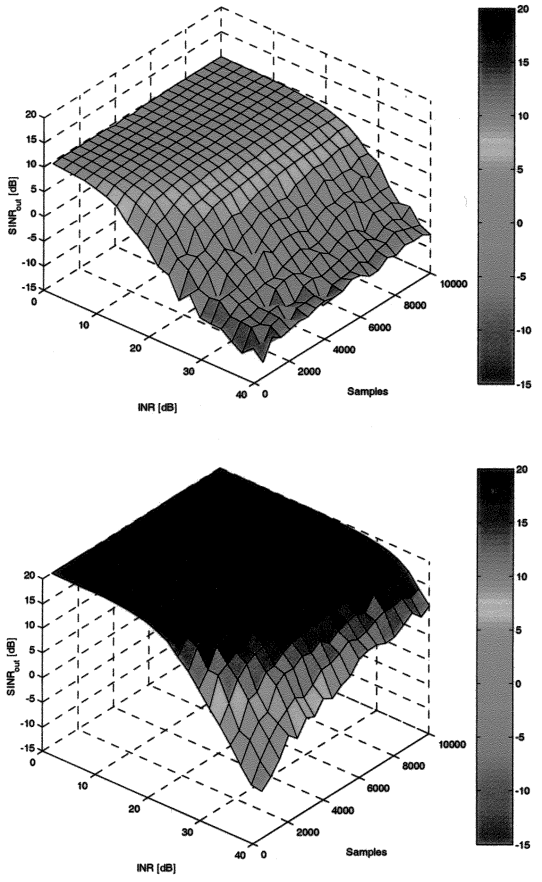


Figure 1: *Cumulant-based beamforming with a BPSK-modulated signal-of-interest and a BPSK-modulated interference. The signal-to-interference-and-noise ratio after beamforming ($SINR_{out}$) is shown as a function of the interference-to-noise ratio (INR) and the number of samples. Scenario 1. The power of the signals-of-interest are 0 dB (upper) and 10 dB (lower).*

Furthermore, the performance of the cumulant-based beamformer is substantially worse when the interference is Gaussian. As shown in Figure 3, the root-mean-square-error of the cumulant estimates for Gaussian signals is larger than for PSK-modulated signals. Thus, Gaussian interference will result in larger cumulant estimation errors than the PSK-modulated signals.

The cumulant-based blind beamformer is compared with the MVDR beamformer in Figure 4 for the case of a single BPSK-modulated signal arriving from 10°. Perfect knowledge of the DOA and array response vector was assumed when performing MVDR beamforming. In *(a)* the SINR, calculated with 100 and 1000 samples, is shown for different SNR's. The MVDR beamformer suffers from signal cancellation when few samples are available. The performance degradation is worse for strong signals. The cumulant-based beamformer clearly outperforms the MVDR beamformer in this scenario. When 1000 samples are available it maximizes the output SINR. However, for 100 samples the performance deteriorates slightly for strong signals.

The output SINR as a function of the number of available samples is shown in *(b)*, for a 0 and 20 dB strong signal-of-interest. For the case of a 20 dB signal-of-interest, the MVDR beamformer performs badly. The bad performance is caused by a mismatch between the theoretic covariance matrix and its sample estimate. This performance degradation is more severe for strong signals. For a 0 dB signal-of-interest the MVDR beamformer needed almost 1000 samples to achieve near optimum performance. In comparison, the cumulant-based beamformer achieves a near optimum SINR for very few samples. For the case of a 20 dB signal-of-interest, the cumulant-based beamformer is less than 1 dB from the optimum SINR with less than 200 samples.

The MVDR beamformer is sensitive to errors in the array response vectors or to erroneous DOA estimates for the signal-of-interest. These errors lead to signal cancellation which can severely degrade the performance. The sensitivity of the MVDR beamformer can be reduced by applying modifications, such as point or derivative constraints, diagonal loading, noise eigenvalue equalization or steering vector projection, see for instance [5] and the references therein. In comparison, the cumulant-based beamformer is robust since it estimates the array response vectors directly from the received data. Hence, it does not suffer from the signal cancellation phenomena; thus, modification such as diagonal loading will not improve its performance.

In military applications the beamforming algorithms must be robust to hardware failure, *e.g.*, antenna element failure. In Figure 5 the performance of the MVDR and cumulant-based beamformers were examined in the presence of element failure. In this simulation, element four contained only noise, *i.e.*, no signals were present for that antenna. No modifications were applied to the beamformers. The cumulant-based beamformer is robust against this type of hardware failure while the MVDR-

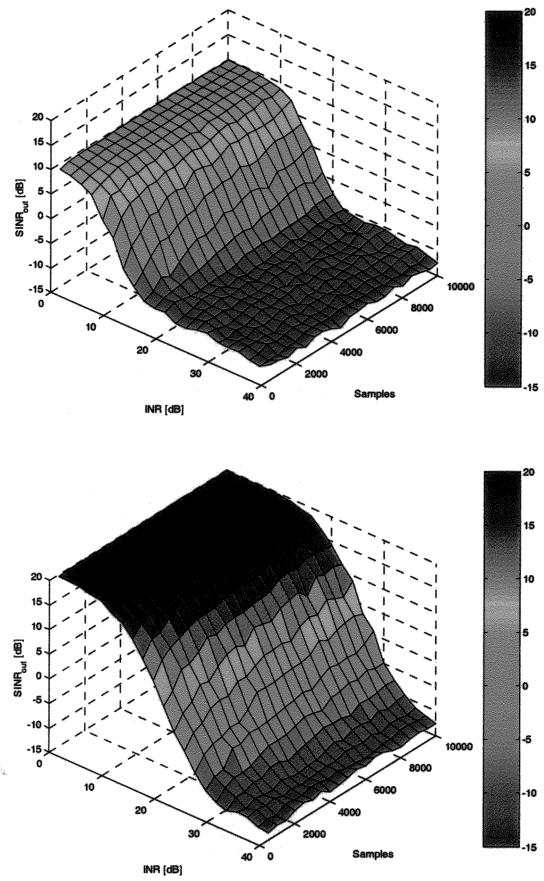


Figure 2: *Cumulant-based beamforming with a BPSK-modulated signal-of-interest and a Gaussian interference. The signal-to-interference-and-noise ratio after beamforming ($SINR_{out}$) is shown as a function of the interference-to-noise ratio (INR) and the number of samples. Scenario 1. The power of the signals-of-interest are 0 dB (upper) and 10 dB (lower).*

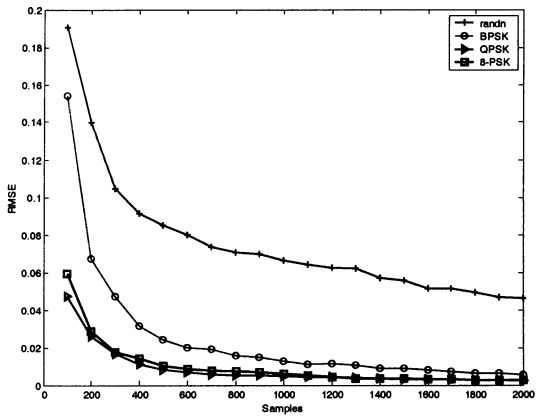


Figure 3: *The root-mean-square-error (RMSE) of the fourth-order cumulant estimates ($cum(s(t),s^*(t),s(t),s^*(t))$ is shown as a function of the number of samples (bits). The cumulants are estimated for BPSK, QPSK, 8-PSK, and Gaussian signals.*

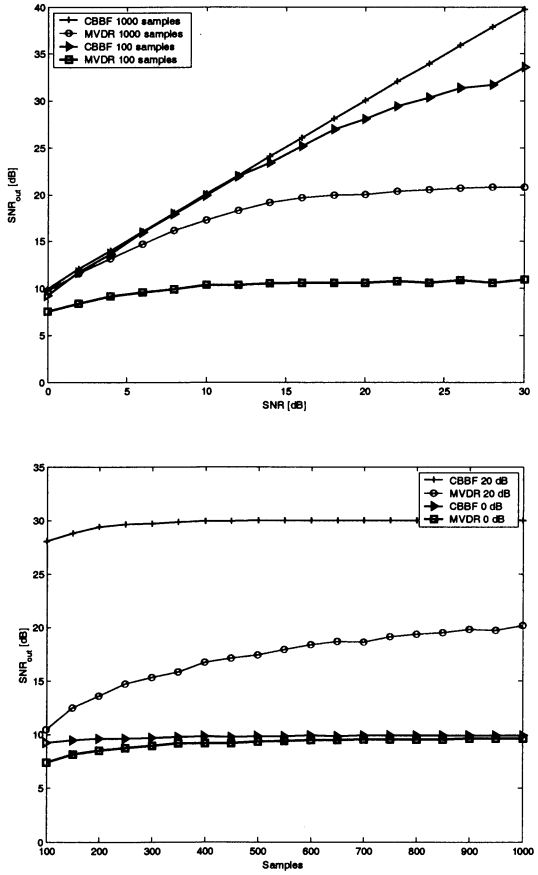


Figure 4: *Comparison between the cumulant-based beamformer (cbbf) and the MVDR beamformer. SINR is shown as a function of SNR for the SOI (upper) and as a function of the number of samples (lower). The SOI is a BPSK-modulated signal which arrives from 10°.*

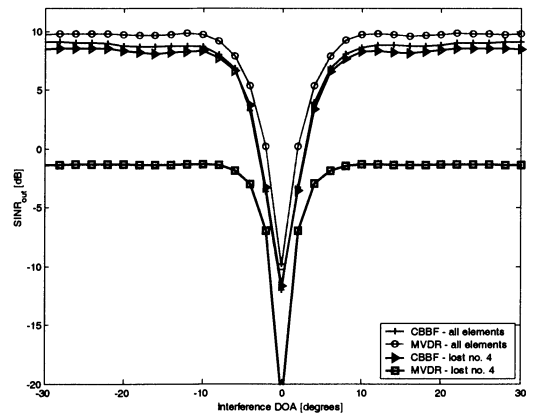


Figure 5: *Performance in scenarios with antenna element failure for the cumulant-based beamformer (cbbf) and the MVDR beamformer. Antenna element 4 contains only noise (circle and square). Scenario 2.*

beamformers performance is unacceptable. For the cumulant-based beamformer it is essential that none of the two reference elements are lost (apply a selection procedure [4]). In cases were it is possible to detect element failure the beamforming can instead be performed for a nine element sparse array.

5 Conclusions

In this paper we have examined the performance of the fourth-order cumulant-based blind beamforming method which was proposed in [4].

The main advantages with the cumulant-based blind beamformer are:

- Performs well when the signal-of-interest is strong.
- No knowledge is required about the array response vector or direction-of-arrival for the signal-of-interest.
- Outperforms the standard MVDR beamformer in scenarios with few samples.
- Robust against unknown antenna element failure.

The main drawbacks are the increase in computational complexity and the performance deterioration that occurs in scenarios with strong interference signals. It also performs substantially worse for Gaussian interference compared to PSK-modulated signals.

References

[1] O. L. Frost III, "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proc. IEEE*, vol. 60, pp. 926-935, August 1972.

[2] R. T. Compton, *Adaptive antennas: Concepts and performance*, Prentice-Hall, 1988.

[3] J. C. Liberti and T. S. Rappaport, *Smart Antennas for Wireless Communications: IS-95 and Third generation CDMA Applications*, Prentice-Hall, 1999.

[4] E. Gönen and J. M. Mendel, "Applications of cumulants to array processing, Part III: Blind beamforming for coherent signals," *IEEE Trans. Signal Processing*, vol. 45, no. 9, pp. 2252-2264, September. 1997.

[5] F. L. T. Eklöf, J. G. J. Rantakokko and L. E. Pettersson, "On broadband adaptive beamforming for tactical radio systems," *IEEE Military Communications Conference*, Washington D.C., USA, October 2001.

[6] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 4-24, April 1988.

[7] M. C. Dogan and J. M. Mendel, "Cumulant-based blind optimum beamforming," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 30, pp. 722-741, July 1994.

[8] R. Roy and T. Kailath, "ESPRIT –Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 37, no. 7, pp. 984-995, July 1989.

Space-Time Code Design for Combined Channel Estimation and Error Protection

Jochen Giese[1] and Mikael Skoglund[1]

[1]Royal Institute of Technology, 100 44 Stockholm, Sweden. E-mail: {giese,skoglund}@s3.kth.se

Abstract

We investigate the design of space-time codes that operate over frequency-selective channels where neither the transmitter nor the receiver has channel state information. A transmission model for this setup is presented and shown to be closely linked to a single-antenna model which has been investigated earlier. Using a randomized search strategy, codes of arbitrary rates for combined channel estimation and error protection over these channels can be designed. Simulations demonstrate the performance of codes designed with the proposed scheme. An example code outperforms an approach based on an extension of the Alamouti scheme.

1 Introduction

The demand for higher data rates in mobile communications has recently spurred enormous research efforts to increase the data throughput of mobile communication links. The fading induced by mobile communication channels was traditionally seen as an impairment to increase data rates. However, in a multiple-antenna setup it was shown in [1] that independent fading between different antennas can actually increase the capacity of the mobile communication link and thus higher data rates can potentially be supported. Various papers have presented efficient signaling structures for either known [2,3] or unknown [4] single-tap channels for multiple antenna systems. The impact of the multi-path characteristic of mobile communication channels resulting in inter-symbol interference (ISI) has recently received attention in e.g. [5–7]. In these papers, the receiver is assumed to have acquired channel state information. This is usually done by the receiver using a known-input channel identification algorithm based on the transmission on some known pilot symbols. The obtained estimates are then considered sufficiently reliable to be used in subsequent equalization and decoding stages. This split approach of channel estimation and channel coding is beneficial from a system conception point of view because estimation and decoding can be optimized separately; however, joint optimization of both channel estimation and coding was demonstrated in [8,9] and [10] to result in superior performance. The codes presented in those references were however designed for single-antenna systems and thus cannot take advantage of the potential capacity increase offered by multiple-antenna systems.

In this paper, we generalize a method proposed in [9] for single-antenna systems in order to design error-correcting space-time codes of arbitrary rates for unknown multi-tap channels. In Section 2, a transmission model describing ISI for multi-antenna systems is introduced. By calculating exact pairwise word-error probabilities for the joint channel and data detector, we can evaluate the union bound on the word-error probabilities as a design criterion in our random search procedure in Section 3. Simulation results illustrate the performance of these codes (Section 4) before we conclude in Section 5.

2 Problem description

2.1 System model and assumptions

The model description is organized as a generalization from a single antenna model to several transmitter antennas, and then to a model with multiple antennas at both the transmitter and receiver side.

Consider the transmission of a BPSK symbol sequence $\{b_t\}$, $t = 0, \ldots, T-1$, $b_t \in \{\pm 1\}$ through a frequency-selective channel described by the channel coefficients $\{h_l\}$, $l = 0, \ldots, L-1$, $h_l \in \mathbb{C}$ with additive white complex Gaussian noise of variance σ^2 denoted by w_t. Assuming sufficiently long guard intervals between the transmission of consecutive sequences such that no inter-sequence interference is present, the received signal y_t can be described as

$$y_t = \sum_{l=0}^{L-1} h_l b_{t-l} + w_t \qquad t = 0, \ldots, T+L-2$$

where we define $b_k := 0$ for $k < 0$ and $k > T + L - 2$. Stacking all observed y_t into a vector $\mathbf{y} = [y_0 \;\; y_1 \;\; \ldots \;\; y_{T+L-2}]^\dagger$ where $[\cdot]^\dagger$ denotes transpose we

get

$$\mathbf{y} = \underbrace{\begin{bmatrix} b_0 & 0 & \cdots & 0 \\ b_1 & b_0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ b_{L-1} & b_{L-2} & \cdots & b_0 \\ \vdots & & \vdots & \vdots \\ b_{T-1} & b_{T-2} & \cdots & b_{T-L} \\ 0 & b_{T-1} & \cdots & b_{T-L+1} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & b_{T-1} \end{bmatrix}}_{\mathbf{B}} \cdot \underbrace{\begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{L-1} \end{bmatrix}}_{\mathbf{h}} + \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{T+L-2} \end{bmatrix}}_{\mathbf{w}} = \mathbf{Bh} + \mathbf{w}. \quad (1)$$

In the case with $M > 1$ transmitter antennas, antenna m transmits the sequence $\{b_t\}_m$ defining the matrix $\mathbf{B}_m$ through the channel denoted by $\mathbf{h}_m$. Stacking all $\mathbf{h}_m$ vectors into one large vector $\tilde{\mathbf{h}} = [\mathbf{h}_1^\dagger \ \ldots \ \mathbf{h}_M^\dagger]^\dagger$ and regrouping all $\mathbf{B}_m$ matrices into one matrix, the transmission model is now

$$\mathbf{y} = [\mathbf{B}_1 \ \mathbf{B}_2 \ \ldots \ \mathbf{B}_M]\,\tilde{\mathbf{h}} + \mathbf{w}.$$

Generalizing further to N receiver antennas leads to

$$\underbrace{[\mathbf{y}_1 \ \ldots \ \mathbf{y}_N]}_{\mathbf{Y}} = [\mathbf{B}_1 \ \mathbf{B}_2 \ \ldots \ \mathbf{B}_M] \underbrace{\left[\tilde{\mathbf{h}}_1 \ \ldots \ \tilde{\mathbf{h}}_N\right]}_{\mathbf{H}} + \underbrace{[\mathbf{w}_1 \ \ldots \ \mathbf{w}_N]}_{\mathbf{W}}$$

where $\mathbf{y}_k$ denotes the symbol sequence received on receiver antenna k with noise component $\mathbf{w}_k$ and $\tilde{\mathbf{h}}_k$ is the vector describing the channels from all M transmitter antennas to receiver antenna k. In order to obtain a similar expression to (1), the columns of $\mathbf{Y}$, $\mathbf{H}$ and $\mathbf{W}$ can be stacked into the vectors $\bar{\mathbf{y}}$, $\bar{\mathbf{h}}$ and $\bar{\mathbf{w}}$ respectively to finally arrive at

$$\bar{\mathbf{y}} = \underbrace{(\mathbf{I}_N \otimes [\mathbf{B}_1 \ \mathbf{B}_2 \ \ldots \ \mathbf{B}_M])}_{\bar{\mathbf{B}}}\bar{\mathbf{h}} + \bar{\mathbf{w}}$$

where $\mathbf{I}_N$ is an $N \times N$ identity matrix and $\otimes$ denotes the Kronecker product.

We assume that the channel coefficients $\bar{\mathbf{h}}$ are *unknown to both transmitter and receiver.* Moreover, the values of these coefficients are assumed constant during the transmission of T symbols and statistically independent of the coefficient values during the next block of T symbols. The number L of channel taps is assumed known or correctly upper-bounded.

2.2 Decoding

Since the channel coefficients $\bar{\mathbf{h}}$ are unknown, we assume that the receiver implements joint Maximum Likelihood estimation of $\bar{\mathbf{h}}$ and detection of $\bar{\mathbf{B}}$ which reduces to the criterion

$$(\hat{\bar{\mathbf{h}}}, \hat{\bar{\mathbf{B}}}) = \arg\min_{\bar{\mathbf{B}},\bar{\mathbf{h}}} \left\{||\bar{\mathbf{y}} - \bar{\mathbf{B}}\bar{\mathbf{h}}||^2\right\}$$

where the minimum is over all $\bar{\mathbf{h}} \in \mathbb{C}^{LMN}$ and all possible codewords $\bar{\mathbf{B}}$. This criterion leads [9] to

$$\begin{aligned} \hat{\bar{\mathbf{B}}} &= \arg\min_{\bar{\mathbf{B}}} ||\bar{\mathbf{y}} - \bar{\mathbf{P}}_{\bar{\mathbf{B}}}\bar{\mathbf{y}}||^2 = \arg\min_{\bar{\mathbf{B}}} ||(\mathbf{I} - \bar{\mathbf{P}}_{\bar{\mathbf{B}}})\bar{\mathbf{y}}||^2 \\ &= \arg\min_{\bar{\mathbf{B}}} ||\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}}\bar{\mathbf{y}}||^2 \end{aligned}$$

where $\bar{\mathbf{P}}_{\bar{\mathbf{B}}}$ is the projection matrix onto the column space of $\bar{\mathbf{B}}$ and $\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}}$ the projection matrix onto the corresponding null space. We emphasize that the mapping described above *combines* channel estimation, equalization and decoding into one *joint* operation. If $\bar{\mathbf{B}}$ has full column rank then $\mathbf{P}_{\bar{\mathbf{B}}} = \bar{\mathbf{B}}\left(\bar{\mathbf{B}}^\dagger\bar{\mathbf{B}}\right)^{-1}\bar{\mathbf{B}}^\dagger$; otherwise the projection matrix can always be calculated via the singular-value decomposition (see e.g. [11]).

2.3 Encoding

The encoder maps the incoming data bits to symbol sequences that are transmitted via the M antennas. Since the modulation alphabet is restricted to BPSK and MT BPSK symbols are transmitted within one codeword, the set $\mathcal{B}$ of all possible codeword matrices has a cardinality of $|\mathcal{B}| = 2^{MT}$. Therefore, every $\bar{\mathbf{B}} \in \mathcal{B}$ can be uniquely identified by the integer $i = i(\bar{\mathbf{B}}) \in \mathcal{C}_{\mathcal{B}}$, $\mathcal{C}_{\mathcal{B}} = \{0, \ldots, 2^{MT}-1\}$ by using the binary expansion

$$i = \sum_{k=1}^{MT} d_k 2^{MT-k}, \quad d_k \in \{0,1\}$$

where the symbol $b_{t,m}$ sent on antenna m at time instant t is $b_{t,m} = 1 - 2d_{(m-1)T+t+1}$.

When a specific *code* of rate R for combined channel estimation and error protection is used, a subset $\mathcal{D} \subset \mathcal{B}$ with $|\mathcal{D}| = 2^{RT}$ or, equivalently, a subset $\mathcal{K} \subset \mathcal{C}_{\mathcal{B}}$ with $|\mathcal{K}| = 2^{RT}$ forms the set of allowed codewords. The encoder can then assign a codeword to each combination of RT incoming data bits.

Note that using a fixed code $\mathcal{K}$, all $\bar{\mathbf{B}}$ with $i(\bar{\mathbf{B}}) \notin \mathcal{K}$ can be discarded at the receiver side and the decoding criterion can be reformulated to

$$\hat{\bar{\mathbf{B}}} = \hat{\bar{\mathbf{B}}}(i') \quad \text{where} \quad i' = \arg\min_{i\in\mathcal{K}} \left\{||\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}(i)}\bar{\mathbf{y}}||^2\right\}$$

where we note that all different matrices $\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}(i)}$ can be calculated off-line and stored for all $i \in \mathcal{K}$.

3 Code Design

In the setup described in Section 2, one specific code is completely described by the set $\mathcal{K}$ with $|\mathcal{K}| = 2^{RT}$. In this section we focus on a design procedure to specify $\mathcal{K}$. This is an integer optimization problem and we apply the randomized search procedure of *Simulated Annealing* [12].

3.1 Design criterion

The aim of a code design is to find a code that has good error-correcting capabilities, or, in other words, a low probability P_e of false detection of the transmitted codeword in the presence of noise. Thus the code design criterion must try to minimize the word error probability $P_e := \Pr(\hat{\bar{\mathbf{B}}} \neq \bar{\mathbf{B}})$. Since an exact expression for P_e appears to be mathematically intractable, we resort to using the union-bound criterion, resulting in the bound

$$\begin{aligned} P_e &= 2^{-RT} \sum_{i \in \mathcal{K}} \Pr(\hat{\bar{\mathbf{B}}} \neq \bar{\mathbf{B}}(i) | \bar{\mathbf{B}}(i) \text{ was transmitted}) \\ &\leq 2^{-RT} \sum_{i \in \mathcal{K}} \sum_{j \in \mathcal{K}, j \neq i} p_{j|i}, \end{aligned}$$

where we define the *pair-wise error probability* $p_{j|i}$ as

$$p_{j|i} = \Pr(||\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}(j)}(\bar{\mathbf{B}}(i)\bar{\mathbf{h}} + \bar{\mathbf{w}})||^2 < ||\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}(i)}(\bar{\mathbf{B}}(i)\bar{\mathbf{h}} + \bar{\mathbf{w}})||^2). \tag{2}$$

In the following, an exact expression for $p_{j|i}$ is given under the assumption that the channel coefficients $\bar{\mathbf{h}}$ for each sequence are drawn according to a Gaussian distribution with mean $\mathbf{m}_{\bar{\mathbf{h}}} = E[\bar{\mathbf{h}}]$ and covariance matrix $\mathbf{C}_{\bar{\mathbf{h}}}$ (assumed real). This model corresponds to frequency-selective Rayleigh ($||\mathbf{m}_{\bar{\mathbf{h}}}|| = 0$) or Rice ($||\mathbf{m}_{\bar{\mathbf{h}}}|| \neq 0$) fading with possibly correlated scattering [13]. Since the channel $\bar{\mathbf{h}}$ and the noise $\bar{\mathbf{w}}$ are Gaussian, the received signal $\bar{\mathbf{y}}$ conditioned on $\bar{\mathbf{B}}$ is also Gaussian with mean $\mathbf{m}_{\bar{\mathbf{y}}} = \bar{\mathbf{B}}\mathbf{m}_{\bar{\mathbf{h}}}$ and covariance

$$\begin{aligned} \bar{\mathbf{S}}_{\bar{\mathbf{y}}}(i) &= E\left[(\bar{\mathbf{y}} - \mathbf{m}_{\bar{\mathbf{y}}})(\bar{\mathbf{y}} - \mathbf{m}_{\bar{\mathbf{y}}})^*\right] \\ &= \bar{\mathbf{B}}(i)\mathbf{C}_{\bar{\mathbf{h}}}\bar{\mathbf{B}}(i)^{\dagger} + \sigma^2 \mathbf{I}_{N(L+T-1)}, \end{aligned}$$

where $[\cdot]^*$ denotes conjugate transpose. In [9], it was shown that the error probability in (2) is equivalent to the probability that a linear combination of independent non-central χ^2-variables is positive

$$p_{j|i} = \Pr\left(\sum_{r=1}^{\rho} \tilde{\lambda}_r \chi^2(2k_r, \eta_r) > 0\right) \tag{3}$$

where $\chi^2(\kappa, \eta)$ denotes a χ^2 variable with κ degrees of freedom and non-centrality parameter η, $\tilde{\lambda}_r$ are the ρ distinct non-zero eigenvalues with multiplicity k_r of the matrix

$$\mathbf{S}_{\bar{\mathbf{y}}}^{\frac{1}{2}} \left(\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}(i)} - \bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}(j)}\right) \mathbf{S}_{\bar{\mathbf{y}}}^{\frac{1}{2}} = \sum_{p=1}^{N(T+L-1)} \lambda_p \mathbf{q}_p \mathbf{q}_p^{\dagger}$$

with eigenvalues λ_p to orthonormal eigenvectors $\mathbf{q}_p$, $p = 1, \ldots, N(T+L-1)$, and where η_r can be obtained as

$$\eta_r = \sum_p |\mathbf{q}_p^{\dagger} \mathbf{S}_{\bar{\mathbf{y}}}^{-1/2} \bar{\mathbf{B}}(i) \mathbf{m}_{\bar{\mathbf{h}}}|^2$$

where the sum is over all p for which $\lambda_p = \tilde{\lambda}_r$. Imhof [14] showed that (3) can be rewritten as

$$p_{j|i} = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin f_1(x)}{x f_2(x)} dx \tag{4}$$

where

$$\begin{aligned} f_1(x) &= \frac{1}{2} \sum_{r=1}^{\rho} \left[2k_r \arctan(\tilde{\lambda}_r x) + \frac{\eta_r \tilde{\lambda}_r x}{1 + \lambda_r^2 x^2} \right] \\ f_2(x) &= \prod_{r=1}^{\rho} (1 + \tilde{\lambda}_r^2 x^2)^{k_r/2} \exp\left[\frac{1}{2} \sum_{u=1}^{R} \frac{\eta_u \tilde{\lambda}_u^2 x^2}{1 + \tilde{\lambda}_u^2 x^2} \right]. \end{aligned}$$

The integral in (4) can be evaluated in a numerically stable fashion for noise powers greater than approximately 10^{-3}; for smaller values, the approximate expressions given in [15] are applicable. Having an explicit formula for $p_{j|i}$, the design criterion based on the union bound can thus be formulated as

$$\begin{aligned} \text{find} \quad & \mathcal{K} = \arg \min_{\mathcal{K} \subset \mathcal{C}_{\mathcal{B}}} 2^{-RT} \sum_{i \in \mathcal{K}} \sum_{j \in \mathcal{K}} p_{j|i} \\ \text{such that} \quad & |\mathcal{K}| = 2^{RT} \end{aligned}$$

where we have defined $p_{i|i} := 0$ for simplicity.

Note that if $\bar{\mathbf{B}} \in \mathcal{B}$ then also $-\bar{\mathbf{B}} \in \mathcal{B}$. Since $\bar{\mathbf{P}}^{\perp}_{\bar{\mathbf{B}}} = \bar{\mathbf{P}}^{\perp}_{-\bar{\mathbf{B}}}$ the codewords $\bar{\mathbf{B}}$ and $-\bar{\mathbf{B}}$ cannot be distinguished at the receiver even in noise-free transmission. Therefore, any $\mathcal{K} \subset \mathcal{C}_{\mathcal{B}}$ will not contain both $i_+ = i(\bar{\mathbf{B}})$ and $i_- = i(-\bar{\mathbf{B}})$ to avoid this identifiability problem.

3.2 Code Design

The optimization method of *Simulated Annealing* [12] for integer programming has proven to yield good results in code design problems [16]. The target function to be minimized is interpreted as the energy ϵ of a physical system with many degrees of freedom in thermal equilibrium. Cooling down (annealing) the system will drive the system to the most probable state of lowest energy when the annealing process is sufficiently slow.

Applied to code design, the state is represented by the code $\mathcal{K}$ and its energy ϵ is described by a scaled version of the union bound $\epsilon(\mathcal{K}) = \sum_{i \in \mathcal{K}} \sum_{j \in \mathcal{K}} p_{j|i}$. The initial state of the system is an arbitrary code $\mathcal{K} = \mathcal{K}_1 \subset \mathcal{C}_{\mathcal{B}}$ chosen at a high temperature τ. Thermal movement is then simulated by randomly exchanging elements $u \in \mathcal{K}$ with $v \in \mathcal{C}_{\mathcal{B}} \setminus \mathcal{K}$, thus changing the energy of the system from ϵ_1 to ϵ_2.

Energy changes $\Delta\epsilon = \epsilon_2 - \epsilon_1 < 0$ reducing the energy are always accepted whereas changes $\Delta\epsilon > 0$ are also accepted, but with probability $p = \exp(-\Delta\epsilon/\tau)$. This probability refers to a Boltzmann factor $\exp(-E/(k_b T))$ (notation see [12]) in statistical mechanics to allow state degradation with a temperature-dependent probability. The possibility of accepting changes increasing the energy allows the system to escape from local energy minima configurations.

The code is "annealed" by slowly decreasing τ (or, equivalently, the probability of code degradation) in order to drive the system to a stable state with minimum energy. Note that the initial code $\mathcal{K}_1$ does not notably influence the final result because the code $\mathcal{K}$ gets randomized quickly at high τ.

4 Simulation Results

Our design approach described in Section 3 includes in a natural fashion the possibility of designing codes of different rates for an arbitrary number of transmitter and receiver antennas.

We restrict our examples to codes for transmission on Rayleigh fading channels of length $L = 2$. The channel taps are assumed uncorrelated and have equal mean powers equal to $1/L = 1/2$.

For a fair comparison of codes with different M, the $\{\pm 1\}$ components in $\bar{\mathbf{B}}$ are scaled to $\{\pm 1/\sqrt{M}\}$ to transmit at each time instant the same energy independently in M. Moreover, we define the signal-to-noise-ratio as $\text{SNR} = 1/(R\sigma^2)$ in order to keep the ratio of signal power per data bit to noise power constant when transmitting at varying rates.

We have designed codes of word length $T = 6$ with rates $R = 1/3, 1/2, 1, 4/3$ for $M = 2$ transmitter antennas as well as a code of rate $R = 1/2$ for $M = 1$ transmitter antenna, where all codes were designed for $\text{SNR} = 10^{-2}$. For example, the code for $M=2$ antennas, $R = 1/2$ is $\{156, 615, 897, 1181, 1273, 1515, 1786, 1994\} \subset \{0, \ldots, 2^{12}-1\}$ and the code for $M = 1$ antenna with rate $R = 1/2$ is $\{1, 7, 10, 12, 16, 22, 27, 29\} \subset \{0, \ldots, 2^6 - 1\}$. We assume $N = 1$ receiver antenna. Performance results in terms of word-error rate (WER) are summarized in Figure 1.

Comparing the codes for $M = 2$, it becomes apparent that the loss in power when transmitting at lower rates is more than compensated for by the increased capability of the code to be decoded correctly. For example, the $M = 2$ code with $R = 1/2$ has a gain of about 3.6 dB compared to the code with $R = 1$ at a remaining WER of 10^{-3}. Moreover, this code offers a significant diversity gain obtained by using $M = 2$ antennas compared to the $M = 1$, $R = 1/2$ code for SNR larger than about 20 dB. It is also interesting to note that the slope of the $M = 2$ curve is not the same for all R: the codes with rate $R \geq 1$ do not have the same diversity gain as the codes for $R < 1$. It is a matter of future research to further investigate this fact. Moreover note that we cannot design codes of rates $R = 1$ or higher for $M = 1$ because of the identifiability problem discussed in Section 3.1.

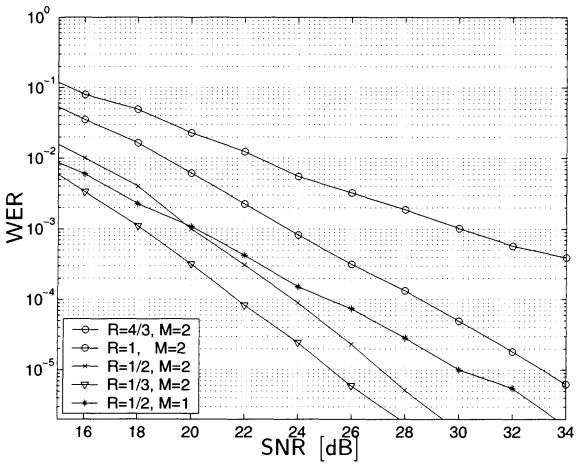


Figure 1: Performance comparison for $T = 6$ with variable rate

In order to compare our scheme with a previously known code, we also constructed a code based on the Alamouti scheme [17]. Simple concatenation of three $T = 2$ codewords of the Alamouti code used with BPSK to a code $\mathcal{K}_{A_3}$ will result in an identifiability problem because for any $\bar{\mathbf{B}}$ with $i(\bar{\mathbf{B}}) \in \mathcal{K}_{A_3}$ it is clear that also $i(-\bar{\mathbf{B}}) \in \mathcal{K}_{A_3}$ (see Section 2.3). Therefore, we construct a $T = 5$, $R = 0.8$ code $\mathcal{K}_{A_2}$ by concatenating two Alamouti codewords and setting the remaining two symbols to fixed values. Simulation results for the code $\mathcal{K}_{A_2}$ as well as for our designed code are summarized in Figure 2.

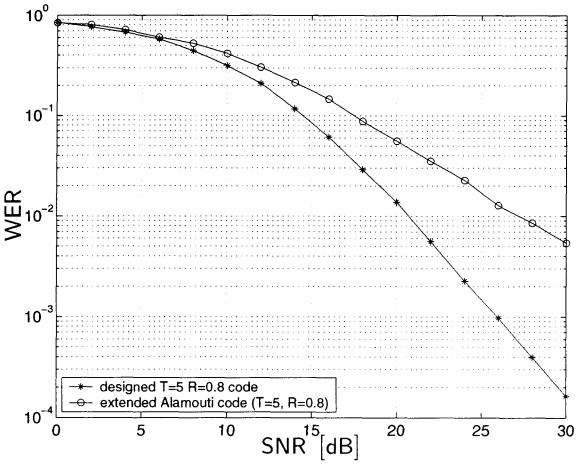


Figure 2: Comparison of designed $T = 5$, $R = 0.8$ code vs. extended Alamouti code

We note that the code specifically designed for $T = 5$, $R = 0.8$ clearly outperforms the extended Alamouti code.

5 Conclusions

We presented a transmission model for space-time codes over unknown multi-tap channels and investigated a design method for codes of arbitrary rates over such channels. The diversity offered by multiple antenna transmission resulted in lower word error probability for high SNR compared to a single-antenna scheme. Comparing codes of different rates, it became apparent that the gain in word error probability using codes with low rates outweighs the loss in power for each transmitted symbol resulting in a combination of diversity and coding gain. Moreover, we observed that the rate influences the diversity gain in our setup.

In this paper, we investigated relatively short codes, but already these codes show promising results. We consider the present work a starting point for further research on space-time codes over unknown channels with an emphasis on extending the design to more antennas for coded transmission over longer channels.

References

[1] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, Nov.-Dec. 1999.

[2] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 744–765, Mar. 1998.

[3] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999.

[4] B. M. Hochwald, T. L. Marzetta, T. J. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Transactions on Information Theory*, vol. 46, no. 6, pp. 1962–1973, Sept. 2000.

[5] E. Lindskog and A. Paulraj, "A transmit diversity scheme for channels with intersymbol interference," in *Proc. ICC*, 2000, pp. 307–311.

[6] S. Zhou and G. B. Giannakis, "Space-time coded transmissions with maximum diversity gains over frequency-selective multipath fading channels," in *Proc. Globecom*, 2001, pp. 440–444.

[7] E. Larsson, P. Stoica, E. Lindskog, and J. Li, "Space-time block coding for frequency-selective channels," submitted to *IEEE Transactions on Wireless Communications*, Nov. 2001.

[8] M. Skoglund and S. Parkvall, "Code design for combined channel estimation and error correction coding," in *Proc. IEEE International Symposium on Information Theory*, Sorrento, Italy, June 2000.

[9] M. Skoglund, J. Giese, and S. Parkvall, "Code design for combined channel estimation and error protection," to appear in *IEEE Transactions on Information Theory*, May 2002.

[10] O. Coskun and K. M. Chugg, "Combined coding and training for unknown ISI channels," submitted to *IEEE Transactions on Communications*, June 2000.

[11] G. Strang, *Linear Algebra and its Applications*, Harcourt Brace & Company, 3rd edition, 1988.

[12] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.

[13] J. G. Proakis, *Digital Communications*, McGraw-Hill International Editions, 3rd edition, 1995.

[14] J. P. Imhof, "Computing the distribution of quadratic forms in normal variables," *Biometrika*, vol. 48, no. 3-4, pp. 419–426, 1961.

[15] M. Brehler and M. K. Varanasi, "Asymptotic error probability analysis of quadratic receivers in Rayleigh-fading channels with applications to a unified analysis of coherent and noncoherent space-time receivers," *IEEE Transactions on Information Theory*, vol. 47, no. 6, pp. 2383–2399, Sept. 2001.

[16] A. A. El Gamal, L. A. Hemachandra, I. Shperling, and V. K. Wei, "Using simulated annealing to design good codes," *IEEE Transactions on Information Theory*, vol. IT-33, no. 1, pp. 116–123, Jan. 1987.

[17] S. M. Alamouti, "A simple transmit diversity technique for wireles communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

Sub-Band and Frequency Domain Double Talk Robust Algorithms

Jiaquan Huo[1], Sven Nordholm[1], Ka Fai Cedric Yiu[2] and Kok Lay Teo[2]

[1]Australian Telecom. Research Inst., Curtin University of Technology Perth, Australia, E-mail: Jiaquan@atri.curtin.edu.au

[2] Department of Applied Mathematics, The Hong Kong Polytechnic University Hong Kong, E-mail: mateokl@polyu.edu.hk

Abstract

Acoustic echo control is of vital interest for all hands-free operation of telecommunications equipment. In a normal office room environment the reverberation time will be several 100 ms, which corresponds to several hundred samples long equivalent discrete-time impulse response at a 8 kHz sampling rate. The long channel will create high complexity in the implementation of an acoustic echo canceller. Thus it is important to get numerically and complexity efficient methods, such as sub-band or frequency domain techniques.

An important property of an acoustic echo controller is its capability to handle double talk situation. An often used strategy is to use a robust adaptive algorithm in combination with a double talk detector and a 2-path model. A robust algorithm allows the double talk detector to get a more reliable double talk decision. In this paper robust frequency domain and sub-band domain algorithms are evaluated.

1 Introduction

Echo arises at various points in a voice communication network. Therefore it can cause significant degradation in conversation quality if proper control is not taken. A common way to address the echo problem is to employ an adaptive filters are [1]. The task of the adaptive filter is to estimate the echo path, when the echo path is known a replica of the echo is created and subtracted from the microphone signal. Since hands-free telephony is characterized by long impulse responses, which implies the the need of high order adaptive filters . An elegant solution to reduce the computational complexity of such high-order filters lies in the sub-band approach.

In the presence of a strong near end signal, which is referred to as a double-talk situation, the adaptation of the filter coefficients becomes questionable. The adaptive algorithm mistakes the near-end signal as un-cancelled echo and adjusts the filter coefficients in an unappropriate manner. This may cause the algorithm to diverge, resulting in unacceptable performance.

The general double-talk handling method is to stop the adaptation whenever a strong near-end signal is detected. Different double-talk detection schemes have been proposed in the literature [2]. In [3], robust NLMS algorithms were proposed for network echo cancellation. In this paper, the robust algorithm in [3] is extended to a frequency domain implementation using a delay-less structure, see Fig. 1. The results indicate that there are a number of clear benefits by using a robust sub-band algorithm implementation, namely low computation complexity, fast convergence enhanced by independent normalization in each sub-band, high robustness to double talk disturbance and fast tracking.

2 Sub-band and Frequency Domain Filtering

The delayless subband echo canceller is illustrated in Fig. 1. In the delay-less sub-band echo canceller, the echo path is modelled in sub-bands with a set of parallel adaptive filters. The sub-band filters are then collectively transformed to a single full-band filter via a weight transform, using the FFT-2 method [4]. This full-band filter then models the acoustic channel. By separating the adaptation and filtering operation, it is possible to maintain the advantages from both. The adaptation will have the desired features of sub-band processing such as signal de-correlation and computational efficiency. The filtering operation will have no extra delay as is the case in conventional sub-band echo cancellation schemes.

The signal $x(n)$ is filtered by a set of L tap analysis filters $\mathbf{a}_m(n)$ and decimated by a factor of D to form a set of M sub-signals $x_m(k)$ as

$$x_m(k) = \mathbf{a}_m^H \mathbf{x}(kD) \tag{1}$$

where

$$\mathbf{x}(kD) = \begin{bmatrix} x(kD) \\ x(kD-1) \\ \vdots \\ x(kD-L+1) \end{bmatrix} \tag{2}$$

The analysis filters are related to a single prototype filter

as

$$A_m(z) = A_0(zW_M^m) \tag{3}$$

where $A_m(z)$ is the transfer function of the mth analysis filter, $A_0(z)$ is that of the prototype filter and $W_M^m = e^{-j2m\pi/M}$. Such a set of filters is referred to as a uniform DFT-modulated filter bank [5] and can be implemented very efficiently.

The delay-less sub-band adaptive filter can be configured in either an open-loop or a closed-loop manner. For both configurations, the sub-band adaptive filters are chosen to minimize the sub-band error signal power. In the open-loop configuration, the sub-band error signal $e_m(k)$ is generated locally in the m:th sub-band as

$$e_m(k) = d_m(k) - \hat{\mathbf{h}}_m^H \mathbf{x}_m(k) \tag{4}$$

where the sub-band reference signal $d_m(k)$ is formed from $d(n)$ in the same way as Eq. 1. The closed-loop structure on the other hand, obtains the sub-band error signals by analyzing the full-band error signal $e(n)$ by $\mathbf{a}_m$ as in Eq. 1. By utilizing the full-band error signal, it is possible for the closed-loop delay-less sub-band adaptive filter to achieve the optimal Wiener solution, free of aliasing distortion.

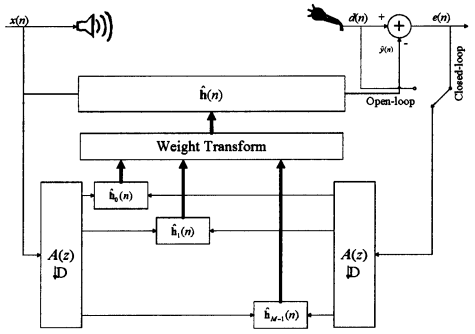


Figure 1: Delay-less sub-band adaptive filter structure

In the study only closed-loop form of the delay-less sub-band adaptive filter architecture is considered. The closed loop implementation has the potential to reach the same suppression level as a full-band implementation, which gives this choice a fair comparison. As mentioned above, echo is cancelled using a N tap FIR full-band filter characterized by the impulse response $\hat{\mathbf{h}}(n)$. Therefore, to get an efficient implementation the full-band filter is implemented in the frequency domain [6, 7].

3 Methods for Robust Adaptive Filtering

The general idea for double talk handling is to use an adaptive algorithm that is robust to double talk errors thus allowing time for a double talk detector to operate [2, 3]. The critical part to study is the size of misalignment and how fast the misalignment occurs in double talk situations. There is a trade-off between this robustness to double talk and convergence and tracking of channel variations. The standard adaptive algorithm used in acoustic echo cancellation is the normalized least mean squares (NLMS) algorithm. To get a more robust adaptive algorithm while maintaining the fast convergence for small errors, thus the capability to track variations in the channel, an error function that is quadratic for small errors and linear for large errors is commonly used [2, 3]. Since derivatives are essential for the algorithm to converge quickly, smooth functions are preferable in measuring the modelling errors. Robust statistics provide several techniques [8, 9] to combine the least-squares measure with a more robust measure for large residuals. Here, we concentrate on the Hubers function, which have been shown to work well. In earlier work by Huo et. al. [10] several different functions have applied to design filters adaptively in the presence of double-talk. In that study it was concluded that the Huber function has the desired properties of trading robustness versus tracking and convergence.

The idea is to apply Hubers function in each subband individually, the subband signals are decimated by a factor of $D = M/2$ and used to adapt an N/D tap adaptive filter $\hat{\mathbf{h}}_m(k)$, where m is the index of subbands, and k is the subband time index, as [3]

$$\hat{\mathbf{h}}_m(k+1) = \hat{\mathbf{h}}_m(k) + \mu(k) sgn(e_m(k))^* \psi(e_m(k)) \mathbf{x}_m(k) \tag{5}$$

where

$$\mu(k) = \frac{\mu}{\| \mathbf{x}_m(k) \|_2^2 + \delta} \tag{6}$$

$$sgn(e_m(k)) = \frac{e_m(k)}{|e_m(k)|} \tag{7}$$

$$\psi(e_m(k)) = \min(|e_m(k)|, s_m(k) k_0) \tag{8}$$

where k_0 and δ are a constants, $s_m(k)$ is an adaptive scale parameter. The scaling parameter $s_m(k)$ is adapted according to

$$s_m(k+1) = \lambda s_m(k) + \frac{1-\lambda}{\beta} \psi(e_m(k)) \tag{9}$$

given the constants λ and β, the scale factor $s_m(k)$ controls the robustness and tracking performance of the algorithm. According to [3, 11], β should be a parameter that is determined by k_0. However, in our experiments, the dependence between these two parameters does not turn out to be critical. It can be deduced from the above equation that β put emphasis on new data and raises the steady state value of $s_m(k)$, assuming $\beta < 1$. Hence, a smaller β implies better tracking but less robustness. The forgetting factor λ plays a similar role. This adaptation should only be carried out during far-end single talk period.

Another important factor in order to handle the double talk is to include energy detectors [12] in each sub-band. These energy detectors are included to inhibit adaptation in sub-bands with weak far-end signal.

4 Simulation Results

In the simulation, the full band filter length is set to $N = 1024$, the number of sub-bands is $M = 128$ with a decimation factor of $D = 64$.

Near-end speech is introduced from sample 80,000 to 100,000. Four situations are studied, see Fig. 2, two closed loop sub-band situations usingtwo different β factors, either $\beta = 1$ or $\beta = 0.60665$ and two robust NLMS situations as described by Gänsler [3, 11] with the same β factors, as above. In the $\beta = 1$ situation both the robust NLMS and the closed loop sub-band method show a high degree of robustness. The closed loop sub-band method is almost unaffected by the double talk. The $\beta = 0.60665$ gives a too high threshold and the behavior will be close to a non-robust algorithm. The faster convergence of the closed loop sub-band method also means that it diverges faster in a non-robust situation. It is thus critical to manage the threshold level of the algorithms. The results for different values of β are shown in Fig. 3.

Echo path variation is simulated by switching the original channel to a new channel at sample 110,000. This new channel corresponds to moving the microphone about 24cm from the original position. The closed loop sub-band methods consistently gives a faster re-convergence, see Fig. 4. The scaling factor is also presented in Fig. 5. Clearly the scaling factor is not much affected by the channel variation.

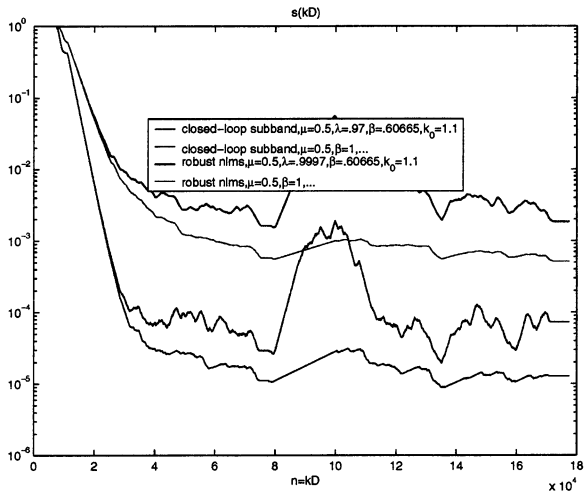


Figure 3: Evolution of scale parameters during double talk.

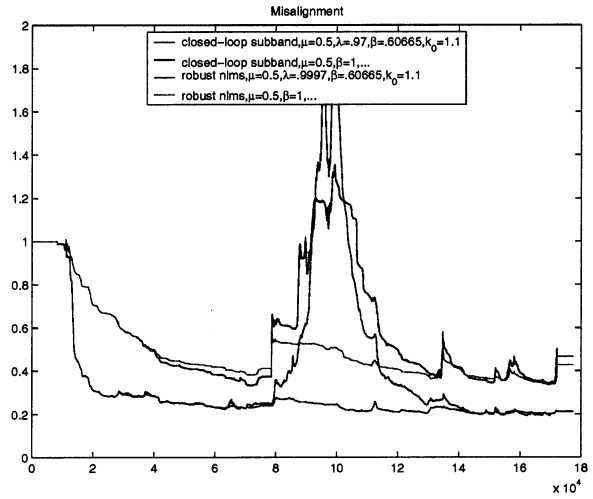


Figure 2: Comparison of robustness during double talk

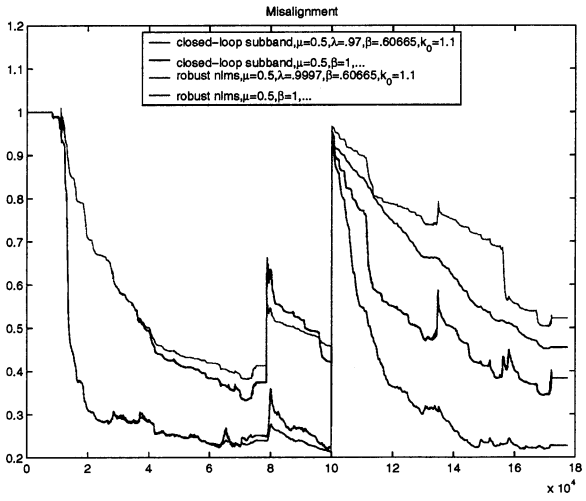


Figure 4: Comparison of tracking

5 Conclusions

The study has shown that it is beneficial to combine a robust algorithm with sub-band processing. This allows us to do time frequency analysis of the involved signals, which improves the performance of the robust algorithm. Sum-

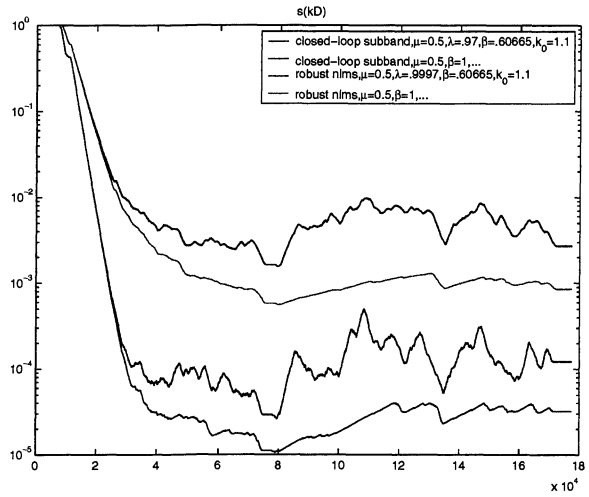


Figure 5: Evolution of scale parameters during tracking.

marizing the benefits from using a robust sub-band algorithm implementation:

- Less computations, there is a saving in computational load due to the decimation in the sub-bands.
- Independent normalization, the NLMS algorithm is normalized individually in each sub-band.
- Energy detection, the input signal energy level in each sub-band is estimated thus enabling control of the weight update depending on the input signal excitation in that sub-band.
- Faster tracking, due to the decoupling of the sub-bands, which allows faster and individual up-date of the scaling parameter determining the robustness.

Further studies is needed to understand performance for time-varying channels in a real environment. The performance of this algorithm in conjunction with a double talk detector and a two-path model is another topic of vital interest.

References

[1] Steven L. Gay. An introduction to acoustic echo and noise control. In S.L. Gay and J. Benesty, editors, *Acoustic signal processing for telecommunication*, chapter 1. Kluwer Academic Pubishers, 2000.

[2] T. Gänsler, J. Benesty, and S.L. Gay. Double-talk dection schemes for acoustic echo cancellation. In S.L. Gay and J. Benesty, editors, *Acoustic signal processing for telecommunication*, chapter 5, pages 81–97. Kluwer Academic Publishers, 2000.

[3] T. Gänsler, S.L. Gay, M.M. Sondhi, and J. Benesty. Double-talk robust fast converging algorithms for network echo cancellation. *IEEE Trans. Speech and Audio Proc.*, 8(6):656–663, 2000.

[4] J. Huo, S. Nordholm, and Z. Zang. New weight transform schemes for delayless subband adaptive filters. In *Globecom 2001*, 2001.

[5] P.P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.

[6] M. Xu and Y. Grenier. Time-frequency domain adaptive filter. In *Proc. ICASSP'89*, pages 1154–1157, 1989.

[7] J. M. P. Borallo and M. G. Otero. On the implementation of a partioned block frequency domain adaptive filter (PBFDAF) for long acoustic echo cancellation. *Signal Processing*, 27:301–315, June 1992.

[8] P. J. Huber. *Robust Statistics*. John Wiley and Sons, 1981.

[9] F. Hampel, P. Rousseeuw E. Ronchetti, and W. Stahel. *Robust Statistics*. John Wiley and Sons, 1986.

[10] J. Huo, K.F.C. Yiu, and K.L. Teo. On the robust filter design for echo cancellation with double-talk. In *Proc. International Conference on Optimisation Techniques and Applications*, volume 1, pages 699–703, 2001.

[11] T. Gänsler. A double-talk resisant subband echo canceller. *Signal Processing*, 65:89–101, 1998.

[12] J. Huo, S. Nordholm, and Z. Zang. Using recursive least-squares algorithms for acoustic echo cancellation. In *Proc. ECMCS'2001*, 2001.

Subband Signal Separation

Jörgen Nordberg[1], Hai Huyen Dam[2] and Sven Nordholm[3]

[2] Blekinge Institute of Technology, 372 25 Sweden. E-mail: Jorgen.Nordberg@bth.se.

[2]Australian Telecommunications Research Institute, Curtin University of Technology, Perth, Australia and Australian Telecommunications CRC. E-mail: dam@atri.curtin.edu.au.

[3] Australian Telecommunications Research Institute, Curtin University of Technology, Perth, Australia and Australian Telecommunications CRC. E-mail: sven@atri.curtin.edu.au.

Abstract

In wireless communication, the transmitted signals arrive at the receiver will be subjected to multipath propagation, i.e. the signals will be received with different time delays and attenuations. It is therefore not enough to use only a spatial receiver since such a receiver can not exploit the multipath properties. A temporal receiver, on the other hand, works well for multipath situations but does not utilize the spatial information which is of vital importance in multiuser systems. Thus, a combined spatial and temporal receiver structure can utilize multipath, spatial distribution and independence properties of the multiple sources. Such a spatial-temporal receiver structure is usually computationally complex. In this paper a subband signals separator is presented to reduce the required complexity without a significantly reduction in performance. The subband signal separator will be compared with its fullband equivalent using performance measures as the complexity and mean-square-error. Simulation results show that up to 82 % reduction in the computationally complexity with only a slightly degradation in mean square error performance.

1 Introduction

The demand of increasing capacity in future and existing wireless communication systems have recently resulted in a lot of interest into smart antennas and multi-input multi-output (MIMO) systems [1]. For MIMO systems, the multipath richness of the channel where several versions of the transmitted signals arrive at the receiver with different delays and attenuations determines the channel capacity. It is not enough to use only a spatial receiver since such a receiver can not exploit the multipath properties. A temporal receiver structure, on the other hand, works well for multipath situations but does not utilize the spatial information which is of vital importance in multiuser systems. Thus, a combined spatial and temporal receiver structure will provide a solution to these demands.

The transmitted signals are mixed in the channel by multi-paths having different attenuations and delays. The receiver takes the signals at each antenna element and un-mixes them as close as possible to the original transmitted signals. The spatial-temporal receiver can utilize multipath, spatial diversity and independence properties of the multiple sources. One of the major concerns in using this type of receiver is the complexity. The computational cost can be substantially reduced by using a multi-rate signal processing receiver.

The multi-rate signal processing technique employed here is a delayless filterbank (FB) procedure [2]. This technique uses the advantage of multi-rate signal processing for the implementation of the algorithm but avoids the filterbank and block processing delays by transforming the un-mixing subband filters to an un-mixing fullband filter. The difference in signal separation performance of subband and fullband solution depends on the filterbank and the subband to fullband transformation.

The delayless signal separation has been investigated in [3]. In this paper, a filter bank design method is considered to improve signal separation performance. This performance depends on both the in-band aliasing and design of the filterbank for a fixed cut-off frequency. Both the in-band aliasing and the filter performance with respect to the least square criterion can be formulated as quadratic functions of the filter coefficients. The design problem becomes to minimize filter performance with respect to some acceptable attenuation in the passband. This problem can be formulated as a semi-infinite quadratic programming problem, which can be solved using recently developed semi-infinite quadratic programming approach

[4] or conventional discretization approach.

Simulation results show that the delayless multirate implementation has approximately the same separation performance as the fullband but with a few percentage of the computational complexity. The computational complexity for the subband and the fullband scheme is compared by using a relative number of floating point operations (RFLOPS) measure. This RFLOPS is based on the number of the fullband floating point operation (FLOPS). The RFLOPS is reduced when the number of subband increases at the expense of the separation performance. The trade-off in performance of the mean square error (MSE) and the RFLOPS for different number of subbands is also discussed.

2 System Model

Consider the system model given in Figure 1. The statistically independent source bits are generated with equal probabilities and modulated using binary phase shift keying (BPSK).

The S transmitting signals are denoted as

$$\mathbf{s}[n] = [s_1[n], \cdots, s_S[n]],$$

where n is the time index through a medium, i.e. air, water or space. At the receiver, an array of I sensors picks up a set of signals

$$\mathbf{x}[n] = [x_1[n], \cdots, x_I[n]],$$

where $x_i[n]$, $1 \leq i \leq I$, is the received signal from the channel. The channel represents the medium which affects on the transmitting signals and the transmitting filters and can be described through propagation models consisting of doppler effects. This channel can be modeled as a time varying FIR channel mixing matrix with different characteristics such as deep nulls and multipath propagation. Denote the mixing matrix $\mathbf{C}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_{11} & \cdots & \mathbf{c}_{1S} \\ \vdots & \ddots & \vdots \\ \mathbf{c}_{I1} & \cdots & \mathbf{c}_{IS} \end{bmatrix} \tag{1}$$

where

$$\mathbf{c}_{ik} = [c_{ik}[0], \cdots, c_{ik}[N-1]]^T, 1 \leq i \leq I \text{ and } 1 \leq k \leq S$$

is the channel impulse response of length N from the k^{th} transmitted signal to the i^{th} sensor.

The received signal at the i^{th} sensor, $1 \leq i \leq I$, in a noise free case is given as

$$x_i[n] = \sum_{k=1}^{S} \sum_{l=0}^{N-1} c_{ik}[l] s_k[n-l]. \tag{2}$$

The signal $x_i[n]$ is then presented to un-mixing filters. These un-mixing filters estimates the inverse of the channel mixing filters which can have non-minimum or minimum phase response. If the channel mixing filters have non-minimum phase response, i.e. some of their zeros are outside the unit circle, their true inverting causal filters are not bounded. However, there exit bounded non-causal inverse filters. A delay τ_D is included in the receiver to enable a finite length approximation of the inverting filters, see Figure 2. Denote

$$\mathbf{w}_{ik} = [w_{ik}[0], \cdots, w_{ik}[L-1]]^T$$

as the impulse response of the un-mixing filter from the sensor signal $x_i[n]$ to the output $u_k[n]$. The filter length L is chosen so that the transmitted signal $s_k[n]$ can be recovered. The un-mixing filters form the following un-mixing matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_{11} & \cdots & \mathbf{w}_{1S} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{I1} & \cdots & \mathbf{w}_{IS} \end{bmatrix}. \tag{3}$$

In this study, the number of transmitted signals S is assumed to be equal to the number of sensors with $S = I = 2$. However, the problem can easily be extended for general situations.

In order to study the performance of the delayless subband signal separation method, the source signals are assumed to be known. The optimum un-mixing filters $\mathbf{W}$ is obtained by using the Wiener correlation method. Denote $\mathbf{R}_{x_i,x_k}$ as an $L \times L$ correlation matrix between the sensor signals $x_i[n]$ and $x_k[n]$ and $\mathbf{r}_{x_i,s_k}$ as an $L \times 1$ cross-correlation vector between the sensor signal $x_i[n]$ and the transmitted signal $s_k[n]$. These correlations are estimated from the data by using ensemble average [5]. The impulse response for the un-mixing filters are obtained by solving

$$\mathbf{R}_{\mathbf{xx}}\mathbf{W} = \mathbf{r}_{\mathbf{xs}}, \tag{4}$$

where

$$\mathbf{R}_{\mathbf{xx}} = \begin{bmatrix} \mathbf{R}_{x_1,x_1} & \mathbf{R}_{x_1,x_2} \\ \mathbf{R}_{x_2,x_1} & \mathbf{R}_{x_2,x_2} \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} \\ \mathbf{w}_{21} & \mathbf{w}_{22} \end{bmatrix},$$

and

$$\mathbf{r}_{\mathbf{xs}} = \begin{bmatrix} \mathbf{r}_{x_1,s_1} & \mathbf{r}_{x_1,s_2} \\ \mathbf{r}_{x_2,s_1} & \mathbf{r}_{x_2,s_2} \end{bmatrix}.$$

These filters also minimize mean squared error between the output signal and the transmitted signal [3]. The output signals from the un-mixing filters are given as follows:

$$u_k[n] = \sum_{i=1}^{2} \sum_{l=0}^{L-1} w_{ik}[l] x_i[n-l], \qquad 1 \leq k \leq 2. \tag{5}$$

3 Delayless Subband Signal Separation

The proposed delayless subband signal separation scheme [2] is presented in Figure 2. The transmitted and the received signals are divided into M frequency bands. The subband un-mixing filters are obtained in each frequency bands by using the optimal Wiener correlation method, see (4). These subband un-mixing filters are transformed into the frequency domain by using a fast Fourier transform (FFT). The frequency response for the fullband un-mixing filters are obtained by stacking different subband frequency responses. The impulse response for the fullband un-mixing filters are the inverse FFT (IFFT) of the stacked fullband frequency response [2].

An uniformly modulated filterbank (UMF) is used to split the signals into different frequency bands. The different subbands filters in the filterbank are the modulated versions of a prototype filter. Thus, the design of the UMF is reduced to the design of a prototype filter with length K and the impulse response $\mathbf{h}$,

$$\mathbf{h} = [h[0], \cdots, h[K-1]]^T \tag{6}$$

where $[.]^T$ denotes the transpose of a vector. The transfer function of the prototype filter becomes

$$H(z) = \sum_{l=0}^{K-1} h[l] z^{-l} = \mathbf{h}^T \phi(z) \tag{7}$$

where $\phi(z) = [1, \cdots, z^{-(K-1)}]^T$.

The transfer function of the m^{th} subband in a UMF is obtained from $H(z)$ as follows

$$H_m(z) = H(zW_M^m) = \sum_{l=0}^{K-1} h[l](zW_M^m)^{-l} \tag{8}$$

where $W_M = e^{-j2\pi/M}$ and $0 \leq m \leq M-1$.

Polyphase decomposition is used to implement the UMF efficiently where the decimation is performed before the subband filtering [6]. To avoid aliasing, the filterbank is oversampled with a factor of two, i.e. the subband signals are decimated by a factor $D = \frac{M}{2}$.

The transformation between subband and fullband depends on the performance of the transfer function $H(z)$ with the corresponding frequency response $H(e^{j\omega})$. In the following section, the design of $H(e^{j\omega})$ will be considered. The desired filter is lowpass with a normalised cut-off frequency $\omega_p = \pi/M$. Since it is desirable to have approximately linear phase in the passband, the desired complex frequency response is specified as

$$H_d(e^{j\omega}) = e^{j\omega\tau}, \quad \forall \omega \in [-\omega_p, \omega_p] \tag{9}$$

where $\tau = (L-1)/2$.

The most common criteria in filter design are least square and min-max [7]. Since it is important to have a flat passband response for the prototype filter, a least square criterion is chosen in the cost function. However, a min-max criterion is also included in the constraints to restrict the maximum attenuation of the filter within some acceptable region.

The least square and the min-max errors are given as

$$e_1(\mathbf{h}) = \frac{1}{2\omega_p} \int_{-\omega_p}^{\omega_p} |H(e^{j\omega}) - H_d(e^{j\omega})|^2 d\omega \tag{10}$$

and

$$e_2(\mathbf{h}) = \max_{-\omega_p \leq \omega \leq \omega_p} |H(e^{j\omega}) - H_d(e^{j\omega})|. \tag{11}$$

It follows from (7) that $H(e^{j\omega})$ is a linear function of the coefficient $\mathbf{h}$. Thus, (10) is reduced to a quadratic function of $\mathbf{h}$

$$e_1(\mathbf{h}) = \mathbf{h}^T \mathbf{A} \mathbf{h} - 2\mathbf{h}^T \mathbf{b} + 1,$$

where $\mathbf{A}$ is a $K \times K$ matrix

$$\mathbf{A} = \frac{1}{2\omega_p} \int_{-\omega_p}^{\omega_p} \phi(e^{j\omega}) \phi^H(e^{j\omega}) d\omega, \tag{12}$$

and $\mathbf{b}$ is a $K \times 1$ vector

$$\mathbf{b} = \frac{1}{2\omega_p} \int_{-\omega_p}^{\omega_p} \mathcal{R}\{e^{j\omega\tau} \phi^H(e^{j\omega})\} d\omega. \tag{13}$$

The operator $[.]^H$ and $\mathcal{R}\{.\}$ denote the Hermitian operation of a matrix and the real part of a complex function, respectively.

The prototype filter is also designed to have a minimum in-band aliasing effect. Since the transfer functions of the subband filters are related by (8), it is sufficient to minimize the energy in the aliasing terms of the first subband. This aliasing error can be formulated as a quadratic function of $\mathbf{h}$ [8],

$$\beta(\mathbf{h}) = \mathbf{h}^T \mathbf{C} \mathbf{h}, \tag{14}$$

where $\mathbf{C}$ is a $K \times K$ matrix,

$$\mathbf{C} = \frac{1}{2\pi D} \sum_{d=1}^{D-1} \int_{-\pi}^{\pi} \phi(e^{j\omega/D} W_D^d) \phi^H(e^{j\omega/D} W_D^d) d\omega.$$

The problem becomes to minimize both the least square error $e_1(\mathbf{h})$ and the aliasing effect $\beta(\mathbf{h})$ with respect to some acceptable attenuation in the passband. A joint cost function can be formulated as

$$e_1(\mathbf{h}) + \beta(\mathbf{h}) = \mathbf{h}^T (\mathbf{A} + \mathbf{C}) \mathbf{h} - 2\mathbf{h}^T \mathbf{b} + 1. \tag{15}$$

Combine this cost function with the min-max constraints (11), we have the following problem

$$\begin{cases} \min\limits_{\mathbf{h}} \mathbf{h}^T (\mathbf{A} + \mathbf{C}) \mathbf{h} - 2\mathbf{h}^T \mathbf{b} + 1 \\ |H(e^{j\omega}) - H_d(e^{j\omega})| \leq \epsilon, \quad \forall \omega \in -[\omega_p, \omega_p] \end{cases} \tag{16}$$

where ϵ is the maximum deviation in the passband. This problem (16) can be solved using: (i) semi–infinite quadratic program technique or (ii) conventional quadratic programming using discretization. In this paper, we use the second method to design the prototype filter. The opitmazation problem is then reduced to a quadratic programming problem, which can then be solved using standard quadratic programming techniques.

4 Simulation Examples

In this section, performance comparison between the subband and the fullband approach is presented. This comparison is based on the MSE defined for all the sources,

$$\varepsilon_i = 10\log_{10}\left(\frac{\sum_n |s_i[n]-u_i[n]|^2}{\sum_n |s_i(n)|^2}\right),\ 1\le i\le 2, \quad (17)$$

where the summation is taken over the total number of transmitted symbols, which is set to 10^5 in all the simulations.

Comparison is also based on RFLOPS which is normalized to fullband (with one subband) FLOPS. The design of the prototype filter is excluded in the calculation of RFLOPS since this filter depends only on the number of subbands. The RFLOPS represents the computational complexity of the different equalization structures. The MSE and RFLOPS are evaluated and compared using a stochastic channel.

Example : Monte Carlo channel mixing matrix

A Monte Carlo simulation is run by using 30 random FIR channel mixing filters with length 10. The impulse responses for the mixing filters are generated from a gaussian distribution with mean 0 and variance 1. The ensemble average MSE are obtained by averaging the RFLOPS and the MSE.

Tables 1-2 compare the performance of the two prototype filters with 256–tap un-mixing filters at the receiver. The length of the prototype filters is four times the total number of subbands. The impulse response for the first prototype filter (Table 1) is obtained by minimizing the least square error criterion in conjunction with the Hamming window [9], while the impulse response for the second prototype filter (Table 2) is obtained by optimizing the filter performance and the in-band aliasing, see (16). Please note, that ϵ in (16) is chosen such that the filter has less than 1 dB attenuation in the passband. The MSE for the two sources is improved significantly by using the second prototype filter. This improvement is more distinct when the total number of subbands is large, e.g. $M = 16$ or $M = 32$.

The Tables also show the trade-off between the MSE and the RFLOPS when the number of subband increases from 4 to 32. The first row corresponds to the performance of the fullband solution ($M = 1$). When $M = 4$, the RFLOPS increases while the MSE remains approximately the same. This increased in computational complexity is due to the use of an oversampled filterbank. Thus, M requires to be larger than four to have a computational reduction. When the number of subbands is increased to 8, the RFLOPS is reduced by 32% at an expense of 0.3 dB reduction in the MSE. The MSE are slightly lower with an increase in the number of subbands to $M = 16$ and $M = 32$.

5 Conclusions

In this paper, the performance of subband and fullband methods are compared for different number of subbands. Simulations have been performed for a stochastic channel. The results show that by properly optimizing the prototype filter, the subband performance is significantly improved. Moreover, the relative number of floating point operations is significantly reduced for more than four subbands at an expense of a small degradation in the MSE.

Future work includes further optimizing the prototype filter and incorporating adaptive estimation for the subband un-mixing filters. The adaptive estimation can be implemented by using a blind separation algorithm, e.g. independent component analysis (ICA) [3]. The advantage of a blind algorithm is that the training data (pilot sequence) is not required at the receiver.

6 Acknowledgement

The authors would like to thank Tech. Licentiate J. M. de Haan for uesfull discussion on filterbank design.

References

[1] G. Foschini, "Layered Space-Time Architecture for Wireless Communication in a Fading Environment when using Multi-element Antennas," Bell Labs Technical Journal, pp. 41-59, Autumn 1996.

[2] D. R. Morgan, and J. C. Thi, "A Delay-less Subband Adaptive Filter Architecture," *IEEE Trans. on Signal Processing*, vol. 43, no. 8, pp. 1819-1830, Aug 1995.

[3] N. Grbic, X. Tao, S. Nordholm, and Ingvar Claesson, "Blind Signal Separation Using Overcomplete Subband Representation," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, July 2001.

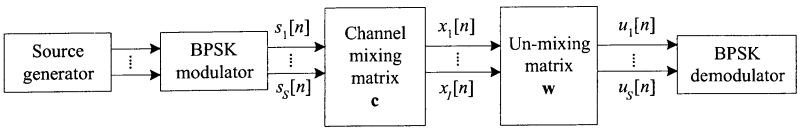


Figure 1: Channel model.

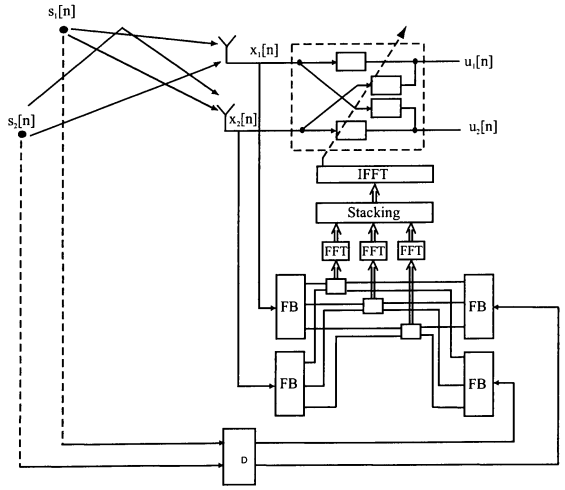


Figure 2: A non-causal delayed subband signal separation scheme.

Number of subbands	Prototype filter 1		
	RFLOPS	ε_1 [dB]	ε_2 [dB]
1	1.0	-28.4	-29.2
4	1.4	-28.0	-28.8
8	0.68	-27.5	-28.0
16	0.33	-25.9	-26.6
32	0.18	-22.6	-23.3

Table 1: Monte Carlo simulation with 256 taps for the un-mixing filters. Prototype filter 1 is obtained by minimizing the least square solution in conjunction with Hamming window method.

Number of subbands	Prototype filter 2		
	RFLOPS	ε_1 [dB]	ε_2 [dB]
1	1.0	-28.4	-29.2
4	1.4	-28.3	-29.0
8	0.68	-28.1	-28.8
16	0.33	-27.6	-28.4
32	0.18	-26.5	-27.1

Table 2: Monte Carlo simulation with 256 taps for the un-mixing filters. Prototype filter 2 is obtained by minimizing the filter performance in conjunction with the in-band aliasing effect.

[4] H. H. Dam, K. L. Teo, S. Nordebo, and A. Cantoni, "The Dual Parameterization Approach to Optimal Least Square FIR Filter Design Subject to Maximum Error Constraints," *IEEE Trans. Signal Processing*, vol. 48, pp. 2314–2320, no. 8, Aug. 2000.

[5] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, 1996.

[6] P. P. Vaidyanathan, *Multirate Systems and Filter Banks* Prentice Hall, 1993.

[7] T. W. Parks, C. S. Burrus, *Digital Filter Design* John Wiley & Sons Inc, 1987

[8] J. M. De Haan, N. Grbic, I. Claesson, and S. Nordholm, "Design of Oversampled Uniform DFT Filter Banks with Delay Specifications using Quadratic Optimization," *In Proc. The 2001 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP-01, vol. VI, pp. 3633–3636, Salt Lake City, USA, May 2001.

[9] Matlab: "Reference Guide," The Math Works, 1996.

Synchronization of Multicarrier Systems on the Twisted Copper Pair Channel

Daniel Landström[1], Niklas Petersson[2], Per Ödling[3] and Per Ola Börjesson[4]

Lund University, Dept. of Electroscience, Box 118, SE-221 00 Lund, Sweden
[1]Daniel.Landstrom@es.lth.se, [2]Niklas.Petersson@es.lth.se, [3]Per.Odling@es.lth.se, [4]Per.Ola.Borjesson@es.lth.se

Abstract

We address time synchronization in an OFDM system and present an ML estimator that is unbiased also for dispersive channels. The estimator uses the redundancy introduced by the cyclic prefix, and knowledge of the channel. Previously presented estimators based on the cyclic prefix are either biased or have high complexity for this environment. In this paper we show that with proper choice of signal model, the complexity can be reduced considerably while maintaining the performance. Part of the novelty lies in the proper decomposition of correlation matrices that appear in the ML estimator. The new estimator structure consists of a linear filter, followed by an auto-correlating filterbank and a maximization.

1 Medea+ INCA A106

The research presented here is a part of a Medea+ project. Medea+ is a sub-program of the pan-European cooperative research and development program EUREKA. Medea+ is aiming at strengthening the electronics and communications industries in Europe, in particular the European semiconductor companies, closing the gap in key technologies with the global competitors from the USA and from Asia / Pacific. Medea+ has 38 projects, one of them being INCA A106. INCA is a collaboration between 13 parties with the main purpouse of developing top-of-the-line DSL modems.

Within this project, we collaborate closely with Telecommunications Research Center Vienna (FTW) and Telia Research AB. We work with Synchronization, RFI problems, Clipping mitigation, in developing signal processing algorithms for the wireline system-on-chip applications.

This work was performed in the Media+ INCA A106 project and is supported by VINNOVA.

This work has i part been presented as D. Landström, N. Petersson, P. Ödling, P. O. Börjesson, "OFDM Frame Synchronization for Dispersive Channels", Proceedings of the 6th IEEE International Symposium on Signal Processing and its Applications (ISSPA 2001), Kuala-Lampur, Malaysia, August 2001.

2 Introduction

In this paper we focus on the estimation of time offset for *Orthogonal Frequency-Division Multiplexing* (OFDM) symbols for synchronization purposes in a dispersive channel environment. OFDM systems are in general quite sensitive to time and frequency offsets [1]. Not only may synchronization errors cause *intersymbol-interference* (ISI), they can also cause the loss of orthogonality between the subcarriers resulting in *intercarrier-interference* (ICI) [2].

OFDM systems often operate in a multipath environment, and thus under a dispersive channel [3]. In [4] a *maximum-likelihood* (ML) estimator for OFDM systems with a *cyclic prefix* (CP) [5] was derived for an *additive white Gaussian noise* (AWGN) channel. Applying this estimator to a system with a dispersive channel results in an error floor in the time and frequency offset estimation. The error floor stems from the estimator being biased in this environment [6]. In the dispersive channel environment the channel will introduce dependency between the samples, and the simple correlation structure of the received signal used in the AWGN model is not valid any more. In [4] the time offset estimate, $\widehat{\theta}$, is taken as the position of the log-likelihood function's peak. The estimate will be biased if the estimator from [4] is used on a signal that has been transmitted over a dispersive channel. This is because the function where we look for the maximum will be

$$\Lambda_{\text{Disp}}(k) = |h(k)|^2 * \Lambda_{\text{AWGN}}(k)$$

where $\Lambda_{\text{AWGN}}(k)$ is the log-likelihood function from [4], and $\Lambda_{\text{Disp}}(k)$ is the same function calculated on a signal that has been transmitted over the dispersive channel $h(k)$.

The difficulties in the dispersive channel has been observed and addressed before. In [6] an ML estimator for dispersive channels is derived. This estimator is complex to implement, but is suggested as a bound on the achievable performance for ML estimation based on cyclic prefix and channel knowledge.

We now present an ML estimator for use with

dispersive channels. This estimator will be unbiased unlike previous cyclic-prefix based ML estimators designed for the AWGN channel. The estimator exploits the redundancy of the received signal introduced by the cyclic prefix and knowledge of the channel impulse response to determine the time offset of the received OFDM symbol. The estimator uses a new structure shown in Figure 1 and can be implemented with lower complexity compared to [6], while essentially maintaining the optimal performance. The basic idea is to have a prefilter collect the information, so that the correlation can be calculated with lower dimension.

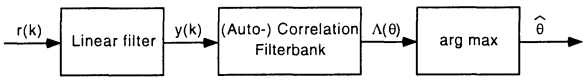


Figure 1: Estimator structure

3 The signal model

In an OFDM system [3], complex data symbols are modulated on a large number of carriers by means of an inverse *discrete Fourier transform* (DFT) of size N. Each block of samples is cyclically extended with a prefix of length L before it is transmitted over the channel [5]. In the receiver the cyclic prefix is removed and the data are demodulated by means of a DFT. If synchronization is correct and the cyclic prefix is longer than the length of the channel impulse response, ISI and ICI is avoided [5].

The sensitivity to a symbol-time offset has been investigated in [7]. The requirements on the time offset estimate $\widehat{\theta}$ is determined by the difference in length between the cyclic prefix and the channel impulse response. To avoid ISI and ICI, the time offset estimate needs to be within an interval so that the samples fed to the FFT are not affected by the previous symbol through channel leakage. The time offset will then only appear as phase offsets of the demodulated data symbols. In a coherent system the phase offsets will be compensated for by the channel equalizer and the system performance depends on the performance of the channel estimator and equalizer. For larger time offsets ISI and ICI occur [7]. Symbol timing requirements can thus be relaxed by increasing the length of the cyclic prefix.

Unlike a real OFDM system, where each symbol is cyclically repeated, we model only one cyclic prefix in our signal model. This way we can derive an estimator that uses information from one symbol only, whereas the observation interval can be longer to handle smearing effects of the dispersive channel. To model our system we assume that the transmitted discrete-time signal $s(k)$ is Gaussian . Because of the cyclic prefix this signal is not white, since L consecutive samples are pairwise correlated with L other

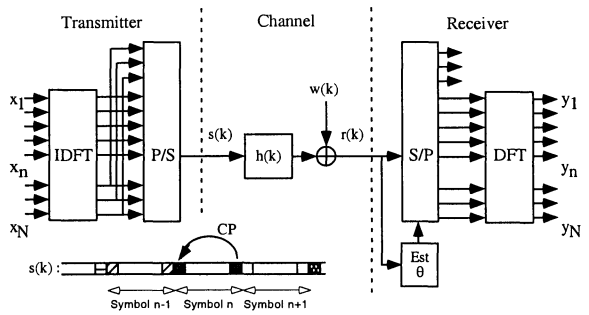


Figure 2: A typical OFDM system. Note the cyclic repetitions in the transmitted signal $s(k)$.

consecutive samples, N samples ahead, see Figure 2. By observing this correlation one may find where the OFDM symbol is likely to start. In our model the transmitted signal $s(k)$ has the covariance function

$$\mathbf{C_s}(\theta)_{i,j} = \begin{cases} \sigma_s^2 & \text{if } j-i=0, \\ \sigma_s^2 & \text{if } j-i=N,\ i \in [\theta, \theta+L-1] \\ \sigma_s^2 & \text{if } i-j=N,\ j \in [\theta, \theta+L-1] \\ 0 & \text{otherwise} \end{cases}.$$

We stress that this model only reflects the appearance of one cyclic prefix in the transmitted signal. The simulations will however evaluate the estimator performance for an OFDM system with a stream of symbols, each containing this redundancy.

The system we are targeting is to operate in a dispersive channel. Furthermore we assume the channel impulse response to be known to our estimator. Given the channel impulse response, $h(k)$, we model the received signal $r(k)$ as

$$r(k) = (h * s)(k-\theta) + w(k), \quad k \in \mathcal{I}_{obs}, \qquad (1)$$

where θ is the time offset, $\mathcal{I}_{obs}$ is the observation interval, and $w(k)$ is additive complex white Gaussian noise with variance σ_w^2, with zero mean. In matrix notation we can write this as

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{w},$$

where $\mathbf{H}$ is a matrix whose entries have the form $\mathbf{H}_{(i:i+H-1),i} = (h(0), h(1), \ldots, h(H-1))$, and $\mathbf{r}, \mathbf{s}$, and $\mathbf{w}$ are vectors with the samples from $r(k)$, $s(k-\theta)$, and $w(k)$, respectively.

Because of the dispersive channel, $h(k)$, the correlation structure of $s(k)$ is not transferred directly to the received signal $r(k)$. The correlation matrix for the received signal, $r(k)$, will be

$$\mathbf{C_r}(\theta) = \mathbf{H}\mathbf{C_s}(\theta)\mathbf{H}^H + \sigma_w^2\mathbf{I}.$$

4 The ML estimator

The ML estimator for model (1) is [6, 8]

$$\widehat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \left\{ -\mathbf{r}^H \mathbf{C}_{\mathbf{r}}^{-1}(\theta)\, \mathbf{r} \right\}. \qquad (2)$$

We can rewrite $\mathbf{C_r}(\theta)$ as the sum of matrices where one is independent of θ and the other is not,

$$\mathbf{C_r}(\theta) = \mathbf{C_{r0}} + \mathbf{C_{rcp}}(\theta),$$

where $\mathbf{C_{r0}} = \sigma_s^2 \mathbf{H}\mathbf{H}^H + \sigma_w^2 \mathbf{I}$ is a band matrix representing the channel and the noise correlation. The second term $\mathbf{C_{rcp}}(\theta)$ is a matrix representing the total correlation of samples located in the repeated parts, *i.e.*, the cyclic prefix. That is, $\mathbf{C_{rcp}}(\theta) = \mathbf{H}\mathbf{C_{cp}}(\theta)\mathbf{H}^H$, where

$$\mathbf{C_{cp}}(\theta)_{i,j} = \begin{cases} \sigma_s^2 & \text{if } j-i=N,\ i \in [\theta, \theta+L-1] \\ \sigma_s^2 & \text{if } i-j=N,\ j \in [\theta, \theta+L-1] \\ 0 & \text{otherwise} \end{cases},$$

as also illustrated in Figure 3.

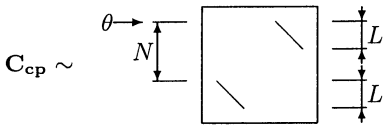


Figure 3: Schematic figure of the $\mathbf{C_{cp}}$ matrix. The elements in the marked diagonals are all σ_s^2.

Note that only $\mathbf{C_{rcp}}(\theta)$ depends on the unknown θ, and that the dependency lies in the positions of the non-zero values, not in the values themselves.

The ML estimator from Equation (2) can be written as

$$\widehat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \left\{ -\mathbf{r}^H \left(\mathbf{C}_{\mathbf{r}}^{-1}(\theta) - \mathbf{C}_{\mathbf{r0}}^{-1} \right) \mathbf{r} \right\}, \qquad (3)$$

since $\mathbf{C_{r0}}$ is independent of θ and does not affect the maximization [6].

In this paper we choose to perform the maximization over this form, rather than over Equation (2). By choosing this form the estimator can be given a new structure, that will not only reduce the complexity, but also make room for new interpretations. Following Appendix A, we can rewrite

$$\mathbf{r}^H \left(\mathbf{C}_{\mathbf{r}}^{-1}(\theta) - \mathbf{C}_{\mathbf{r0}}^{-1} \right) \mathbf{r} = \mathbf{y}^H \mathbf{Q}(\theta) \mathbf{y},$$

where

$$\mathbf{y} = \mathbf{H}^H \mathbf{C}_{\mathbf{r0}}^{-1} \mathbf{r},$$
$$\mathbf{Q}(\theta) = -\mathbf{C_{cp}}(\theta) + \mathbf{C_{cp}}(\theta) \left(\mathbf{H}^H \mathbf{C}_{\mathbf{r}}^{-1}(\theta) \mathbf{H} \right) \mathbf{C_{cp}}(\theta).$$

The estimator can be written

$$\widehat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \left\{ -\mathbf{y}^H \mathbf{Q}(\theta)\, \mathbf{y} \right\}. \qquad (4)$$

The properties of $\mathbf{Q}(\theta)$ is important for the complexity of the estimator. The first term is just the negative of the $\mathbf{C_{cp}}(\theta)$ matrix. The second term consists of three parts. By multiplying the middle expression with $\mathbf{C_{cp}}(\theta)$ from both sides we will cut out four parts of the middle expression. The $\mathbf{Q}(\theta)$ matrix will be zero in all positions except for in the four areas, see Figure 4.

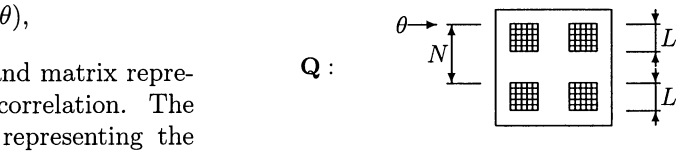


Figure 4: Schematic figure of the $\mathbf{Q}(\theta)$ matrix. All elements outside the marked areas are zero.

The location of the non-zero values in the $\mathbf{Q}$ matrix shows that only limited (length $L+L$) intervals of the $\mathbf{y}$ vector has to be known to calculate one point in the log-likelihood function. However, we need to calculate $\mathbf{y} = \mathbf{H}^H \mathbf{C}_{\mathbf{r0}}^{-1} \mathbf{r}$ to know the $L+L$ values. Since the inverse $\mathbf{C}_{\mathbf{r0}}^{-1}$ is not limited to a few diagonals, as the $\mathbf{C_{r0}}$ matrix is, we will need to know $\mathbf{r}$ in a larger interval.

Although $\mathbf{C_{r0}}$ is Toeplitz, $\mathbf{C}_{\mathbf{r0}}^{-1}$ is not. However, for infinite observation interval, $\mathcal{I}_{obs} \in (-\infty, \infty)$, the matrix multiplication $\mathbf{y} = \mathbf{G}\mathbf{r} = \mathbf{H}^H \mathbf{C}_{\mathbf{r0}}^{-1} \mathbf{r}$ could be expressed as a convolution, $y(k) = g(k) * r(k)$, with a non-causal filter represented by

$$g(k) = \mathcal{F}^{-1} \left\{ \frac{H^*(f)}{\sigma_s^2 |H(f)|^2 + \sigma_w^2} \right\}, \qquad (5)$$

where $H(f) = \mathcal{F}\{h(k)\}$.

For infinite observation interval it can be shown that the θ-dependency in $\mathbf{Q}(\theta)$ lies in the positions of the non-zero values, not in the values themselves. As the values are independent of θ, we can use a fixed matrix and slide the $\mathbf{y}$ vector to find the log-likelihood function. We write the quadratic form as

$$\begin{aligned} \Lambda(\theta) &= -\mathbf{y}^H \mathbf{Q}(\theta)\, \mathbf{y} \\ &= -\sum_k \sum_j \mathbf{y}_k^* \mathbf{Q}_{k,j}(\theta)\, \mathbf{y}_j \\ &= -\sum_i \sum_k \mathbf{y}_k^* \mathbf{Q}_{k-\theta,k-\theta+i}(0)\, \mathbf{y}_{k+i} \\ &= -\sum_i \sum_l \mathbf{y}_{l+\theta}^* \mathbf{Q}_{l,l+i}(0)\, \mathbf{y}_{l+\theta+i}, \end{aligned}$$

where $\mathbf{Q}(0)$ is the θ-independent matrix $\mathbf{Q}(\theta)|_{\theta=0}$. This result is important because one can now implement the estimator with a bank of time-invariant filters $\widetilde{q}_i(k)$, whose outputs we add.

$$\Lambda(\theta) = \sum_i \sum_k \widetilde{q}_i(\theta - k) y^*(k) y(k+i)$$

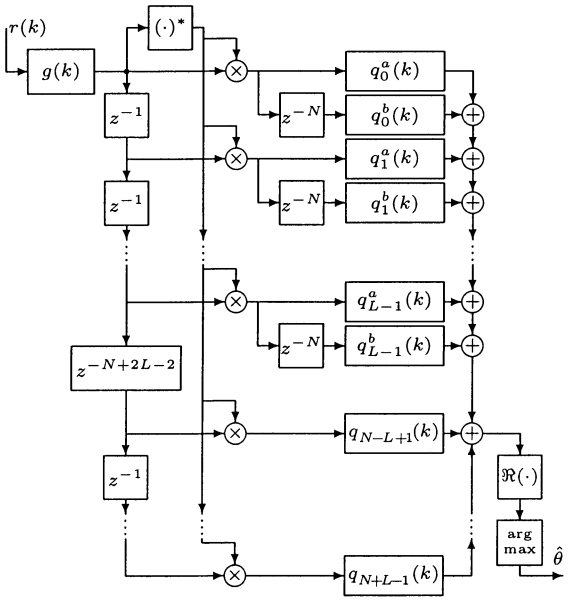


Figure 5: Structure of the estimator for dispersive channels.

where $\widetilde{q}_i(k)$ is the negative of the ith diagonal of $\mathbf{Q}(0)$, time-reversed.

Two properties of $\mathbf{Q}(0)$ are important for further reducing the complexity. First, it is symmetric so that only the upper triangular part of the matrix needs to be evaluated. Secondly, some of the diagonals are zero, and need not to be considered. The estimator can be expressed as

$$\widehat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} \left\{ \sum_{i \in \mathcal{I}} \gamma_i(\theta) \right\}, \qquad (6)$$

where $\mathcal{I} = \{0, \ldots, L-1, N-L+1, \ldots, N+L-1\}$,

$$\gamma_i(m) = \Re \left\{ \sum_{k=-\infty}^{\infty} q_i(m-k) y(k) y^*(k+i) \right\}, \qquad (7)$$

and

$$q_i(k) = \begin{cases} \widetilde{q}_i^*(k) & \text{if } i = 0 \\ 2\widetilde{q}_i^*(k) & \text{otherwise.} \end{cases}$$

The $q_i^a(k)$ and $q_i^b(k)$ in Figure 5 refers to the first and the second half of the $q_i(k)$ filters.

Estimator (6) suggests the implementation structure shown in Figure 5. First the products $y(k)y^*(k+i)$ are formed. These are fed into the filterbank where the outputs are summed, and the real part gives the log-likelihood function. The peak in the log-likelihood yields the time offset estimate.

5 Simulations

To compare the performance of the different estimators, a number of simulations have been performed. Although the estimator was derived with only one cyclic repetition, we now simulate it in an OFDM system with multiple cyclically extended symbols. The simulation system is an OFDM system with symbol length $N = 128$ and cyclic prefix length $L = 16$. The observation interval $\mathcal{I}_{obs}$ is of length $4N + 2L$, because θ has $N + L$ possible values, $\mathbf{Q}$ has dimension $N + L$ and the we choose a prefilter length of $2N$. In each simulation 50000 symbols are transmitted over a channel $h(n)$, which is assumed to be time-invariant, with impulse response

$$h(n) = \tfrac{1}{\sqrt{2.13}} \left(\delta(n) + 0.7\delta(n-6) + 0.8\delta(n-12) \right)$$

which has normalized energy.

We have simulated three estimators. As a reference we have the estimator from [4] which is designed for an AWGN channel (denoted AWGN in the figure). The second estimator is from Equation (3) which does the full matrix multiplication in the quadratic form (denoted Disp in the figure). The last estimator, from Equation (6) and shown in Figure 5, uses a prefilter and has a filterbank implementation of the quadratic form. For the simulations, the prefilter in Equation (5) is calculated by a $2N$-point FFT, and thereby constrained to a total length of $2N$ samples. The impulse response is delayed with N samples, making the first N samples of the impulse response correspond to the non-causal filter taps.

In Figure 6, the errors in the estimated time offsets, $\widehat{\theta} - \theta$, are shown in histograms. The top subfigure shows the ML estimator for an AWGN channel [4]. The estimate $\widehat{\theta}_{\mathrm{AWGN}}$ shows a considerable bias because of the dispersive channel. The two tops at -25 and 25 represent outliers that are more than 24 samples from the true θ. The center subfigure shows the estimates from Equation (3) (Disp). This estimator shows an unbiased behaviour. The bottom part, which shows the estimator in Figure 5, also shows an unbiased behaviour.

The mean square error, MSE, of the estimates for the different estimators is shown in Figure 7. Simulations were performed with SNR in the range 0 to 20 dB. The AWGN-based estimator has as expected the highest MSE. The two dispersion-based estimators shows a considerable smaller MSE, and are almost identical in performance up to SNR values of about 16 dB. There the prefilter-based estimator seems to reach an error floor, whereas the matrix-based estimator continues to give better performance. For such high SNR, where $\sigma_w^2 \ll \sigma_s^2$, the prefilter $g(k)$ in Equation (5) approaches the inverse filter to $h(k)$. Since the filter, $g(k)$, is limited to a length of $2N$, the impulse response is cut off at a point where the filter coefficients are not negligible, which causes the degradation in performance.

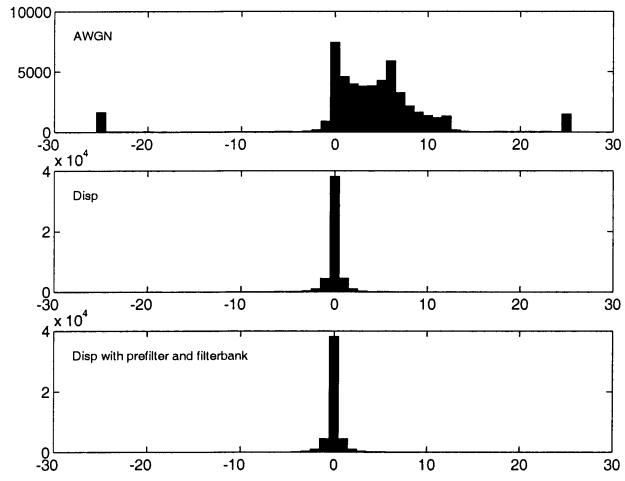


Figure 6: Histogram over the time offsets at SNR = 10 dB. Please note the difference in the vertical scale.

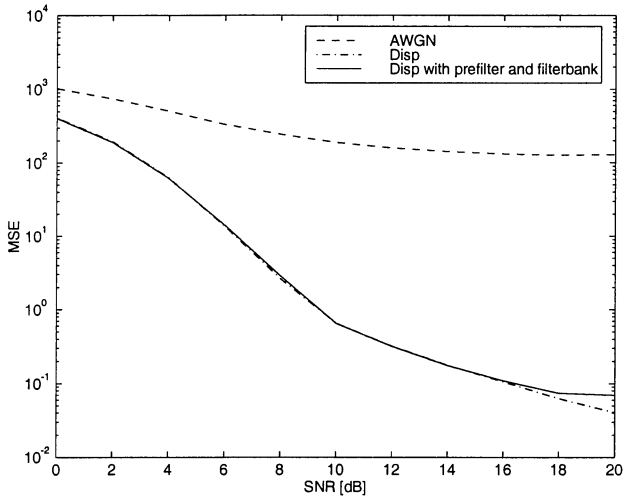


Figure 7: Estimators mean-square error as a function of SNR.

6 Summary

In this paper we derive the ML estimator for time offsets in a dispersive channel environment. The estimator is based on the redundancy introduced by the cyclic prefix and assumes knowledge of the channel impulse response.

By the proper structuring of the calculations and some approximations where the observation interval goes to infinity, we end up with a new estimator structure. The new structure consists of a linear prefilter and a correlating filterbank. By introducing a prefilter the dimension of the estimation problem can be reduced. Letting the observation interval go to infinity suggests a filterbank implementation of the estimator's quadratic form.

Simulations show that although we make some approximations, the performance of the new estimator structure is virtually indistinguishable from the performance of the estimator without approximations, for a large range of SNR's.

A Appendix

Inverse of $(\mathbf{A}+\mathbf{B})$

$$(\mathbf{A}+\mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\left(\mathbf{B} - \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\right)\mathbf{A}^{-1}$$

Proof: (Note that only A needs to be invertible.)

$$\begin{aligned}
&(\mathbf{A}+\mathbf{B})\left(\mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{B} - \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B})\mathbf{A}^{-1}\right) \\
&= \mathbf{I} + \mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1} + \mathbf{B}\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1} \\
&= \mathbf{I} + \mathbf{B}\mathbf{A}^{-1}(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B}\mathbf{A}^{-1} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} \\
&= \mathbf{I}
\end{aligned}$$

References

[1] T. Pollet, M. van Bladel, and M. Moeneclaey, "BER sensitivity of OFDM systems to carrier frequency offset and Wiener phase noise," *IEEE Trans. Commun.*, vol. 43, no. 2/3/4, pp. 191–193, Feb/Mar/Apr 1995.

[2] T. Pollet and M. Moeneclaey, "Synchronizability of OFDM signals," in *Proc. Globecom*, Singapore, Nov. 1995, vol. 3, pp. 2054–2058.

[3] J. G. Proakis, *Digital communications*, Prentice-Hall, New York, USA, 3rd edition, 1995.

[4] J.-J. van de Beek, M. Sandell, and P. O. Börjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Trans. Signal Proc.*, vol. 45, no. 7, pp. 1800–1805, July 1997.

[5] A. Peled and A. Ruiz, "Frequency domain data transmission using reduced computational complexity algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Denver, CO, USA, 1980, pp. 964–967.

[6] J.-J. van de Beek, P. O. Börjesson, M. L. Boucheret, D. Landström, J. Martinez Arenas, P. Ödling, and S. K. Wilson, "Three non-pilot based time- and frequency estimators for OFDM," *Elsevier Signal Processing*, vol. 80, no. 7, pp. 1321–1334, July 2000.

[7] M. Gudmundson and P.-O. Andersson, "Adjacent channel interference in an OFDM system," in *Proc. IEEE Vehic. Technol. Conf.*, Atlanta, GA, USA, Apr. 1996, vol. 2, pp. 918–922.

[8] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 1993.

Up- and Downlink Performance Evaluation Using Space and Polarization Measurements of UTRA-FDD Channels

Peter Eneroth[1], Christian Bergljung[2], Peter Karlsson[3], Mark Beach[4], and Sze Ern Foo[5]

[1,2,3]Telia Research AB, Box 94, 201 20 Malmö, E-mail: {Peter.B.Eneroth, Christian.S.Bergljung, Peter.C.Karlsson}@telia.se
[4,5]University of Bristol, Queen's Building, Bristol BS8 1TR, E-mail: M.A.Beach@bristol.ac.uk, S.E.Foo@bristol.ac.uk

Abstract

The use of adaptive antennas in mobile communication systems has in several studies shown great promise to increase the system capacity. In a frequency division duplex (FDD) system, the uplink and downlink channel may differ significantly. The following question arises: To what extent can the uplink channel information be used for downlink beam-forming? This paper presents a study of data from a measurement campaign, where the UTRA uplink and downlink channels have been measured simultaneously using a dual polarized 8-element array in urban and suburban environments. The estimated gain was in the range 6-8 dB for uplink and downlink.

1 Introduction

Spectrum efficiency is always a strong factor in the selection of wireless technology and adaptive antennas can provide an increased spectral efficiency in, for example, WCDMA systems. With the use of adaptive antennas, the energy can be directed towards the desired user and thus reduce the interference experienced by other users in different spatial locations. In an interference-limited system, like the UTRA, this will enable us to increase the number of users, the data rate, the coverage, as well as a combination of these parameters.

An adaptive antenna is typically implemented as an array of antenna elements and a processing unit for beam-forming by linearly combining the signals from the different elements. Usually, the distance between any two of the antenna elements is approximately half a wavelength; this corresponds to the range 69-78 mm in the UTRA FDD uplink and downlink allocation. Further, due to the space and the power consumption, the adaptive array will usually only be implemented on the base-station (BS) side, leaving the mobile terminal (MT) with a normal omni-directional antenna.

In the uplink situation, the adaptive antenna can estimate the spatial channels both from the desired user and undesired interferers. This can either be done by estimating the direction of arrival (DOA) of the desired signal and then point the beam in that direction, or directly deriving an optimal beam which maximize the signal to noise and interference ratio (SNIR). In both situations, the actual beam-forming is performed by multiplying the received signal from each antenna element with a complex number, the beam coefficients. The difficulty lies within the determination of these coefficients, using various DOA algorithms or other means.

In a time division duplex (TDD) system, the uplink and downlink channels share the same frequency band, and hence the downlink channel can be accurately approximated from the knowledge of the uplink channel. That is, we may use the same beam coefficients for uplink and downlink if only the time variation of the channel is small enough. In a FDD system, the uplink and downlink channels are transmitted on different frequency bands. Here the two spatial channels can be quite different, and we will not be able to use the same beam coefficients for the uplink and the downlink channels. Nevertheless, the direction to the user, determined from the uplink channel, can be used to calculate the downlink beam coefficients.

In this paper we will study the data from a measurement campaign, where the UTRA uplink and downlink channels are measured simultaneously in an array. We will show how different candidate adaptive antenna architectures can improve the carrier to noise and interference ratio in this scenario.

2 Measurements and Models

The measurements were conducted in urban and suburban environments in the city of Bristol, UK, during the autumn of 2001. See maps in Figure 1 and Figure 2. In the campaign, a wideband vector channel sounder was used to estimate the channels (almost) simultaneously. The channels had a bandwidth of 20 MHz and were centred as in the UTRA FDD bands, namely at 1.92 and 2.12 GHz. The bandwidth of each channel was later reduced to 5 MHz, as this is used in the UTRA. The transmitter was equipped with an omni-directional antenna and the receiver with an 8-element wideband dual polarised array. The array, manufactured by Allgon Systems, had an element spacing of 80 mm, and each element was a standard sector antenna with a 120 degrees wide beam. In the urban environment measurements were conducted from 12 static positions, and 3 driving routes. In the suburban environment, 12 static and 2 driving routes were meas-

ured. The driving routes had a length of 3 to 13 minutes and were measured every 1.7s.

In the evaluation of the measured data, it is required to estimate the angle and time delay of each impinging wave. We will next study the model to be used.

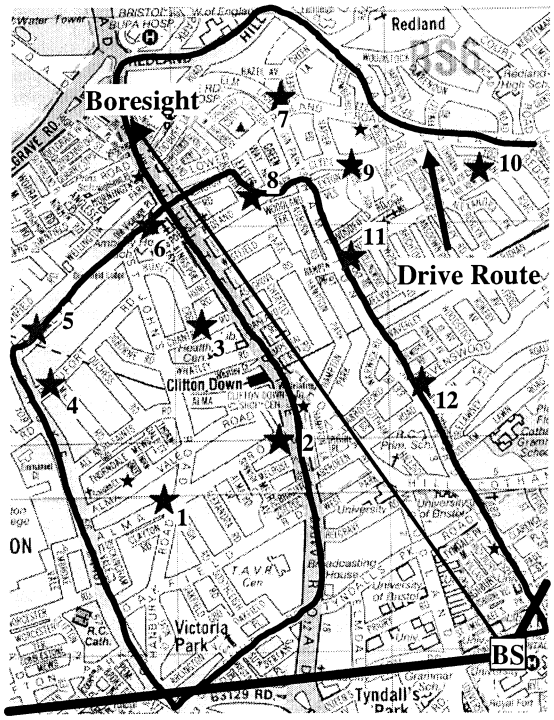


Figure 1: *Map over suburban environment. The Base station is in the lower right corner, directed according to the "Boresight" arrow. The stars represent static measurement points. On the driving route, measurements were conducted every two seconds.*

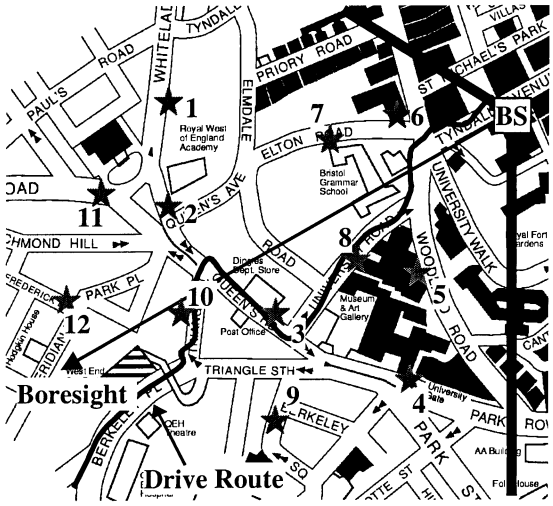


Figure 2: *Map over urban environment. Figure notations as in Figure 1.*

2.1 Channel Model

In a scattering environment several paths exist between the transmitter and the receiver, and the received signal is the vectorial sum of signal components. Each signal component, l, has experienced a different path environment, which will determine the DOA ϕ_l, amplitude α_l and time delay τ_l. In general, these parameters are time-varying. When analyzing data from the measurement campaign, these parameters are estimated in accordance to the classical channel model for spatial channels [1],

$$\mathbf{x}(t) = \sum_{l=0}^{L(t)-1} \mathbf{a}(\phi_l(t))\alpha_l(t)s(t-\tau_l(t)) + \mathbf{n}(t). \quad (1)$$

Here, $L(t)$ denotes the number of impinging waves, $\mathbf{a}$ the steering vector, s the transmitted signal, $\mathbf{n}$ the vector of independent noise components and $\mathbf{x}$ the received signal. The row vector sizes correspond to the number of receiver antenna elements.

2.2 DOA Algorithms

A fundamental and difficult part in the analysis of measured spatial channel data is to estimate the delay, angle and amplitude of the individual impinging radio waves. This is still an active research field, aiming for accurate high-resolution algorithms. During the analysis, we have used a range of algorithms including the conventional method of steering a beam in different directions and measure the received power, to subspace methods as root-MUSIC [2], unitary ESPRIT [3], and estimating the maximum likelihood (ML) estimate with the SAGE algorithm [4]. Even if the results differ between the different methods, they were all considered to be consistent, and unitary ESPRIT will be used in the following. A channel estimation example can be seen in Figure 9.

The subspace methods above work on the premise that the signals impinging on the array are not heavily correlated. As several versions of the transmitted signal with different paths are received, this premise is not fulfilled in a multipath environment. Therefore, the correlation has been reduced using spatial smoothing [2].

In an operating UTRA system, the power control unit will limit the dynamic range of the channel. Therefore, we have chosen also to only use wave estimates that are maximum 10, 15, or 20 dB smaller than the largest wave. This will also reduce the problem of separating small true estimates from noise components.

2.3 Measurement Data Statistics

One way to characterize the measurement environment is to calculate the path loss exponent. In

$$PL(d) = PL(d_0) + 10n\log_{10}\left(\frac{d}{d_0}\right), \quad (2)$$

the path loss is described as a function of the distance d and the path loss exponent n [2]. Taking all data from the measured driving routes in the suburban and urban environments, we get a path loss exponent of 3.1 and 3.3 respectively. In [2], an urban environment is described as an environment with an exponent between 2.7 and 3.5.

Other interesting parameters in the analysis of spatial channels are the root mean square (rms) delay spread (DS) and the rms angular spread (AS). They indicate the spread of the impinging waves, and typically increase as the environment gets more cluttered, such as an urban environment. Defined as,

$$DS = \sqrt{\mathrm{E}[\tau^2] - \mathrm{E}^2[\tau]}, \quad (3)$$

$$AS = \sqrt{\mathrm{E}[\phi^2] - \mathrm{E}^2[\phi]}, \quad (4)$$

where τ and ϕ are the delay and azimuth angle of an impinging wave, respectively. Using data from the two driving routes, we estimate the impinging waves using the unitary ESPRIT algorithm. The angular distribution, see Figure 4, resembles the results presented in [5].

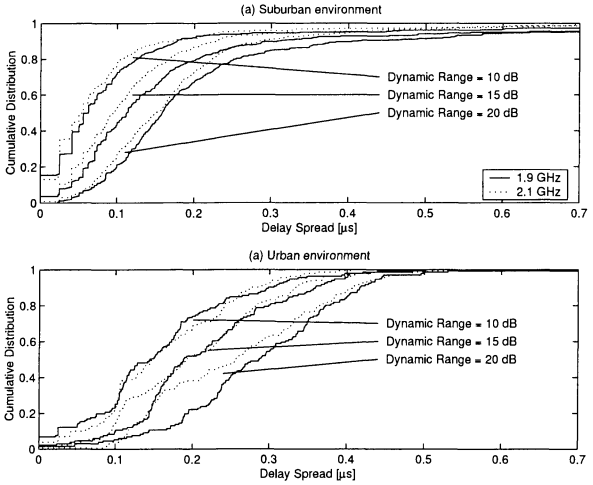


Figure 3: *The cumulative distribution of the delay spread. (a) Measurements performed in suburban environment, (b) urban environment.*

In a basic version of an adaptive array system, it would be possible direct the uplink beam in the direction where the power is maximized. Then the same beam direction could be used for the downlink beam. The first question would then be how big angular difference there is between the uplink and downlink channels? In Figure 5, the offset between the true direction (GPS data based) and the direction of the beam, which maximizes the power, is shown for both the 1.9 and 2.1 GHz channels. The maximum offset between the frequencies during these particular driving routes are 11 degrees in the suburban environment and up to 50 degrees in the urban environment, however in most cases are the offsets rather small. This offset needs to be considered when designing the outage of a given Tx lobe width using adaptive antennas.

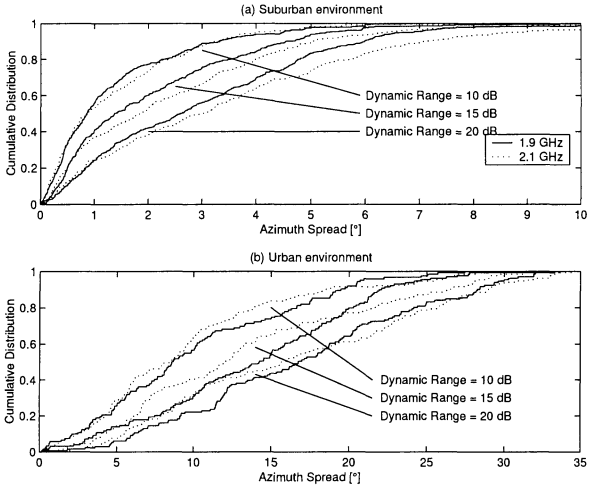


Figure 4: *The cumulative distribution of the angular spread.*

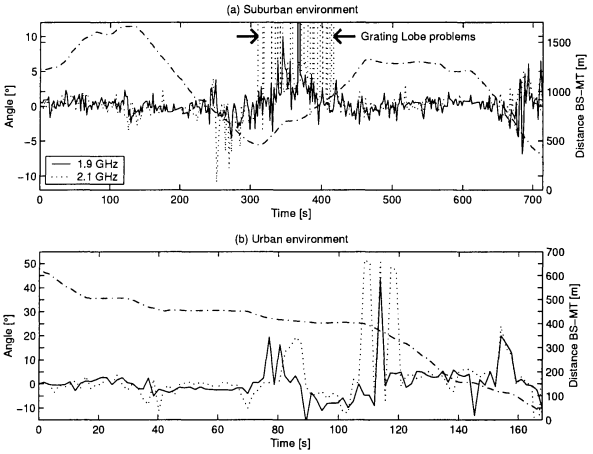


Figure 5: *Angular offset for the beam angle with maximal energy for 1.9 (solid line) and 2.1 GHz (dotted line). GPS data is used as the true direction. The distance between the MT and BS is shown with dash-dotted line.*

3 Channel Enhancements using an Adaptive Array

In this section we will calculate the carrier to interference and noise ratio (CINR) enhancement when the BS is equipped with an adaptive antenna. We will use the measurements described in Figure 1 and Figure 2, where data from the driving route will be used as the desired user, and the 12 static measured points will be used as interfering users. A synthetic white Gaussian background noise of -103 dBm and -100 dBm will also be added to the uplink and downlink, respectively. In both the downlink and uplink, a power control (PC) device will set the transmitted power on the individual channels using a BS with only one element, i.e. the transmitted power is set so that CINR targets for the used services for all users are achieved. Then we will apply the array in different formations, and calculate the CINR improve-

ments for the investigated user. Statistics from the whole driving route is presented.

As UTRA is a CDMA based system, several users may use the same frequency and are separated with different spreading codes. As the receiver de-spreads a signal, other users and the background noise is attenuated a factor G, denoted spreading gain. By using different spreading gains, it is possible to alter the data rate of the channel, and this data-rate flexibility is also one advantage of the UTRA system. As the target CINR for a given service is given as a ratio between the information bit and noise energy, E_b/N_0, we will use a generalized form of the spreading gain, defined as $W/R = d_c/d_b$, where d_c and d_b are the chip and information bit rate, respectively. Different services needs different data rates, and in our simulations, interference number 2, 4, and 6 are using a 384 kbit/s data service with W/R = 10, and target CINR at 2.6 dB in the uplink and 2.5 dB in the downlink. All other users are using a speech channel at 8 kbit/s, with W/R = 480 and the target CINR is 2.4 dB in the uplink and 1.9 dB in the downlink [6].

We will investigate an adaptive array, which is directed towards the impinging waves, with 1 degree of direction resolution. The array may have 1, 2, 4 or 8 elements, where each element is a sector antenna. The more elements used, the sharper beams may be formed, as is shown in Figure 6. The uplink channel is used to estimate the direction from which the maximum power is received for each user, and these directions are then used for both uplink and downlink beam forming. In the downlink, the total transmitting power is adjusted to be independent of the number of antenna elements.

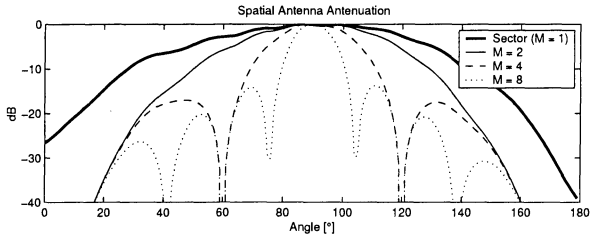


Figure 6: *Antenna beams, with different number of antenna elements (M), directed towards boresight.*

From each measurement point on the driving route and each user, we will transmit a power controlled data sequence. Here, transmit equals filtering with the measured channel. The *CINR* is then calculated as,

$$CINR = \frac{W/R \cdot P_{User}}{P_{Noise} + \sum_{\forall Int.} P_{Int.}}, \quad (5)$$

where P denotes received power, W/R spreading gain, and *Int.* interferers. Since ideal PC is used, the *CINR* value will always be identical to the target value. When beam forming is applied, without recalculating the PC values, the *CINR* value is (almost always) improved. The cumulative distribution of this improvement, in dB, for different number of elements is shown in Figure 7 and Figure 8. The median gain is >9 dB for the uplink channel in the suburban environment. The median gain is reduced to 7 dB in the urban part. When the desired user is in the same direction as one of the high data-rate users, the adaptive antenna gains is reduced. However, we can expect a significant gain in at least 90% of the situations, as indicated by the figures.

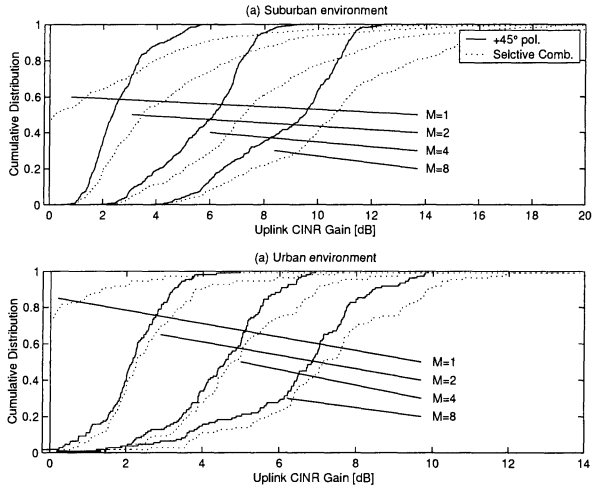


Figure 7: *Uplink improvements compared to one element, single polarization BS antenna.*

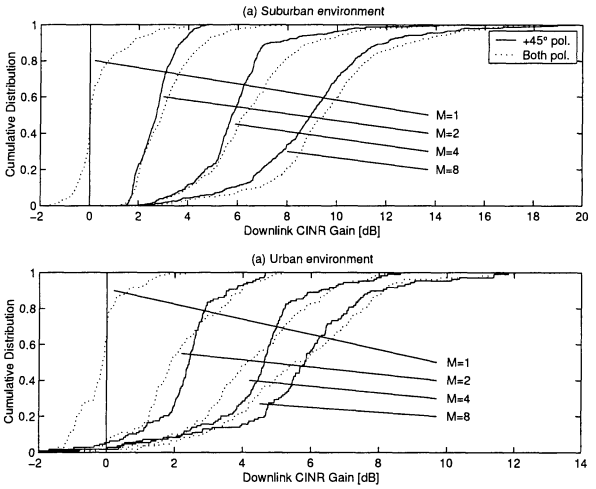


Figure 8: *Downlink improvements compared to one element, single polarization BS antenna.*

The measurements were conducted with cross polarization antenna elements, allowing us to include polarization gains in the investigation. In the uplink situation, we have chosen "selective combining", where the best polarization is always used. If the same technology would be deployed in the downlink, feedback information from the MT would be needed in order to choose polarization. This is probably too complex and instead we have chosen a feasible solution, where half the transmitting power is used on each polarization in the downlink. Therefore, the dual polarization solution will always be advantageous in

the uplink, whereas it is not always true in the downlink. However, the number of major fading dips may be reduced using dual polarization in the downlink.

It should be noted, that in this study we have only considered the performance in one cell. In an operating UTRA system, interference from adjacent cells and adjacent channels will have a significant impact on the interference level.

4 Antenna Grating Lobes

If the distance between the antenna elements in the array is larger than one half wavelength, spurious wave components will be present, a spatial version of the sampling theorem. For larger antenna element distances, the following relation exists between the maximum useable beam width and antenna element distance [7],

$$\frac{d}{\lambda} < \frac{1}{1+|\sin\phi|}, \qquad (6)$$

where d is the distance between antenna elements, ϕ the angle from boresight, and λ the carrier wavelength. Since the array used for our measurements had a spacing of 80 mm, the maximum useful bean angle is 145° and 101°, for the uplink (1.9 GHz) and downlink (2.1 GHz) respectively. Due to a sector width of approximately 120°, see Figure 6, grating lobe problems will only occur in the downlink. The grating lobe problem in UTRA FDD was discussed in [9]. Shown in Figure 9, is an estimation of the impinging waves for a user at the lower left corner of the driving route in Figure 1. Unitary ESPRIT was used for the estimation. Here the uplink is estimated correctly, whereas the downlink is estimated to be where the grating lobes appear. In an operating UTRA FDD system, the downlink energy would be directed towards both 40° and 160°, reducing the efficiency of the adaptive antenna.

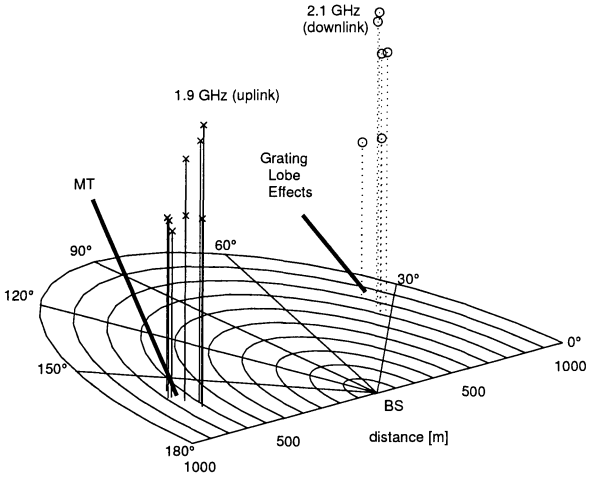


Figure 9: *Estimation of impinging waves using unitary ESPRIT.*

5 Conclusions

We have seen that the introduction of adaptive antennas can improve the CINR significantly, and thereby reduce the total load factor in the cell. The CINR improvement in the suburban environment is on average approximately 8 dB in the uplink and 7 dB in the downlink, for an eight element array. In the urban environment, where the angular spread is larger, is the CINR improvement in average approximately 1 dB less.

By using cross polarization antenna elements further improvements are possible in the uplink when selective combining is used, whereas the major advantage in the downlink is that major fading dips in one polarization can be compensated with the other polarization.

If the distance between the antenna elements is larger than half the wavelength, impinging waves from the edges will be "mirrored" at other directions, and thereby reducing the possible improvements with the adaptive antenna.

Acknowledgement

The measurement campaign was conducted in cooperation with Allgon Systems AB who provided the antenna array.

References

[1] R. B. Ertel et el., "Overview of Spatial Channel Models for Antenna Array Communication Systems", *IEEE Personal Communications*, Feb. 1998, pp. 10-22.

[2] J. C. Liberti and T. S. Rappaport, Smart Antennas for Wireless Communications: IS-95 and Third Generation CDMA Applications, Prentice Hall PTR, 1999.

[3] M. Haardt and J. Nossek, "Unitary ESPRIT: How to Obtain Increased Estimation Accuracy with a Reduced Computational Burden," IEEE Trans. Signal Processing, pp. 1232 -1242, May 1995.

[4] B. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, K. Ingeman Pedersen, "Channel parameter estimation in mobile radio environments using the SAGE algorithm," IEEE J. on Selected Areas in Communications, pp. 434-450, March 1999.

[5] K. I. Pedersen, P. Mogensen, and B. Fleury, "A stochastic model of the temporal and azimuthal dispersion seen at the base station in outdoor propagation environments", IEEE Trans. Vehicular Technology, pp. 437-447, March 2000.

[6] "Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS", ETSI TR 101 112 V3.2.0 (UMTS 30.03), Attachment 5, 1998.

[7] R. C. Johnson, Antenna Engineering Handbook, McGraw-Hill Inc. 1993.

[9] Mythri Hunukumbure, Mark Beach, Ben Allen, Peter Karlsson & Paul Fletcher, "Smart Antenna Performance Degradation due to Grating Lobes in FDD Systems", COST260, Gothenburg, April, 2001.

Suppressed ambiguity in range by phase coded waveforms

Sune R.J. Axelsson

Swedish Defence Research Agency
Saab Bofors Dynamics
S-581 88 Linköping, Sweden
Email: Sune.Axelsson@dynamics.saab.se

Abstract

In Doppler radar for surveillance and Remote Sensing, the ambiguities in range/Doppler are usually solved by varying the pulse repetition frequency. In this paper, two alternative methods are discussed based on phase-coding and orthogonal waveforms. The first one distributes a phase code over a pulse sequence with only one phase step per pulse. The received pulses are correlated with the phase code sequence, and by varying the delay of the phase code, the subintervals in range are scanned and reflector responses outside the focused range interval become highly suppressed.

The alternative method that was studied applies a set of near-orthogonal phase codes, which are modulating the pulses transmitted. In the receiver, the different subintervals in range are scanned, or detected in parallel, by correlating the signal from the scene by the delayed code sequence of the pulse transmitted. Because orthogonal codes are used, a strong suppression is achieved for signals originating outside the focused subinterval in range.

1 Introduction

Coherent pulse Doppler radar is used for the estimation of range and velocity of moving targets. In the measurement of rapid targets at far range, some ambiguities may occur, however.

A standard method of solving the ambiguities in range is to send out pulse sets with varying repetition frequency (staggered prf). By investigating the varying target range mapping with respect to prf, the true target range can be estimated [1].

In this paper, some alternative methods are discussed based on phase-coding and orthogonal waveforms.

2 Ambigutities in range and doppler

Conventional pulse radar uses a fixed pulse repetition frequency (f_r), and non-ambiguous range measurements can be performed if

$$R < c/(2\, f_r). \qquad (1a)$$

The Doppler shift generated by a moving target is given by

$$f_d = 2V_{rel}/\lambda \qquad (1b)$$

A non-ambiguous Doppler measurement is limited by $f_d < f_r$ if both I- and Q-components are detected.

As a consequence, a high prf is necessary against rapid targets to avoid ambiguity in Doppler, but will reduce the length of the non-ambiguous range interval.

There are two main methods to perform non-ambiguous measurements in Doppler and range over extended intervals. One is staggered prf [1]. The other one discussed in this paper is based on phase coding or noise modulation. This method gives a shorter time of measurement compared with staggered prf, but requires a far more complex signal processing.

3 Phase codes

Pulse compression radar is usually modulated by linear frequency modulation (chirp or stepped FM), or by using random phase modulation [1].

In the receiver, a correlation procedure, or a matched filter, demodulates the received signal from the scene. The correlation output is built up to a narrow, strong peak, when the time delay of the reference signal is the same as the delay of the signal received from the target. For other time delays, the output becomes highly suppressed.

3.1 Barker codes

An optimum code is the Barker code, which is known up to the length of thirteen and assumes pseudo-random phase jumps of 180 or 0 degrees.

The Barker code has a maximum sidelobe/mainlobe ratio of PSL=N^{-2} (power), which means –22.3 dB for Barker 13. The integrated sidelobe level (ISL) is higher: –11.5 dB.
Both PSL and ISL can be highly improved by proper filtering technique [1]-[2].

3.2 Combined Barker codes

The search for Barker codes when N>13 has given negative results [3] . A number of different non-optimum codes have been suggested, however.

Such a solution is based on combined Barker codes. As an example, we define a code of length N=4x5=20 by starting from N=4 (+ + - +) that is repeated five times. Then we multiply all subgroups (with four elements in each) by the elements of Barker N=5 (+ + + - +).

The maximum length of a combined Barker code is N=13x13=169 and the correlation response for that code is shown in Figure 1.

The combined Barker codes have several advantages. By combining Barker codes in different ways, we can generate a large number of codes, which can be used in different modes of operation in a radar system.

The computation procedure for estimating the coefficients of an ISL-optimised filter is highly simplified, because a combination of filters can be used, which have been optimised for the individual Barker codes. Similarly, the sidelobe suppression (PSL/ISL) can be studied simpler by combining the performances of the subcodes.

3.3 PN-sequences

Common forms of binary phase codes are the PN-sequences generated by a shift register with feedback connections. By changing those, new codes are easily generated. By using n shifts, codes of length N= 2^n-1 can be formed. The PN-sequences have relatively high ISL-levels, which makes the code less appropriate in applications with many targets or a high clutter level.

3.4 Polyphase codes

In the general case, a polyphase code contains M different phase values $\varphi_m=2\pi(m/M)$ for m=0, 1, ..., M-1. A binary code has M=2.

A useful code type is Frank's polyphase, which is a discrete approximation of the phase variation of a linear FM-waveform.

A Frank-code has for large M a side lobe suppression, which can be at least 10 dB better than that of pseudo-random codes of the same length.

3.5 Random phase codes

For long code lengths, a random choice of phase values gives fairly good performance. Figure 1 shows, as a comparison between the combined Barker code of N=13x13 and the corresponding correlation output for random sampling with equal probability. See also [4]-[7] for a more detailed analysis of the performance of different codes.

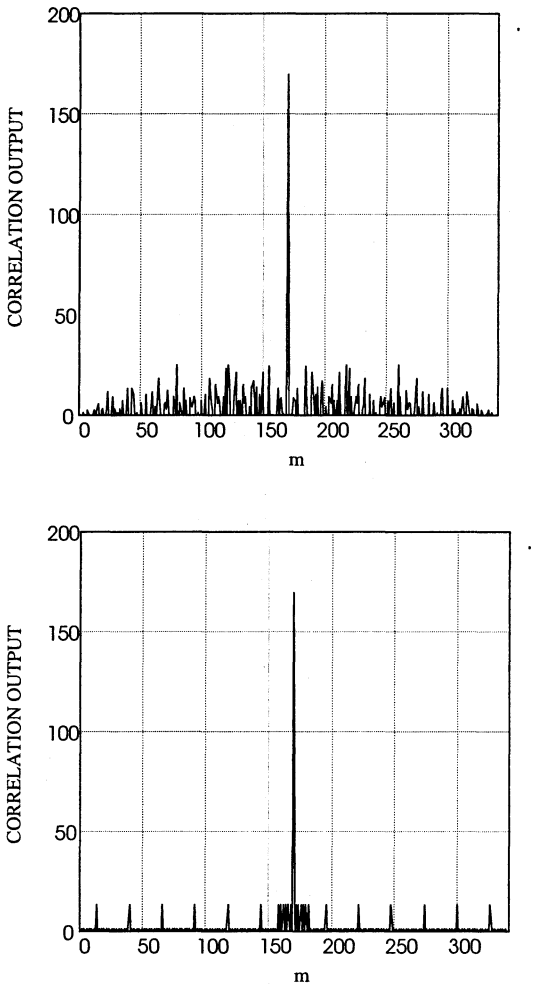


Figure 1: Correlation response of a combined Barker 13x13 (top). As a comparison below, random phase jumps between 0 and 180 degrees with equal probability.

4 Coded phase jumps between pulses

As mentioned above, a pulse Doppler radar using a fixed prf gives an ambiguous range estimation when $R>c/2f_r$.

The ambiguity can be removed, if we make the phase jump 0 or 180 degrees randomly between each outgoing pulse

$$S_0(t)= p(t)\exp[j2\pi f_0 t+j\theta(t)] \quad (2)$$

where p(t) denotes the pulse amplitude and $\theta(t)$ is the phase function.

The received signal from a point reflector positioned at range R and delayed T=2R/c is then given by

$$S(t)= S_0(t-T)= p(t-T)\exp[j2\pi f_0(t-T)+j\theta(t-T)] \quad (3)$$

The Doppler shift as defined by (1b) is hidden in the time variation of T, when there is a relative movement between the target and the radar platform. After down-conversion and coherent detection (I and Q), the received signal can be written as

$$S(t) = I + jQ = p(t-T)\exp\{j[\theta(t-T)+2\pi f_d t+\varphi_0]\} \quad (4)$$

We divide the search in range into subintervals, each with a length corresponding to the requirement for non-ambiguous range indication ($R<c/2f_r$).

The different range intervals are scanned by multiplying S(t) by $\exp\{-j\theta(t-kT_r)\}$ for k=0 ... K and $T_r= 1/f_r$, giving

$$S(t,k) = p(t-T)\exp\{j[\theta(t-T)-\theta(t-kT_r)+2\pi f_d t]\} \quad (5)$$

For a reflector positioned in the k:th range interval, defined by $kT_r<2R/c<(k+1)T_r$: $\theta(t-T)=\theta(t-kT_r)$. From (5), we find

$$S(t,k) = p(t-T)\exp\{j2\pi f_d t]\} \quad (6)$$

This means that the true Doppler signal is extracted and the frequency shift can be detected by spectral analysis (FFT).

If the reflector is positioned outside the k:th range interval, $\theta(t-T)\neq\theta(t-kT_r)$ and the phase coherence of the Doppler signal is destroyed. As a result, the signal power is spread out in the Doppler spectrum.

Hence, by detecting samples of (5) from many pulses followed by FFT, the target can be extracted by narrow-band Doppler filtering. Signals from range intervals outside the focused one are not matched and become highly suppressed in both range and Doppler.

This effect is displayed in Figure 2. The upper graph shows the range/Doppler matrix for 169 received pulses without phase coding. A strong ambiguity in range occurs. Each subinterval in range consists of 300 range cells and totally 3900 range cells of the totally possible 169x300 =50 700 are shown in the image. The target reflector is positioned at range cell number 1700 and Doppler cell number 70 of totally 169, and the true Doppler frequency is then 70/169 of the prf.

The significant improvement after phase coding is shown in the lower graph of Figure 2. A distinct target response is obtained in the right range/Doppler cell.

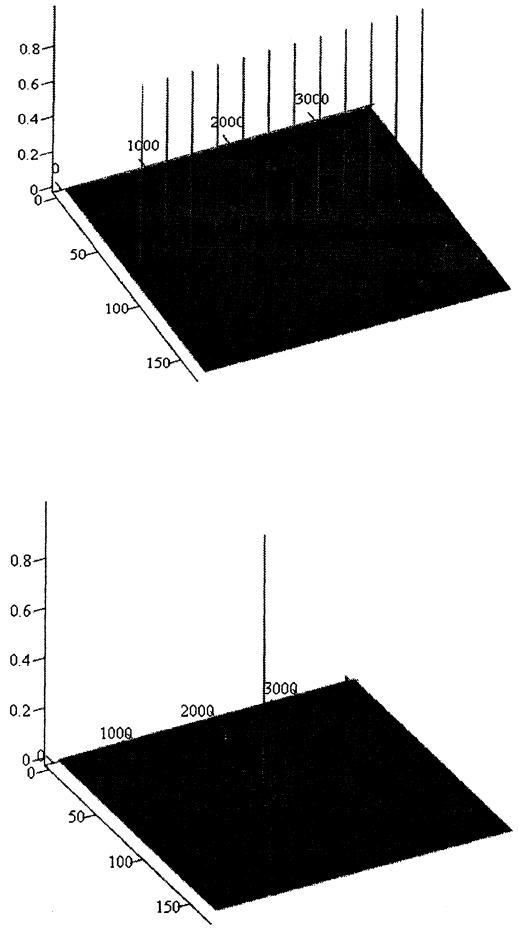


Figure 2: Upper: Range/Doppler response for 169 received pulses without phase coding, showing strong ambiguity in range. Lower: After phase-coding the transmitted pulses.

5 Pulse compression radar

The above method was based on adding a phase jump of 0 or 180 degrees to each transmitted pulse according to a given code pattern. This method works for pulse compression radar and FMCW as well.

For pulse compression radar, the transmitted signal is

$$S_0(t)= p(t)\sin[2\pi f_0 t+\psi(t)+\theta(t)] \quad (7)$$

where the pulse amplitude p(t) is extended in time, $\psi(t)$ defines a rapid phase or frequency modulation, and $\theta(t)$ is the slow phase modulation, which is added to suppress the range ambiguity as shown above.

The received signal $S_0(t-T)$ is demodulated by eliminating the rapid phase function $\psi(t-T)$, which is performed by matched filtering or correlation techniques.

The resulting output can then be written as (4) for short pulse radar

$$S(t,T)= I + jQ = p_c(t-T)\exp\{j[\theta(t-T)+ 2\pi f_d t+\varphi_0]\} \quad (8)$$

where $p_c(t-T)$ represents the pulse function after compression. The remaining phase code $\theta(t-T)$ is then removed as described above for short pulse radar.

6 Orthogonal phase codes and noise

The method above assumed repetitive phase or frequency modulation $\psi(t)$ and the slow phase modulation $\theta(t)$ was added to suppress range ambiguity.

A more obvious approach is to change the rapid modulation function $\psi(t)$ between each pulse, giving a suppressed correlation output when the received pulse is correlated with the wrong transmitted pulse code. In that case, the slow phase modulation $\theta(t)$ is no longer required. This method involves also deterministic waveforms such as frequency chirp, where the modulation sweep can be randomly changed, or delayed between each pulse. For phase coded waveforms, the code is changed between each transmitted pulse, and the selected codes should be near-orthogonal on purpose to suppress the side lobes outside the focused range interval.

In that case, the transmitted signal is given by

$$S_0(t;n)= p(t)\sin[2\pi f_0 t+\psi_n(t)] \qquad (9)$$

where $\psi_n(t)$ for n=1 to N form a set of near-orthogonal phase codes. Different types of codes, which have particular importance in telecommunication, are discussed in [4]- [7].

Randomly generated signals work very well, when very long sequences are used, and that is the reason behind the increasing interest for noise modulated radar [8]-[9]. In statistical sense, non-overlapping time sequences of noise are orthogonal. The side lobe suppression depends on code length as 1/N.

7 Conclusions

Non-ambiguous radar measurements can be achieved by staggered prf [1]. An alternative method, discussed in this paper, adds a pseudo-random phase step to each transmitted pulse. After pulse compression, the pulse samples received are correlated with the phase code sequence transmitted.

By varying the delay of the reference code, the ambiguous subintervals in range are scanned and responses outside the focused range interval are highly suppressed.

Another solution is to transmit pulses, which are modulated by orthogonal phase codes. A further step is the use of long random noise sequences as applied in noise radar.

Acknowledgements

This work was performed at the Swedish Defence Research Agency (FOI) for the Swedish Defence Materiel Administration (FMV) within a radar project, which was co-ordinated by Staffan Jonsson (FOI) and Tomas Larsson (FMV).

References

[1] J.L. Eaves and E.K. Reedy: *Principles of Modern Radar*, Van Nostrand, 1987.

[2] M.H. Ackroyd and F. Ghani: Optimum mismatched filters for sidlobe suppression, IEEE, vol. AES-9, March 1973, pp. 214-218.

[3] M.N. Cohen, M.R. Fox and J.M. Baden: Minimum peak sidelobe pulse compression codes. IEEE International Radar Conference 1990, pp. 633-638.

[4] H.F. Harmuth: *Transmission of Information by Orthogonal Functions*. Springer-Verlag, 1972.

[5] H.D. Lyke: *Korrelationssignale*. Springer-Verlag 1992.

[6] M.R. Schroeder: *Number Theory in Science and Communication*. Springer-Verlag 1986.

[7] P. Fan and M. Darnell: *Sequence Design for Communications Applications*. Research studies Press, John Wiley, 1996.

[8] I.P. Theron et al: Ultra-Band Noise Radar in the VHF/ UHF Band, IEEE AP-47, June 1999, pp. 1080-1084.

[9] Sune Axelsson: On the Theory of Noise Doppler Radar. Proc. IGARSS 2000, Honolulu, 24-28 July 2000, pp.856-860.

A Research Agenda for Distributed Policy-based Management

Mads Dam[1,2], Gunnar Karlsson[1], Babak Sadighi Firozabadi[2], and Rolf Stadler[1]

[1] KTH IMIT, Isafjordsgatan 39, SE-16440 Kista, Sweden, E-mail: {mads, karlsson, stadler}@imit.kth.se
[2] SICS, Box 1263, SE-16429 Kista, Sweden, E-mail: {mfd, babak}@sics.se

Abstract

Policy-based management is based on defining a set of global rules, according to which a network or distributed system must operate. In the last few years, policy-based management has begun to emerge as the dominant paradigm for developing network and systems management functions, primarily, since it can reduce complexity in management applications. Although attempts are underway to standardize policy-based management, significant research challenges remain. At KTH and SICS, a joint activity has been started to focus on some of the key issues. The paper outlines the research agenda for this activity.

1 Introduction

The concept of using policies in management applications has been applied selectively for a quite some time. Only recently, however, attempts are being made at applying this concept to virtually all management functions, as well as to the development of a generic architectural framework, protocols, data models, etc., for this purpose. A recent activity of the IETF WG on policy-based management [1] has created a strong interest in the subject domain, and has inspired numerous projects in research labs.

A group at Imperial College started systematic research in this field in the early 90s [2]. While this work focuses on distributed systems management for organizations, the primary goal of the IETF initiative is to support new Internet services, such as Diff Serv, Int Serv, MPLS, IP telephony and VPNs, starting at the device level. Both approaches are to a large extent complementary.

Although standardization efforts are underway, key aspects in the area of policy-based management still neded to be been addressed. For instance, the impact of QoS policies on the network state is barely understood. Also, realizing effective mappings of high-level policies into lower-level policies remains unsolved in many cases. Finally, the IETF framework does not provide answers to the question of how to best distribute the components of a policy-based management system in medium and large-size networks.

This short paper describes our research agenda in the field of policy-based management. Supported by VINNOVA, we have started three projects that focus on different aspects of distributed policy-based management. These projects are discussed in the following sections.

2 Policy-based management for large-scale dynamic networks

This project aims at developing engineering solutions for policy-based management in large-scale dynamic networks. We focus on networks with at least 1000s of nodes, which have a dynamic state and a frequently changing topology, and which do not have a centralized database with up-to-date topology and state information. Today, such networks exist in the form of Autonomous Systems, i.e. Internet domains under independent administration, or cable networks that distribute digital TV. In the future, when the technology for ad-hoc networks will have matured and networks that are part of pervasive computing environments will become a reality, large-scale dynamic networks are expected to become even more widespread.

Such networking environments pose specific challenges for the introduction of policy-based management. We are addressing some of the key issues in this area, such as:

- Policies with end-to-end significance are formulated by network operators and are entered into the management system at one or more access points. To take effect, they need to be translated into lower-level policies and disseminated to network elements. Therefore, an architecture needs to be engineered that allows for *translating and disseminating policies*. This architecture is part of the management system and, for large networks, must be distributed to allow for efficient operation.
- In reaction to changes in the network state and network topology, lower-level policies need to be re-computed for the network to remain operational and to efficiently deliver services with guaranteed quality.

In large-scale networks, it is infeasible to perform this type of processing in a centralized way. For this reason, policies must be re-computed locally, within smaller network areas, on a time scale that depends on the frequency of changes. The question that needs to be addressed is: Which architecture and which mechanisms support best the *local, autonomous re-computation of policies*?

- In large networks, we must assume that, at any given time, some of the network elements or other network components are out of service or malfunctioning. A policy-based management system needs to be *robust* enough to keep network services running during fault conditions and to *initialize network components* with appropriate policies as they become operational.

We are specifically interested in QoS policies and performance aspects. Global policies in this context define end-to-end objectives for packets or flows that traverse a network. They can be formulated on the packet level, such as defining a limit on the end-to-end delay of a specific class of packets, or on the level of flows, such as specifying an upper bound for blocking rates for certain classes of flows.

The task of the management system is to break down the above global policies into lower level policies that apply to network regions, which, in turn must be translated into policies on the device level. As the example of the upper bound on an end-to-end delay shows, there are generally many solutions to implement such a global policy. (An end-to-end delay of, say, 50ms, can be divided in many ways into smaller delay bounds that can be enforced on the device level.) At the same time, every solution that is implemented will change the network state and, therefore, the way the network is controlled. Generally speaking, it is a hard problem to devise a system that *concurrently* supports several QoS policies in an efficient way.

Additional problems we plan to investigate in this work include the dissemination of QoS policies and the re-computation of local policies, triggered by topology changes or network faults, in the context of large networks.

We intend to address these issues from the perspective of network architectures, protocols, algorithms and software architectures. As far as aspects of network architectures are concerned, we plan to work within the IETF framework on policy management. Since that framework does not address the issues of large-scale systems, our work will likely be complementary to the IETF activities.

In terms of protocols, algorithms and software architectures, we plan to base the research on the application of pattern-based management, an approach for managing large networks, which has been developed in our group [3]. Our plan is to simulate and evaluate the developed architecture and the algorithms on high-end PCs. Such a platform will allow us to evaluate different design decisions against each other and to gain experimental results for large network configurations.

Further, in order to understand and solve some of the problems of applying our solution to today's networks, we plan to implement a policy-based QoS management on lab, testbed which contains 12 Cisco routers.

2 Decentralized Contract Management

We will address the issue of managing access to disparate resources that are not under control of a single network administrator/provider.

In a multi-domain network, a number of individuals and/or service providers interact in a collaborative environment to provide certain services organized according to a set of rules and policies that define how their resources can be shared among them.

A multi-domain network is usually a composition of heterogeneous and independently designed domains with no centrally controlled enforcement of the policies. Consequently, there is no guarantee that policies will be followed as they are prescribed: members of a network may fail to, or choose not to, comply with the rules. If there is no way of practical (physical) enforcement of policies, then it would be useful to have a *normative control mechanism* for their *soft enforcement* [4]. By soft enforcement we mean that even if service providers are practically able to avoid complying with the organization policies, they can still be subject to sanctioning and remedial action as the consequence of their behavior.

Consider following scenario: Alice wants to make a network reservation from her computer in domain A (the source domain) to Charlie's computer in domain C (the destination domain). Alice has a service level agreement (SLA) with its service provider (domain A) to get some level of service defined in terms of quality of service (QoS) parameters. The path between Alice and Charlie includes some other intermediary domains. An end-to-end reservation for Alice requires a chain of SLAs between her domain and Charlies domain. Beside these SLAs, each domain may have its own local policy that does not necessarily comply with the SLAs that the domain has signed for.

The issues are, how can Alice's request for bandwidth reservation be supported by a set of credentials that shows her right for that request? and what happens if she does not get the required QoS as she is *entitled* to?

The approach we are suggesting is that Alice's request for bandwidth reservation is submitted with evidence for her right for that in terms of digital credentials. Now it is up to each domain to decide whether they are, according to their SLAs, obliged to provide the service she requires. If Alice does not receive the service she is entitled to, then it must be possible to find the point of failure, and, perhaps, some sanctions should be imposed on those accountable for that failure.

3 Secure Policy Management

The flexibility provided by policy-based network management comes with some serious hazards, as illustrated by the recent SNMP vulnerability uncovered by Oulu university's secure programming group.

The vulnerabilities include:

- Denial of service attacks. By messing up routing policies, networks can be rendered inoperable.
- Hijacking and misuse of network equipment.
- Unauthorized access to corporate resources, such as intranets.

These types of threats are less concerned with security at the level of network elements than with the secure management of the policy base itself. A comprehensive solution, which can scale to large, decentralized networks must address at least the following issues:

- A secure and decentralized management framework for access, authorization, and auditing. This includes a rich structure for groups, roles, and delegation, to provide efficient organizational support for the management task.
- An intrusion detection component, which will monitor network operation to determine, whether and to which extent hosts and network elements have been compromised.
- A framework for secure and efficient policy distribution, including secure methods for managing the distribution mechanism itself.

We base our work on the Constrained Delegation framework developed in the context of the Amanda project at SICS [5, 7]. Our starting point is the Simple Public Key Infrastructure (SPKI) framework for authorization in distributed systems [6]. We are extending this to constrained delegation and specific network policy formats. The goal is that the framework should be flexible and efficient enough to support small footprint implementations, for instance in the context of smart cards.

Secure distribution of policies is a multifaceted problem. One aspect is privacy. Since policies describe the way a network behaves, its structure and management, and possibly even its response to attacks, quite a lot of information can be extracted from policy traffic by a potential adversary. In the context of static networks of policy servers, privacy can be supported by standard end-to-end solutions, but for large, autonomous networks, more dynamic solutions are called for. A second set of issues concerns the management of the policy engine itself. Such an engine must be able to handle a wide array of concerns, including traffic-aware mechanisms for policy updates, constraint resolution and pre-computation, and policy-control of the updating traffic itself. Finally, the need for policy protection must be weighed against the need for intrusion detection and compliance monitoring. Knowing that it is difficult at best to protect a network against attacks of various kinds, how can we ensure that a compromised network element (for example a policy repository) does not compromise the entire network? What is involved in doing this efficiently and in a scalable way?

A serious stumbling block towards the adoption of policy-based control on a larger scale is policy enforcement and integration. Policies should, as far as possible, be independent of the platforms on which they are enforced. However, to be realized, the policies must be translated into forms understandable by the individual network elements. Thus, to be able to leverage all the functionality provided by the network elements, quite a lot of generality is required of the policy framework. For instance, for access policies, the framework must know how to render a given "object-level access policy" into any of the target systems supported by the management system. Hard-coding the representation into the policy management framework can do this. A better solution, however, would be to find a way of specifying the target access mechanism as a policy object itself, albeit on the meta-level. New target architectures can then more easily be accommodated, by simply accounting for its access control regime at the level of meta data, rather than having to implement a complete new translation algorithm.

4 Acknowledgements

This research is supported by the Swedish Agency for Innovation Systems (VINNOVA) under the project name "Policy-based Network Management."
Thom Birkeland contributed to this paper.

5 Bibliography

[1] http://www.ietf.org/html.charters/policy-charter.html

[2] J. Moffett, M. Sloman, "Policy Hierarchies for Distributed Systems Management," IEEE Journal on Selected Areas in Communications, Vol. 11 No. 9, Dec. 1993, pp. 1404-1414.

[3] R. Stadler: "Decentralized and adaptable management based on active networking technology," The First International Workshop on Active Network Technologies and Applications (ANTA 2002), Tokyo, Japan, March 25-26, 2002.

[4] Babak Sadighi Firozabadi and Marek Sergot, "Contractual Access Control - Position Paper", to appear in the proceeding of Security Protocol, 10th international workshop, Cambridge, UK, April 2002.

[5] O. Bandmann, M. Dam and B. Sadighi Firozabadi, "Constrained Delegation" Proc. 23rd annual IEEE Symposium on Security and Privacy, Oakland, 2002.

[6] C.M. Ellison, B. Frantz, B. Lampson, R. Rivest, B.M. Thomas and T. Ylonen, "SPKI Certificate Theory" RFC 2693, 1999, expired, URL: ftp://ftp.isi.edu/in-notes/rfc2693.txt.

[7] O. Bandmann and M. Dam, "A Note on SPKI's Authorisation Syntax" Proc. 1st annual PKI Research Workshop, NIST, Gaithersburg, 2002.

A New Approach to Monitoring in IP Networks

Thomas Lindh[1]

[1]KTH Syd, Marinens väg 30, 136 40 Haninge, Sweden. E-mail: Thomas.Lindh@syd.kth.se.

Abstract

This paper presents an approach for performance monitoring in IP networks that combines active and passive methods. Traffic meters and dedicated monitoring packets are the main parts of the method. These monitoring functions could be a part of the ordinary network elements, or stand-alone systems. A prototype implementation in Linux-based routers is also presented in the paper. The main application is monitoring in IP-based virtual private networks.

1. Introduction

Measurements and estimation of performance parameters in IP networks are becoming increasingly important for today's operators. There are several reasons to develop efficient and reliable methods in this field. Deployment of differentiated services in IP networks will require effective but also simple methods for measurement of relevant performance parameters to support and verify service level agreements with customers. A monitoring system must also support the daily operation, traffic control and planning of an operator's network with relevant and timely measurements and estimates. Accurate measurements of usage are also the foundation for accounting and billing. Measurements are undoubtedly important for a network operator, albeit they are not a goal but means to achieve certain objectives. In this paper we describe a prototype implementation of a method for performance monitoring of QoS parameters in IP networks, especially virtual private networks. Our approach is to use an appropriate combination of active and passive methods in order to measure and estimate performance parameters in the actual user traffic.

2. Background and Related Work

Methods for measuring and monitoring network performance parameters are traditionally divided into passive and active techniques. The idea behind passive methods is to capture packets in order to store and collect information from various fields within the packet header. Unlike active methods, passive monitoring systems are non-intrusive and do not load the network. Although all passive methods share that basic characteristic, there exist quite different variants. While traffic meters usually minimize the amount of stored data by e.g. merely incrementing counters, other passive tools are designed to record some header fields from each packet.

The method described in this paper can be seen as an integration of passive components (traffic meters) and active components (monitoring packets). Previous work in the project has been presented in [1] and [2]. Some relevant related work in this field are RIPE NCC Test Traffic Measurements [3], NIMI [4], Surveyor [5], AMP [6], NetFlow [7] and Service Assurance Agent from Cisco [8], NeTraMet [9] and several tools based on the RMON management information base. As in the RMON case, SNMP (Simple Network Management Protocol) is a frequently used standard to organize and manage measurement data. Besides these well-known frameworks and tools, a test project at IBM [10] and the EURESCOM project QUASIMODO [11] are also related to our work.

Measurements may also be classified in terms of the level of communications at which they are performed. Besides the network level, measurements of services and applications, e.g. end-to-end between servers and clients, are often equally important. Measures of response times and availability for servers are often valuable complements to measures of packet losses, delays and throughput between network edge nodes. However, this paper is focused on performance monitoring at the network level (IP) between edge nodes that surround a core network. The rapidly growing interest in management of service level agreements and support for traffic engineering in IP networks is reflected in a growing number of products and tools in the area. This development emphasizes the importance of the standardization efforts from IETF (IPPM Working Group), ITU, ETSI and other international bodies. The SLA Handbook from TeleManagement Forum [12] is meant as a common framework to be used by operators and their customers in the entire process of creation and management of service level agreements.

3. The Monitoring Method

3.1 The Context

Of course a monitoring method does not live a life of its own. It has to be seen in a broader context and be subordinated a network or service operator's policy and objectives regarding performance monitoring. There is no purpose in carrying out extensive measurements and monitoring schemes simply for their own sake. An operator has to consider its entire service production and business process, in relation to the customers' needs and requirements.

The goal is to achieve a proper balance between performance measurements between edge nodes at the network level, and between end nodes at the application level (a vertical view). Today's measurement systems and tools are used for different purposes, such as monitoring of service level agreements, network operations and traffic control, and accounting (a horizontal view). Although this paper focuses on measurements on the IP level between edge nodes, the broader context should not be forgotten. There are several advantages to be gained from a vertical as well as a horizontal integration approach.

3.2 The Main Characteristics of the Method

The implemented method has the following goals and main characteristics [1].

- The goal is to measure and estimate losses, delays, delay variations, and throughput using a single procedure instead of several different methods.
- The method is based on an in-service technique that reflects the performance of actual user traffic.
- It should be possible to differentiate the results with respect to different customers, types of traffic, service classes, and virtual private networks.
- It should also be possible to obtain high resolution rather than merely long-term averages for the entire measurement periods, e.g. the distribution in time of losses and throughput levels.
- A combination of traffic meters and dedicated monitoring packets are used to obtain these goals. The method can either be implemented as an integral part of existing network nodes (embedded monitoring) or as stand-alone systems.
- The distance between the monitoring packets, i.e. the size of the monitoring block, determines the level of resolution and the accuracy of the measurement results. This parameter can be used to adjust the measurements to an operator's monitoring policy.

3.3 Virtual Private Networks

This study focuses on IP-based virtual private networks since they represent a case where more elaborate performance monitoring is clearly motivated. The goal is to measure and estimate losses, delays and throughput for IP traffic in virtual private networks between provider edge nodes or customer edge nodes, based on the general topology in Figure 1. This is further explained in [1].

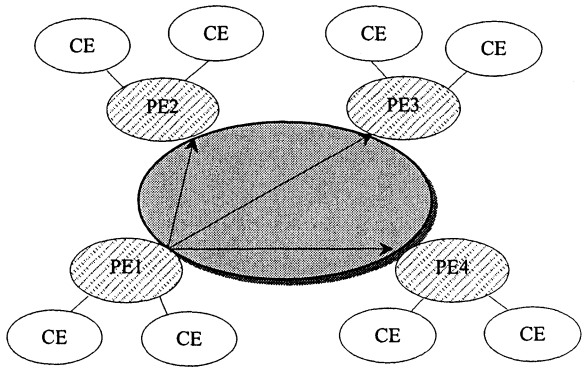


Figure 1: A topology for virtual private networks. The provider edge node 1 (PE1) acts an entry-monitoring node for the outgoing traffic and the receiving nodes act as exit-monitoring nodes for the incoming traffic. Corresponding monitoring functions may also be implemented in customer edge nodes (CE) as well as in individual hosts.

3.4 Requirements for Monitoring in IP Networks

IP's connectionless nature is a major problem in designing monitoring systems. To use monitoring packets in IP networks requires that an entry-monitoring (sending) node can find out which of the exit-monitoring (receiving) nodes the packet will traverse, based solely on its destination IP address. Similarly, a receiving node has to determine from which sending node an arriving packet was sent. In addition, the receiving node must be able to measure or estimate the packet size for the monitoring blocks, in order to compute throughput and utilization.

The solution for a general IP network requires efficient look-up functions in order to determine the correct exit (and entry) monitoring nodes. However, it turns out that in several important cases the complexity of the problem is considerably reduced. In virtual private networks, implemented by means of point-to-point tunnels, the address of the exit node is actually carried in the additional IP header. In a more general case a look-up function that returns the correct exit (entry) monitoring node given an IP destination (source) address is needed.

Besides IP-based VPNs in fixed or mobile networks there are several other obvious cases where the method is applicable such as other tunnel-based services, destination-based services and the virtual leased lines (expedited service in DiffServ) [13]. Even though this paper focuses on monitoring between edge nodes that surround a core network, the method is not limited to this case. Measurements in access networks and server networks are two other important applications for the method.

3.5 Traffic Meters and Monitoring Packets

The basic idea is to use traffic meters in conjunction with monitoring packets, which are inserted into the user traffic. These functions could either be implemented in the existing edge routers or as stand-alone systems. Traffic meters act as packet filters designed to count packets (and bytes) belonging to the kind of traffic subjected to monitoring by the operator. These filtering and counting functions are coordinated with the generation of dedicated monitoring packets, inserted between blocks of ordinary data packets as shown in Figure 2. Alternatively the distance between the monitoring packets could be implemented as time intervals.

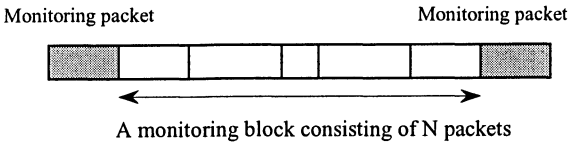


Figure 2: Two monitoring packets enclose a monitoring block that consists of N user packets on the average.

The sending monitoring edge router (or stand-alone system) generates monitoring packets, for example between every N user packets on average or at specific time intervals. These monitoring packets can be used in two different ways: either to carry monitoring information (counter values and timestamps), or to synchronize storage of these values in the corresponding monitoring edge routers (or stand-alone systems). In the latter case the critical synchronization process between the counter functions and the generation of monitoring packets is not necessary.

The sending monitor inserts the number of packets (and bytes) sent so far and a timestamp into the monitoring packet (or stores these values locally). The receiving monitor detects the monitoring packets, inserts (or stores locally) the number of packets (and bytes) received so far and a timestamp.

3.6 Measurement Results

It is possible to obtain the following results using this method.

- The number of lost packets between sending and receiving nodes and the packet loss ratio during the measurement period.
- The length of the loss-free periods and the loss periods expressed in terms of the number of consecutive monitoring blocks that contain lost packets and the number of monitoring blocks that are loss-free. Alternatively this could be expressed as loss-free and loss seconds.
- Samples of the packet delays, and delay variations, between sending and receiving nodes.
- An estimate of the utilized capacity (throughput) between each pair of sending-receiving edge nodes. This requires that the packet length can be estimated.

The results could be based on continuous per day measurements or limited to busy periods.

3.7 Simulations Based on Traffic Data

The described method has also been simulated in Matlab® based on traffic from the RIPE-NCC test traffic boxes [3]. The traffic data are recordings of traffic between test boxes in Copenhagen, Stockholm and Helsinki during 10 days in November 2000. The simulations show that the method provides acceptable estimates of the average delay and delay variations, packet losses and throughput [1].

4. Implementation

4.1 A Prototype for Embedded Monitoring

In this section we describe an implementation of embedded monitoring where these functions constitute a part of a router. A proof of concept rather than to optimize performance has been the main goal.

The main idea is to implement counters related to packet filters that keep track of the number of incoming and outgoing packets, and insert dedicated monitoring packets between blocks of N data packets on the average. These functions could be realized either on the network interface cards or in the kernel.

The test environment in Figure 3 consists of two hosts that send and receive UDP packets, two edge routers and a computer that emulates losses and delays in a core network. The edge routers communicate by means of an IP tunnel. PC-based routers running the Linux operating system have been used. The work is documented in [14].

The embedded monitoring functions within the edge routers consist of a packet filter, counting functions, a process that sends (receives) monitoring packets and facilities the storage and processing of monitoring data. Iptables in Linux [15] has been configured as a filter that selects the relevant packets to be monitored. Netfilter [16] is then used to forward the packets from the kernel to a queue in user space, where the monitoring functions are implemented as a C program.

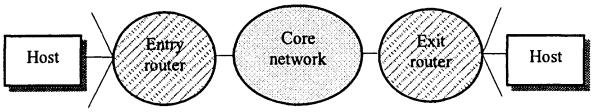


Figure 3: The figure shows the test environment for the prototype implementation. Losses and delays in a core network are emulated using Nistnet.

The counters keep track of the cumulative number of sent (received) packets that pass through the filter. The

monitoring packets are inserted between blocks of N user data packets. These dedicated packets carry timestamps, the number of sent and received packets so far, and a sequence number. Packet losses and delays in the core network are emulated using Nistnet [17].

4.2 The Monitoring Functions

Figure 4 shows a brief flow diagram of the monitoring procedure in an entry-monitoring node. The main functions in respective modules are explained below.

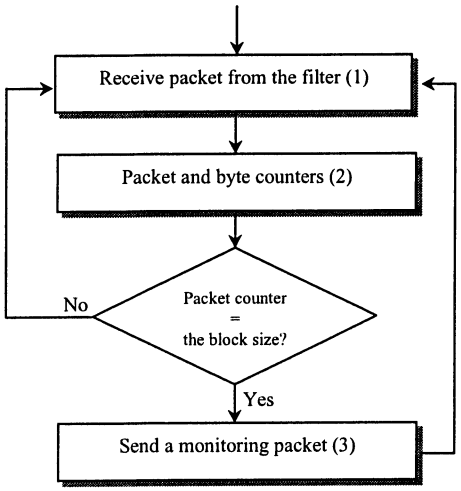


Figure 4: A flow diagram that shows the structure of the monitoring functions in an entry-monitoring node. The numbers within brackets refer to the comments below.

1. Packets that match the filter conditions in the entry router, implemented in Iptables, are sent from the kernel to a program in user space that handles the monitoring functions.
2. The cumulative values of the packet counter and the byte counter are incremented with respect to the incoming packet.
3. If the packet counter has reached N (the size of the monitoring block) a monitoring packet is generated. A timestamp, the current values of the packet counter, the byte counter and a sequence number are inserted into the monitoring packet. The data packet and the monitoring packet are returned to the kernel, and forwarded according to the ordinary routing functions.

The flow diagram in Figure 5 illustrates the monitoring functions in an exit-monitoring node.

1. Packets that match the filter conditions in the exit router, implemented in Iptables, are sent from the kernel to a program in user space that handles the monitoring functions.
2. The cumulative values of the packet counter and the byte counter are incremented according to the incoming packet.
3. If the exit router receives a monitoring packet a timestamp and the current value of the counters are inserted, and the packet is sent back to the entry router.

If the clocks in the entry and exit nodes are synchronized the one-way delay can be calculated. Otherwise, as in current version of the prototype, the timestamp at the exit node registers the arrival times for the monitoring packets. When the monitoring packet has returned to the entry node a timestamp is inserted, in order to compute the round-trip time.

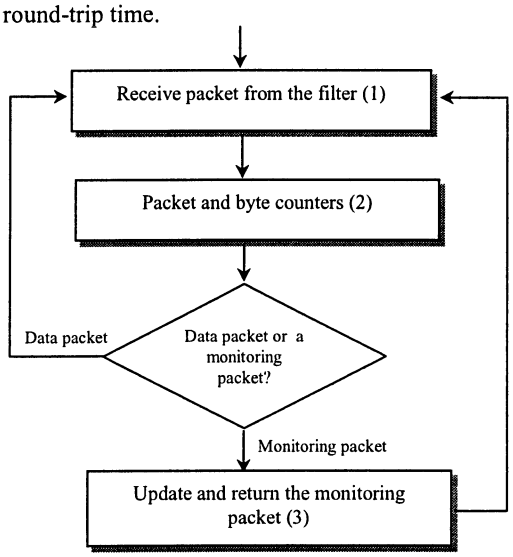


Figure 5: A flow diagram of the monitoring functions in an exit node. The numbers in brackets refer to the comments above.

The Figures 6-8 illustrate examples of measurement results that can be obtained using the prototype. The size of the monitoring block is 1000 packets in all referred cases. Figure 6 shows the distribution of the round-trip times and Figure 7 shows the distribution of the length of the loss-free periods. The distribution of the throughput can be seen in Figure 8. The maximum throughput is around 310 kb/s and the mean value is 260 kb/s in this example.

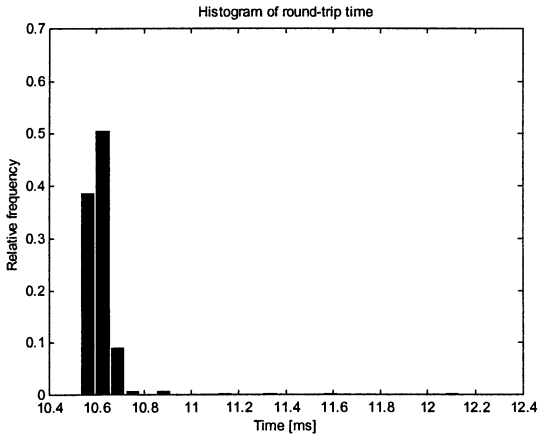


Figure 6: Distribution of the round-trip times with a mean value of 10.6 ms and a maximum of 12.1 ms.

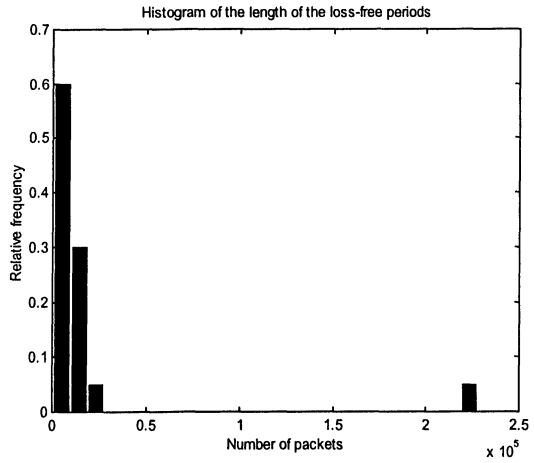


Figure 7: Distribution of the length of the loss-free periods.

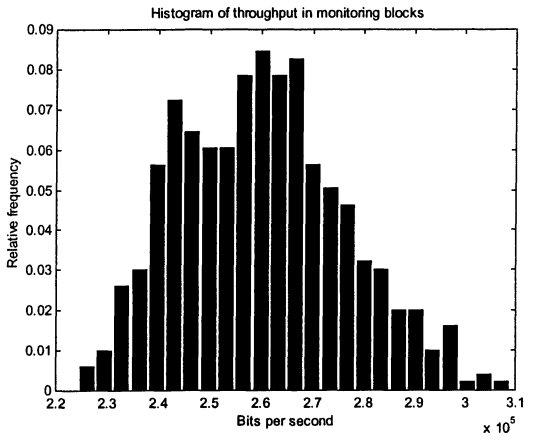


Figure 8: Distribution of the throughput per monitoring block.

4.3 Monitoring of Service Level Agreements

Performance monitoring is of special concern when the traffic is divided into different service classes with certain constraints and guarantees regarding packet losses, delays and throughput. The monitoring functions (packet filters and meters) can be configured to distinguish between different types of traffic and service classes. Some examples of guarantees for quality of service in SLAs that are feasible to verify are given below. The measurements can be configured per service class and per VPN.

- A service class with guarantees for low levels of packet losses, e.g. the packet loss ratio $< x$ % and the loss-free periods (or blocks) $> y$ seconds (or # packets) on the average.
- A service class with guarantees for low levels of packet delays, e.g. a mean delay $< z$ ms (with specified confidence level and interval), a maximum delay below u ms (below the 95 percentile), and a mean delay variation for the monitoring packets less than v ms (below the 95 percentile).
- The utilized capacity can be measured for all service classes, with different degrees of resolution.

5. Conclusions and Future Work

The paper presents a simple prototype implementation of a method for performance monitoring, especially in IP-based virtual networks, which can be seen as a combination of the traditional passive and active approaches. The results using this method are estimates of losses (the overall packet loss ratio as well as the dispersion of losses), delays and delay variations, and throughput between edge nodes. The size of the monitoring block is an important parameter that determines the accuracy and degree of resolution of the measurements. Besides the embedded approach stand-alone implementations will also considered for future work. Monitoring of streaming media from server networks and IP traffic in mobile internet are also interesting applications for the method.

References

[1] Lindh T.: "An Architecture for Embedded Monitoring of QoS Parameters in IP-Based Virtual Private Networks", Proceedings of Passive and Active Measurements (PAM) 2001, Amsterdam, 23-24 April 2001.

[2] Lindh T.: "Embedded Monitoring of QoS Parameters in IP-Based Virtual Private Networks", Proceedings of Integrated Network Management VII (IM2001), Seattle, 14-18 May 2001.

[3] Georgatos F., Gruber F., Karrenberg D., Santcroos D., Susanj A., Uijterwaal H., Wilhelm R.: "Providing Active Measurements as a Regular Service for ISP's", Proceedings of Passive and Active Measurements (PAM) 2001, Amsterdam, 23-24 April, 2001.

[4] Paxson V., Adams A.K., Mathis M.: "Experiences with NIMI", Proceedings of Passive and Active Measurements (PAM) 2000, Hamilton, New Zealand, April 3 and 4, 2000.

[5] Kalidindi S., Zekauskas M.: "Surveyor: An Infrastructure for Internet Performance Measurements", INET'99, San Jose, June 1999.

[6] McGregor T., Braun H-W., Brown J.: "The NLAR Network Analysis Infrastructure", IEEE Communications Magazine, Vol. 38, No. 5, May 2000.

[7] White Paper - NetFlow Services and Applications: http://www.cisco. com/warp/public/cc/pd/iosw/ioft/neflct/tech/napps_wp.htm.

[8] Service Assurance Agent: http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120t/120t5/saaoper.htm.

[9] NeTraMet home page: http://www2.auckland.ac.nz/net/ NeTraMet/

[10] Beige M., Jennings R., Verma D.: "Low Overhead Continuous Monitoring of IP Network Performance", Proceedings of Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS'99, Chicago, July 1999.

[11] "Quality of Service Methodologies and solutions within the service framework: Measuring, Managing and Charging QoS", EURESCOM Project P906, January 2001.

[12] "SLA Management Handbook", TeleManagement Forum, June 2001.

[13] "The challenge of enabling QoS in IP networks", Operax AB, October 2001.

[14] Nordén P.: "Monitoring of Performance Parameters in IP Networks", Degree project, KTH, 2001.

[15] Linux 2.4 Packet Filtering HOWTO, http://netfilter.filewatcher.org/unreliable-guides/packet-filteringHOWTO/index.html

[16] The Netfilter Project: Packet Mangling for Linux 2.4: http://www.netfilter.kernelnotes.org.

[17] Nistnet home page: http://www.antd.nist.gov/itg/nistnet/.

A distributed admission approach based on marking mechanism over Bluetooth best-effort network

Le Thanh Son, Henrik Schiøler and Ole Brun Madsen
Distributed Real-Time Group
Department of Control Engineering
Aalborg University
Fredrik Bajersvej 7C, 9220 Aalborg E, Denmark
lts,henrik,obm@control.auc.dk

Abstract

The end-to-end Quality of Service delivered in Bluetooth networks depends on a large number of parameters at different levels, e.g. link capacity, packet delays, etc, which are requested in certain patterns and controlled by various algorithms. In this paper, a method of adaptive distributed admission with end-to-end Quality of Service (QoS) provisions based marking information for real time and non real time traffics in Bluetooth networks is highlighted, its mathematical background is analyzed and a simulation with bursty traffic sources, Interrupted Bernoulli Process (IBP), is carried out. The simulation results show that the performance of Bluetooth network is improved when applying the distributed admission method.

1 Introduction

With the current development of mobile devices, short range wireless communications have become more and more popular, and many research on short range wireless communications, such as Bluetooth, have gained special interests, especially in industry. Bluetooth is a wireless communication technology using the unlicensed Industrial-Scientific-Medical (ISM) frequency band, supporting synchronous voice traffic and asynchronous IP-based data traffic [1]. The primary focus with Bluetooth is cable replacement. However, future uses as mobile ad hoc network for Bluetooth are awaited and supported by specifications.

The end-to-end Quality of Service (QoS) delivered to end users in Bluetooth networks depends on a large number of parameters at different levels. The mobility of the "thin" nodes, scarce resource and the limited power available make them even more difficult to handle efficiently compared to fixed networks. The current Bluetooth network does not give any QoS guarantees and in [5, 6] the issues of QoS was mentioned, like packet delay, packet losses and congestion. There are some methods to address these issues, e.g. Random Early Discard (RED) [8] or Measurement based admission control (MBAC) [11]. A problem nevertheless remains with these methods is how decisions should be made on an end-to-end QoS basis. There is another approach, where the connection will decide whether or not it should accept a Bluetooth packet to enter the network. The Explicit Congestion Notification (ECN) [7] provides a mechanism to convey information back through an ECN marking bit. Recent research [9, 10] show how virtual queue marking scheme allows a broad QoS framework for flows. In this paper, an approach of the distributed admission control with the end-to-end Quality of Service (QoS) provisions for real time and non real time traffics for Bluetooth networks is highlighted. The key idea is to consider the acceptance of a new Bluetooth packet, and this depends on the transient behavior of the network. Some mathematical background is analyzed and a simulation with bursty traffic sources, IBP, is carried out. The simulation results show that the performance of Bluetooth network is improved when applying the adaptive distributed admission with QoS provisions approach.

2 Bluetooth technology

Bluetooth [1], a new wireless communication technology, operates at 2.4 GHz in globally available, license-free ISM band. The operating band is divided into 1 MHz-spaced channels with the chosen modulation scheme of GFSK (Guasian Frequency Shift Keying), this equates to 1 Mb/s. The channel is divided into consecutive time slots, each time slot lasts 0.625 ms. Every Bluetooth device has a unique Bluetooth device address. All Slaves use the Master's clock and address to synchronize to the Master's frequency hop sequence. The Master allows Slaves to transmit by allocating slots for voice or data traffic.

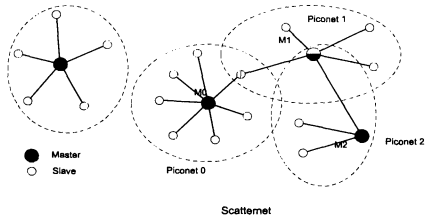


Figure 1: Bluetooth piconet and scatternet

Bluetooth packet is sent in one slot, three slot, or five slots on single hop channel. Between master-slave pairs, both Synchronous Connection Oriented (SCO) link and Asynchronous Connectionless link (ACL) are supported. A collection of multiple piconets can be linked into a larger ad hoc network, which called scatternet, where some devices are member of more than one piconet, which called bridge node. Figure 1 shows an example of a piconet and a scatternet, which includes three piconets. The bridge node can have a role as master in a piconet and slave in another or slave in both. It requires carefully packet scheduling for the node participating in multiple piconets. Bluetooth also supports power-saving techniques. A slave in a piconet may be in active, parked, sniff or hold mode.

3 Adaptive distributed admission approach with QoS provisions in Bluetooth network

3.1 Model analysis

Consider a Bluetooth network with a set of links L where a link $l \in L$ has capacity c_l. The network has a set of traffic sources R that connect to nodes in the network. Associated with each source $r \in R$ is a route also denoted by r, which consists of a subset of L. The source generates traffic at rate x_r. The packets are transmitted through the network and received with rate $y_r \leq x_r$. The received rate could be less than transmitted rate not only because of congestion in Bluetooth network, but of many other factors, such as transmission errors, which are significant in wireless network. The loss rate for source r on a link $l \in r$, if no feed back from the network, is given by:

$$PL_{r_l} = x_r p_l(c_l, \sum_{s:l\in s} x_s) \tag{1}$$

where p_l is the total loss probability at link l, when total arrival rate into the link is $\sum_{s:l\in s} x_s$. It could be assumed the total loss is distributed among users in proportion to their rates and the nodes in Bluetooth network transmit packets with no transmission errors, i.e p_l depends on the capacity of link, which causes congestion at nodes in the network, and total arrival rate into the link.

The total loss rate for a traffic source due to congestion is:

$$PL_r = \sum_{l:l\in r} x_r p_l(c_l, \sum_{s:l\in s} x_s) \tag{2}$$

If $U_r(x_r)$ is utilization function of rate x_r, the optimal rate for the network can be obtained by solving the following maximization problem [13]:

$$\max_{x_r} \sum_r \left\{ \Delta_r U_r(x_r) - \beta \sum_{l:l\in r} x_r p_l(c_l, \sum_{s:l\in s} x_s) \right\} \tag{3}$$

where $\Delta_r > 0$.

3.2 Distributed marking mechanism

The distributed marking mechanism has a bufferless model. At each node j on the link l, Bluetooth packets, which come to the node, are marked when the arrival packets at the node in the link exceed virtual queue level $\tilde{b_j}$, where $\tilde{b_j} = kb_j$, $k \leq 1$ and b_j is real buffer size of the node. The information on the marked packets is periodically sent back to the source nodes. The buffer occupancy process of the node j of link l on route r is a function of arrival rate $\lambda_j^a = \sum_{r:l\in r} x_r$ and service rate λ_j^d.

The virtual queue of the node has service rate ρ_j^d in place of λ_j^d:

$$X_j = \Phi(\lambda_j^a, \rho_j^d) \tag{4}$$

If a_j is acceptance probability for node j on route r, the acceptance probability depends on the buffer occupancy of the node Q_j. When a Bluetooth packet enters a node in the network, it requests a route. The node probes and accepts the entering packet with probability:

$$\overline{a_r} = \prod_{j\in r} a_j(Q_j) \tag{5}$$

The product form acceptance probability corresponds to the assumption of independence of the packet level dynamics at each sources, conditional on the current load, a packet must be accepted by each node along its route link:

$$a_j = P(X(Q_j) < \tilde{b_l}) \tag{6}$$

If queue delays are negligible, the deterministic fluid flow approximation is considered, then from [10], a

differential equation for the accepted rate x_r of user r in the time evolution:

$$\frac{d}{dt}x_r(t) = \kappa_r \left(w_r - x_r \sum_{j \in r} \mu_j(\lambda_j^d, \sum_{s:j \in s} x_s(t)) \right) \quad (7)$$

where $\mu_j(\lambda_j^d, \sum_{s:j \in s} x_s(t))$ is the probability that a packet at node j to be marked at time t, w_r is target number of marks per unit time and κ_r is positive gain parameters. If we assume that marking is independent from node to node, then the probability that a packet on route r receives a mark at time t is $1 - \prod_{j \in r}(1 - \mu_j(\lambda_j^d, \sum_{s:j \in s} x_s(t)))$.

The messages containing marking information of links on routes and time stamp are periodically sent back to source nodes. The updating period can be optimized for a particular traffic pattern to give a good measure of buffer occupancies, which expressed in the number of marks, at the nodes in network.

3.3 The adaptive distributed admission control approach with QoS provisions

The adaptive virtual queue based distributed admission with QoS provisions for Bluetooth network is described in figure 2. From [10], the acceptance probability of a user r of a node can be estimated through the packet marking information from each link during period T_u, and is represented by packet marked probability:

$$Pr(\sum_{i \in T_u} W_i^l < \widetilde{w}_p^l) < p_{mark}^l \quad (8)$$

where W_i^l is the number of packets marked, returning periodically to the source node from the node j on the link l of the route r, $\widetilde{w}_p^l$ is threshold of the packets marked receiving during the probed period.

As mentioned from above, the marking information is periodically updated to the source node. Instead of computing the exact probability, it could, for example, estimate from a history records marking packets received during current probed period, previous periods and traffic estimation to get the mean $\widehat{\mu}_j^l$ and standard deviation $\widehat{\sigma}_j^l$, then use the granting criterion for each packet, which arrives to the node and makes a request of route r:

$$\widehat{\mu}_j^l + Q\widehat{\sigma}_j^l < \theta_{mj}^l \quad (9)$$

where θ_{mj}^l is the mark threshold for link l of the routes from the node j and Q is some fixed value depending on the allowed mark probability p_{mark}^l. When the node has measured large enough change in the estimated packet marked probability for a link, it will start capacity reallocation for the source routes of the link.

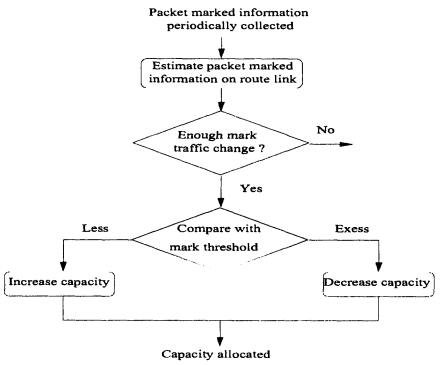


Figure 2: Adaptive virtual queue based distributed admission approach with QoS provisions for Bluetooth network

The next step is to adaptively change capacity, which can be done through the acceptance probabilities of the sources at the node, which require the routes r comprised in the link l according to the estimated packet mark probability and the granting criterion.

4 Simulations

4.1 Simulation model

The Bluetooth network in consideration is a simple scatternet, as depicted in figure 3, which consisted of two inter-connected piconets. The piconet provides full-duplex transmission using time slots, where each slot is 0.625 ms long. The polling in Bluetooth can be done in many different ways. The difference between the polling mechanisms is related to the order in which slaves are polled and the service disciplines used to serve a slave. In this simulation, we use the Fair Exhaustive Polling (FEP) [3] as the polling scheme in the piconet. The FEP gives high bandwidth efficiency with a fair allocation and still being simplistic. Each piconet has a master and N slaves, $0 \leq N \leq 7$ and its traffic can be modeled as in [5]. The inter-piconet traffic is transfered through a bridge node ($S07/S17$, fig. 3). It is possible to send packets in multiple slots, which may be either 3 or 5 slots long. In this study, only the single slot data packet transmission between nodes in the network is considered. Further, we assume that there is no transmission error and packet loss only occurs when the buffers are overflown. Moreover, the piconets were assumed being synchronized, which means there were no guard frames when the bridge node switched between the piconets.

The homogeneous traffics of packet arrivals can be modeled as Interrupted Bernoulli Process (IBP)

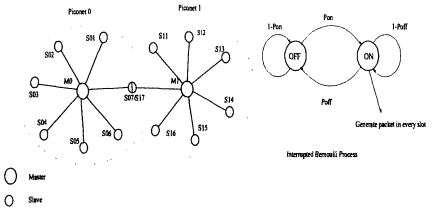


Figure 3: The modeled Bluetooth scatternet with two inter-connected piconets and IBP traffic model

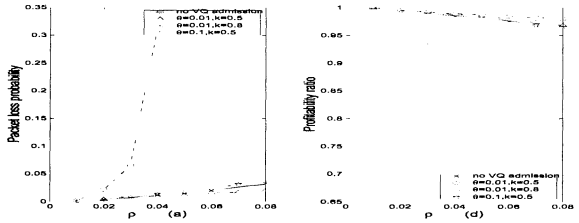


Figure 4: Performance of Bluetooth scatternet, with and without adaptive virtual queue distributed admission control with QoS provisions, in terms of packet loss performance and profitability ratio

sources. The probability of transitions between ON and OFF states could be characterized by p_{ON} and p_{OFF}[2]. The packet arrival is aligned with the time slots in the modeled piconets.

The modeled scatternet uses the predictive inter-piconet scheduling [6], with the scheduled inter-connected window period T_s of the bridge node is set to 400 time slots and is fairly allocated to the piconets, which the bridge node is connected to. The allocated active intervals of the bridge node on each piconet, which it belongs to, are dynamically varied depending on inter-piconet traffics and to/from the bridge node.

To verify working ability of the adaptive virtual queue based distributed admission with QoS provisions approach in the Bluetooth simulated network, the output buffer sizes on each node are set to 40 packets, the virtual queue levels of each node are set to 0.5 and 0.8 and the marking thresholds θ^l_{mj} are kept to 0.01 and 0.1. The acceptance probability varies according to the estimated marking probability in a simple additional increase/multiplicative decrease steps, which is very popular in TCP flow control. In here, the multiplicative ratio is set to 2 and the step is 0.05. The simulation length is 10^6 time slots.

4.2 Numeric results

Figure 4 shows comparison of the modeled scatternet, in terms of the packet loss performance and profitability, with and without the adaptive virtual queue based distributed admission control with QoS provisions, different marking thresholds and virtual queue levels. The profitability is ratio between handled traffic without overhead and traffic admitted to the network. An admitted packet arriving to a node is considered to be lost if the buffer of the node is full. With increasing traffic load, the probability of packet loss was kept very low and profitability was kept high when distributed admission control was applied, comparing to the case when the distributed admission control was not applied, which suffered very high packet loss and steep decreasing of profitability.

Figure 5 shows the performance of the modeled network, when the adaptive distributed admission was applied and different marking thresholds and virtual queue levels. It can be seen from figure 5.a, when increasing traffics of nodes in the network, the probability of the marked packets was increased. Higher packet marking threshold means higher packet marked probability. In figure 5.b, it also shows the variations of the probability of packet loss corresponding to the traffic variations. It indicated that when the traffic load was increased, the probability of packet loss was also increased. But it was apparent that the probability of packet loss was far lower comparing to the case that not applying the adaptive virtual queue based distributed admission control with QoS provisions. It also indicated that when increasing the marking threshold or virtual queue level, the probability of packet loss was increased.

Figures 5.c and 5.d show the performance of the modeled network, in terms of the packet blocked probability and the process gain ratio. The process gain ratio is defined by ratio of handled traffic, without overhead, and arrival traffic. The increment of traffic loads of the nodes in the network was followed by the increment of the packet blocked probability and decrement of process gain. When the packet marking threshold was increased or virtual queue level was increased approaching to the real buffer size, the process gain was higher. It means there was an increment of throughput. However, from figures 5.a and 5.b, the packet loss was also higher.

5 Conclusions and future development

In this paper, the adaptive virtual queue based distributed admission control with QoS provisions approach for Bluetooth network was presented and analyzed. The virtual queue fundamentals were studied. The virtual queue marking scheme and the adaptive distributed admission approach, which satisfied constraints of Bluetooth network, e.g 'thin' node, scarce resource, etc, were proposed and analyzed. The sim-

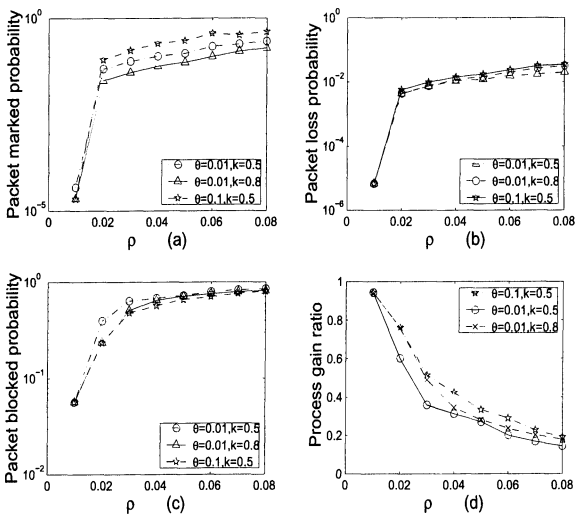


Figure 5: Performance of Bluetooth scatternet, with adaptive virtual queue based distributed admission control with QoS provisions, different marking thresholds and virtual buffer sizes

ulation of behaviors of Bluetooth network with and without the adaptive distributed admission control with QoS provisions were carried out for bursty traffic sources. The results showed the necessity of the adaptive distributed admission control approach, which can offer high capacity, while still satisfies QoS requirements. The simulation also showed some characteristics of the approach and indicated relationship between QoS requirements, such as packet delay, packet loss, and throughput.

The working ability of Bluetooth network, when the adaptive distributed admission control approach is applied, will be significantly improved with constructing appropriate capacity estimation and optimizing marking scheme according to traffic characteristics. The complexity of the approach is on the management of marking information (i.e. delays, estimations, timeouts). These request further researches on marking scheme as well as capacity estimation in various dynamic traffic patterns, for both real time and non real time traffics, especially Internet traffic.

Acknowledgements

We would like to express our special thanks to Department of Control Engineering, Aalborg University and Department of Telecommunications and Signal Processing, Blekinge Institute of Technology, for supporting to the research.

References

[1] Bluetooth Special Interest Group. *Specification of Bluetooth System.* Documentation available at http://www.bluetooth.com, May 2001

[2] N. Johansson, M. Kihl and U. Korner. *TCP/IP over Bluetooth Wireless Ad-hoc Network.* IFIP Networking 2000, Paris, May 2000

[3] N. Johansson, U. Korner and P. Johansson. *Performance Evaluation of Scheduling Algorithms for Bluetooth.* Proc. of IFIP TC6 Fifth International Conference on Broadband Communications '99, Hongkong, November 1999

[4] P. Johansson *et al Inter-piconet communication in Bluetooth networks.* November 2000

[5] H. Schiøler, J.D. Nielsen, O.B. Madsen and S. Le Thanh *Communicating cooperative robots.* Proc. WPMC'01, Aalborg, September 2001

[6] S.T. Le, H. Schiøler, O.B. Madsen *Predictive scheduling approach in Inter-piconet Communications.* Proc. WPMC'01, Aalborg, September 2001

[7] S. Floyd *TCP and Explicit Congestion Notification.* ACM Computer Communication Review,24:10-23, 1994

[8] S. Floyd and V. Jacobson *Random Early Detection gateways for congestion avoidance.* IEEE/ACM Transactions on Networking, 1:397-413, 1993

[9] R.J. Gibbens and F.P. Kelly *Distributed congestion acceptance control for a connectionless network.* Proc. of the 16th International Teletraffic Congress-ITC 16, Elsevier, June 1999

[10] F.P. Kelly, Peter B. Key and Stan Zachary *Distributed admission control.* IEEE JSAC 2000, unpublished

[11] T.K. Lee and M. Zukerman *Simple Measurement based connection admission control for heterogeneous traffic sources.* Proc. ITC 98, 1998

[12] Harold J. Kushner *Heavy Traffic Analysis of Control Queueing and Communication Networks.* Springer, 2001

[13] S. Kunniyur and R. Srikant. *A time-scale decomposition approach to adaptive ECN marking"* to appear in the IEEE Transactions on Automatic Control, Anchorage, Alaska, 2001.

[14] Costas A. Courcoubetis *et al Providing bandwidth guarantees over a best-effort network: call-admission and pricing"* Infocom '01, Anchorage, Alaska, USA.

On Router Control for Congestion Avoidance

Marine Auvert*, Håkan Hjalmarsson*, Karl H. Johansson*, and Gunnar Karlsson†

* Dept. of Signals, Sensors & Systems, Royal Institute of Technology, kallej@s3.kth.se
† Dept. of Microelectronics and Information Technology, Royal Institute of Technology, gk@imit.kth.se.

Abstract

This short paper deals with active queue management for computer networks. The goal is to develop control mechanisms for routers in heterogeneous networks that reduce traffic fluctuations. The proposed control strategy operates with local information (such as estimated arrival rates) and actively use the buffers to smooth traffic, and thus it avoids the buildup and propagation of traffic bursts.

1 Introduction

A continually increasing use of the Internet and its widening set of offered communication services increase the demands on the control of the network. The control objective is often to improve traffic throughput and to accommodate different service demands better. Traffic congestion in statistically multiplexed computer networks is a major concern. Today's congestion control for the Internet is implemented as an end-to-end protocol, namely the transport control protocol (TCP) [1, 2, 3, 4]. The service policy in routers is predominantly work conserving: packets are sent out at the full rate of the output link. This allows congestion to propagate to downstream routers and might create a wide area of congestion. There are (at least) two reasons for why it is difficult to attenuate this behavior with existing congestion control mechanisms such as TCP: (1) congested network traffic has often both long-range and short-range dependencies, and is therefore hard to control at the single time scale of the end-to-end protocol; and (2) the actual control actuation of the end-to-end protocol on the congested area is implicit and complex, since there are several protocol instances that interact. Park [5] points out the need for control strategies working over longer time scales compared to what is the case today. Recent models capturing the self-similar nature of the traffic [6, 7] open up new possibilities for the use of model-based control. Our idea is to use feedback control to smooth the pattern of departing packets of individual routers. The importance of smooth network flows has been emphasized by Massoulié and Roberts [8].

Initial work on router control strategies for reducing traffic fluctuations is presented in this paper. The main contribution is to introduce a particular idea based on a so called virtual service rate, which can be implemented in today's routers. Some preliminary simulation results are presented. Smoothing has been used in network control before, but then on an end-to-end control basis. For example, smoothing has been proposed for reducing bit rate variability for video transmission [9], which seems not to have been studied from a feedback control point of view.

2 Congestion Control

Traffic congestion in communication networks can be reduced for example by introduce new control mechanisms in the routers or by conventional end-to-end control protocols such as TCP.

2.1 Router Control

By actively using the buffers in the routers to smoothen traffic, it is possible to improve the overall efficiency of a packet-switched network. If each router sends out packets at the full rate, the utilization of the individual link is maximized at each time instant. Such a local optimization might not be globally optimal, however. This is due to limited buffer capacity in each router of the network. If one of the routers overflows, and thus packets are dropped, retransmission leads to decreased overall throughput of the network. A better strategy is then to spread out the temporary storage of packets to several buffers. A globally optimal solution needs a centralized control structure where information about the traffic throughout the network is transferred and where the control decisions are taken. This is not realistic for networks such as the Internet. However, it is possible to improve the performance also with local information and control.

Although routers forward packets at the full rate of the output links, it would be possible to use the output buffers to smoothen the pattern of departing packets to avoid aggregation of bursts. Each packet has, of course, to be sent at the bit rate of the link. But after a packet has been sent, the server could wait before serving the next packet. The length of the introduced idle period could depend on such dynamic information as the fullness of the buffer, the past departure pattern, and the predicted arrival pattern. The hope is that the regulated traffic will make congestion

more rare, which in turn means that losses are reduced and delays are less varying. This would in turn lead to higher throughput and less capacity wasted on retransmissions. The control law should determine the lengths of the idle periods between packets, so that the variance of the departure process is minimized.

2.2 End-to-End Control

A fundamental question is how much can be gained by using feedback information to regulate flows in a network. The end-to-end control in TCP uses acknowledgements from the receivers as the only explicit feedback information whereas the available bit rate (ABR) service in an ATM network relies on information from the switches. Similar feedback is suggested also for the Internet protocol by means of explicit congestion notification [10]. Explicit feedback increases both the network complexity and the requirements on the nodes. We therefore believe that it is of interest to quantitatively assess the performance benefits of explicit feedback information from the network. This would for instance tell how conservative it is to separate the traffic flow control problem into local controls in the nodes and end-to-end control. The ideal objective is to produce curves of performance gain vs. feedback bit rate. Work in this area is yet sparse [11]. The issue of what type of feedback information to use also needs to be addressed. An interesting suggestion by Kelly and co-workers [12] is that the shadow prices for resources should be used. It is critical to consider the delay dependence in this analysis since the benefits of feedback diminish as the network delays increase.

3 Network Flaps

It is possible to influence the distribution of the outgoing traffic from a router by adding an idle time period to the service time.

3.1 The Idea

Packets are sent out at the maximum transmission rate in routers today. This leads to that traffic bursts propagate downstream in the network. To avoid the accumulation of bursty traffic, which may create a congested area of the network, the proposed idea is to smooth out the distribution of the departure times of the routers. This idea can be compared to the action of rubber flaps on a conveyer belt, where the flaps even out the distribution of the goods. Therefore, we call the mechanism added to the routers *network flaps*.

3.2 Prototype System

Consider a simple subnetwork consisting of two routers in series modeled as the queueing system in Figure 1. This is

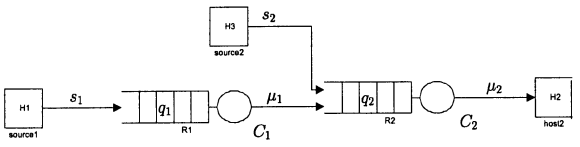


Figure 1: Prototype network with two buffers.

the simplest configuration to illustrate the idea of network flaps and is based on a bottleneck topology. The queues contain $q_i(t)$ packets, $i = 1,2$, at time instant t, where $0 \leq q_i(t) \leq q_{i\max}$. The incoming traffic comes from two virtual sources having time varying source rates $s_i(t)$, $i = 1,2$. The output link capacities are fixed and equal to C_1 and C_2, respectively. By assuming that $C_1 > C_2$, the second link represents a bottleneck, and possibly the two flows aggregating at the second buffer create an overflow. The output rates $\mu_i(t)$, $i = 1,2$, are limited by the corresponding capacity. The arrival rates of the two buffers are $\lambda_1(t) = s_1(t)$ and $\lambda_2(t) = s_2(t) + \mu_1(t)$.

It is common in most applications to let the departure rate attain its maximum value $\mu_i = \mu_{i\max}$ as soon as there are packets to send, i.e., packets in buffer i are sent out at maximum rate. This leads to maximal utilization of the link at each time instant. When the traffic is bursty and the load of the network is high, this approach however may lead to performance losses. The evolution of the queue lengths are shown in Figure 2 for the system in Figure 1, when the inter-arrival times for the source rates have exponential distribution and $\mu_i = \mu_{i\max}$. The first queue never builds up due to that the capacity C_1 is basically high enough to let all packets be served as soon as they arrive. The second link clearly creates a bottleneck since the capacity C_2 is not high enough. Therefore q_2 grows. In this particular example $q_{1\max} = q_{2\max} = 50$, so the bottleneck leads to loss of packets, as shown in the lower plot of Figure 2.

3.3 Virtual Service Time

We suggest that bursty traffic should be smoothed by delaying packets. This can be interpreted as considering the output rate μ_i as a control variable, which belongs to the interval $[0,\mu_{i\max}]$. An equivalent interpretation is to introduce the notion of a *virtual service time* T_v, which is equal to the sum of the actual service time T_s and a controlled delay T_c, i.e., $T_v = T_s + T_c$. Of course, the control signal $T_c \geq 0$ only takes positive values, and for a conventional router we have $T_c \equiv 0$.

A simple illustration of the effect of a virtual service time is to let T_c be a positive constant. The result of this is shown in Figure 3. Note how the first buffer is utilized to store packets, so that the total number of dropped packets in this case is less than in Figure 2. It is reasonable to let T_c be chosen such that T_v is approximately constant over a

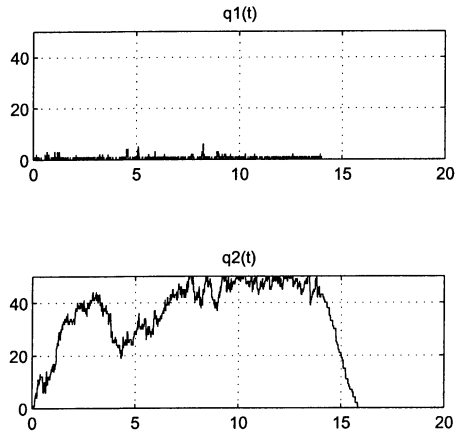


Figure 2: Simulation of the system in Figure 1. The departure rates are limited by the fixed capacities of the links.

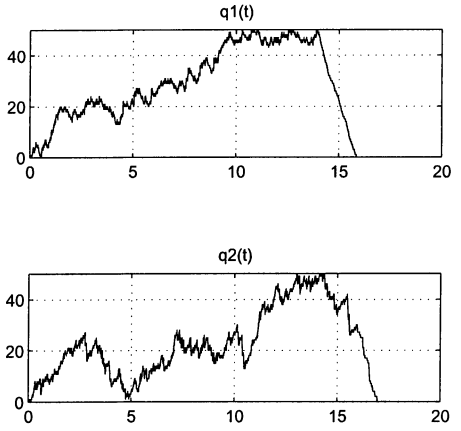


Figure 3: Introducing a virtual service time distributes the packet load over both buffers and reduces the total loss of packets.

certain time interval. A natural choice is to let this constant be equal to the (estimated) mean of the inter-arrival times.

Note that the control strategy above should be applied when the load of the (sub)network is high. When the load is low, it is always optimal to use the maximum sending rate. This is captured by the switched control strategy with three states illustrated in Figure 4. We assume that only local information is available to the router, i.e., the router may estimate the arrival rate λ and measure the queue length q. If λ is small, it can be interpreted as if there is no risk for congestion and thus $T_c = 0$. If λ is large and $q \ll q_{\max}$, one should use the control actuation available in T_c to smooth the distribution. If both λ and q are large, one should use less control actuation to avoid the own queue to become full.

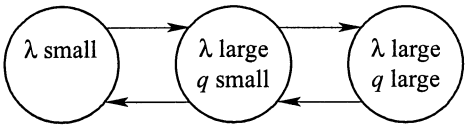


Figure 4: Finite state machine that illustrates the three states of a router controller with network flaps. The controller state depends on the arrival rate λ and the queue length q.

3.4 Preliminary Analysis

The following heuristic analysis gives hint on when network flaps are useful.

One bottleneck: Consider a configuration consisting of two routers in series. Assume Router 1 has an infinite size buffer ($q_{1\max} = \infty$), while the buffer of Router 2 is finite ($q_{2\max} < \infty$). Assume the link connecting the two routers has infinite capacity, but that the outgoing link from the second router has finite capacity C packets/s. Router 1 can be seen as the aggregate of a complete network. The downstream router and its outgoing link will act as a bottleneck.

Note first that the average transmission rate is limited by C and, hence, that if the sender transmits an even flow of F packets/s, where $F < C$, then there is no gain in using network flaps. Under this condition there is no extra delay introduced by the network.

Let now the transmitted flow be bursty. We will consider the extreme case that the sender transmitts FT packets in one second every T seconds, with T being large. The upstream router will buffer these packets and then subsequently deliver them to the downstream router.

Using network flaps, ideally the transmission can be made controlled such that no packets are lost in Router 2. Hence, the receiver will receive the packets with the rate C packets/s and the average packet delay will be $FT/2C$. Clearly, this is optimal.

Without network flaps, Router 1 will (with its infinite buffer size) deliver all packets at once to Router 2. If $FT > q_{2\max}$, this will imply that $FT - q_{2\max}$ packets will be lost. Define LAT, the loss alert time, to be the time it takes for the sender to become aware of the packet loss and to initiate a retransmission of lost packets. Once the sender is notified of the packet loss, the lost packets are immediately sent out. This means that, as long as the bottleneck buffer has not become empty so that the router has become idle, no loss in performance is obtained compared to using network flaps. If, however, $LAT > q_{2\max}/C$, Router 2 will become empty before the re-sent packets arrive and this will cause a loss of performance. It is easy to see that under the assumption that FT is much larger than $q_{2\max}$ (i.e., the peaks of the bursts are sufficiently high), the average packet delay in the network becomes

$$\frac{FT}{2C} \cdot \frac{LAT}{B/C}$$

which obviously is larger than the optimal packet delay $FT/(2C)$ obtained by network flaps. Note that the loss alert time relative to the bottleneck delay B/C plays a role for when network flaps are useful. The loss alert time can be approximated by the round-trip time.

Two bottlenecks: We will now show that network flaps may increase the capacity of a network. Consider a network with three routers. Router 1 and Router 2 are connected to Router 3 by two separate links with capacity C_1 and C_2, respectively. The outgoing link of Router 3 has capacity C_3. Suppose that $C_1 + C_2 > C_3$, so that the downstream router with corresponding outgoing link is the bottleneck when both Router 1 and Router 2 are active. Suppose also that $C_1 < C_3$, so that the first link is the bottleneck when only Router 1 is active. Consider again a bursty scenario, when suddenly senders transmit packages above the network capacity.

Using network flaps, ideally Routers 1 and 2 would adapt their rates to their fair shares (relative their rates) $C_3C_1/(C_1+C_2)$ and $C_3C_2/(C_1+C_2)$, respectively, in order to not overload the bottleneck. The overall rate would then be C_3.

Now, if the upstream routers transmit at full rates but Router 3 favours Router 2, so that this router is allocated more than its fair share of the capacity, then the transmission from Router 2 will finish before the transmission from Router 1. In the instant Router 1 will become the bottleneck, the transmission rate will be reduced from C_3 to C_1. The worst-case scenario is when Router 3 first transmits all packets from Router 2 before the packets from Router 1 are considered. In this case the overall capacity of the network becomes $C_3(C_1 + C_2)/(C_3 + C_2)$, which clearly is less than C_3.

3.5 Simulations

Consider the system in Figure 1 again. Let the inter-arrival times for the source rates s_i, $i = 1,2$, be Pareto distributed with distribution function

$$F(x) = P(X \leq x) = 1 - (\alpha/x)^{\beta}, \qquad x \geq \alpha$$

where α, β are positive constants. Figure 5 shows the response of the system when no control is applied. Note how the burstiness of the arrival rates affects the queue lengths. Also, note that the second queue fills up quite early, although the first is hardly used.

We may now ask if by applying network flaps, it is possible to smooth the traffic. A naive approach is to simply use linear control, based on the queue lengths. Figure 6 shows a control structure where the output rate μ_1 of the first router is controlled using linear feedback from the difference between the queue length q_1 and a set-point value $q_{1\text{sp}}$. The gain of the controller is chosen according to a linear quadratic control design (e.g., [13]). Figure 7 shows

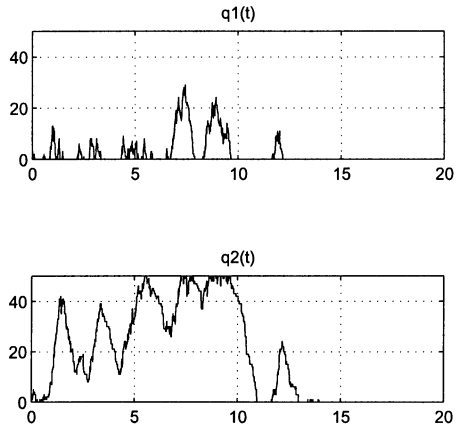


Figure 5: Simulation of protype network in Figure 1 with (bursty) Pareto distributed sources. No control is applied, but the departure rates are maximal.

the result of this control. Note how the first queue builds up. This reduces the load on the second.

Figure 8 shows a control structure where the output rate μ_1 of the first router is controlled using linear feedback from $q_2(t) - q_{2\text{sp}}$. Again the gain of the controller is chosen according to a linear quadratic control design. Figure 9 shows the result of this control.

Both the decentralized and the centralized control structure improve the performance (lower the total number of dropped packets). The second strategy gives the better behavior, but note that it requires the sending of information from the second router to the first (which might not be desired from an implementation point of view).

4 Conclusions

We have discussed router control for congestion avoidance. Our aim is to address the problem by a combination of router controls and end-to-end controls. In this paper we focus on the former. Some promising initial results was shown for the network flaps, our proposal for router control. However, more investigation is needed. We believe that this research area presents several challenging problems from a control theory as well as from a computer network point of view and we hope to present more conclusive results in the near future.

References

[1] V. Jacobson. Congestion avoidance and control. In *Proc. of SIGCOMM*, pages 314–329, 1988.

[2] J. Walrand and P. Varaiya. *High-performance communication networks*. Morgan Kaufmann, 2nd edition, 2000.

[3] L. L. Peterson and B. S. Davie. *Computer networks: a systems approach*. Morgan Kaufmann, 2nd edition, 2000.

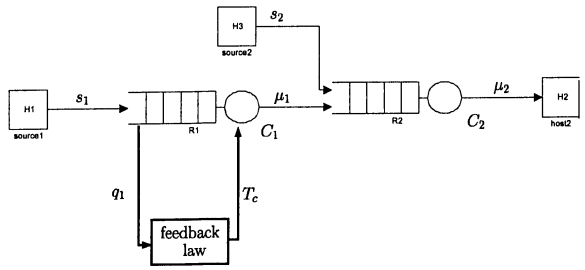


Figure 6: Control structure where the departure rate μ_1 of the first router is controlled using information about q_1. This corresponds to decentralized control, where only local information is used in the controller implementation.

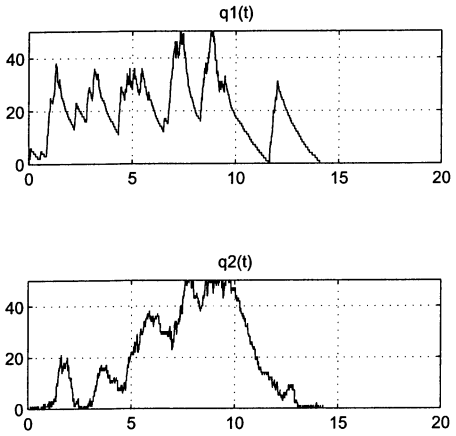


Figure 7: Simulation with control of the departure rate μ_1 of the first router using information about the queue length q_1. The control structure is shown in Figure 6.

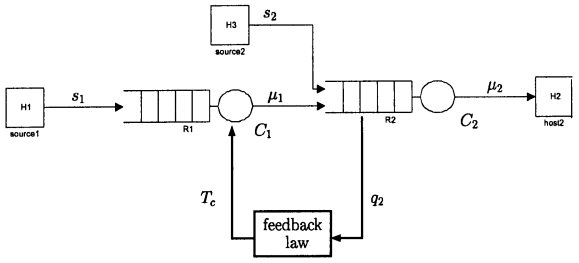


Figure 8: The departure rate of the first router is controlled using information about q_2. This requires information from the second router being send to the first, which corresponds to a centralized control structure.

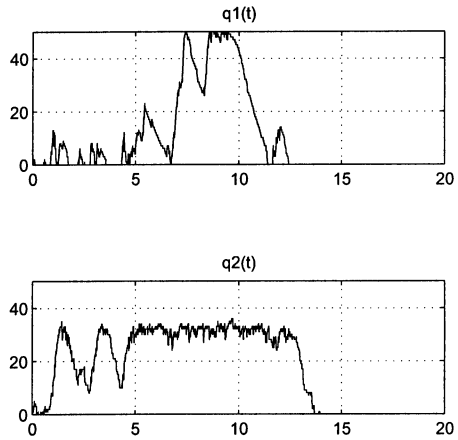


Figure 9: Simulation with control of the first router using information about the queue length q_2 as shown in Figure 8.

[4] S. Mascolo. Congestion control in high-speed communication networks using the Smith principle. *Automatica*, 35(12):1921–1935, 1999.

[5] K. Park. Future directions and open problems in performance evaluation and control of self-similar network traffic. In *Self-Similar Network Traffic and Performance Evaluation*. Wiley-Interscience, 2000.

[6] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, 1994.

[7] V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Trans. on Networking*, 3(3):226–244, 1995.

[8] L. Massoulié and J. Roberts. Bandwidth sharing: objectives and algorithms. In *Proc. IEEE INFOCOM*, pages 1395–1403, New York, NY, March 1999.

[9] Z. Jiang and L. Kleinrock. A general optimal video smoothing algorithm. In *Proc. of IEEE INFOCOM*, San Francisco, CA, 1998.

[10] S. Floyd. TCP and explicit congestion notification. *ACM Computer Communication Review*, 24(5):10–23, 1994.

[11] F. Paganini, J. Doyle, and S. Low. Scalable laws for stable network congestion control. In *IEEE CDC*, Orlando, FL, 2001.

[12] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. *Journal of Operational Research Society*, 49:237–252, 1998.

[13] K. J. Åström and B. Wittenmark. *Computer-Controlled Systems*. Prentice Hall, Englewood Cliffs, N.J., 3rd edition, 1997.

PBAC: Probe–Based Admission Control

Ignacio Más[1] and Gunnar Karlsson[2]

IMIT/KTH, S-164 40 Kista (Sweden).

[1]E–Mail: nacho@imit.kth.se

[2]E–Mail: gk@imit.kth.se

Abstract

Different end–to–end measurement based admission control (MBAC) schemes have recently been proposed to support quality of service for real–time data. All these designs share the idea of decentralizing the admission decision by requiring each end host or gateway to probe the network before sending data. This probing provides certain measurements on the network status that can be used to accept or reject the incoming flow. In this article a probing procedure to perform an admission decision for a controlled load service (CLS) is proposed. The admission control offers a reliable upper bound on the packet loss for the new session even with short probing phase durations (e.g. half a second). A simple explanation of the procedure is given, as well as some simulation results, both for unicast and multicast cases.

1 Introduction

Today's real–time communication on the Internet requires a better and more predictable service quality than what is possible with the available best–effort service. The IETF has proposed two different approaches to provide quality of service (QoS) guarantees: Integrated Services (IntServ) [1] and Differentiated Services (DiffServ) [2]. Both IETF schemes have different advantages and disadvantages. Some recent efforts have focused into a combination of both ideas, trying to provide the superior scalability of DiffServ with the strict quality of services assurances of IntServ.

In recent years, new proposals have appeared to provide quality of service. All these proposals share the common idea of endpoint admission control: a host send probe packets before starting a new session and decides about the session admission based on statistics of probe packet loss [3, 4], explicit congestion notification (ECN) marks [5, 6], delay or delay variation [7]. These schemes aim at providing a service similar to the controlled–load service in the IntServ architecture [8], as they ensure a probabilistic upper bound on the packet loss and delay by limiting the system load, but they do not provide strict or per–flow performance guarantees. A detailed comparison on the different endpoint admission control proposals is given in [9].

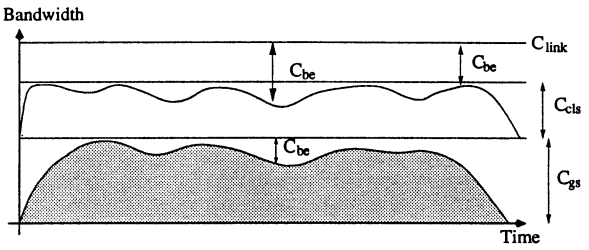


Figure 1: The separation of the service classes.

We propose a probe–based admission control (PBAC) scheme fitting into a controlled–load service class, which is part of a reduced service–set architecture currently under development in our group [3, 10, 11]. This architecture includes three different service classes: a guaranteed service with no packet loss and deterministic bounds on the delay, a controlled–load service with an upper bound on packet loss and low delay, and a best–effort service without QoS guarantees (see Fig. 1). Both the guaranteed service and the controlled–load service are allocated fixed parts of the link capacity. The parts taken together should be less than the total capacity to avoid starvation of the best–effort service. The best–effort traffic uses the capacity that is not allocated to the two other service classes, as well as the allocated but temporarily free capacity. This architecture only requires that the routers differentiate among three different classes, and have three service class queues (see Fig. 2). The low delay for the guaranteed and the controlled–load service is ensured by using small, packet–scale buffers in the routers. The CLS service uses the PBAC scheme to avoid network overload and thus to limit the packet loss probability. The purpose of the probing scheme is to prevent new sessions from degrading the QoS of ongoing sessions below some pre–established level. PBAC, thus, prevents congestion from occurring, rather than

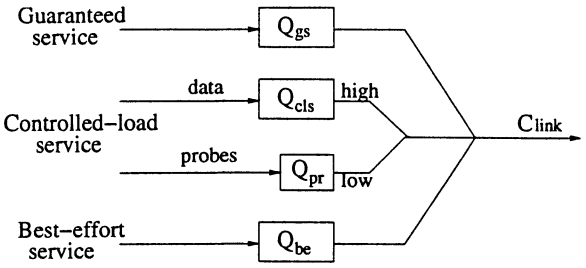


Figure 2: The queueing system.

resolving it after it happens.

2 Procedural description

The phases of the session establishment in the PBAC scheme are as follows: when a host wishes to set up a new flow, it starts by sending a constant bit rate probe at the rate the data session will maximally require. The length of the probe and the packet size is set to allow loss measurement with enough accuracy. The receiving host counts the number of received and lost packets. When the probing period finishes it compares the estimated probe loss with a target value and sends back a packet accepting or rejecting the incoming session. Finally, when the sending host receives the acceptance decision, it starts the session, or, in the case of a rejection, backs off for a certain amount of time, before starting to probe again.

The target value for the acceptable packet loss probability is the same for all sessions, otherwise the highest value would determine the network load and the packet loss probability for all the accepted sessions. The PBAC scheme requires a high degree of statistical multiplexing at the routers, thus a limitation on the maximum allowed data rate is needed (no more than 1% of the link capacity) in those links where multiplexing is used for dimensioning. This limitation does not apply to the local loop or over dimensioned access links where there is no statistical multiplexing. The procedure relies on a common trust between the hosts and the network. The particular means to ensure the proper host behavior is based on public/private key cryptography.

In order to avoid disturbing ongoing data sessions, probe packets are sent in a strict lower priority than data packets. Two different buffering strategies are possible: in the first one the router has two transmission queues, one with high priority for the traffic of accepted sessions and one with low priority for probes, while in the second approach, one queue is used with a discard threshold for the probes.

2.1 Multicast support

Our quality of service scheme for the multicast traffic is based on admission control for both multicast senders and receivers. There are two multicast groups: one for probe packets and one for the session itself. All that the probing procedure assumes is a sender based multicast routing protocol with a root node (*rendez–vous* point). The probe from the sender is continuously sent to the root node, and it will forwarded along the multicast tree whenever receivers have joined the probe group.

The receiver first joins the probe group and measures the probe packet loss for a certain period of time. It chooses the amount of time that it will measure the probe packet loss, as a longer admission period gives a higher accuracy of the probe packet loss and a lower blocking probability. Once the receiver has performed the estimation of the packet loss probability and compared it with the target loss, it leaves the probe group. If the admission decision is positive, the receiver immediately joins the data group, while in the case of a negative decision, it needs to back off for a period of time before trying to join again. It is important to note that there are probe packets forwarded on the multicast probe tree only when there are receivers joined to the group.

All receivers must be able to recognize special 'increase of rate' packets in the data group. These packets are sent by the multicast senders when the peak rate of the group needs to be increased. When a host receives one of these packets, it should immediately join the probe group, because the peak rate of the data group is increased and a new admission decision is therefore necessary.

The admission procedure for multicast senders is performed at the root node. If we have multiple senders that share the same reservation and do not transmit simultaneously, then a new sender only needs to perform a unicast probe to the root node. In case an increase of the peak rate is necessary, the sender notifies all the receivers by special data packet ('increase of rate' packets), as mentioned above.

3 Performance evaluation

The performance of PBAC, both in the unicast and the multicast cases has been evaluated by simulation. Different figures have been obtained for blocking and packet loss probabilities.

3.1 Simulation methodology

All the simulations performed have been carried out with the network simulator NS–2[1]. The different traf-

[1] http://www-mash.cs.berkeley.edu/ns/

fic sources used for all the simulations are summarized in Table 1.

3.2 Evaluation of the queueing scheme

In order to understand the effect that the two proposed queueing schemes would have for our admission control, we performed a set of tests with both queues and compared the results. The buffer sizes of the double queues were two packets for the probe queue and eight packets for the data queue. The threshold queue had eight packets of buffer space in total, with a threshold value of two packets. Figure 3 shows the results for the Par1 type of sources (See Table 1).

All the figures obtained show that there is similar behavior in both probe loss curves, the main difference being the magnitude of the loss that we get for the probing. As expected, the probe loss in the case of the threshold queue is a little bit higher than the one we would achieve for the double queue system, since the two buffer positions below the threshold are shared between high and low priority packets. The loss levels are however always in the same order of magnitude for all load levels we have tested. We conclude that the design decision of choosing one queueing system or the other does not affect the feasibility of our admission control, and both queueing systems could be used together in a network, as long as we take into account the increased loss for the threshold.

3.3 Evaluation of the admission decision

To perform the acceptance decision, we define a target loss probability p_{pr} and measure the empirical loss probability p_{me} from the probe. The accuracy of the measurement, as we have previously said, depends on the number of probe packets, so we should use the smallest packet size possible. The admission decision rule assumes a normal distribution of the probe loss ratio in order to calculate a confidence level on the estimated session loss probability. A session is acceptable if: $p_{me} + z_R\sqrt{\frac{p_{me}\times(1-p_{me})}{s}} \leq p_{tg}$, given that $p_{tg} \times s > 10$ [10]. In this, s is the number of probe packets, R is the confidence level we want to have and z_R is the $1-(1-R/2)$–quantile of the normal distribution. The second condition ensures that we have sufficient number of samples for the estimation.

With the performance evaluation of the admission control, we are trying to show the existence of a general relationship between probe and session loss probabilities. We have performed different simulations and compared the different probe/session loss figures obtained. The results are plotted in Fig. 4. This figure has been obtained for Type1 scenarios with 70 active sources of the different types over 2 Mb/s links. It shows that there is a gap between the expected session loss and the probe loss that depends on the type of sources being used. In general there is one order of magnitude higher probe loss than session loss. Figure 5 contains the results obtained by changing the source types among the first set of sources (Type1) over a 2 Mb/s link. From these figures we see that the actual probe to session loss ratio does not depend heavily on the type of sources, and that there is a nearly linear relationship between the loss types.

3.4 Blocking probability for multicast

One of the most important reference values for the performance analysis of the admission control scheme is the blocking probability for new receivers. It depends primarily on the network load, the target packet loss probability, and the length of the probe period. From these three variables, the multicast receiver can only modify the last one, as the network load is given and the target loss probability is fixed for the controlled load service.

Several simulations have been performed for the different source types in Table 1, with probe durations of up to ten seconds. The blocking probability for each scenario was obtained for more than two thousand session establishment attempts, while varying the network load.

Figure 6 shows the different values for the blocking probability versus the normalized link load for different values of the target probability. The sources used for this figure were Par1.

An important aspect of the admission control is that the blocking probability depends on the load on the different links of the multicast tree, so that different receivers will get different blocking probabilities. To show this, we created a more complex topology in which an intermediate router is receiving some extra background load (an additional 10 % of the link capacity). The different link loads for the first set of receivers and the second set of receivers makes the blocking probability of the latter much higher. The values can be seen in Table 2, where 'First receivers' are the ones that do not suffer the increased load, and 'Second receivers' are the ones behind the second bottleneck link, which carries more background load. The difference in blocking probability is greater in the case where we want to have a lower packet loss probability. The figures in the table have been obtained for Par1 type of sources with a probe length of five seconds.

4 Conclusions

We have presented an admission control method based on probing that provides a well–defined upper limit

Table 1: Parameters of the different test sources

Source	Type	Average On Time	Average Off Time	Peak Rate
Exp1	Exponential	325 ms	650 ms	64 kb/s
Par1	Pareto (β=1.5)	325 ms	650 ms	64 kb/s
Mixed1	Both	325 ms	650 ms	64 kb/s
Exp2	Exponential	20 ms	35.5 ms	128 kb/s
Par2	Pareto (β=1.5)	20 ms	35.5 ms	128 kb/s
Mixed2	Both	20 ms	35.5 ms	128 kb/s
Video	Traces	-	-	360 kb/s (64 kb/s avg.)

on the packet loss probability for a flow. We have offered simulation results that show a clear relationship between the probe packet loss and the expected session loss. The admission control scheme works equally well in the unicast and multicast cases. The packet loss probability for new multicast receivers is bounded and depends on the network load and the target loss probability. These two factors influence the blocking probability for new senders and receivers. By increasing the setup delay of new multicast receivers, the blocking probability can be reduced and a smaller target loss probability can be enforced. To summarize, we offer an admission control scheme that is simple and offers reliable performance under different network conditions.

References

[1] R. Braden, S. Clark, and S. Shenker, "Integrated services in the internet architecture," RFC 1633, IETF, June 1994.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W Weiss, "An architecture for differentiated services," RFC 2475, IETF, December 1998.

[3] G. Karlsson, "Providing quality for internet video services," in *Proc. of CNIT/IEEE IT-WoDC 98*, Ischia, Italy, September 1998, pp. 133–146.

[4] V. Elek, G. Karlsson, and R. Rönngren, "Admission control based on end-to-end measurements," In *Proc. of IEEE INFOCOM 2000* [12], pp. 623–630.

[5] F. P. Kelly, P. B. Key, and S. Zachary, "Distributed admission control," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 12, pp. 2617–2628, 2000.

[6] T. Kelly, "An ECN probe–based conection acceptance control," *ACM Computer Communication Review*, vol. 31, no. 3, pp. 14–25, July 2001.

[7] G. Bianchi, A. Capone, and C. Petrioli, "Packet management techniques for measurement based end-to-end admission control in IP networks," *Journal of Communications and Networks*, vol. 2, no. 2, pp. 147–156, June 2000.

[8] J. Wroclawski, "Specification of the controlled-load network element service," RFC 2211, IETF, September 1997.

[9] L. Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang, "Endpoint admission control: Architectural issues and performance," in *Proc. of ACM SIGCOMM 2000*, Stockholm, Sweden, August/September 2000, vol. 30, pp. 57–69.

[10] I. Más Ivars and G. Karlsson, "PBAC: Probe–based admission control," in *Proc. of QofIS 2001*, Coimbra, Portugal, September 2001, Springer, vol. 2156 of *LNCS*, pp. 97–109.

[11] E. Ossipov and G. Karlsson, "A simplified guaranteed service for the Internet," in *Proc. of PfH-SPN 2002*, Berlin, Germany, April 2002, (To be published).

[12] *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, Tel Aviv, Israel, March 2000.

Table 2: Blocking probability for different links in the multicast tree.

P_tg	First receivers		Second receivers	
	link load (%)	Block prob. (%)	link load (%)	Block prob. (%)
0.001	60	0.67	70	5.7
0.01	60	0.57	70	5.1
0.1	60	0.4	70	3.9

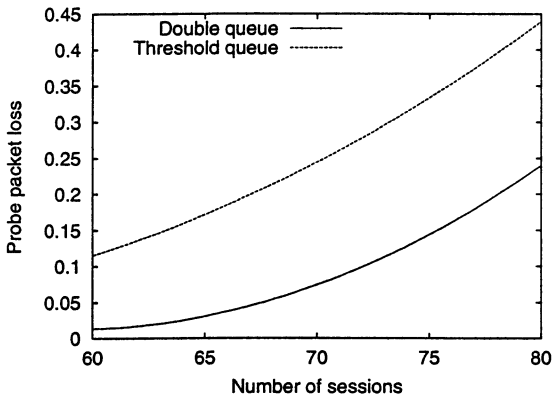


Figure 3: Probe loss with the two different queue schemes for the Par1 sources.

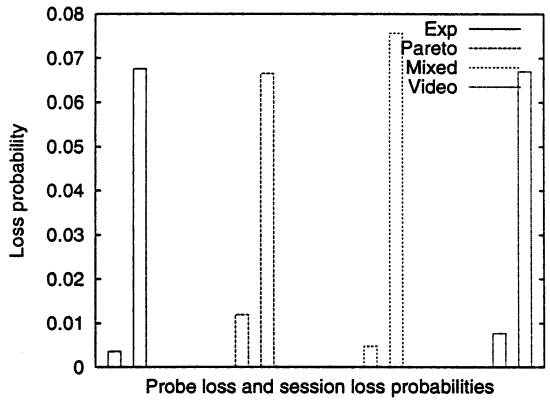


Figure 4: Probe and session loss probabilities for different scenarios.

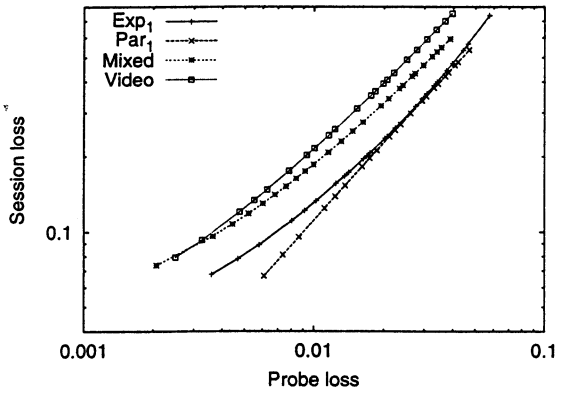


Figure 5: Sessions versus probe packet loss for the different source types in the Type1 scenarios.

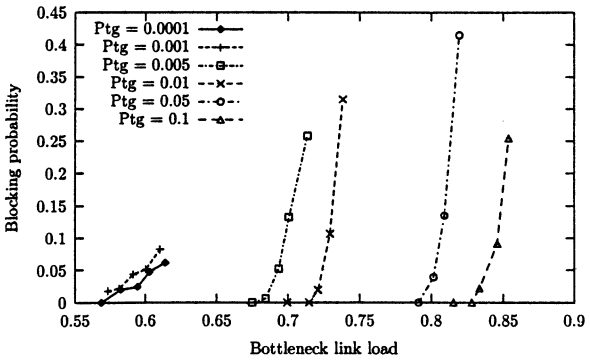


Figure 6: Blocking probability versus link load for different loss target probabilities. Probe rate=64 kbit/s, probe length= 10s. Par1 sources

Scheduling Algorithms for LET in Spatial Reuse TDMA

Jimmi Grönkvist

Adress: Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping. E-mail: jimgro@foi.se.

Abstract

Spatial reuse TDMA is a fixed assignment access scheme for multi-hop radio networks. The idea is to let several radio terminals use the same time slot when the interferences caused are not too severe. Most previous methods for reuse scheduling assign transmission rights to nodes or alternatively to links, i.e., transmitter/receiver pairs. Recently, a novel assignment method have been proposed, link assignment with extended transmission rights (LET), which have the advantages of both methods.

However, the performance of LET is dependent on the link schedule which it is based on. In this paper, the design of the algorithms that generate the link schedules is studied and we suggest how to expand a basic link assignment algorithm in order to take better advantage of this new assignment strategy so as to decrease the packet delay.

1 Introduction

We consider a radio network where a number of radio units are spread out in some terrain. If the received signal power is sufficient in relation to noise and interferences, it is assumed that any two radio units can communicate, i.e., establish a *link*. In a multihop network, the power consumption can be kept low, and area coverage is achieved by letting messages be relayed over one or several intermediate nodes. Distributed multihop radio networks are often referred to as ad hoc networks.

One problem in a radio network is the interferences caused by simultaneously transmitting nodes. These conflicts occur if the received signal is too weak compared with the interfering signals. An important issue is therefore to design access schedules that control the use of the channel. A widely used access schedule is time division multiple access (TDMA). For sparsely connected networks, however, this is usually inefficient. To increase capacity one can instead use spatial reuse TDMA (STDMA), which is an extension of TDMA where the capacity is increased by spatial reuse of the time slots. An STDMA schedule describes the transmission rights for each time slot.

Different algorithms for generating STDMA schedules have been proposed [1, 2, 3, 4, 5]. These algorithms generally are of two types; transmission rights have been scheduled either to nodes or links. In the first case, only the sender is determined in advance, in the second, both sender and receiver. Link assignment schedules can achieve higher throughput due to higher spatial reuse of the channel, but at the cost of higher delay at low traffic loads [6].

Recently a new assignment strategy has been proposed [7], LET, which achieves the higher throughput of link assignment without the cost of higher delay for low traffic loads. This strategy is based on a link schedule, but where transmissions rights are extended in the following way: Assume that a node is scheduled to transmit to a specific neighbor according to the link schedule. Now, if the node does not have a message in queue destined for this specific neighbor, the node is permitted to use the scheduled slot for transmitting a message to another neighbor. Such transmissions will not always be conflict free.

This results in as high degree of spatial reuse as link assignment, but at low traffic the same behavior as node assignment.

However, the performance of LET is dependent on the link schedule which it is based on. In this paper we study how the choice of link schedules affects the delay in the network. We see that a poor choice of link schedule causes poor performance when using LET. We also suggest how to expand a basic link assignment algorithm in order to take better advantage of the new assignment strategy so as to decrease the packet delay. With this procedure, the slots assigned to each node will be spread evenly over the period, resulting in a decreased network delay when using LET.

We use an interference based model of the network as suggested for STDMA scheduling in [8].

The network performance is evaluated using simulations in terms of end-to-end delay for different network connectivity.

2 Network Model

Here we describe the interference based model of a radio network. The network is represented by a set of nodes $\mathcal{V}$ and the basic path-loss $L_b(i,j)$ between any two distinct nodes (v_i, v_j), $i \neq j$.

In the following we will assume isotropic antennas and that all nodes use equal transmission power for simplicity.

For any two nodes, (v_i, v_j) where v_i is the *transmitting* node and $v_j \neq v_i$, we define the signal-to-noise ratio

(SNR), Γ_{ij}, as follows

$$\Gamma_{ij} = \frac{P_i}{L_b(i,j)\,N_r}, \tag{1}$$

where P_i denotes the power of the transmitting node v_i, $L_b(i,j)$ is the basic transmission path-loss between nodes v_i and v_j, and N_r is the noise power in the receiver. For convenience, we define $\Gamma_{ii} = 0$ corresponding to the physical situations of a node not being able to transmit to itself.

We say that a pair of nodes (v_i, v_j) form a *link*, (i,j), if the signal-to-noise ratio (SNR) is not less than a threshold, γ_C. That is, the set of links in the network, $\mathcal{K}$, is defined:

$$\mathcal{K} = \{(i,j) : \Gamma_{ij} \geq \gamma_C\} \ . \tag{2}$$

For a set of links, $K \subseteq \mathcal{K}$, we define the *transmitting nodes*:

$$V_T(K) = \{v_i : (i,j) \in K\} \ .$$

For any link, $(i,j) \in K$, we define the *interference* as follows

$$I_K(i,j) = \sum_{v_k \in V_T(K) \setminus v_i} \frac{P_k}{L_b(k,j)}. \tag{3}$$

Furthermore, we define the *signal-to-interference ratio* (SIR):

$$\Pi_K(i,j) = \frac{P_i}{L_b(i,j)(N_r + I_K(i,j))}. \tag{4}$$

We assume that any two radio units can communicate a packet without error if the SIR is not less than a threshold, γ_R. A schedule S is defined as the sets X_t, for $t = 1, 2, \ldots, T$, where T is the period of the schedule. The sets X_t contain the links assigned time slot t. A schedule is called conflict free if the SIR is not less than the threshold γ_R for all receiving nodes in all sets X_t.

Furthermore, we assume that a node cannot transmit more than one packet in a time slot and that a node cannot receive and transmit simultaneously in a time slot.

3 LET

We say that a link (k,l) is *adjacent* to link $(i,j) \in K$ iff $\{i,j\} \cap \{k,l\} \neq \emptyset$. Furthermore we define $\Psi(K)$ as the union of all adjacent links to the links in K. We assume that a node cannot transmit more than one packet in a time slot and that a node cannot receive and transmit simultaneously in a time slot. Alternatively, we say that a set of links K and the set of its adjacent links $\Psi(K)$ must be disjoint:

$$K \cap \Psi(K) = \emptyset. \tag{5}$$

The signal-to-interference criteria (4) gives the following condition

$$\Pi_K(i,j) \geq \gamma_1 \ \forall \ (i,j) \in K. \tag{6}$$

If the above two conditions, (5) and (6), hold for a set of links $K \in \mathcal{K}$, we say that the links in K can *transmit simultaneously*.

Notice that the interference term in (3) only depends on which nodes that are transmitting and not on the nodes that are receiving messages. Assume that a node is assigned as a transmitter in a slot, i.e., an outgoing link of the node is assigned the slot. If this node redirects the transmission to a node, other than the assigned receiving node, the inequality in (6) still holds for all links originally assigned to the slot. This means that the interference level of the other simultaneously receiving nodes will not change. (Recall the assumption of omni-directional antennas.) The redirected transmission in itself can of course not always be guaranteed to be conflict free.

LET is based on these observations and we conclude this section by giving a more formal description of the LET algorithm that we use in our simulations.

This is done in each node:

Step 1 If there is a packet in the scheduled link's queue:

- **1.1** Transmit this packet on all links that it is supposed to be transmitted on.
- **1.2** Remove the packet from all link queues where this transmission was successful.

Step 2 Otherwise if there is no such packet. Choose one of the node's other outgoing links that can be exchanged conflict-free in the schedule that has a packet in queue:

- **2.1** Transmit this packet on all links that it is supposed to be transmitted on.
- **2.2** Remove the packet from all link queues where this transmission was successful.

Step 3 Otherwise if there is no such packet. Choose one of the rest of the node's outgoing links that has a packet in queue:

- **3.1** Transmit this packet on all links that it is supposed to be transmitted on.
- **3.2** Remove the packet from all link queues where this transmission was successful.

Step 4 Otherwise, do nothing.

4 Performance Evaluation

The relaying of traffic in multihop networks causes a considerable variation of the traffic load on the links and nodes in a network. To achieve large capacities efficient traffic controlled schedules have to compensate for this problem. In a traffic controlled schedule, links or nodes can use several slots, see [4], according to the traffic load.

The algorithms we use for generating the schedules for the simulations are fully traffic compensated. The assignment algorithm is described in [9].

This basic algorithm is a greedy algorithm that assigns slots according to priority. The priority of a link is the traffic load through a link multiplied with the time since it was last assigned a time slot.

In short the algorithm works as follows, choose the link with highest priority which has not yet been checked in the time slot. Assign it to the time slot if possible. If all links with time slots left have been checked, continue to the next slot. Proceed until all links have been scheduled their guaranteed time slots.

In our evaluations we use the *average packet delay* as a performance measure. This is a commonly used performance measure for multi-hop networks. *Packet delay* is the time, in time slots, from the arrival of a packet at the buffer of the origin node v_k to the arrival of the packet to the destination node v_l. Due to the relaying of packets the statistical properties of this random variable is complicated and an exact analytical analysis of the expected value is difficult [10]. In our analysis we have chosen computer simulations as evaluation method.

In our traffic model we assume both point-to-point traffic, i.e. a packet entering the network has only one destination, and broadcast traffic, i.e. a packet entering the network has all other nodes as destination. We assume a uniform traffic model.

5 Problems with LET

Although any link-assigned schedule may be used to extend transmission rights, some link schedules may give LET more or less desirable properties. Here we discuss what effect the link schedule has on delay.

At low traffic load, the nodes will normally only have at most one packet in queue at a time. In this case, LET behaves as node assignment, with the node-assigned schedule as the transmitting nodes in the link-assigned schedule. A problem when generating STDMA schedules is how to determine which slots to give a node or link that is going to receive more than one slot, since delay through the node depends on which specific time slots the node is given.

For example, assume that a node has received two time slots and these are spread in such a way that the distance between them is approximately half the frame length. Then for low traffic loads, the delay will be at most half the frame length. If the node is given two consecutive time slots, the maximum delay might be the entire frame length. That is, it is usually efficient to spread a node's time slots evenly over the frame. This problem gets worse if nodes receive many time slots, especially since a large part of the traffic usually flows through these nodes.

In some algorithms, see e.g. [9], an effort is made to spread the time slots a node or link is given equally over the frame. However, even if the link schedule has perfectly spread time slots, this may not be the case for the transmitting nodes. Therefore, LET might give considerably higher delay than a node-assigned schedule if the link schedule tends to give the transmitting nodes consecutive time slots.

Any assignment algorithm, especially if it is of a greedy type, has a set of rules to determine which link to assign to a time slot. One method used is node ID, i.e. the lower the ID number, the higher the priority. To assign links, the pair of node IDs of the transmitting and receiving nodes can be used. A sorted list according to link priority would then give the outgoing links from a node consecutive places in the list. Even if another system for link priority is used, node ID is eventually used if priority is equal for several links.

Now, assume we have a fully connected network, i.e. all nodes can communicate with all other nodes without relaying. Furthermore, in this case we can assume that there is no spatial reuse, although in a link schedule using an interference-based model, this might be possible due to capture. In this case the assigned schedule would be the sorted list described above, which gives high network delay.

One way to avoid this problem is to give the links a link ID which is random, although different, for each link. This gives a fairly efficient solution.

However, this still does not spread the transmitting nodes in the link-assigned schedule evenly over the frame.

In the next section we will describe an algorithm that attempts to do that.

6 Double Priority

We will expand the basic algorithm described so that it uses double priority. The links are here assigned time slots according to two priority lists. The primary priority of a link is the number of time slots passed since the transmitting node was previously assigned a slot multiplied by the relative traffic of the transmitting node. The secondary priority of a link is the number of time slots passed since the link was previously assigned a slot multiplied by the relative traffic of the link. When we determine which link to test next, we first check which link has the highest primary priority. If two or more links have equal priority, the secondary priority is used to determine which of these to test first.

With this procedure, the slots assigned to each node will be spread evenly over the period, resulting in a decreased network delay when using LET.

In the following, we describe the traffic sensitive algorithm obtained with the above ideas.

Step 1 Initialize:

- **1.1** Enumerate the links
- **1.2** Create a list, A, containing all of the links and an empty list B.

1.3 Set t to zero.

1.4 Calculate the number of slots each link is to be guaranteed and set h_{ij} to Λ_{ij}.

1.5 Set τ_i^1 to zero for all nodes.

1.6 Set τ_{ij}^2 to zero for all links.

Step 2 Repeat until list A is empty:

Step 2.1 Set $t \leftarrow t + 1$ and $V_t \leftarrow \emptyset$.

Step 2.2 For each link (i, j) in list A:

2.2.1 Set $L_t \leftarrow L_t \cup (i, j)$.

2.2.2 If the set of links in L_t can *transmit simultaneously*:

- If $h_{ij} = 1$, remove the link from list A and add to list B.
- Set $h_{ij} \leftarrow h_{ij} - 1$, and set τ_i^1 and τ_{ij}^2 to zero.

2.2.3 If the set of links in L_t cannot *transmit simultaneously*: set

$$L_t \leftarrow L_t \setminus (i, j).$$

Set $\tau_{ij}^2 \leftarrow \tau_{ij}^2 + 1$.

Step 2.3 For each link (i, j) in list B but not in L_t:

2.3.1 Set $L_t \leftarrow L_t \cup (i, j)$.

2.3.2 If the links in L_t can *transmit simultaneously*, set τ_i^1 and τ_{ij}^2 to zero.

2.3.3 If the links in L_t cannot *transmit simultaneously*, set

$$L_t \leftarrow L_t \setminus (i, j)$$

and set $\tau_{ij}^2 \leftarrow \tau_{ij}^2 + 1$.

Step 2.5 Set $\tau_i^1 \leftarrow \tau_i^1 + 1$ for nodes not in $V_T(L_t)$.

Step 2.6 Reorder lists A and B according to priorities,

$$\tau_i^1 \Lambda_i \text{ and } \tau_{ij}^2 \Lambda_{ij}$$

with the highest priority first.

7 Evaluation Method

To generate realistic link behaviors, a terrain-data based ground wave propagation model, Vogler's five knife-edge model, has been used for the calculation of the basic transmission path-loss, see [11] for more details.

For each of the generated networks average delay have been determined by using simulations. In the simulations, these additional assumptions have been made:

Shortest route, i.e., packets sent between two nodes will always use the path which requires the least number of transmissions. If several routes of the same length exists, all packets between two specific nodes will always use the same route. When estimating packet delay, packets are generated by a pseudo Poisson process with intensity λ and with a uniform traffic distribution.

Furthermore, a perfect feedback channel with no delay is assumed and that the link ID is random.

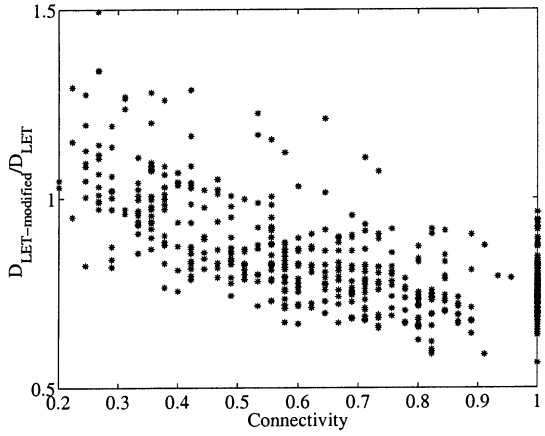


Figure 1: The figure shows the ratio between network delay for LET and network delay for the modified LET with unicast traffic. The ratio is plotted for 500 networks of size 20 nodes.

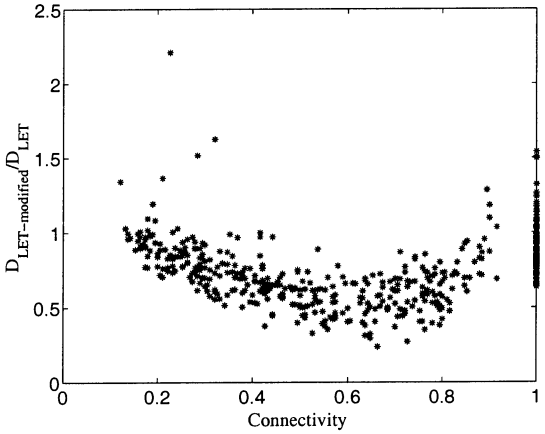


Figure 2: The figure shows the ratio between delay for the modified LET and delay for LET with broadcast traffic. The ratio is plotted for 500 networks of size 20 nodes.

8 Simulation Results

In figure 1, we plot the ratio between network delay for the modified LET and network delay for ordinary LET. We see in this figure that for most of the networks the modification improves the performance of LET. This is especially the case for high connectivity. For low connectivity it is less efficient, probably due to rather inefficient distribution of time slots in these cases.

Since the double priority method for LET was developed for unicast traffic, it is interesting to see how it behaves for broadcast traffic.

In figure 2 the ratio is plotted for networks with broadcast traffic of size 20 nodes. We can see that for most networks the modification gives a considerable improvement, especially for medium connectivity networks. We can conclude that the modification works well also for broadcast traffic, even better than for unicast traffic.

9 Conclusions and Comments

In this paper we have briefly introduced how network performance is affected by the link-assignment algorithm we base LET on. The use of double priority gives very efficient schedules for a fully connected network. However, for low connectivity it is a less efficient way of spreading time slots than the single priority method used in previous chapters. More research is necessary if we want to fully exploit the advantages of LET.

The double priority method is a good start, though. For most networks it gives an improvement compared to the single priority method in terms of delay both for unicast traffic and broadcast traffic.

References

[1] A. Ephremides and T. Truong, "Scheduling broadcasts in multihop radio networks," *IEEE Trans. Commun.*, vol. 38, no. 4, pp. 456–460, Apr. 1990.

[2] I. Chlamtac and S. Kutten, "A spatial reuse TDMA/FDMA for mobile multi-hop radio networks," in *Proceedings of IEEE INFOCOM '85*, 1985, vol. 1, pp. 389–394.

[3] I. Chlamtac and A. Lerner, "A link allocation protocol for mobile multi-hop radio networks," in *GLOBECOM '85, IEEE Global Telecommunications Conference, Conference Record*, 1985, vol. 1, pp. 238–242.

[4] J. Shor and T. Robertazzi, "Traffic sensitive algorithms and performance measures for the generation of self-organizing radio network schedules," *IEEE Trans. Commun.*, vol. 41, no. 1, pp. 16–21, Jan. 1993.

[5] S. Ramanathan and E. Lloyd, "Scheduling algorithms for multihop radio networks," *IEEE/ACM Trans. Networking.*, vol. 1, no. 2, pp. 166–177, Apr. 1993.

[6] J. Grönkvist, "Assignment methods for spatial reuse TDMA," in *MobiHOC 2000*, 00.

[7] J. Grönkvist, "Node and link assignment in spatial reuse schemes," in *Nordic Radio Symposium*, 2001.

[8] O. Somarriba, *Multihop Packet Radio Systems in Rough Terrain*, Tech.lic. thesis, Radio Communication Systems, Department of S3, Royal Institute of Technology, SE-100 44 Stockholm, Sweden, Oct. 1995.

[9] J. Grönkvist, "Traffic controlled spatial reuse TDMA for multihop radio networks," in *Personal, Indoor and Mobile Radio Communications*, 1998, pp. 1203–1207.

[10] R. Nelson and L. Kleinrock, "Spatial-TDMA: A collision-free multihop channel access protocol," *IEEE Trans. Commun.*, vol. 33, no. 9, pp. 934–944, Sept. 1985.

[11] B. Asp, G. Eriksson, and P. Holm, "Detvag-90® — Final Report," Scientific Report FOA-R–97-00566-504–SE, Defence Research Est., Div. of Command and Control Warfare Tech. Linköping, Sweden, Sept. 1997.

The design and application of an IP rate reservation scheme

Evgueni Ossipov[1], Gunnar Karlsson[2]

[1,2] Kungliga Tekniska Högskola, KTH/IMIT, Sweden.
[1]E-mail: eosipov@imit.kth.se.
[2]E-mail: gk@imit.kth.se.

Abstract

The growth of the Internet that we have witnessed in the last years has made the network an essential part of everyday life. The Internet of the 21st century will provide transport not only for data communication but also of voice and multimedia. However, the success of the IP technology has also led to two major problems facing the Internet today: Lack of sufficiently fast switching equipment and of support for real-time services. The first problem comes from an increasing demand of communication capacity. The second comes from the fact that the Internet was not designed to transmit real-time data. Our research concerns the architectural design for quality of service (QoS) provisioning, and the development of fast routers that are able to handle the increasing traffic volumes of the Internet.

1 Introduction

The major trend in the Internet today is to provide different levels of service for data traffic and real-time traffic. The ideas of quality of service differentiation in the Internet have been under intensive research during the last decade. The integrated services (intserv) [1] and differentiated services (diffserv) [2] architectures are the two most known approaches in this area. The intserv architecture classifies network traffic into three classes: *guaranteed service*, *controlled load service* and *best effort*. In diffserv, the traffic is assigned to specific behavior aggregates.

For traffic with guarantees, intserv provides reservation of bandwidth and buffers by using signaling between network nodes. An example of such signaling is RSVP [3]. Intserv keeps per-flow soft state in each network node. In comparison, diffserv avoids per-flow states in the routers, and ingress nodes perform traffic metering and admission control on the flows. The services offered to customers are statically described by service level agreements.

From the point of view of performance and scalability, intserv appeared to be a cumbersome architecture for large IP networks. The performance is basically limited by a router's ability to process signaling messages and maintain the set of per-connection states. Diffserv, is however free of these limitations since a router serves all flows of the same per-hop behavior together. On the other hand, it cannot regulate the resource provisioning in a scalable manner and the quality levels achieved by the service differentiation are unclear. The limitations of both architectures stated above, have motivated our work on combining the principles of the two approaches in our service architecture. The work presented in this article develops the ideas of a QoS architecture proposed by Karlsson and Orava [4].

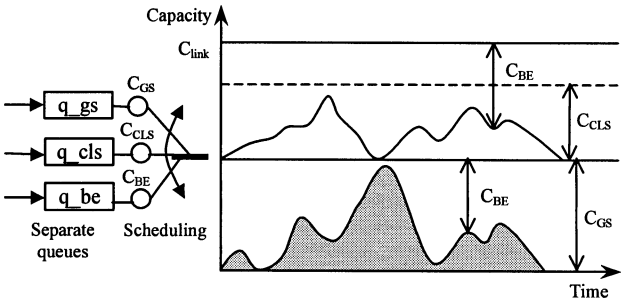


Figure 1: The service architecture

We consider a service architecture consisting of three quality classes, as illustrated in Fig. 1. The *guaranteed service* (GS) class provides deterministic guarantees, namely guaranteed throughput, absence of packet loss due to queue overflows in routers, and tightly bounded delay.

The end-to-end guarantees are provided by means of explicit resource reservation, using a simple signaling protocol and non-work-conserving scheduling in the core routers, proposed in [5]. Flows of the *controlled load service* (CLS) class have bounded packet loss and limited delay. The third class of service is the customary *best effort* (BE). The flows of this class have no guarantees, they obtain leftover capacity in the routers, and may be lost due to congestion

In this article we describe the basic aspects of our work on the GS class and its application for the resource provisioning in the backbone network. More detailed description of our GS architecture is provided in [6]. The work in our group that is related to CLS is described in [7], [8].

The remainder of the document is organized as follows. Section 2 describes our simplified signaling. In Section 3 the description of the scheduling is presented. Section 4 describes the application of our service architecture for resource provisioning in the backbone network.

2 Signaling protocol

We have the following design goals for our signaling protocol: It should not require per-connection soft-states in the routers; the number of signaling messages per connection should be low, and the time to process a signaling message should be small. The protocol should be robust and tolerate loss of certain types of messages.

Our signaling protocol assumes the reservation state to be purely additive: The reserved rate of an outgoing link of a router equals the sum of the incoming rates for the link. Each GS connection is described solely by a peak rate as traffic descriptor.

The following messages are needed for the signaling:

- Reservation message *mR(r)* for the amount of capacity *r*. Messages of this kind may be discarded by the routers.
- Confirmation message *mC(r)* from the initiating party for the reserved capacity. Messages of this type may not be lost.
- Tear down message *mT(r)*, which indicates that the initiating party finishes the connection and releases the reserved capacity. Messages of this type may not be lost.

The message *mR(r)* is carried in the available capacity from which it reserves a share; *mC(r)* and *mT(r)* use the established reservation.

In order to establish a guaranteed service connection, the initiator sends *mR(r)* towards the destination. After issuing this message a waiting timer is started with a value *T*. If no acknowledgment arrives during *T* seconds, the initiator sends a new reservation request. The diagram in Fig. 2 represents the behaviors of the sender and the receiver.

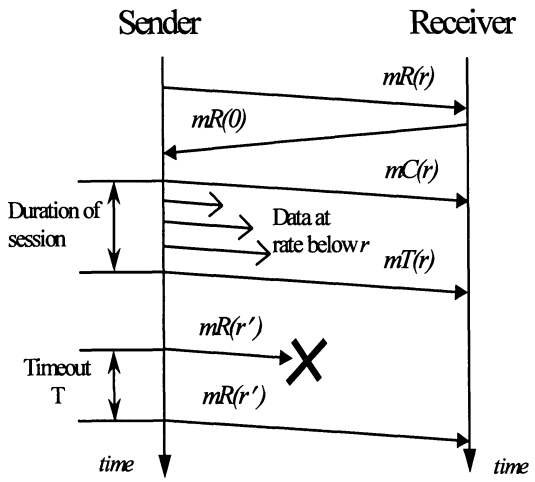


Figure 2: Setup and tear down of reservations

In all routers, we keep track of available capacity for each network link in a variable C_{GS}, and have two public state variables for *requested* and *confirmed* capacity, R_r and R_c. All users are able to modify these variables by means of the signaling protocol. To make a reservation the router first performs admission control, comparing the requested rate (r) in the signaling message with available capacity on the link. If the requested rate exceeds the available capacity ($r > C_{GS} - R_r$) the router blocks this reservation. When a reservation is admitted the router adds the reserved rate to the requested capacity variable. This permits the

requestor to send data with a rate below r. The router does the reverse to release capacity when the end system terminates the connection by an explicit tear down message.

The reservation message is discarded when a router blocks a reservation request. This means that there will be pending reservations upstream along the path of the message. Thus, we need some garbage collection process to re-establish the reservation levels in the routers that the message has successfully passed.

Since a message of the type *mC()* will not be lost, all successfully acknowledged reservations are expected to be confirmed. The garbage collection starts when all GS capacity is reserved $R_r=C_{GS}$. The router will therefore expect a confirmation for the reserved capacity within a waiting time $W>T$. If some of the requested capacity is not confirmed after W, the output port will free the capacity R_r-R_c.

3 Scheduling

In order to satisfy the properties of our service model, a router needs a special scheduling algorithm to enforce additivity of the reservation state and to ensure absence of packet loss for the GS traffic in the network. It should work with variable length packets up to some maximum size L. The algorithm should not prevent best effort traffic from being sent for long periods. The buffer space for the GS traffic in the routers should be enough to avoid packet loss, taking into account the worst permissible arrival pattern.

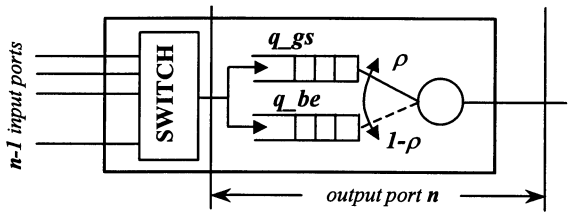


Figure 3: An output port of a router

Consider a router in the network as shown in Fig. 3. The router has n ports. Traffic for output n enters the router through $n-1$ input ports. After the route lookup, packets come to the switching fabric and are switched to the appropriate output ports. At an output port, all incoming packets are sorted according to the service class and placed into the corresponding queue: one for guaranteed service (q_gs) and one for best effort service (q_be).

The scheduler will serve one or more packets from the GS queue (a *GSBurst*), then it will schedule an idle period long enough to preserve the outgoing reserved rate of GS traffic. During the idle it will serve packets from the best effort queue. Note that proposed scheduling scheme is non-work-conserving only with respect to GS packets. Best effort traffic gets all the capacity that is not used by GS. The schematic in Fig. 4 shows this scheduling. Thus we define the reservation ratio of an outgoing link as

$$\rho=\frac{GSBurst}{Idle+GSBurst}. \quad (1)$$

where *GSBurst* is the smallest amount of GS traffic so that ρ is maintained and the following idle period is sufficiently long to transmit one BE packet of maximum size.

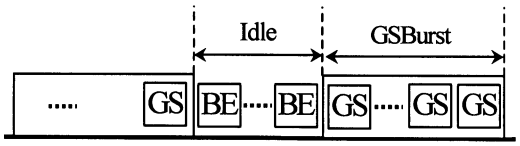


Figure 4: Scheduling

4 Performance evaluation

We have implemented the signaling protocol and our scheduling in the network simulator ns-2 [9]. For the signaling our goal was to observe the performance of our protocol in a real router; we therefore measured the actual time the processor spends on processing every signaling message and not the local ns-2 time, since it is independent of the computer power. In a series of simulations on a processor Intel Pentium III 500 MHz we obtain the following results. The time of processing one reservation message is 35 microseconds, a tear down message takes 32 microseconds, and a confirmation message 17 microseconds. Although the operations of reservation and tear down are similar, recall

that the router performs admission control for reservations. This explains the slightly different processing times for these types of messages. The average rate of message processing is therefore 30×10^6 messages per second.

Our simulation results for the scheduling show that our scheduling scheme provides delay bounds acceptable for real time applications. The delay bound depends on number of input ports on a router and not on number of hops traversed by a flow. Moreover, in our architecture the BE traffic gets certain guarantees to be transmitted even in case of high arrival rate of GS packets [6]. The developed router architecture does not allow clumping of GS bursts, thus making it easier to dimension the buffers needed in a router.

4 Application

Our simple GS service can be used to provide QoS for VPN, VoIP or other kinds of traffic with firm requirements on quality of service. To generalize the ideas of our service architecture for the whole Internet, we develop the following hierarchical structure. We separate the set of services which will exist inside a domain from the set in the backbone network. Inside an autonomous system (AS) the traffic will be classified into our three classes, while the backbone routers will see the aggregated traffic from different autonomous systems as belonging to one single service class, the GS class. In other words GS tunnels will be established across the backbone network between different autonomous systems. The backbone router of an AS will reserve the capacity on the path to other AS using our signaling protocol. Then the router will divide the reserved capacity into three classes and offer them to the users inside its autonomous system. These modifications in the Internet architecture will allow dynamic capacity allocation on the links between different Internet service providers, which in turn implies easier management of the resources in the backbone network. From the economical point of view this architecture provides a possibility to have flexible and simple charging for Internet service based on the rate reserved between different providers and between a provider and the end user.

References

[1] R. Braden, D. Clark, S. Shenker, "Integrated services in the Internet architecture: an overview", RFC1633, IETF, June 1994.

[2] S. Blake, D.Black, M.Carlson, E.Davies, Z. Wang, W. Weiss, "An architecture for differentiated services", RFC 2475, IETF, December 1998.

[3] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource reservation protocol (RSVP)", RFC 2205, IETF, September 1997.

[4] G. Karlsson, F. Orava, "The DIY Approach to QoS", *in Proc. Seventh International Workshop on Quality of Service (IWQoS 1999)*, 1999, pp. 6-8.

[5] M. Mowbray, G. Karlsson, and T. Köhler, "Capacity reservation for multimedia traffics", *Distributed Systems Engineering*, Vol. 5, 1998, pp. 12-18.

[6] E.Ossipov, G.Karlsson, "A simplified guaranteed service for the Internet", *in Proc. of Seventh International Workshop on Protocols for High Speed Networks (PfHSN 2002)*, 2002.

[7] V.Elek, G.Karlsson, R. Rönngren, "Admission control based on end-to-end measurements", *in Proc. of INFOCOM 2000*, pp. 623-630.

[8] I. Mas Ivars, G. Karlsson, "PBAC: probe based admission control", *In Proc.of Second International Workshop on Quality of Future Internet Services (QoFIS 2001)*, pp. 97-109, 2001.

[9] Network Simulator ns-2, http://www.isi.edu/nsnam/ns/

The Effectiveness of Traffic Shaping in Networks with Small Buffers

György Dán and Viktória Fodor[1]

[1]KTH/IMIT, Electrum 229,16440 Kista,Sweden. E-mail: {gyuri,viktoria}@it.kth.se.

Abstract

In this work the problem of shaping live video streams is considered. We propose low complexity shaping algorithms adequate for real-time operation and supporting applications with a wide range of delay tolerance. The effect of shaping is investigated considering video streams multiplexed at an output link with a small buffer to absorb packet scale congestion. The advantage of using small buffers when transmitting video streams is the limited delay and delay variation. Consequently, we concentrate on the loss characteristics to evaluate the performance of the proposed solutions. We present mathematical analysis based on fluid flow modeling and the theory of large deviations and confirm the results with simulation.

1 Introduction

The transmission of live video traffic over the Internet is a fundamental problem of network design, since many video applications require limited end to end packet loss, delay and delay variation. It is generally accepted that traffic control functions must be employed to guarantee these service requirements at a reasonably high network load. The introduction of new control functions in the Internet, however, is a critical issue. First, a very high number of networking devices has to be updated or replaced, second, the complexity of the control functions may limit the span and the transmission capacity of the network.

Recently research efforts address this question by proposing probe based endpoint admission control (PBAC) solutions [4, 2] to provide quality of service guarantees in a way that the functionalities of the routers are kept simple and traffic control functions are placed into the hosts – or edge gateways – only. In the PBAC schemes a host, before transmitting traffic with QoS requirements, probes the network's transmission capability by sending a sequence of probe packets and decides about the transmission based on the statistical quantities of the probing process.

While the admission control ensures that the load of the network stays reasonably bounded, additional control functions can be applied at the hosts to increase the acceptable load, like traffic shaping to decrease the burstiness of the traffic streams and thus decrease the packet loss at the multiplexing nodes and forward-error correction to recover from packet losses.

In this work we investigate how source shaping can increase the efficiency of live video transmission considering MPEG video streams multiplexed at an output link with small buffer. The advantage of using small buffers is that the delay and delay variation is strictly limited by the buffer size an thus only the packet loss has to be controlled. Source shaping provides the following favorable properties: i) it does not require global decision or any modification in the network, ii) can improve the service quality of streams with different QoS requirements and traffic characteristics and iii) can be introduced in the network gradually.

We propose two low complexity solutions to shape live video streams without packet loss and with given delay bound. We evaluate the performance of the proposed schemes with simulation and with analytical methods based on fluid flow approximation applying the theory of large deviations [1].

2 Related work

In this section we survey previous research results that inspired our work on shaping and transmission of online video traffic over the Internet.

The transmission of video streams requires limited packet loss, end to end delay and delay variation. Obviously, these values depend on the size of the buffers at the routers. Depending on the size of the buffer one can differentiate between packet scale buffering and burst scale buffering. In the first case only a small buffer is provided to absorb packets arriving simultaneously, thus the packet loss probability might be high while the delay is strictly limited. In the second case the buffer provides enough space to absorb larger bursts and consequently, limits the loss probability while the control of delay and delay variation becomes a complex issue [9]. The use of packet scale buffering for transmitting delay and loss sensi-

tive data like coded video streams have been proposed in [10, 4], showing that low packet loss probability and high network utilization can be achieved if the peak rate of the streams is low compared to the link capacities. In [10] packet scale buffering is proposed together with source shaping. It is proved that a single buffer leaky bucket is an optimal shaper in this scenario.

The shaping of stored variable bit rate video streams is widely analyzed in the literature. Recent solutions are based on network calculus e.g., [3] or the bounding interval dependent (BIND) characterization of the streams [5].

In the case of live video streams there is a tradeoff between the delay introduced and the available information on the traffic to be transmitted. Only few works address this question. In [5] a BIND based solution is proposed that continuously adjusts the model parameters. In [7] the authors derive the shaper rate for statistically identical, peak rate controlled and leaky bucket shaped sources feeding a buffered multiplexer. The solution, however, can not handle multiplexed sources with different characteristics.

Algorithms for lossless shaping of individual streams are presented in [8] and [6]. In these works the statistical quantities of the shaped streams are analyzed, but network scenarios are not considered. In [8] shaping with delays in the range of 2-30 seconds is proposed, an adequate solution for broadcasting applications. The algorithm shown in [6] works with shaper delays less than a second. To avoid the fluctuation of the shaper rate, the algorithm uses past frame sizes to predict future traffic intensity, and needs to know the length of a GOP in advance.

Our contribution to this line of works is the definition of shaping algorithms with very low complexity that support the transmission of live video streams with a wide range of acceptable shaper delay and the analysis of the effect of shaping when the video streams are multiplexed at an output link with small buffer.

3 Model description

The system model considered in this paper includes traffic sources, source shapers and a multiplexing node with a single output link (figure 1).

The sources generate MPEG coded streams, the most commonly used encoding scheme for the storage and transmission of video information. Compression is achieved by eliminating the spatial and temporal redundancy of the information in the frames. Spatial redundancy is decreased by intraframe coding of the individual frames, while temporal redundancy is reduced by interframe coding between subsequent frames. Thus the sequence of frames consists of intraframe coded frames (I frames), and interframe

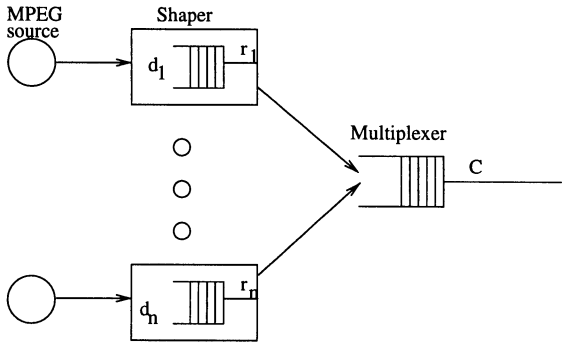


Figure 1: The considered system with MPEG sources, source shapers and a multiplexer.

coded predicted (P frames) and bidirectionally predicted frames (B frames). The subsequent frames between two consecutive I frames form a group of picture (GOP).

As a consequence of the coding scheme, the loss of data in an I frame propagates forward through the next GOP and backward to the last P frame. Meanwhile, the loss of data in a B frame only affects that particular frame.

The intraframe coding compresses complex scenes with less efficiency, and consequently, the frame sizes change on the long term at the scene changes. The interframe coding leads to short term frame size fluctuation, since I frames are usually significantly larger than P frames, and P frames are larger than B frames.

Shapers are used at the sources to decrease the frame to frame fluctuation of the coded video stream. Frames leaving the encoder are stored in the shaper buffer and are transmitted with a given transmission rate. The shaper is designed to introduce limited delay and provide lossless transmission, that is, no data can be lost due to buffer overflow or delay limit violation. To achieve this, the shaper transmission rate has to be adjusted depending on the size of the arriving frames and the buffer size has to be large enough to store the frames waiting for transmission.

The shaped video streams are multiplexed at a network node providing packet scale buffering, the size of the buffer is in the order of the ratio of the output link transmission capacity to the peak rate of the video streams.

4 Source shaping algorithms for live video streams

In this section two algorithms are proposed to control the rate r of the source shaper when transmitting live MPEG video streams. The algorithms aim to minimize the maximum and the variance of the transmis-

sion rate. The transmission delay in the shaper never exceeds the predefined maximum shaper delay and the shaping is lossless. The algorithms are simple in terms of the complexity of the shaper rate calculation and the amount of information considered, a requirement to assist real-time operation. Furthermore they provide an efficient solution for shaping with a large range of shaper delays. There is no need for any apriori information on the encoding scheme of the video, i.e., the number and sequence of I, P and B frames.

Both of the algorithms assume that the shaper can detect the type of the arriving frame. The shaper rate can be changed at any time t when a new frame is generated by the source and placed into the shaper buffer. The shaper rate is increased if the new frame (type I, P or B) can not be transmitted within the delay limit d, to provide lossless shaping, and it is decreased if the new frame is of type I, and all the traffic in the shaper buffer can still be transmitted within the delay limit. To decrease fluctuation, the new shaper rate is calculated as the average of the current rate and the minimum rate allowed by the delay limit.

The two proposed algorithms described differ on one point. The first algorithm is optimal in the sense, that the minimum shaper rate is calculated considering the residual acceptable transmission delay (RTD) for all the frames waiting in the shaper buffer, in the second, simplified solution the shaper rate calculation is based on the buffer content (BC), i.e. the total amount of data waiting in the shaper buffer. To avoid delay bound violation for frames stored in the buffer, the shaper rate can be decreased only if the buffer is empty before the new frame arrival.

As a consequence, the simplified BC algorithm follows the decreasing intensity of the stream with some delay compared to the RTD solution.

5 Performance evaluation

In order to assess the effectiveness of the proposed source shaping solutions we consider the statistical quantities of the shaped video traces and packet loss statistics in the case of multiplexing video streams at a single node with packet scale buffering.

The analysis is based on the fluid flow modeling of the traffic streams and uses results of the theory of large deviations to approximate probabilities of rare events. Simulations were done using ns-2.

The considered scenario is shown in figure 1. We present results for an MPEG-4 encoded video trace from a talk show with an average rate of 540kbps. We consider a single outgoing link at the multiplexer node with a capacity of 22.5Mbps. The multiplexing buffer can store up to 15 packets.

First we consider the statistical properties of a single video stream before and after shaping for different values of maximum shaper delay d.

Figure 2 shows the number of transmitted bits in a frame time for the original and shaped trace. The maximum shaper delays are 40 ms and 120 ms, the shaping is performed using the BC shaper algorithm. Even the relatively small shaper delay of 40 ms allows a significant reduction of the rate fluctuation, the maximum transmission rate is decreased from 3.1Mbps to 2Mbps.

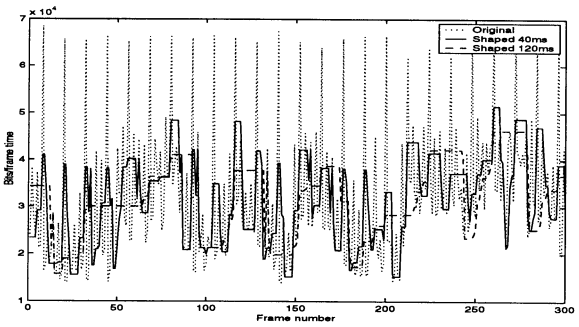


Figure 2: Number of bits transmitted in a frame time without shaping and with shaping for d=40 ms and d=120 ms, using the BC shaper algorithm.

Figure 3 shows the coefficient of variation (CoV) of the trace. The CoV is calculated using the moving average over one GOP time as mean value. This way the CoV reflects the frame to frame rate fluctuations without the rate variation due to the scene changes. The two curves in the figure show the CoV values for the two proposed shaping solutions.

For small values of shaper delay d the CoV decreases very rapidly, reflecting that a delay of a couple of frame times (20-80 ms) allows the smoothing of the transmission rates of consecutive I, P and B frames. At larger delays the marginal gain decreases significantly. The two shaping methods result in similar changes in the CoV values. As expected, the RTD method decreases the rate fluctuations better. For large values of d the difference is around 20% in the terms of CoV reduction. However, the difference is not significant, and due to its simplicity the BC method may provide the adequate shaping solution.

Figures 4 to 7 focus on the average packet loss probability of multiplexed video streams as a function of the average load, defined by the ratio of the sum of the mean rates of the streams to the link transmission capacity.

We show results for shaping with the BC method. To see the effect of shaping two scenarios are considered. In the first one only one stream is shaped, while all the other multiplexed streams are transmitted unshaped. In the second scenario half of the streams are shaped at the source.

Figure 4 shows the loss probability of the shaped and unshaped sources for a shaper delay of 40 ms in the case of 1 shaped source. Figure 5 shows the results for the scenario where half of the sources are shaped. Figures 6 and 7 show the relative loss probabilities of the shaped and unshaped sources compared to the average loss probability for a shaper delay of 120 ms.

The results show that the gain achieved by shaping the sources increases as the average load, and thus the loss rate decreases. Since the desired loss probability of video streams is in the order 10^{-5} or even smaller, the gain can be up to 35%, even if only one stream is shaped. Comparing figures 4 and 5 it can be seen that shaping half of the sources decreases the overall loss probability by roughly one order of magnitude.

In addition to the average packet loss probability of the video streams it is worthwhile to evaluate the packet loss probability in individual frame types, since it affects the percepted visual quality.

Figure 8 shows the loss probability of the shaped and unshaped I, P and B frames relative to the average loss probability of the shaped and unshaped sources respectively for the scenario when half of the sources is shaped with a maximum shaper delay of 120 ms. The figure shows that while in the case of unshaped sources the loss probability of the I frames is the highest, up to 100% above the average loss probability and that of the B frames is the lowest, in the case of shaped sources the loss probabilities of the individual frame types are roughly the same. Consequently, as losses in the I frame have a significant effect on the visual quality, the positive effects of the shaping include not only lower loss probability but also the improved distribution of these losses among the frame types.

6 Conclusion

In this paper we proposed and evaluated solutions that assist live video transmission over the Internet.

We proposed computationally simple shaping algorithms that are adequate to shape live video streams with a wide range of delay tolerance at the shaper and provided analytical method to evaluate the efficiency of the proposed solutions.

The performance evaluation based on the proposed analytical method and extensive simulations proved that i) a simple shaper algorithm based on buffer occupancy and delay limit results in efficient shaping in many cases; ii) even shaping with very low delay bound, adequate for real-time applications, improves the performance in terms of packet loss probability and the packet loss distribution among the different frame types; and iii) shaping provides a means to improve the quality of individual video transmissions even if not all the hosts shape their traffic.

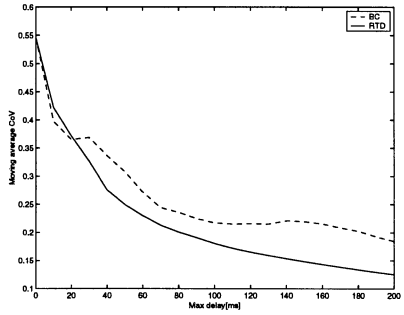


Figure 3: CoV of the shaped trace versus maximum shaper delay d, considering the BC and RTD shaping algorithms.

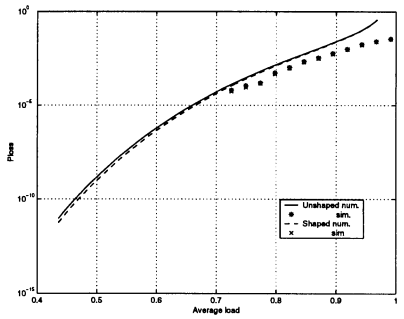


Figure 4: Loss probability of shaped and unshaped streams for $d = 40$ ms and 1 shaped stream. Simulation validates the analytical results.

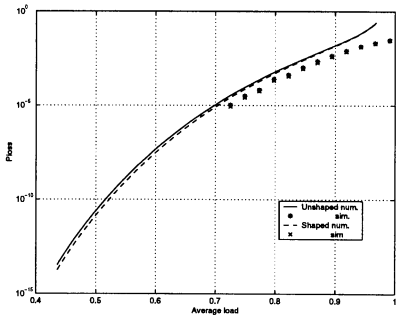


Figure 5: Loss probability of shaped and unshaped streams for $d = 40$ ms and half of the streams shaped. Simulation validates the analytical results.

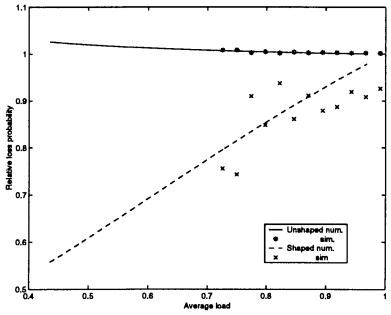


Figure 6: Relative loss probability of shaped and unshaped streams for $d = 120$ ms and 1 shaped stream. Simulation validates the analytical results.

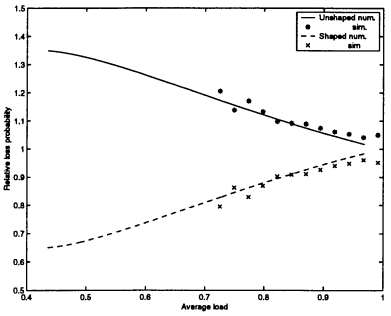


Figure 7: Relative loss probability of shaped and unshaped streams for $d = 120$ ms and half of the streams shaped. Simulation validates the analytical results.

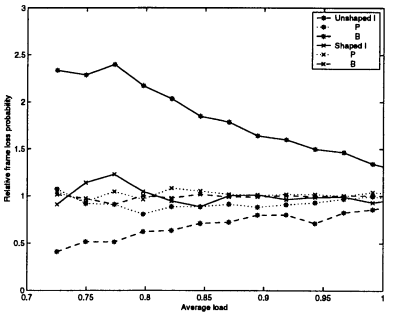


Figure 8: Relative loss probability of I,P,B frames of the shaped and unshaped streams for $d = 120$ ms, half of the streams shaped. Simulation results.

Finally, as ongoing work we further investigate the methods to assist video transmission over the Internet. Specifically, we are interested in the efficiency of source shaping versus buffering at the routers and buffering versus forward error correction.

References

[1] "Broadband network teletraffic, Final report of action COST 242," Springer, 1996.

[2] L. Breslau, E. W. Knightly, S. Shenker, I. Stoica and H. Zhang, "Endpoint Admission Control: Architectural Issues and Performance," in *Proc. of ACM SIGCOMM 2000*, 28 Aug. - 1. Sept., 2000, pp. 57-69.

[3] R. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation," *IEEE Transactions on Information Theory,* vol. 37, no. 1, Jan. 1991, pp. 114-131.

[4] V. Elek, G. Karlsson, R. Ronngren, "Admission Control Based on End-to-end Measurements," in *Proc. of IEEE INFOCOM 2000,*, March 26–30, 2000, pp. 623–630.

[5] H. Zhang and E. W. Knightly, "RED-VBR: A Renegotiation-based Approach to Support Delay-Sensitive VBR Video," *ACM Multimedia system Journal,* May 1997.

[6] Simon S. Lam, Simon Chow, David K. Y. Yau,"An Algorithm for Lossless Smoothing of MPEG video," *ACM SIGCOMM Computer Communication Review,* vol. 24, no. 4, Oct. 1994, pp. 281-293

[7] N. Modani, P. Dube and A. Kumar, "Measurement Based Optimal Source Shaping with a Shaping+Multiplexing Delay Constraint" *Proc. of IEEE Infocom 2000*, March 26–30, 2000,pp. 1807-1816.

[8] J. Rexford, S. Sen, J Dey, W. Feng, J. Kurose, J. Stankovic and D. Towsley, "Online Smoothing of Live, Variable-bit-rate Video" *Proc. of International Workshop on Network and Operating Systems Support for Digital Audio and Video*, May 1997, pp. 249-257

[9] J. W. Roberts, "Traffic Theory and the Internet," *IEEE Communications Magazine,* Jan. 2001., pp.94-99.

[10] M. Reisslein, K. W. Ross, S. Rajagopal, "Guaranteeing Statistical QoS to Regulated Traffic: The Single Node Case," in Proc. *IEEE Infocom'99,* pp. 1061-1072, 1999.

On Delay Estimations in Random Medium Access Protocols

Erika Johansson[1] and Ulf Sterner [2]

[1] Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping, Sweden. E-mail: erijo@foi.se.

[2] Swedish Defence Research Agency (FOI), P.O. Box 1165, SE-581 11 Linköping, Sweden. E-mail: ulfst@foi.se.

Abstract

For many of the services that an ad hoc network should support, it is not enough to know the networks average capacity when determining if it fulfils the user demands. Here we investigate how the network throughput is affected by applying a firm demand on the fraction of data that must reach its destination within a certain time. Furthermore we study the covariance of the packet delays.

To look into these issues, we compare different network set-ups for the IEEE 802.11b MAC protocols. Our results show that the application of firm demands can lead to severe reductions in network throughput, while at the same time reducing the interpacket delay covariance.

1 Introduction

In future military engagements, the need to quickly acquire and assimilate information at all levels of the command hierarchy is foreseen. In many of the command and control systems projected, low level units are expected to make well informed and autonomous decisions. Combat information and sensor data must thus be available "on the move" even at the level of the individual soldiers.

In crucial situations, e.g. communication on the battlefield itself, there will be a need for high performance wireless communication without the support of a pre-deployed infrastructure. In these cases a radio network should be able to be successfully deployed in unknown terrain and with a minimum need of network planning. Furthermore, the networks should utilize distributed network control to increase robustness e.g. in the event of node annihilation. Such networks are often referred to as *ad hoc* networks. The network also has to remain intact as the units move rapidly into new terrain. Network protocols must therefore be able to cope with rapid and unexpected topology changes caused by the movements of the radio units.

A military ad hoc network must also support a wide category of services [1, 2], e.g. group calls, situation awareness data and fire control data. The different services will all have different *Quality of Service (QoS)* demands, i.e. different demands on delay, packet loss ratio, throughput, etc.. Here we compare the network capacity resulting from the introduction of a firm demand with the average network capacity. The firm demand we here choose to concentrate on is that at least a certain fraction of the intraffic should arrive at its destination within a specific time limit.

As a measure of the network capacity, we choose to use the throughput-delay characteristics. To exemplify the effect of the introduction of firm demands we thus estimate the throughput-delay characteristics with and without the firm demand for a set of example networks.

We perform this comparison for the two basic Medium Access Control (MAC) protocols defined in the IEEE 802.11b standard [3], and for different network sizes and set-ups. Furthermore, we look at the effects of using different intraffic models.

Our simulation results show that even a modest firm demand results in considerable losses in throughput, compared with the throughput possible if we only consider the mean delay. The demands on the maximum packet loss and the maximum delay are thus important when determining the throughput possible in a network.

When deciding what packet delay a certain service can withstand, it is important to know what packet loss ratio the service can endure but also to know how independent the packet delays are. To lose one out of a hundred packets might be acceptable for a service, while losing ten packets in a row might be a problem even if the average loss is only one out of a thousand packets. To study this issue, we estimate the interpacket delay covariance. Our simulations show that there is indeed a dependency between the packets with regard to their delays. This dependency is reduced however, by the application of the firm demand used here.

We continue our paper with a presentation of the used performance measures. We then give a short overview of our network model and the simulation set-up. Then our results are described. Finally, we present some conclusions from our results.

2 Performance Measures and Firm Demands

A military network should provide a variety of services (e.g. voice, data communications and situation awareness data) both as point-to-point communication and as multicast. These services result in different demands on delay, throughput, packet loss ratio etc.. The demands on each parameter will be very different depending on the service,

e.g. the delay demand will vary from minutes (file transfers), to a few hundreds of a second (fire-control data).

In a fixed network, it might be possible to allocate resources such that a service can be guaranteed a certain amount of the network capacity and, through this, a certain maximum delay can be guaranteed. However, in a mobile radio network in quickly changing terrain and with limited resources, no such guarantees can be made. As a realistic example of a firm demand, as mentioned earlier, we choose that

> *a fraction α of the packets must reach their destinations within a specified time d.*

This means that the packet loss ratio, including the packets that are delivered too late, is less than $1 - \alpha$.

2.1 Throughput-Delay Characteristics

A classic method to analyze network capacity is to study, by analytical calculations or simulations, the throughput-delay characteristics or the maximum throughput [4, 5, 6, 7]. We here denote the average delay by

$$\mu_d = \frac{1}{M} \sum_{i=1}^{M} d_i \tag{1}$$

where M is the number of packets in the simulation and d_i is the delay for packet i. The packets are here numbered in the order they enter the network, without regard to which node they enter the network at.

To calculate the fraction of intraffic that reaches its destination within a specified time d_{max}, we define an indicator function

$$I_i(d_{max}) = \begin{cases} 0 & \text{if } d_i > d_{max}, \\ 1 & \text{if } d_i \leq d_{max} \end{cases} . \tag{2}$$

Once we specify the fraction α of the traffic that must reach its destination within a certain time, we can use this indicator function to find the smallest delay d_α for which α is upheld. This can be expressed as

$$d_\alpha = \inf_d \frac{1}{M} \sum_{i=1}^{M} I_i(d) \geq \alpha \tag{3}$$

where M is the number of packets in the simulation and d is the maximum delay.

2.2 Interpacket Delay Covariance

Another interesting issue is whether the packet delays are independent or not. If the delays are independent: the late arrival of one packet can be considered an isolated event. If not: the late arrival of one packet presage the late arrival of other packets. In the first case, the packet delay distribution completely describes the delay behavior of the network. However, if the packet delays are correlated, i.e. has a high interpacket delay covariance, other means must be used to fully describe the delay behavior.

To investigate whether the packet delays are independent or not, we estimate the interpacket delay covariance as

$$R_d(j) = \frac{1}{M-1-j} \sum_{i=1}^{M-j} (d_i - \mu_d)(d_{i+j} - \mu_d). \tag{4}$$

where M is the number of packets in the simulation, d_i is the packet delay for packet i and μ_d is the average delay. In the same manner, we estimate the interpacket delay covariance in the presence of the firm demand as

$$R_I(j) = \frac{1}{M-1-j} \sum_{i=1}^{M-j} (I_i - \mu_I)(I_{i+j} - \mu_I). \tag{5}$$

Here I_i is the indicator function in equation (2) for a specified maximum delay d_{max}, corresponding to the wanted packet delivery ratio α. Furthermore, μ_I is given by

$$\mu_I = \frac{1}{M} \sum_{i=1}^{M} I_i. \tag{6}$$

3 Simulation Set-up

To investigate the effects of applying a firm demand on the packet delivery ratio and the maximum delay, as well as the issue of delay correlation, we perform a number of simulations using OPNET [8]. In this section we present the network model and the parameters we vary in our simulations.

The capacity of an ad hoc network is influenced by a great number of parameters, e.g. the users service requirements, the routing and MAC protocols available, the terrain, the network topology, the number of users etc.. Hence, this type of network enters a large number of "unknown" factors into the simulations.

To focus on the evaluation of the difference in throughput-delay characteristics, with and without the presence of a firm demand, we use single-hop networks in our simulations. Furthermore, the units in a network are assumed to be stationary and we use free-space propagation to calculate the path loss. We also limit our study to networks with 4 and 24 nodes respectively.

We conduct simulations using the two basic MAC protocols described in the IEEE 802.11b standard, CSMA (Carrier Sense Multiple Access) and RTS/CTS (Request-To-Send/Clear-To-Send). The IEEE 802.11b standard also specifies data rates in the range 1-11 Mbit/s. Here we fix our data rate to 1 Mbit/s.

The classic choice of traffic model is a Poisson model. However, many of the services we are interested in (e.g. a phone call) generates traffic that is better approximated by a deterministic model. Hence, we perform simulations for both Poisson distributed traffic and deterministic traffic.

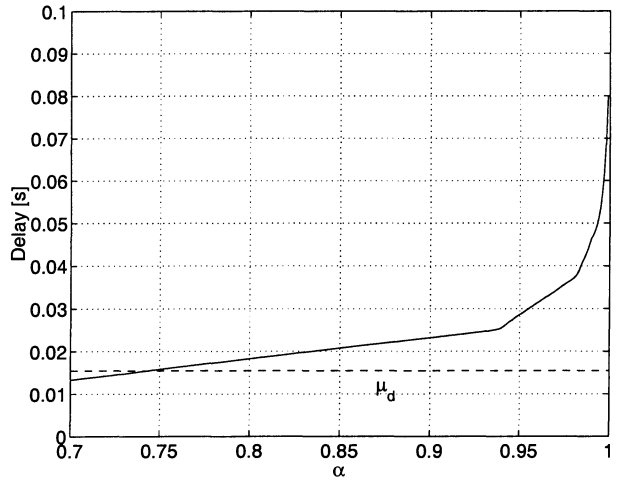


Figure 1: *The delay d_α as a function of different demands on α for a network with traffic load $S = 0.30$ Mbit/s.*

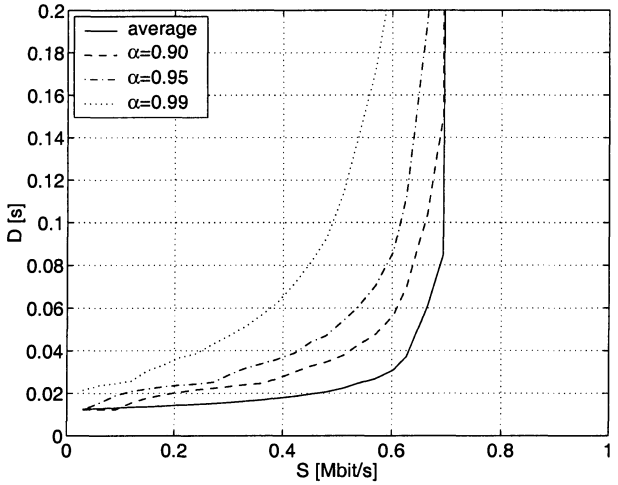


Figure 2: *The average throughput-delay and the throughput-delay resulting from the application of different demands on α.*

The different services that a military network should support also results in packets of varying size. We thus do simulations for two different packet sizes. Really small packets consisting of 53 byte and rather large packets consisting of 1500 byte.

Each node in the network generates traffic according to the chosen traffic model. The destination for the generated packet is then randomly chosen among the other nodes in the network. In our capacity calculations, we however choose to look at the total traffic flow in the network and not at the traffic flow in the individual nodes.

4 Results

In this section we present the results of our simulations. In the first part, we compare the average throughput-delay characteristics with the throughput-delay characteristics resulting from the application of a firm demand. In the second part, we look at the delay covariance for the packets, with and without the presence of the firm demand. We do both these comparisons for different network set-ups. Additional results can be found in [9].

4.1 Capacity in the Presence of a Firm Demand

Once we have chosen a demand on α, we can calculate d_α by using equation (3). In Figure 1, we have plotted the delay d_α as a function of α in a network with a traffic load $S = 0.30$ Mbit/s. The network in question, which we will refer to as *the basic network*, has 24 nodes, uses the CSMA MAC protocol and has Poisson distributed intraffic.

For the chosen intraffic and demand on α, we can find the smallest delay d_α for which the demand is fulfilled. For $\alpha = 0.95$, this means that we have to accept a packet delay $d_\alpha = 0.029$ s. This can be compared with an average delay $\mu_d = 0.016$ s.

In Figure 2, we can observe the difference in throughput-delay with and without the application of firm demands for $\alpha = \{0.90, 0.95, 0.99\}$. We can here see that if we demand that the maximum average delay μ_d should not exceed 0.029 s, we get a maximum average throughput of 0.58 Mbit/s. However, if we apply the firm demand $\alpha = 0.95$ and demand $d_\alpha \leq 0.029$ s, we get a maximum throughput of 0.30 Mbit/s. We can thus conclude that the application of a firm demand in this case reduces the network capacity with almost 50%.

As we can see, the loss in throughput caused by applying a firm demand can be quite significant. For the network in Figure 2, the loss is in the order of 30%−75% for a fixed delay demand $d = 0.029$ s, depending on the demand α. This means that, for a network where firm demands will be present, other measures than the average delay and the average throughput must be used to estimate the network capacity.

In Figure 3, we compare the throughput-delay characteristics for a firm demand $\alpha = 0.95$, while we vary some of the simulation parameters. The solid line represents our basic network. For each of the other curves, one parameter has been changed. The dashed line represents a network that uses RTS/CTS, the dash-dotted line a network with deterministic intraffic, and the dotted line a network with only 4 nodes.

As can be seen in Figure 3, the RTS/CTS protocol gives a better performance, i.e. a lower delay, for high throughputs. This could be expected since the transmitted packets are rather large, 1.5 Kbytes. This is in agreement with the results for the average throughput found in [7] where the distributed coordination function for the IEEE 802.11 protocol is analyzed. In Figure 3, we can also observe that a smaller network results in a higher throughput, as could be

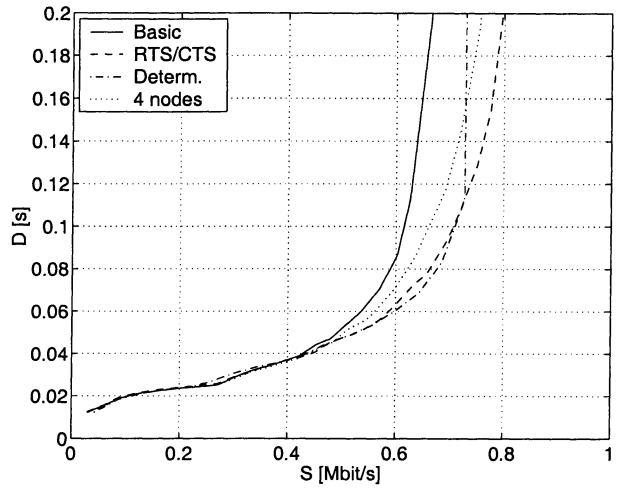


Figure 3: *The throughput-delay characteristics for* $\alpha = 0.95$ *for different parameter changes.*

expected.

If we compare the curves for the different traffic models in Figure 3, we find that for low throughputs, there is very little difference between the traffic models. For higher throughputs however, there is a distinct difference between the two traffic models. The deterministic traffic model can be observed to have higher throughput for a given delay.

4.2 Interpacket Delay Covariance

To investigate whether the packet delays are independent or not, we look at the interpacket delay covariance functions defined in equation (4) and (5). In Figure 4, we have plotted the interpacket delay covariance in the basic network, when no firm demands are present, for different traffic loads. As we can see in the figure, the interpacket delay covariance increases when the traffic load increases. This is a result of an increasing number of packet collisions when increasing the intraffic and also that, when collisions occur, more packets are affected in a network with high traffic load.

If we fix the traffic load, we can compare the interpacket delay covariance with and without firm demands. For the basic network and a traffic load $S = 0.30$ Mbit/s, we have plotted the interpacket delay covariance with the demands $\alpha = \{0.90, 0.95, 0.99\}$ and without any demand in Figure 5. The solid line in the figure represents the interpacket delay covariance when no firm demands are applied.

As can be seen in Figure 5, the interpacket delay covariance decreases when a firm demand is applied. This decrease is a result of the packets being classified as either on time or not. Many packets with large delays will then still have delays lower than d_α and so will be considered delivered on time. Hence, long bursts of packets with high delays, that results in high covariance for the case without

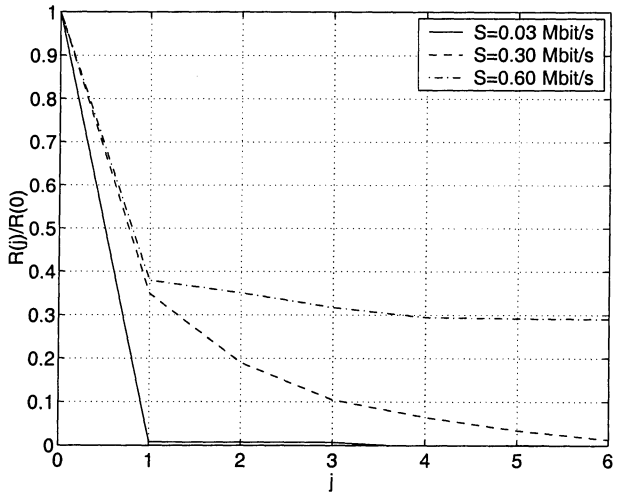


Figure 4: *The interpacket delay covariance for different traffic loads when no firm demands are applied.*

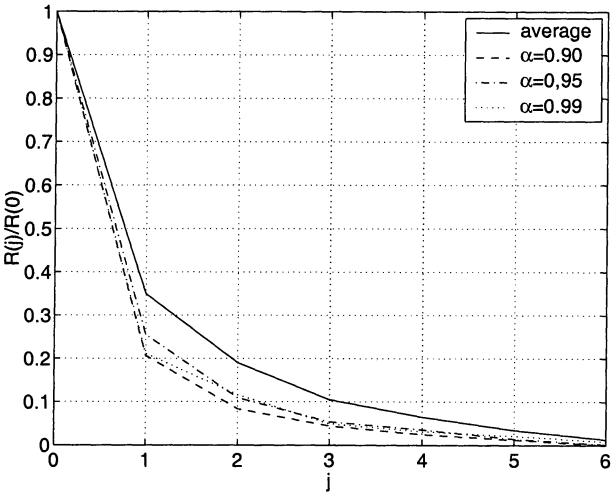


Figure 5: *The interpacket delay covariance for a traffic load of* $S = 0.30$ *Mbit/s for different demands on* α*, compared with the interpacket delay covariance without demands.*

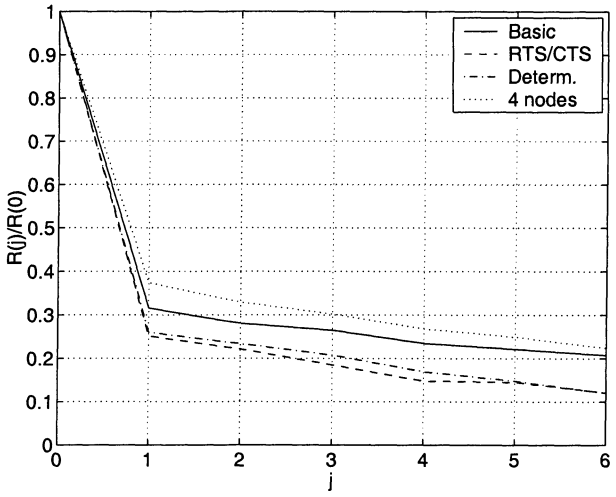


Figure 6: *The interpacket delay covariance for* $\alpha = 0.95$ *and for different parameter changes.*

demands, are broken down into shorter bursts that have less effect on the delay covariance. Furthermore, we can observe in Figure 5 that the difference in interpacket delay covariance for different demands on α is small.

In Figure 6, we compare the interpacket delay covariance for a firm demand of $\alpha = 0.95$, while we vary some of the simulation parameters. The solid line represents the basic network. For each of the other curves, one parameter has been changed. The dashed line represents a network that uses RTS/CTS, the dash-dotted line a network with deterministic intraffic, and the dotted line a network with only 4 nodes.

As can be seen in Figure 6, the RTS/CTS protocol yields slightly lower interpacket delay covariance than the CSMA protocol. This is partly due to that the RTS/CTS protocol is further from the point of saturation than the CSMA protocol, see Figure 3. Furthermore, we find that the deterministic traffic model, compared with the Poisson traffic model, gives a similar lowering of the covariance.

In Figure 6 we can also observe that a smaller network results in higher covariance, which was expected as fewer nodes means that a larger portion of the nodes are affected each time a collision occurs.

5 Conclusions

From our simulations, we can see that there is a large decrease in expected channel capacity when comparing a networks average throughput-delay characteristics to the throughput-delay characteristics resulting from the application of a firm demand. It is thus important to use other measures than the average throughput-delay characteristics for networks where firm demands must be met.

Furthermore, we can note from our results that interpacket delay covariance for high traffic loads is significant. In many cases, it is thus not possible to consider the packet delays to be independent. The interpacket delay covariance however, is reduced by the application of the firm demand used in this paper. This is because the firm demand helps breaking up bursts of packets with long delays into shorter bursts, resulting in a smaller covariance.

The results of our simulations also show that although the choice of MAC protocol, the number of nodes in the network and the traffic model affects the actual results, the trends in throughput-delay characteristics and interpacket delay covariance remains the same whether a firm demand is present or not.

References

[1] F. Eklöf and B. Johansson. Position distribution service for mechanised units. User Report FOA-R–00-01734-504–SE, Defence Research Est., Div. of Command and Control Warfare Tech. Linköping, Sweden, December 2000. In Swedish.

[2] A. Hansson, J. Nilsson, M. Sköld, and U. Sterner. Tactical Radio Access Networks - A Comparision of Cellular and Ad Hoc Network Concepts. Technical Report FOI-R–0086–SE, Swedish Defence Research Agency, Div. of Command and Control Warfare Tech. Linköping, Sweden, March 2001.

[3] IEEE. IEEE std 802.11 - wireless LAN medium access control (MAC) and physical layer (PHY) specifications. 1997.

[4] L. Kleinrock and F. A. Tobagi. Packet Switching in Radio Channels: Part I-Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics. *IEEE Transactions on Communications*, COM-23(12):1400–1416, December 1975.

[5] S. S. LAM. Delay Analysis of a Time Division Multiple Access (TDMA) Channel. *IEEE Transcations on Communications*, COM-25(12):1489–1494, December 1977.

[6] C. L. Fullmer. *Collision Avoidance Techniques for Packet-Radio Networks*. PhD thesis, University of California, 1998.

[7] G. Bianchi. Performance Analysis of the IEEE 802.11 Distributed Coordination Function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, 2000.

[8] OPNET Technologies, Inc. http://www.opnet.com.

[9] E. Johansson and U. Sterner. On the quality of service capacity of random access protocols. FOI Memo 01-4081, Defence Research Agency, Div. of Command and Control Warfare Tech. Linköping, Sweden, Dec 2001.

Radiofrequency exposure from wireless equipment – European limits and test standards

Christer Törnevik

Ericsson Research, Ericsson AB, SE-164 80 Stockholm, Sweden. Email: Christer.Tornevik@era.ericsson.se

Abstract

This paper summarizes the status of the ongoing standardization in Europe in the area of radiofrequency exposure from wireless telecommunications equipment. European standards for mobile telephones have recently been completed and published, and standards for radio base stations are under development. The purpose of these standards is to show that the radiofrequency exposure from wireless products does not exceed the restrictions specified in an EU recommendation, and that the equipment is in conformity with health and safety related requirements in the EU directive covering radio products.

1 Introduction

Mobile telephones, radio base stations and other wireless communications equipment marketed and used within the European Union is covered by the Radio & Telecommunications Terminal Equipment (R&TTE) Directive [1]. This directive specifies essential requirements, including aspects related to the protection of the health and safety of the user and any other person. One such aspect is the human exposure to radiofrequency (RF) electromagnetic fields (EMF). Within the European Union, maximum RF exposure levels are specified in the Council Recommendation 1999/519/EC [2]. In 2000, the European Commission submitted a mandate to the European Committee for Electrotechnical Standardization (CENELEC) regarding development of harmonized standards for wireless equipment and other products covered by the R&TTE directive [3]. The EMF mandate was formally accepted by CENELEC in March 2001. The purpose of the harmonized standards is to show that the RF exposure from the R&TTE devices is in compliance with the limits in the Council Recommendation, and that the equipment thereby meet the health and safety related essential requirements of the directive. The CENELEC Technical Committee TC106x, which deals with electromagnetic fields in the human environment, has an extensive work program for the development of the standards asked for in the mandate. European standards applicable for mobile telephones have already been completed and published, and standards for radio base stations are under development.

2 Radiofrequency exposure limits

In 1999, the EU Council Recommendation on the limitation of exposure of the general public to electromagnetic fields in the frequency range from 0 Hz to 300 GHz was published [2]. This document provides recommendations to the EU member states, and it is based on the guidelines developed by the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [4]. The recommended limitations are a result of an extensive review and analysis of several thousand scientific papers related to effects to human health from exposure to electromagnetic fields. The limits have been set with wide margins to protect all persons from established adverse effects from both short-term and long-term exposure to electromagnetic fields. It is the conclusion of ICNIRP and several other expert groups and authorities, that there are no established health effects below the recommended limits.

In the frequency range from 10 MHz to 10 GHz, which includes the frequency bands used by wireless communications, basic restrictions expressed as Specific Absorption Rate (SAR) levels are specified. SAR, which has the unit W/kg, is a measure of the rate of energy absorption in the body of a person exposed to RF fields. The SAR limits are set to ensure that whole-body or localized adverse heating effects will not occur due to the exposure. In Table 1, the basic SAR restrictions for the general public in the EU recommendation are listed.

Since SAR in many real exposure situations is difficult to measure or calculate, the ICNIRP guidelines and the EU recommendations also specify reference levels. In the frequency range of wireless communications these are given as frequency-dependent electric field strength E (V/m), magnetic field strength H (A/m), and power density S (W/m^2) levels. These should be assessed in air without the presence of the exposed person. Table 2 lists the reference levels for the GSM and UMTS frequency bands.

Table 1: Basic restrictions for the general public (10 MHz – 10 GHz)

Averaging mass	SAR (W/kg)
Whole-body	0.08
10g, head and trunk	2
10g, limbs	4

Table 2: Reference levels for the general public (GSM and UMTS frequency bands)

Frequency (MHz)	E (V/m)	H (A/m)	S (W/m^2)
900	41	0.11	4.5
1800	58	0.16	9
2100	61	0.16	10

The reference levels have been derived from the basic restrictions and include extra margins. This means that compliance with the reference levels ensures compliance with the SAR limits. However, exceeding the reference levels does not necessarily mean that the basic restrictions are exceeded. For example, the field strength levels close to a mobile phone can exceed the reference levels although the maximum localized SAR in the head tissue of the user is in compliance with the basic restriction. The reference levels are normally used in the case of RF exposure compliance assessments of radio base stations, while the mass averaged SAR limit is applicable for mobile phones and other wireless terminals used in close proximity to the body.

3 Compliance assessment standards

3.1 Test standards for wireless terminals

CENELEC has completed the standardization of a SAR test procedure for handheld mobile telephones used at the ear. This basic standard was published as a European Standard (EN 50361) [5] in July 2001. CENELEC has also published a European Standard (EN 50360) [6] that specifies the criteria for evaluation of compliance with the SAR limits in the EU recommendation using the test procedures in EN 50361. This product standard has been listed as a harmonized standard under the R&TTE directive. Testing is excluded for mobile phone products with an average emitted power of 20 mW or less.

The basic standard provides specifications on the measurement system, the SAR test protocol and the uncertainty assessment procedures. The measurement system consists of a phantom model representing a human head, a mobile phone holder, SAR measurement equipment and a scanning system. The holder is used to position the device under test in two different usage positions defined in the standard. The scanning system is needed to move a miniature electric field probe inside the phantom, which is a thin low-loss shell filled with head tissue simulating liquid. From the obtained electric field strength distribution, the localized SAR is calculated. Figure 1 shows a picture of a SAR measurement system that meets the specifications in the standard.

Technical Committee TC106 of the International Electrotechnical Commission (IEC) is also in the process of standardizing SAR measurement procedures for wireless terminals. The coming international standard will be harmonized with the CENELEC document, but it will be extended also to include body-worn devices, two-way radios and palmtop/desktop terminals. The IEC standard is estimated to be published in 2004 and will most probably be adopted by CENELEC as a European standard.

3.2 Test standards for base stations

CENELEC is developing standards for the compliance assessment of base station products. As for mobile phones, both product and basic standards will be published. Draft standards aimed for use by base station equipment manufacturers [7-9] have been developed and will probably be approved and published during 2002.

The basic standard (prEN50383) specifies different methods for the calculation and measurement of the radiofrequency exposure from base stations in order to assess the compliance distances according to basic restrictions or reference levels. For low-power (= 1 watt) base stations with small antennas, a method to measure SAR in a body model at different distances is described. For other base stations, the draft standard specifies experimental, analytical and numerical procedures that should be used to determine three-dimensional field strength distributions around the base station antenna.

The operators that install the base stations and take them into service also need standards. Work is currently ongoing in CENELEC to complete specifications for the assessment of compliance distances for antennas in their real operational environments.



Figure 1: *Photograph of a setup for SAR measurements of mobile phones. A miniature probe is used for measurements of the electric field induced in a head phantom filled with tissue simulating liquid.*

4 Conclusions

The science-based restrictions on human exposure to electromagnetic fields developed by ICNIRP have been adopted in an EU recommendation. The limits are being implemented in EU requirements for radio equipment through harmonized standards under the R&TTE directive. These standards are developed by CENELEC as a part of a mandate received from the European Commission. Standards for mobile phones have already been completed and published, and standards for base stations are under development.

5 References

[1] 1999/5/EC, Directive of the European Parliament and of the Council of 9 March 1999 on radio equipment and telecommunications terminal equipment and the mutual recognition of their conformity (R&TTE Directive), *Official Journal of the European Communities*, L 91/10, April 1999.

[2] 1999/519/EC, Council Recommendation of 12 July 1999 on the limitation of exposure of the general public to electromagnetic fields (0 Hz to 300 GHz), *Official Journal of the European Communities*, L 199/59, July 1999.

[3] M/305, Harmonised standards covering protection from electromagnetic fields (0 Hz to 300 GHz) generated by apparatus included in the scope of either the Low Voltage Directive 73/23/EEC, (LVD) or the Radio Equipment and Telecommunications Terminal Equipment (R&TTE) Directive 1999/5/EC. Standardisation mandate addressed to CEN, CENELEC and ETSI in the field of electrotechnology, information technology and telecommunications, European Commission, November 2000.

[4] ICNIRP, Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz), International Commission on Non-Ionizing Radiation Protection (ICNIRP), *Health Physics*, vol. 74, no. 4, pp. 494-522, April 1998.

[5] EN 50361, Basic standard for the measurement of Specific Absorption Rate related to human exposure to electromagnetic fields from mobile phones (300 MHz – 3 GHz), European Committee for Electrotechnical Standardization (CENELEC), European Standard, July 2001.

[6] EN 50360, Product standard to demonstrate the compliance of mobile phones with the basic restrictions related to human exposure to electromagnetic fields (300 MHz – 3 GHz), European Committee for Electrotechnical Standardization (CENELEC), European Standard, July 2001.

[7] prEN 50383, Basic standard for the calculation and measurement of electromagnetic field strength and SAR related to human exposure from radio base stations and fixed terminal stations for wireless telecommunication systems (110 MHz - 40 GHz)", European Committee for Electrotechnical Standardization (CENELEC), European Standard (Draft), November 2001.

[8] prEN 50384, Product standard to demonstrate the compliance of radio base stations and fixed terminal stations for wireless telecommunication systems with the basic restrictions or the reference levels related to occupational exposure to radio frequency electromagnetic fields (110 MHz - 40 GHz)", European Committee for Electrotechnical Standardization (CENELEC), European Standard (Draft), November 2001.

[9] prEN 50385, Product standard to demonstrate the compliance of radio base stations and fixed terminal stations for wireless telecommunication systems with the basic restrictions or the reference levels related to general public exposure to radio frequency electromagnetic fields (110 MHz - 40 GHz)", European Committee for Electrotechnical Standardization (CENELEC), European Standard (Draft), November 2001.

Recommended limits of electromagnetic exposure for the general public

Lars-Erik Paulsson[1]

[1]Statens strålskyddsinstitut, 171 16 Stockholm. E-mail: lars.erik.paulsson@ssi.se.

Abstract

The Swedish Radiation Protection Authority (SSI) will implement the recommendations on the limitation of exposure to the general public to electromagnetic fields (0 Hz to 300 GHz) as issued by the European Council in 1999. The SSI will issue the recommendations as a General Advice upon which mandatory regulations for specific areas can be based. Recommended limits are given as basic restriction values (e.g. 2 W/kg SAR for mobile phone exposure) and as derived more easily measurable reference values (e.g. 100 μT for 50 Hz magnetic fields). When this text is written the General Advice is only available as a draft for comments. The information given here might not in all details reflect the final version.

1 Background

Biological effects of electromagnetic fields and safety precautions against harmful effects of these fields have been given much attention latest years. The issue is however not new and the present author gave his first presentation of this subject to RVK in 1975. Mandatory regulations for occupational exposure were enforced by the National Board for Occupational Safety and Health in 1976. Recently the rapid growth of mobile communications has brought these issues to the attention also of the general public.

Since mobile communication systems and other radio-related devices are highly standardised in Europe and partly globally it is naturally that safety regulations are harmonised internationally as far as possible. The Council of the European Union issued in 1999 recommendations on the limitation of exposure of the general public to electromagnetic fields [1].

The EU recommendation has been adopted for use in Sweden by the Swedish Radiation Protection Authority (SSI) as a general advice issued under the radiation protection act. When this is written the advice is sent out as a draft for comment. It is planned to be printed in final form within a couple of months. If need should arise in specific areas, mandatory regulations based on the recommendations can be issued later.

The recommendation is based on a comprehensive background material, which has been discussed in the political organs of the European Community. Some of it is condensed and summarised here.

- Members of the general public shall be protected against established adverse health effects that may result from exposure to electromagnetic fields.

- The protection should be based on a commonly agreed framework and should ensure consistency throughout the Community. The International Commission on Non-Ionising Radiation Protection (ICNIRP) [2] has given such a framework.

- Actions on limiting the exposure to the general public to electromagnetic fields should be balanced with other health, safety and security benefits that devices emitting electromagnetic fields bring to the quality of life, in such areas as telecommunications, energy and public security.

- The recommendation applies particularly to areas where the general public spend significant time in relation to the biological effects covered.

- In order to assess compliance with the recommended limits of exposure the standardisation bodies (e.g. Cenelec) should develop standards for design and testing of electromagnetic field emitting equipment.

- Interference problems and effects on medical devices are not within the scope of the recommendation but are dealt with in the legislation for electromagnetic compatibility and medical devices.

- The progress of scientific knowledge in this area should be reviewed regularly e.g. by competent international organisations (e.g. ICNIRP).

2 Recommended exposure limits

The recommended limiting values are presented as two different sets of numbers. The fundamental values are called basic restrictions and they are directly based on the relevant biological interaction processes e.g. current density or energy deposition in the body tissue. Measurements to verify compliance with the basic restrictions are normally not possible in a real exposure situation. Therefore a set of more easily measurable reference values is given. They are derived from the basic restrictions and if an exposure situation can be verified to give exposures below the reference values then the basic restrictions can normally not be exceeded. However, it is possible that an exposure above the reference values can still be shown to comply with the requirements in the basic restrictions. If so, such exposures are regarded to be below the recommended limits. This will typically occur with highly localised sources e.g. mobile phones and the verification process will in these cases involve computer calculations or measurements inside phantom models.

A set of reference values for induced currents and contact currents has not been included in the Swedish general advice partly because they are of limited relevance for the exposure conditions of the general public and partly because of overlap with other standards.

2.1 Basic restrictions

The exposure limits are in accordance with international recommendations set to about 2 percent of the levels where negative health effects are established.

Table 1 Basic restrictions for electric, magnetic and electromagnetic fields

Frequency range	Magnetic flux density mT	Current density (effective value) mA/m^2	SAR (Whole body (mean value) W/kg	Local SAR (head and trunk) W/kg	Local SAR (arms and legs) W/kg	Power density S W/m^2
0	40	-	-	-	-	-
> 0 – 1 Hz	-	8	-	-	-	-
1 - 4 Hz	-	8/f	-	-	-	-
4 Hz - 1 kHz	-	2	-	-	-	-
1 - 100 kHz	-	f/500	-	-	-	-
0,1 - 10 MHz		f/500	0.08	2	4	-
0,01 - 10 GHz	-	-	0.08	2	4	-
10 – 300 GHz	-	-	-	-	-	10

f is the frequency expressed in Hz.

Current density bears upon the mean value over 1 cm^2 perpendicular to the direction of the current.

All SAR-values (Specific absorption rate) bear upon the mean value in 6 minutes' period of time.

Local SAR bears upon the mean value over a mass of 10 g contiguous tissue.

2.2 Derived reference values

Table 2 Derived reference values for exposure limits for the general public

Frequency range	E-fields (V/m)	H-fields (A/m)	B-fields (μT)	Equivalent power density for a plane wave S_{eq} (W/m^2)
0 - 1 Hz	-	$3.2 \cdot 10^4$	$4 \cdot 10^4$	-
1 - 8 Hz	10000	$3.2 \cdot 10^4/f^2$	$4 \cdot 10^4/f^2$	-
8 - 25 Hz	10000	4000/f	5000/f	-
25 - 800 Hz	$2.5 \cdot 10^5/f$	4000/f	5000/f	-
0,8 – 3 kHz	$2.5 \cdot 10^5/f$	5	6.25	-
3 – 150 kHz	87	5	6.25	-
0,150 - 1 MHz	87	$7.3 \cdot 10^5/f$	$9.2 \cdot 10^5/f$	-
1 – 10 MHz	$8.7 \cdot 10^4/f^{1/2}$	$7.3 \cdot 10^5/f$	$9.2 \cdot 10^5/f$	-
10 – 400 MHz	28	0.073	0.092	2
0,4 - 2 GHz	$\frac{1.375 \cdot f^{1/2}}{1000}$	$\frac{0.0037 \cdot f^{1/2}}{1000}$	$\frac{0.0046 \cdot f^{1/2}}{1000}$	$f/2 \cdot 10^8$
2 - 300 GHz	61	0.16	0.20	10

f is the frequency expressed in Hz.

In the frequency range 100 kHz to 10 GHz, S_{eq}, E^2, H^2 and B^2 regard the mean value over a 6 minutes' period of time.

At frequencies higher than 10 GHz, S_{eq}, E^2, H^2 and B^2 are calculated as mean values over a period of time that is $68/(10^9 \cdot f)^{1.05}$ minutes, where f is expressed in Hz.

The mean values for E-fields, H-fields and B-fields are calculated as the square roots of the mean values of the squares of the quantities respectively over the period of time in question.

3 Practical use

The General Advice is not applicable to occupational exposure to electromagnetic fields. For such exposures regulations from the Swedish Work Environment Authority apply. For the general public the most relevant areas will be covered by standards implemented under different directives of the EU (e.g. mobile phones). Procedures or equipment not covered by standards or directives may require separate legislation (e.g. the Swedish regulation on the use of microwaves for drying water leakages in buildings [3]).

4 References

1 European Council, "Council Recommendation of 12 July 1999 on the limitation of exposure of the general public to electromagnetic fields (0 Hz to 300 GHz)", Official Journal of the European Communities L 199, 30.7.1999, p. 59-70

2 ICNIRP, "Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz)", Health Physics, 74: 494-522, 1998

3 Statens strålskyddsinstitut, "Statens strålskydds-instituts föreskrifter om torkning med mikrovågor", SSI FS 1995:3.

A detailed dosimetric study of an inner ear exposed to mobile phone radiation

R. Kopecký[1], Y. Hamnerius[1], H. Nellström[2], M. Persson[1]

[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden.
E-mail: elfmp@elmagn.chalmers.se
[2]Medorad AB, Göteborg, Sweden

Abstract

The Specific Absorption Rate (SAR) distribution is calculated using a detailed inner ear model developed from an existing head model obtained from the Visible Human Project[1], with a 1 mm resolution has been improved in the ear region giving increased anatomical resolution. Modeling of the electromagnetic field from a mobile phone is performed using the Finite-Difference Time-Domain Method (FDTD).

1 Introduction

A head model from Visible Human Project (VHP) has 1 mm resolution and consists of 25 different biological tissues. However inner ear tissues are missing. We have now added the missing tissues of inner ear with the same 1 mm resolution 1 mm.

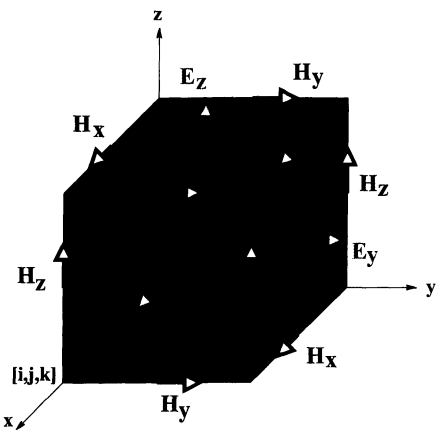


Figure 1: Computational cell with H components on boundary between different materials.

We have done simulations at 1900 MHz using a mobile phone modeled as a $\lambda/4$ monopole antenna above conductive 1 x 5 x 8 cm case covered by PVC. The dimension of the computational space is 225 x 260 x 235 cells. The dielectric properties of the various tissues in the model are obtained from the Brooks airfield base [2] (ε_r, σ) at 1900 MHz. As the model was reclassified, new biological tissues were added and their dielectric properties were added.

For the simulations We use the FDTD method with the Yee cell shown in Figure 1, which is suitable for modeling in biological tissues. As absorbing boundary conditions we use the Berenger's Perfectly Matched Layers (PML)[2].

2 Calculation of SAR distribution

The SAR distribution is calculated in each location i, j, k of the computational space using

$$SAR_{[i][j][k]} = \frac{\sigma_{[i][j][k]} E^2_{ef[i][j][k]}}{\rho_{[i][j][k]}} \quad [\mathrm{W/kg}] \qquad (1)$$

where
$\sigma_{[i][j][k]}$ is the conductivity [S/m].
$\rho_{[i][j][k]}$ is mass density [kg/m^3].
$E_{ef[i][j][k]}$ is rms value of electric field component [V/m].

2.1 Power normalization

The calculated SAR distribution is normalised to the reference radiated power of the mobile phone ensuring total output of 125 mW. The source of the electromagnetic field is modeled by specifying E_z component between the case of the mobile phone and the antenna as

$$E_{z[X0][Y0][Z0]} = Amp \cdot \sin(2\pi f \cdot t) \quad [\mathrm{V/m}] \qquad (2)$$

To achieve steady state the simulation is run for 10 periods of the source signal. Then all real and

[1]http://www.nlm.nih.gov/research/visible/visible_human.html

[2]http://www.brooks.af.mil/AFRL/HED/hedr/reports/dielectric/Report/Report.html

imaginary parts of all field components are found and the impedance of monopole antenna and input power of the antenna is calculated. The normalisation factor between the desired input power of 125 mW and real input power of the antenna is then used for the SAR distribution calculation.

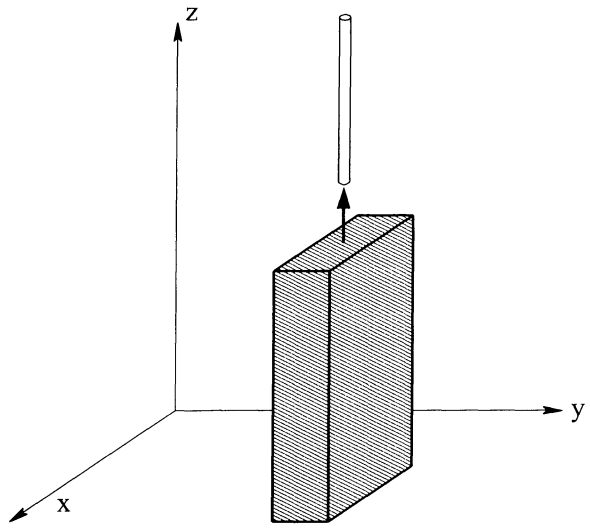


Figure 2: The mobile phone model with $\lambda/4$ long monopole antenna. The arrow represents the source of electromagnetic field in the location of $X0, Y0, Z0$.

3 Simulations

The simulations are done with the original VHP head model as well as the VHP head model with the improved inner ear model. In Table 2 is shown the biological tissues of the original head model an Table 1 are shown the biological tissues of inner ear which were added into the original head model.

biological tissue	ρ[kg/m^3]	ε_r	σ[S/m]
compact bone	1990.0	11.71	0.29
temporal bone	1990.0	11.71	0.29
hearing bone	1990.0	11.71	0.29
cochlea	1007.2	67.06	2.99
vestibullum	1007.2	67.06	2.99

Table 1: Added biologic tissues of the inner ear.

4 Results

For this presentation the interest is focused to a vertical cut referenced z=96. This cut is of interest as it contains the added tissues and the ear nerves nervus cochlearis and nervus vestibularis inferior. The head model, the SAR distribution and the detailed SAR distribution is shown in Figures 3-5 for the original VHP model and in Figures 6-8 for the edited head model.

The power absorbed in the VHP head model is 65 percent of the total radiated power from the mobile phone. The maximum SAR value is 13.9 W/kg and appears in skin of auricle.

biological tissue	ρ[kg/m^3]	ε_r	σ[S/m]
air	1.3	1.00	0.00
blood	1058.0	59.19	2.11
blood vessel	1040.0	43.21	1.11
body fluid	1010.0	68.52	2.09
bone (cancelous)	1920.0	19.21	0.61
bone (cortical)	1990.0	11.71	0.29
bone (marrow)	1040.0	5.35	0.07
cartilage	1097.0	39.98	1.35
cerebellum	1038.0	45.88	1.76
cerebral spinal fluid	1007.2	67.06	2.99
eye (aqueous humor)	1008.9	68.52	2.09
eye (cornea)	1076.0	52.57	1.92
eye (lens)	1053.0	45.24	1.19
eye (sclera)	1026.0	53.42	1.66
fat	916.0	5.34	0.08
glands	1050.0	57.99	1.56
gray matter	1038.0	49.88	1.45
ligaments	1220.0	44.08	1.27
lymph	1040.0	57.99	1.56
muscous membrane	1040.0	43.68	1.28
muscle	1046.8	53.42	1.40
nerve or spine	1038.0	30.74	0.88
skin (dermis)	1125.0	38.71	1.22
tooth	2160.0	11.72	0.29
white matter	1038.0	36.87	0.96

Table 2: Biologic tissues of the Visible Human Head model at 1900 MHz.

biological tissue	max.SAR^1	max.SAR^2
compact bone	79.2	76.3
temporal bone	93.6	105.6
hearing bone	143.3	311.3
cochlea	89.9	28.7
vestibullum	86.6	62.6

Table 3: Max.SAR^1[mW/kg]: maximum values in the biological tissues of the inner ear. Max.SAR^2[mW/kg]: maximum values in area of missing tissues of the inner ear in the original head model.

In the Table 3 is shown the maximum SAR values in the biological tissues of inner ear. For comparison the corresponding values in the locations of the missing inner ear tissues are also shown for the original head model. We can see that the SAR distribution in the soft tissues (cochlea, vestibullum) are increased.

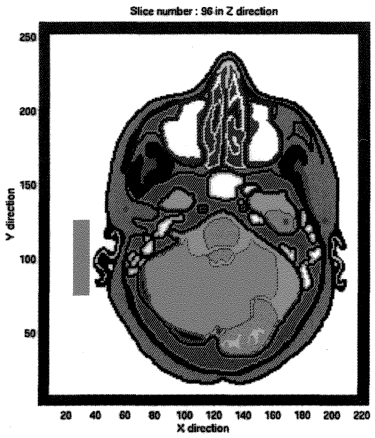


Figure 3: The original VHP head model. Slice z=96.

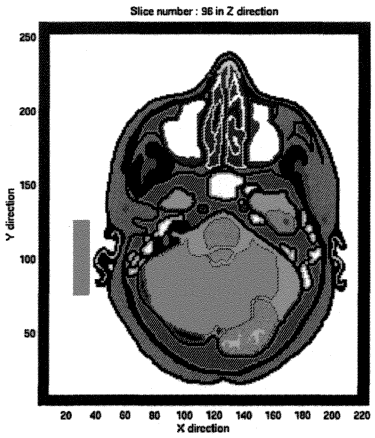


Figure 6: The reclassified VHP head model. Slice z=96.

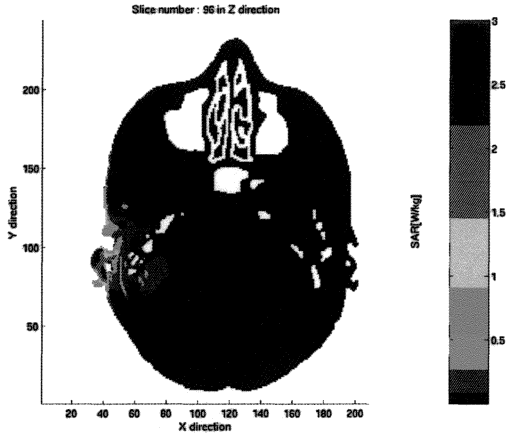


Figure 4: The SAR distribution in slice z=96.

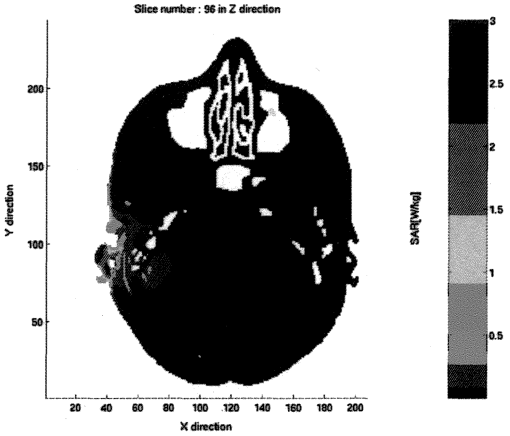


Figure 7: The SAR distribution in slice z=96.

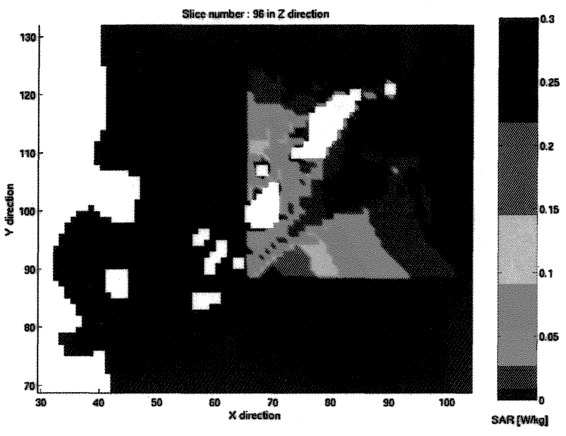


Figure 5: The SAR distribution in area of the missing inner ear in the original head model.

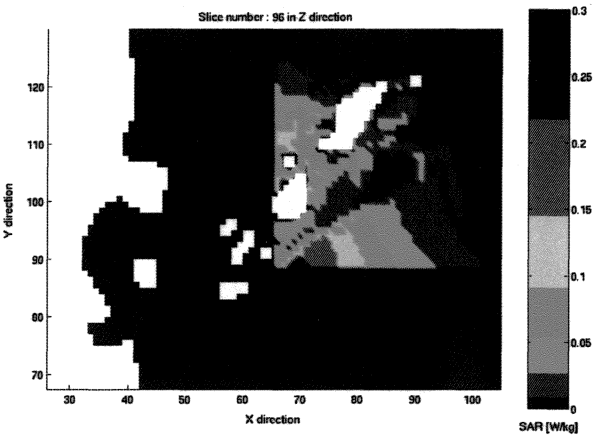


Figure 8: The SAR distribution in the area of inner ear.

5 Conclusion

We have presented results of a dosimetry study of the inner ear using a detailed model of the inner ear at 1 mm resolution and reported that the SAR distribution in the soft tissues are increased. In a further study we will create a more detailed inner ear model at even finer resolution.

References

[1] Taflove, A., "Computational Electrodynamics, The Finite-Difference Time-Domain Method", 1995, Artech House, Inc.

[2] Berenger, J. P., "A perfectly matched layer for the absorption of electromagnetic waves", *J. Computational Physics*, vol. 114, 1994, pp.185-200.

[3] Kunz, K. S.,Luebbers, R. J., "The Finite Difference Time Domain Method for Electromagnetics, 1993, CRC Pres, Inc.

Determination of RF exposure levels and compliance distances of cellular radio base stations

Lennart Hamberg, Nina Lövehagen, Martin Siegbahn and Christer Törnevik

Ericsson Research, Ericsson AB, SE-164 80 Stockholm, Sweden. Lennart.Hamberg@era.ericsson.se, Nina.Lovehagen@era.ericsson.se, Martin.Siegbahn@era.ericsson.se, Christer.Tornevik@era.ericsson.se

Abstract

Procedures for determination of RF exposure from mobile radio base stations by measurements of the Specific Absorption Rate or by calculations of the free space electromagnetic field strengths are presented. The procedures have been applied on different types of base station antennas and distances in compliance with the ICNIRP basic restrictions and reference levels have been assessed. The compliance distance for a micro cell base station antenna was found to be 76 mm at 30 dBm output power and less than 3.5 m for two types of macro base station antennas at output power levels of up to 44 dBm. A cylinder shaped boundary is suggested to show the compliance of RF exposure from macro cell panel antennas with the free space near field reference levels.

1 Introduction

Radio frequency (RF) exposure from radio transmitters, such as cellular radio base stations, are subject to national and international regulations and standards, which are generally based on recommendations from the World Health Organization (WHO) and the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [1]. In Europe, maximum RF exposure levels are specified by the Council Recommendation 1999/519/EC [2] and currently CENELEC TC211 is developing European standards for determination of human exposure to electromagnetic fields from radio base stations.

Basic restrictions for RF exposure in the frequency range used by telecommunication equipment are stated in terms of the Specific Absorption Rate (SAR), which is expressed in units of watt per kilogram (W/kg). Both localized SAR, which is the average SAR in a small volume of 10 g, and the average SAR over the whole body must be in compliance with the basic restrictions.

Assessment of SAR is generally performed by measuring the electric field strength induced in a phantom filled with tissue simulating liquid or by numerical calculations using anatomically realistic human body models. Since determination of SAR might be a very complex procedure, especially for large antenna structures, reference levels are provided for the purpose of comparisons with exposure properties in air. The reference levels are expressed as electric field strength, magnetic field strength and power density. Compliance with the reference levels ensures compliance with respect to the basic restriction but if the reference levels are exceeded, this does not necessarily mean that the basic restrictions are exceeded.

Due to the generally fast decrease of the RF exposure with respect to the distance to the source, compliance distances where the RF exposure meets the basic restrictions or the reference levels can be defined. Similarly, three-dimensional compliance boundaries can be obtained as closed surfaces around base station antennas.

This paper gives a presentation of procedures for assessment of RF exposure from radio base station antennas, which can be used for compliance testing. The procedures are applied on three different base station antennas.

2 Measurements of SAR

The European draft standard CENELEC prEN 50383 [3] defines a protocol for measurements of localized 10 g averaged SAR induced in a homogeneous rectangular phantom filled with tissue simulating liquid. The procedure is applicable to small radio base station antennas in the frequency range 800 MHz to 3 GHz. Measurements of whole body averaged SAR is difficult to perform but compliance with the ICNIRP whole body basic restriction is automatically achieved if the output power of the tested device is less than 1 W [3]. Hence, the measurement protocol is limited to low output power devices.

The draft standard specifies a near field scanner to be used for measuring the electric field (E) in a phantom. SAR is then evaluated by means of the conductivity (σ) and the density (ρ) of the liquid, according to the formula:

$$SAR = \frac{\sigma E^2}{\rho} \text{ (W/kg)}. \quad (1)$$

The size of the phantom is 800 mm × 500 mm × 200 mm and the dimensions are chosen to resemble a human torso. The measurements are performed in a shielded anechoic chamber at the lowest, highest and the center frequency of the tested band, e.g. GSM 900. By measuring SAR for different separations between the

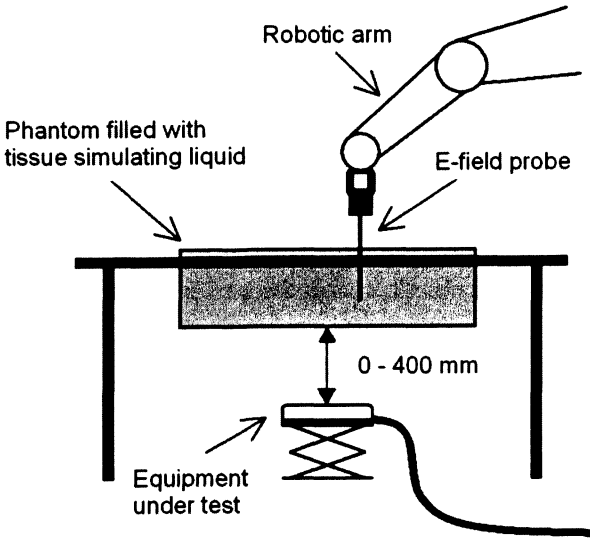


Figure 1: *Example of the measurement setup for SAR measurements of base station antennas. A near field probe is used for measurements of the electric field induced by the equipment under test in a phantom filled with tissue simulating liquid.*

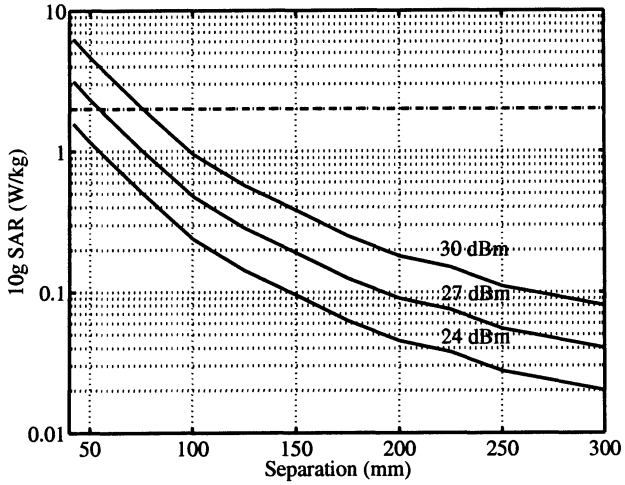


Figure 2: *Maximum 10 g averaged SAR for the tested micro cell antenna. The results were scaled to the output power levels 24 dBm, 27 dBm and 30 dBm. The dashed line corresponds to the ICNIRP basic restriction of 2 W/kg.*

antenna and the phantom, it is possible to find the compliance distance for different configurations of transceivers and output power levels. In Fig. 1, an example of the proposed measurement setup is shown.

The method of flat phantom measurements has been validated by comparison of measured and calculated SAR in a flat phantom with calculated SAR in an anatomically realistic human model [4].

Measurements of maximum 10 g averaged SAR for the micro cell antenna Kathrein 738573 have been performed according to the procedures described in [3]. The dimension of the tested antenna is 300 mm × 150 mm × 30 mm and the gain is approximately 2 dBi. Measurements were performed at the GSM 900 downlink band and the maximum of the results found at the three tested frequencies are shown in Fig. 2. The compliance distances were found to be 40 mm, 55 mm and 76 mm at the output power levels 24 dBm, 27 dBm and 30 dBm, respectively.

3 Field strength calculations

Free space electromagnetic field strengths in the close proximity of a base station antenna can be determined by several numerical calculation methods. Such methods are generally based on solving Maxwell's equations in time or frequency domain using segmented models and in the draft standard [3], the finite-difference time-domain (FDTD) method and the method of moments (MoM) are listed as two applicable techniques. Depending on the method and the computational resources, simple wire mesh models or detailed models based on CAD data can be used to represent the antennas in the calculations.

Based on three dimensional free space field strength results a compliance boundary, where the field strengths meet the reference levels, can be defined. This boundary might have a complex shape and for practical reasons it is convenient to define a simpler geometry at which surface the fields are in compliance with the reference levels. Such geometry could be used in manuals and installation guidelines for radio base stations as an information of the RF exposure from the particular equipment.

The electromagnetic field distributions around a typical GSM 900 antenna have been calculated following the requirements in [3]. The calculations were performed using the software SEMCAD from Schmid & Partner Engineering AG, which is based on the FDTD method and supports creation and imports of CAD objects. The antenna model used for the calculations is shown in Fig. 3a. In Fig. 4, the calculated electric field strength in a horizontal plane is shown. Based on the results for the electromagnetic fields and the power density, RF exposure compliance distance in front of the antenna was found to be less than 3.5 m.

Calculations of the electromagnetic fields around an UMTS antenna have been performed using the MoM software MININEC from EM Scientific Inc. The antenna, which is a cross polarized panel antenna, was modelled by five dipoles and a metallic plate, see Fig. 3b. Since the isolation between the two polarizations is more than 30 dB, only one polarization was considered. The electromagnetic field strength and power density distributions in a volume of 6 m × 4 m × 4 m were calculated at an output power of 44 dBm and in Fig. 5, the compliance boundary for the electric field strength regarding the ICNIRP reference level of 61.4 V/m is shown. In Fig 6 a simpler form of boundary is shown, which can be used as information and guidance for base

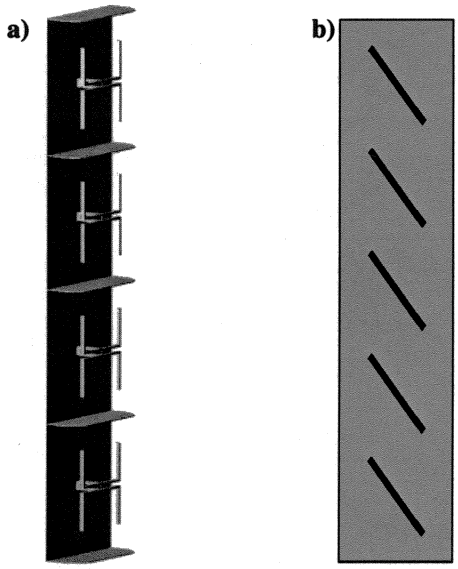


Figure 3: *Models of the tested antennas. The GSM 900 antenna (a) was modelled with the FDTD software SEMCAD while the UMTS antenna (b) was represented by five dipoles in front of a metallic plate. The dipoles were tilted 45 degrees in order to simulate one of the two polarizations of the antenna.*

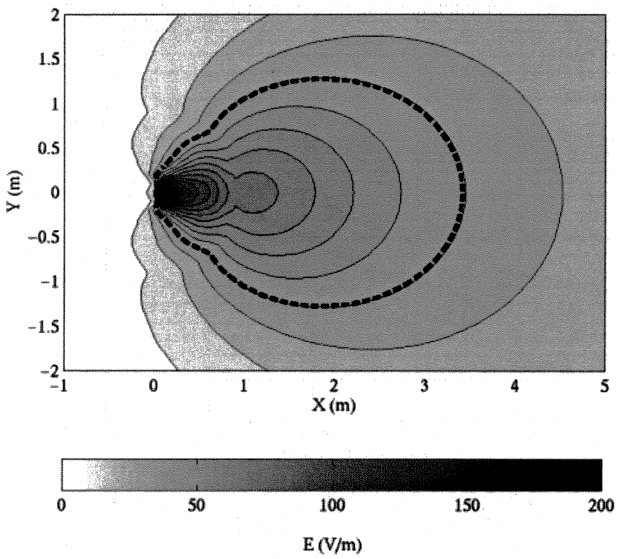


Figure 4: *Electric field strength in the horizontal plane for the GSM 900 antenna with an output power of 43 dBm. The dashed line shows the compliance boundary regarding the ICNIRP reference level of 41 V/m.*

station installations. The boundary is in the form of a cylinder with a diameter of 3 m and with a height that equals the height of the antenna plus 0.2 m. The cylinder is placed so that the distance from the back side of the antenna to the boundary is 0.1 m. The boundary is valid for this particular antenna with 44 dBm output power but the cylinder surface can be applied to panel antennas in general.

4 Conclusions

Measurements of maximum 10 g averaged SAR for a micro cell base station antenna have been performed following the procedures in the European draft standard CENELEC prEN 50383. The compliance distance with respect the ICNIRP basic restriction of 2 W/kg was found to be 76 mm for 30 dBm output power.

Calculations of the electromagnetic fields around two typical base station antennas for the GSM 900 and UMTS downlink bands have been performed using the methods FDTD and MoM, respectively. The compliance distance with respect to the ICNIRP reference levels was found to be less than 3.5 m for the GSM 900 antenna at 43 dBm output power and less than 3.0 m for the UMTS antenna at 44 dBm output power.

A simplified boundary in the form of a cylinder is suggested to show the compliance of RF exposure from macro cell panel antennas with the free space near field reference levels.

5 References

[1] ICNIRP, Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz), International Commission on Non-Ionizing Radiation Protection (ICNIRP), *Health Physics*, vol. 74, no. 4, pp. 494-522, April 1998.

[2] Council Recommendation 1999/519/EC of 12 July 1999 on the limitation of exposure of the general public to electromagnetic fields (0 HZ to 300 GHz) *(Official Journal L 197 of 30 July 1999).*

[3] CENELEC prEN 50383, Basic standard for the calculation and measurement of electromagnetic field strength and SAR related to human exposure from radio base stations and fixed terminal stations for wireless telecommunication systems (110 MHz - 40 GHz)", Technical Committee 211, European Committee for Electrotechnical Standardization (CENELEC), European Draft Standard, November 2001.

[4] L. Hamberg, N. Lövehagen, M. Siegbahn and C. Törnevik, Determination of RF Exposure Compliance Distances of 900 MHz and 1800 MHz Micro Cell Base Station Antennas Using SAR Measurements, submitted to *IEEE Transactions on Microwave Theory and Techniques,* November 2001.

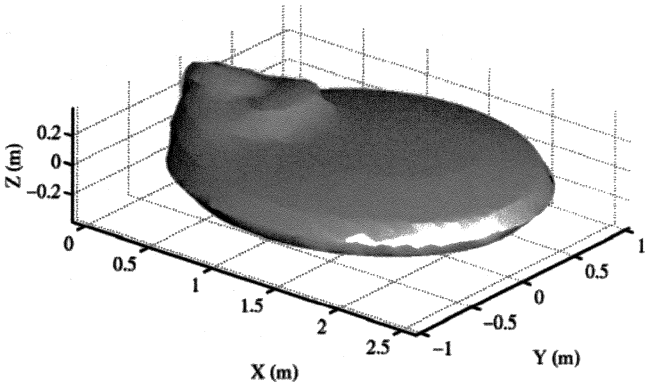


Figure 5: *The compliance boundary outside of which the electric field strength is in compliance with the ICNIRP reference level of 61.4 V/m for the UMTS antenna. The output power was set to 44 dBm.*

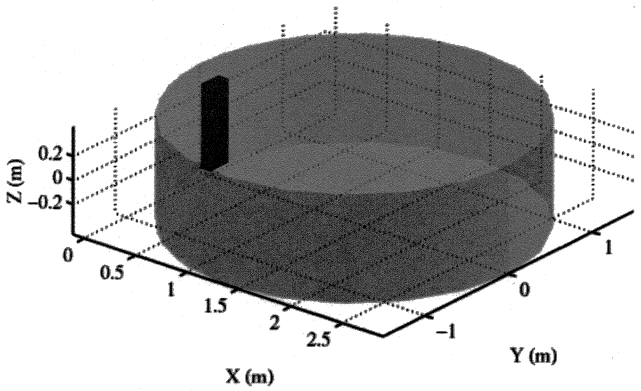


Figure 6: *A simplified boundary for the UMTS antenna outside of which the electromagnetic filed strengths and the power density are in compliance with the ICNIRP reference levels at 2140 MHz. The height of the cylinder is 0.9 m and the diameter is 3 m.*

Image reconstruction in a microwave tomograph for mammography

A. Danielsson[1], M. Persson[1], J. Howard[2]
[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden.
E-mail: andreas.danielsson@elmagn.chalmers.se
[2]Plasma Research Laboratory, Australian National University, Canberra A.C.T 0200, Australia.

Abstract

In the future mammography using tomographic microwave imaging techniques might become useful in the detection of breast cancer tumours. Here we describe the method and the physical basis for it. We specify some demands that has to be put on the image reconstruction algorithm, in the microwave tomograph, in order to be useful for breast cancer tumour imaging.

1 Introduction

All over the world breast cancer is a serious health problem for women. Early diagnosis is crucial for successful treatment and impacts long–term survival rate. The routine method for breast cancer screening is x–ray mammography, which is well recognised and acknowledged for its ability to detect and characterise breast tumours. However there are several limitations, such as high false positive detection rate, and the false negative rate is also not insignificant. This is due to the fact that x–ray mammography images a tumour of soft tissue, in the background of other soft tissue, which gives a low contrast. Therefore there is a need for complementary and/or alternative efficient detection methods. Ultrasound and contrast–enhanced MRI are already used and are effective in the diagnostic evaluation of mammographically detected breast lesions. Unfortunately these methods turn out to be either not sensitive/specific enough or too costly for mass screening purposes.

To improve the breast cancer imaging technologies it has been proposed that one could use a microwave imaging technique. There are several potential advantages with microwave imaging compared to x–ray mammography: Due to the high contrast in dielectrical properties between the normal and malignant tissue, the specificity to distinguish between the tumour and normal tissue is high, therefore it may also be possible to detect smaller tumours at an earlier stage than with conventional mammography. A summary of measurements of the dielectric properties of breast tissue and tumours can be found in Hagness et al, [1]. Microwave imaging uses non–ionising radiation and the microwave radiation levels needed are well within safety standards. Finally the patient comfort is higher since no breast compression is required. On illuminating the breast, microwave radiation is both scattered and absorbed due to the spatial variation in the dielectric permittivity and conductivity of the breast tissue. Measurement of the scattered radiation (amplitude and phase) are taken at several different transmitter/detector locations surrounding the breast. From these measurements one attempts to unfold the dielectric properties of the scattering medium using appropriate inverse methods.

2 Tomography configuration and measurement

In figure 1 we se the principal configuration of a microwave tomograph. The object to image with unknown dielectric parameters, ϵ and σ, (the shaded area) is surrounded by transmitters and receivers (the black dots). A tomographic measurement consists of

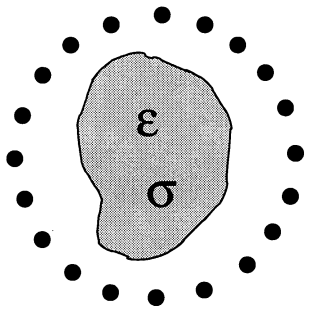


Figure 1: *The microwave tomography configuration. The black dots represents transceivers and the shaded area the object to be imaged.*

several independent illuminations of the object from different directions, i.e. successively one antenna at the time is used as transmitter and the others are used

to register the scattered radiation. This is repeated until all antennas has been used as transmitters in order to get a sufficiently large number of measurements to use in the reconstruction procedure.

3 Inverse algorithms

Many papers have been published on inverse algorithms for the Maxwell equations. The majority of them are analytical or semi–analytical algorithms utilising different approximations, such as the Born approximation, in the inverse step, see for instance Zhang and Liu, [2], and Chew and Wang, [3]. These methods are efficient and valuable when imaging weakly scattering objects but due to the approximations used they are of limited interest in breast cancer imaging. In order to benefit from the fact that there is such a high contrast between the tumour and the surrounding tissue instead a general algorithm is required that is independent of the dielectric properties of the scattering objects.

3.1 Finite difference time domain based inversion algorithm

An algorithm that is better suited and able to image strongly scattering objects is described by Gustafsson, [4]. This algorithm relies on finite difference time domain (FDTD) simulations of the wave propagation in the object and the unknown dielectric properties are successively updated to get a better fit of the simulated signal to the measured signal.

We assume that the measurements have been conducted, as described previously, giving a set of electric and magnetic fields on the receiver locations, i.e for each transmitter location m,

$$E^{(m)}(\mathbf{x},t,m) \quad \text{and} \quad H^{(m)}(\mathbf{x},t,m). \tag{1}$$

In the reconstruction procedure an initial dielectric distribution is assumed, $\alpha(\mathbf{x}) = \{\epsilon(\mathbf{x}), \sigma(\mathbf{x})\}$, that the material is non–magnetic with $\mu = \mu_0$ which is true in general for biological tissue, and the propagation of the field is simulated using the same source field as in the measurement. Thus the image reconstruction procedure is the same as minimising the functional J,

$$J(\alpha) = \sum_{m=1}^{M} \int_0^T \int_{\partial\Omega} \Phi\Big(|E(\mathbf{x},t,m) - E^{(m)}(\mathbf{x},t,m)|^2 + |H(\mathbf{x},t,m) - H^{(m)}(\mathbf{x},t,m)|^2\Big) dS dt, \tag{2}$$

where the integration dS is over the boundary of the imaging area, Φ is a weight function and M is the number of independent measurements. In order to minimise the functional a gradient based optimisation algorithm is used. Solving the adjoint- or dual-problem of Maxwell's equations expressions for the gradients are found. The gradient is

$$\mathbf{G}(\mathbf{x}) = \{G_\epsilon(\mathbf{x}), G_{\sigma/\langle\sigma\rangle}(\mathbf{x})\} \tag{3}$$

where

$$G_\epsilon(\mathbf{x}) = -2\sum_{m=1}^{M} \int_0^T \tilde{E}(\mathbf{x},t,m) \cdot \partial_t E(\mathbf{x},t,m) dt \tag{4}$$

$$G_{\sigma/\langle\sigma\rangle}(\mathbf{x}) = -2\langle\sigma\rangle \sum_{m=1}^{M} \int_0^T \tilde{E}(\mathbf{x},t,m) \cdot E(\mathbf{x},t,m) dt, \tag{5}$$

$E(\mathbf{x},t,m)$ is the solution of the ordinary FDTD simulation, $\tilde{E}(\mathbf{x},t,m)$ is the solution to the adjoint- or dual-problem and $\langle\sigma\rangle$ is a scaling constant to make the solution more robust. In the update of the dielectric parameters it is possible to invoke *a priori* information, such as specifying the range of values where the parameter could be expected to be found within and smoothness conditions.

If the functional is not small enough another iteration is performed, a new simulation is done with the updated set of dielectrics and the gradients are recomputed and the functional is minimised. This process is repeated until the functional has reached the desired minima in order to give the desired accuracy in the reconstructed image.

3.2 Chirp–pulse microwave computed tomography

In order to get an idea of the performance and usefulness of the above mentioned algorithm reconstructions from the above mentioned algorithm are compared to reconstructing the same object using the chirp–pulse microwave computed tomography (CP–MCT) algorithm described by Bertero et al, [5].

This algorithm attempts to repost the microwave imaging problem to the well known problem of reconstructing the image in a x–ray computerised tomograph where the filtered backprojection algorithm, [6], can be used to reconstruct the image from Radon transforms of the x–ray attenuation. The CP–MCT algorithm takes Radon transforms of the attenuation of microwaves propagating straight through the object and uses the filtered backprojection algorithm to create the image.

To create the Radon transforms of the object's attenuation of microwaves a method called microwave time-delay spectroscopy (MTDS) is used, [7]. The

MTDS uses a frequency chirped input pulse and the output is the frequency component that has the shortest propagation time through the material. The assumption made here is that this is also the shortest way, i.e. the straight path between the transmitter and the receiver, thus it is possible to obtain the Radon transform of the microwave attenuation. This of course also put a limitation on the algorithm since the approximation that the fastest way is also the shortest only holds for a weak scatterer.

The Radon transform, f_θ, is not measured directly by the CP-MCT method but something, g_θ, that is related to the Radon transform by a convolution,

$$g_\theta = K * f_\theta, \tag{6}$$

K is a unknown blurring point spread function (PSF). In order to determine the Radon transform of an object we see that it is necessary to determine the PSF, K, which can be done by making a measurement on a known object, with known Radon transform, and then calculate the PSF. The Radon transform of a unknown object can thus be computed by an inverse convolution of eq. (6).

4 Results

Results from simulations of a few simple test cases will be shown on the poster session. Comparison between the two described algorithms for the test cases will be shown.

References

[1] S. C. Hagness, A. Taflove and J. E. Bridges: Three–Dimensional FDTD Analysis of a Pulsed Microwave Confocal System for Breast Cancer Detection: Design of an Antenna–Array Element. *IEEE Trans. Antennas Propagat.*, 47(5):783-791, May 1999.

[2] Z. Q. Zhang and Q. H. Liu: Reconstruction of axisymmetric media with an FFHT-enhanced extended Born approximation. *Inv. Prob.*, 16(5):1281-1295, Oct. 2000.

[3] W. C. Chew and Y. M. Wang: Reconstruction of Two–Dimensional Permittivity Distribution Using the Distorted Born Iterative Method. *IEEE Trans. Medical Img.*, 9(2):218-225, June 1990.

[4] M. Gustafsson: *Wave splitting in direct and inverse problems.* Doctoral Thesis, Lund Institute of Technology, Lund, 2000.

[5] M. Bertero, M. Miyakawa, P. Boccacci, F. Conte, K. Orikasa and M. Furutani: Image Restoration in Chirp–Pulse Microwave CT (CP–MCT). *IEEE Trans. Biomed. Eng.*, 47(5):690-699, May 2000.

[6] A. C. Kak and M. Slaney: *Principles of Computerized Tomographic Imaging*, IEEE Press, 1999.

[7] J. H. Jacobi and L. E. Larsen: Microwave interrogation of dielectric targets–Part 2: By microwave time delay spectroscopy, *Med. Phys.*, 5:509-518, 1978

The importance of Specific Absorption Rate for subjective symptoms among mobile phone users

Jonna Wilén, Monica Sandström and Kjell Hansson Mild

National Institute for Working Life, 907 13 Umeå,
Sweden, E-mail: mild@niwl.se

Abstract

In a previous epidemiological study [1] we found as a side-finding higher prevalence of some subjective symptoms among people with longer calling times per day. In this extrapolative study we have selected four of the most commonly used GSM devices from the epidemiological study and measured SAR_{1g} in three defined volume in the head. The aim is to evaluate the importance of SAR and two new exposure parameters; Specific Absorption per Day (SAD) and Specific Absorption per Call (SAC) for the prevalence of subjective symptoms such as headaches, warmth sensation and fatigue.

The standards concerning exposure to radiofrequency fields [2,3] are exclusively based on acute thermal effects. Heating effects due to absorption of radiofrequency field emitted by commercially available mobile phone (MP) devices are less likely to occur. On the other hand the scientific data today is not enough to ensure that there are no possible non-thermal effects.

In an epidemiological study [1,4] were we studied the prevalence of subjective symptoms among mobile phone users we found as an interesting side-finding that the prevalence of many of the subjective symptoms increased with increasing calling time and number of calls per day. This was valid even in an adjusted analysis where confounding factors such as psychosocial workload, gender and occupation were taken into consideration.

Since a certain amount of radiofrequency energy will be absorbed by the user during a MP call it was of interest to go one step further and study whether the prevalence of subjective symptoms increases with increasing rate of absorption and the total absorbed energy per call and per day. In this extrapolative study, we have selected people from the epidemiological study who used any of the four most common MP devices. We make use of the information about prevalence of symptoms, calling time per day and number of calls per day from the epidemiological study and combine it with measurements of Specific Absorption Rate (SAR) of those most commonly used MP devices.

Different MP devices deposit the energy at different anatomical localisation. The antenna, the chassi as well as the internal structure of the device will affect the absorption. We have therefore defined three volumes in the head (Figure 1) and measured the maximum SAR averaged over a cube of one-gram tissue (SAR_{1g}) in each volume for the four GSM devices. The SAR distributions for the four devices included in this study are shown in Figure 2. Two new exposure parameters Specific Absorption per Day (SAD) and Specific Absorption per Call (SAC) have been introduced. These reflect the total amount of radio frequency energy absorbed in the volume.

The results indicates that SAR, especially for longer calling times per day, might be an relevant exposure parameter for the prevalence of subjective symptoms. In addition, SAD that reflects the total energy absorbed during a day also seems to be a exposure parameter of interest for the prevalence of subjective symptoms.

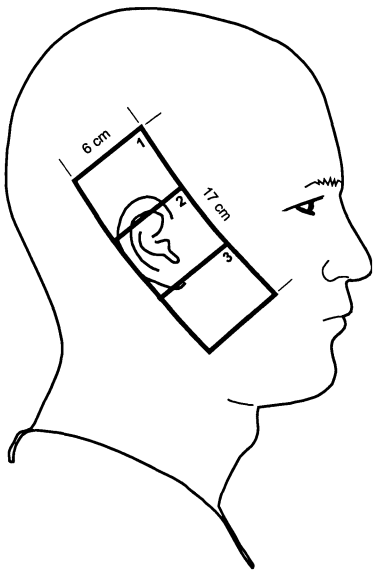


Figure 1. Definition of the three volumes used, above the ear (Vol. 1), on the ear (Vol. 2) and below the ear (Vol. 3).

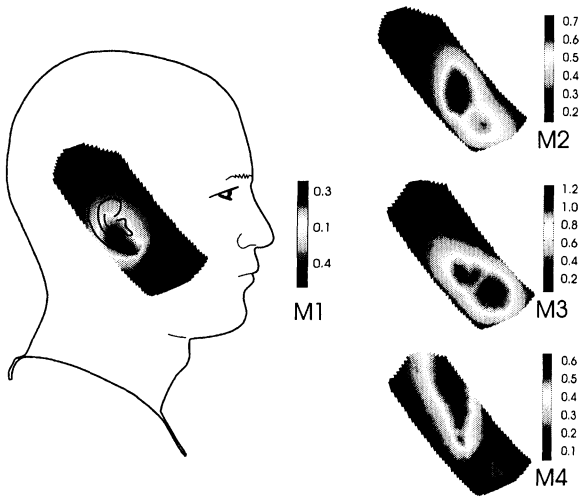


Figure 2. Schematic drawing of the distribution of SAR (W/kg) for the devices used in the study (M1-M4). The figures show the measured SAR on the right hand side of the phantom.

[1] Oftedal G, Wilén J, Sandström M, and Hansson Mild K. 2000. Symptoms experienced in connection with mobile phone use. Occup. Med.-Oxf. 50:237-245.

[2] ICNIRP 1998. Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). International Commission on Non-Ionizing Radiation. Health Phys 74:494-522.

[3] IEEE 1999. IEEE Standard for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz. Standard. C95.1-1999. New York.

[4] Sandström M, Wilén J, Oftedal G, and Hansson Mild K. 2001. Mobile phone use and subjective symptoms. Comparison of symptoms experienced by users of analogue and digital mobile phones. Occup. Med.-Oxf. 51:25-35

Wave-Propagation from Medical Implants – Influence of Arm Movements on the Radiation Pattern

Anders J Johansson[1], Anders Karlsson[2]
[1]Dept. of Electroscience, Lund University, PO. Box 118, 221 00 Lund. E-mail: anders.j.johansson@es.lth.se
[2]Dept. of Electroscience, Lund University, PO. Box 118, 221 00 Lund. E-mail: anders.karlsson@es.lth.se

Abstract
The influence of a patient's arm movements on the radio channel from a medical implant to an outside transceiver has been simulated. This as a part of a study of the feasibility of using a radio-link between the implant and a controller/programmer at the hospital. The result shows that the influence of the arms on a medially placed transmitter is limited. The radiation pattern from a simplified implant model has a shape that is very attractive for the medical application, giving a good communication link in both the upright and prone position of the patient.

1 Medical implant communication

1.1 Pacemaker communication
Implanted pacemakers are today routinely used for correction of heart arrhythmias [1]. A modern pacemaker is programmable from the outside in order to set parameters such as pace rhythm, output levels, sensitivity for motion etc. Data is also recorded and stored in the pacemaker for later retrieval. This data may consist of performance data, battery status and measurements of the intracardial ECG of the patient. Today the communication link used is a magnetic coupling between two coils: one inside the pacemaker and one placed at the chest of the patient. This link has two main drawbacks: the outside coil must be placed on the correct spot with relative high precision and the transmission speed is slow. A higher transmission speed and a simpler interface from the view of the operator would give many benefits for the patient. Faster access to data in the pacemaker gives a shorter time in the doctor's office and allows larger amounts of data to be retrieved from the device in an easy way which would give the patients physician better data to act on when adjusting the pacemaker and prescribing additional treatment. A communication link that reaches 1-10m instead of only working at touch distance would also give a larger flexibility when programming and monitoring the pacemaker, both at the hospital and at home.

1.2 Transmission through the human body
In order to design establish a reliable communication channel between two points there are a number of variables that have to be known. One fundamental parameter is to accurately know the radio channel. There are well-known channel-models that describe the conditions outside the human body such as those in the hospital, as this case is no different from mobile communications channels. What we today have very limited knowledge of is the transmission channel from the implant to the free air space surrounding the patient. The human body is not a good media in which to transmit an electromagnetic wave. It has a high electric conductivity which gives rise to a large path loss within the body.

1.3 MICS
The new standard Medical Implant Communications Service (MICS) for communication with medical implants use the 403-405 MHz band [2]. This band is previously utilized by weather balloons and is now shared between the two applications. This gives very stringent demands on the medical application as this should not impede the metrological use. Output power is thus set to a maximum of 25 uW EIRP and the maximum bandwidth used is set to 300kHz. EIRP should be measured from the whole radiating system, i.e. the radiated power from a body with an implanted pacemaker. For test and certification purposes there is a cylindrical human torso phantom specified. Details about the transmit protocol, modulation etc. is not specified and is left open for the user to decide. The band is available only for communication with medical implants – further relaying of data should be done utilizing other standards.

2 Simulations

In the simulations presented below the impact of the patient moving his limbs is investigated. The patient is simulated as a 180 cm tall male mannequin (figure 1).

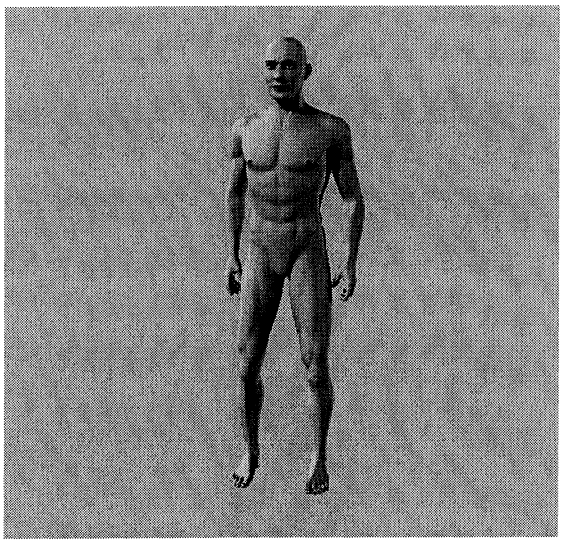


Figure 1. The mannequin generated by Poser.

The shape of the body is taken from the "Poser" program [3], and can be taken as a realistic human body, albeit a bit idealistic. The positions of all the mannequins' limbs and joints are adjustable in the program. Eight different positions of the arms were generated this way, which can be seen in figure 2.

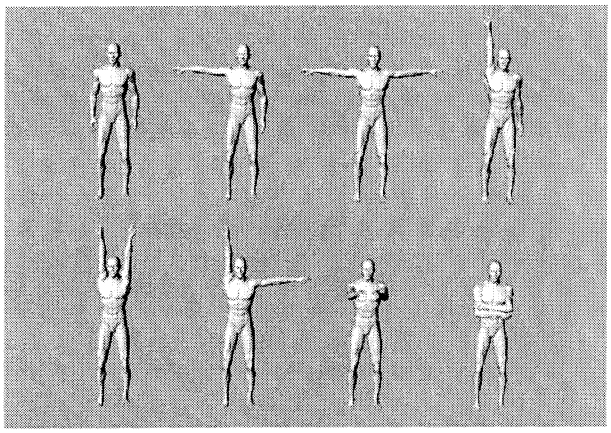


Figure 2. The eight different arm positions that were simulated.

These shapes were then transferred to the SEMCAD FDTD-simulator [4]. The shapes were exported as hollow shells. In the simulator these were then filled with a homogenous matter with the same electrical properties as muscles (Table 1). The parameters are taken from the parametric model of [5]. Muscle tissue was chosen as the influence of the arm positions was the focus of the investigation, and the arms consist mainly of muscle tissue. A simplified pacemaker with a wire antenna is placed inside the mannequin, in the medial line towards the front (figure 3). The antenna is simulated as a perfectly conducting wire without any isolation. The length of the wire is 40 cm. This is taken from the approximate actual length of a pacemaker lead. This is not a resonant length of a wire antenna placed in muscle tissue operating at 403 MHz, but the mismatch will be very low. This as a wire antenna in a conductive medium has a logarithmic damping of the currents along the wire. The damping is due to losses to the surrounding media, as shown in [6]. The pacemaker was simulated as a perfectly conducting short cylinder.
The discretization was non-uniform with a maximum voxel size of 1/10 of a wavelength at all places.

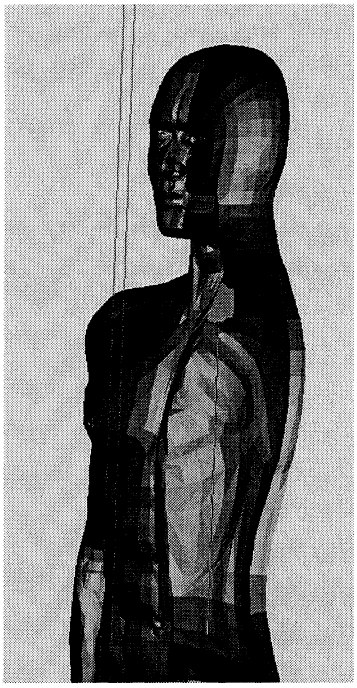


Figure 3. Placement of simplified pacemaker model.

Tissue	Permittivity	Electric Conductivity
Muscle	57.1	0.80

Table 1.

3 Results

3.1 Farfield.

The far-field of the system were studied. As can be seen in figure 6 and 7 the influence of arm movements is very limited when the arms are not placed were the greatest currents on the body are. Figure 4 shows the currents on the surface of the mannequin. When the arms are stretched forward or folded across the breast the far-field radiation pattern is affected in a larger degree.

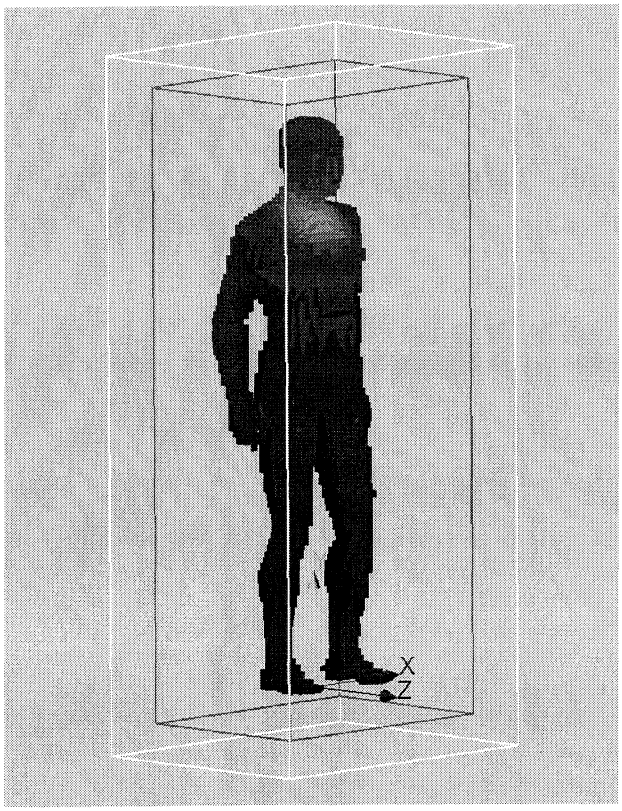


Figure 4. Intensity of currents on the body surface.

In figure 5 the three-dimensional radiation pattern can be seen. Note that we have two dominant minimas in the pattern: one in the posterior direction of the neck and one at the bum as seen from the pacemaker. The full frontal sphere is approximately uniformly illuminated. This is a very good pattern from the view of MICS, as we will have a good channel both when the patient is standing or sitting up facing the outside transceiver, and when the patient is laying on his back in a bed.

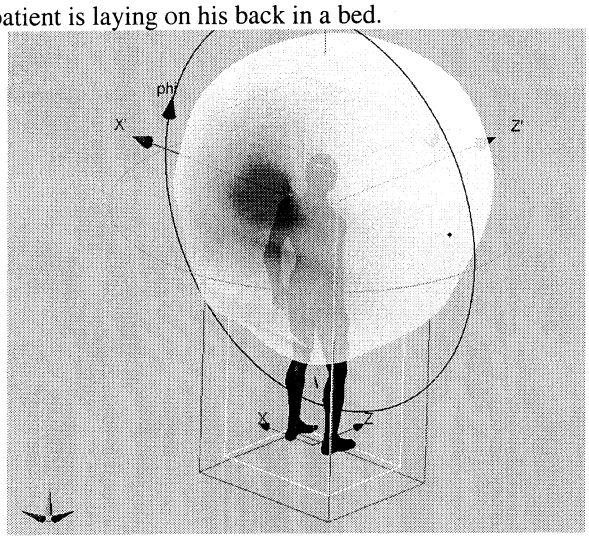


Figure 5. 3D-farfield pattern superimposed on the phantom.

The calculated gain in the upper half of the frontal half-sphere is –28dBi +0/-10dB depending on arm position (figure 6).

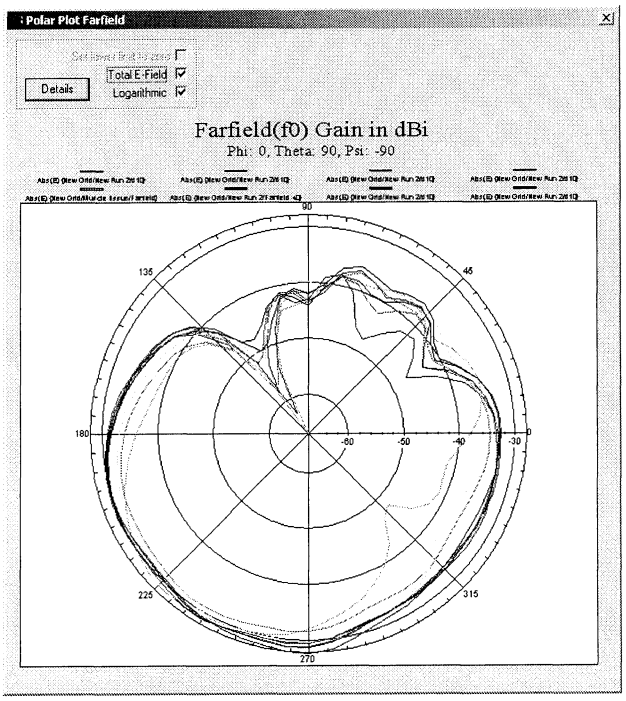


Figure 6. Plot of the gain patterns for the eight arm positions in the sagittal plane. Superior at 180°, Anterior at 270°.

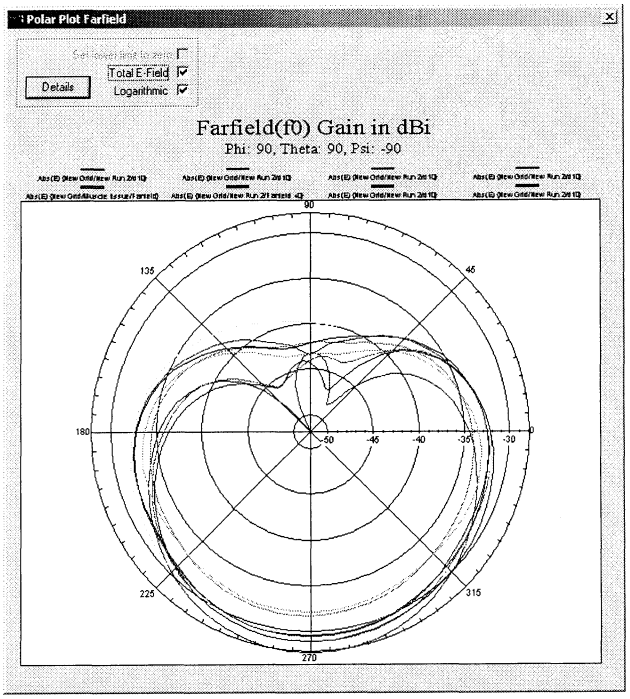


Figure 7. Plot of the gain patterns for the eight arm positions in the transversal plane. Anterior at 270°, Dexter at 180°.

4 Acknowledgment

This investigation was done in the Competence Center for Circuit Design at the department of Electroscience at Lund University.

5 References

[1] J. G. Webster, ed., "*Design of Cardiac Pacemakers*", IEEE Press 1995, ISBN 0-7803-1134-5

[2] Code of Federal Regulations, USA, Title 47 Telecommunication, Part 95, Available at www.acess.gpo.gov

[3] POSER, Curious Labs Inc. 655 Capitola Road, Suite 200, Santa Cruz, CA 95062, www.curiouslabs.com

[4] SEMCAD, Schmid & Partner Engineering AG, Zeughausstrasse 43, 8004 Zurich, Switzerland www.semcad.com

[5] C. Gabriel. *Compilation of the Dielectric Properties of Body Tissues at RF and Microwave Frequencies*, Final report for the period 15 December 1994 – 14 December 1995, Prepared for AFOSR/NL Bolling AFB DC 20332-0001Physics Department , Kings College, London WC2R 2LS, UK

[6] R. W. P. King and G. S. Smith, *Antennas in Matter*, the MIT Press, Cambridge, 1981

Towards a 4G IP-based Wireless System Proposal

Tony Ottosson[1], Mikael Sternad[2], Anders Ahlén[2] Arne Svensson[1] and Anna Brunström[3]

[1] Dept. Signals and Systems, Chalmers University of Technology, SE-41296 Göteborg, Sweden
[2]Signals and Systems, Uppsala University, PO Box 528,SE-751 20 Uppsala, Sweden.
[3]Dept. Computer Science, Karlstad University, SE-65188 Karlstad, Sweden

Abstract

A packet switched wireless cellular system with wide area coverage, high throughput and high spectral efficiency is proposed. Smart antennas at both base stations and mobiles improve the antenna gain and improve the signal to interference ratio. The small-scale fading is predicted and a slotted OFDM radio interface is used, in which time-frequency bins are allocated adaptively to different mobile users, based on their predicted channel quality. This enables efficient scheduling among sectors and users, as well as fast adaptive modulation and power control. We here estimate the spectral efficiency of the suggested downlink. The resulting channel capacity grows with the number of simultaneous users and with the number of antenna elements in terminals. A high efficiency, around 4 bits/s/Hz at 16dB SNR is attained already for a moderate numbers of users and terminal antennas. An outline is given of research pursued within the PCC Wireless IP Project to improve and investigate this type of system.

1 Introduction

With the current trend towards integration of systems and services, wireless data services for mobile vehicular users will require throughputs and Quality of Service (QoS) levels that approach those obtainable in present WLAN's. For economic reasons, this challenge cannot be met solely by an expansion of the bandwidth. A much higher spectral efficiency will be a key feature of any acceptable future 4G solution.

The Wireless IP project [1] within the Swedish research program for Personal Computing and Communications (PCC) studies problems that are important in the evolution of UMTS toward higher data rates, as well as in future 4G technologies for mobile systems. Our goal is to improve the spectral efficiency for packet data, in particular IP traffic, with sufficient quality of service for various traffic classes.

We here outline a radio interface for a spectrally efficient and highly flexible packet switched system that provides high throughput in both the downlink and the uplink. This concept aims at a much higher spectral efficiency than obtained by 3G systems and their enhancements, such as the HSDPA mode of WCDMA, release 5 [2]. Our ultimate goal is a capacity of $50-100$ Mbit/s per sector or beam of a base station, with a somewhat lower capacity in the uplink. This would provide wireless and mobile users with a capacity that corresponds to the present wireline Ethernet.

Our work so far, summarized here, has resulted in methods that can be expected to provide a downlink spectral efficiency of around 4 bits/s/Hz. This corresponds to 20Mbit/s per base station sector/beam for 5MHz channels.

What are the difficulties involved in obtaining a significantly improved spectral efficiency? The transmission of packet data to/from mobile vehicular users encounters several complications:

- A significant fraction of the users will be on the move, and encounter channel properties that vary on all time-scales.
- Uplink transmissions will not be precisely synchronized[1].
- A constrained transmit power in particular in the uplink will limit the transmission range and/or the data rate.

To overcome these difficulties, we develop and combine tools and methods that are outlined below in Section 2. The properties of the proposed downlink are specified in Section 3 and its capacity is analyzed in Section 4. Some open issues under present study are discussed in Section 5. For instance, our effort to improve the efficiency of lower layers are complemented by efforts to improve the transport layer for wireless links.

2 Elements of a 4G Radio Interface

To obtain high data rates in wireless packet data transmission is, in itself, not particularly difficult, but to do this with wide area coverage, without a very dense and expensive infrastructure, is more challenging. To provide high data rate services also to vehicular users with reasonable quality of service is more challenging still, due to their rapidly varying channels. To attain a high spectral efficiency for vehicular users is the toughest challenge of all.

[1]Exact synchronization among terminals and base stations is a very difficult task, that would impose tremendous requirements on the hardware and on the algorithms of the system.

Many types of solutions could be considered, but very few are able to meet all these requirements simultaneously. Unsynchronized packet data systems, perhaps with a collision avoidance scheme, would encounter problems in heavily loaded situations. The same is probably true for ad-hoc radio networks in unregulated frequency bands, at least when wide area coverage, heavy usage and adequate quality of service are to be attained simultaneously.

We have therefore focused on slotted cellular systems in (new) regulated frequency bands, systems that partly build on the 3G philosophy, but that in important aspects depart radically from it.

UMTS and earlier cellular systems are based on averaging, which by necessity limits its spectral efficiency. We instead propose an approach that utilizes feedback control and fast optimization of the resources at hand. Coding overhead is to be minimized. Significant steps in this direction are today taken within 3GPP with the aim of improving the downlink data rate of UMTS [2], and one of our intents is to investigate how far this road might lead.

One of the many difficulties is to avoid ending up with a system with an overwhelming demand for feedback control bandwidth. There is also the risk of developing advanced estimation algorithms that estimate parameters that are in the end of little relevance for the system performance. Such aspects force us to consider the system design as a whole, rather than investigating different algorithms in isolation.

The main purpose of this paper is to outline our provisional target system proposal, so that it can be submitted to discussion, criticism and improvement.

We propose a cellular slotted FDD system and the use of multiple access via OFDM in both uplink and downlink. Uplink transmissions from different mobile users are assumed to be only roughly synchronized and slotted FDD avoids the interference problems encountered in asynchronous TDD systems. Bursty but delay-insensitive data, and streaming audio/video, are expected to comprise major parts of the traffic. Delay-sensitive traffic will also be present. The use of slotted OFDM enables a more fine-grained resource optimization, in both time and frequency, as compared to the use of either TDMA or CDMA. Within such systems, we are studying the application and co-optimization of several methods, introduced below, for improving the spectral efficiency.

2.1 Channel Quality Feedback, Scheduling and Link Adaptation

The variation of the received power with time and frequency is illustrated by Figure 1 for one particular user and fading pattern. Channels to and from different users will fade independently. With many potential users present, the channel could ideally almost always be allocated on a time-frequency slot basis to users who encounter favorable conditions within these particular slots. This effect is sometimes denoted multiuser diversity [3]. The exploitation of this effect requires *channel predictors* (in the mobiles for the downlink and at the base station for the uplink), that predict the short-term fading for each user. It also requires the feedback of the resulting slot quality estimates to a *scheduler*, which then allocates time-frequency slots among the users of each base station sector. The allocation is based on SNR estimates for each user, taking goals and constraints such as throughput, user satisfaction, delay and QoS into account [4, 5].

In the downlink, the time-frequency bins to be allocated are proposed to have a duration of 0.667ms and a width of 200kHz. Their duration corresponds to the slot length of UMTS, but a bandwidth of 5 MHz will be partitioned into 25 bins. The channel will be approximately time and frequency invariant within these time-frequency bins.

The uplink allocation will be complicated by the need for significant guard-bands, due to lack of exact clock synchronization between mobile transmitters. We would here share each 0.667ms time-slot among fewer users, possibly by the use of adaptive bandwidth allocation, and use a few narrow-band channels for short packets [6].

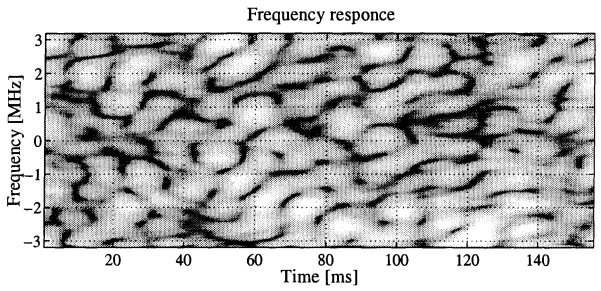


Figure 1: Time-frequency representation of an estimated channel obtained from real measurement data on a 6.4MHz channel. White color denotes high power whereas dark color denotes low power. The dynamic range and the speed of the mobile is approximately 40 dB and 50 km/h, respectively. The coherence bandwidth is 0.6 MHz.

The channel quality feedback enables *fast link adaptation* (adaptive modulation, adaptive coding and power control). Robustness against the use of too high modulation formats due to channel estimation errors can be obtained, at little coding overhead. One tool is Hybrid type-II ARQ in link level retransmission of delay-insensitive traffic [4]. The SNR limits for using different modulation formats can furthermore be optimized with the prediction accuracy taken into account [7]. The system efficiency might be improved further by using power control by waterfilling [8] among the users in different frequency bins within a time slot, but this remains to be investigated.

2.2 Prediction and Predictive Scheduling

The attainable throughput, and in particular also the fulfillment of QoS constraints, can be improved if the scheduling can be performed over a time horizon comprising several slots [5]. Predictors and predictive schedulers can be implemented at reasonable computational complexity [5].

From extensive investigations on broadband measurement data, we have found that the received signal power can be predicted with reasonable accuracy for time intervals corresponding to a movement of up to half a carrier wavelength [9, 10]. Half a wavelength at 1900MHz would at vehicular speeds of about 50 km/h correspond to 5ms, or about 7 of our proposed 0.667ms slots.

2.3 Slot Scheduling among Sectors

Some pairs or groups of sectors, within a cell or in neighboring cells, may create a large amount of mutual interference. In order to reduce this co-channel interference, time-frequency bins are allocated exclusively within these groups of sectors, according to the traffic load in each sector. Information on the traffic load is exchanged between base stations via an inquiry procedure.

This interference reduction could enable the use of aggressive power control schemes in combination with the adaptive modulation, if this is deemed advantageous. It also improves the channel estimation accuracy and reduces the problem of predicting time-varying interference levels.

2.4 Use of Multiple Antenna Elements

To improve the received power, reduce interference and attain sufficient range of the broadband transmission, we propose smart antennas to be used at the base station and preferably also in the terminals. In this first outline, we assume only straightforward robust techniques of low computational complexity, but the use of more sophisticated schemes for array signal processing and spatial multiplexing will be investigated as the project continues.

In the downlink, base stations are assumed to use multiple antennas to form S sectors (beams), that are more narrow than standard 120^o sectors. The terminals may use L antennas/diversity branches for Maximum Ratio Combining (MRC).

In the uplink, base station receivers are furthermore assumed to use MRC of several diversity branches within each sector. The mobiles may there use a few antenna elements to form crude lobes.

3 The Downlink

We assume each time-frequency bin of 0.667ms $\times$200kHz to be partitioned into 6 symbols on 20 subcarriers. Each symbol is 100μs long, with a cyclical prefix of 11μs. This is adequate to cover the delay spread in most urban environments. Of the 120 symbols per bin, five are known pilots that are used for training.[2]

A 5 MHz system will thus use 500 subcarriers (or in practice 512, with 12 idle) and the symbol rate on each subcarrier is about 9000 symbols/s.

The design target vehicular speed is 100 km/h. The system should work efficiently for populations of users with this and lower speeds.

Each mobile user predicts the quality of all bins an appropriate number of time-slots ahead. All mobiles then signal their predicted quality estimates on an uplink control channel, by transmitting the suggested appropriate modulation levels to be used in the different frequency bins of the predicted time slot. A scheduler located close to the base station then allocates time-frequency bins to the different users, based on the predicted channel capacities of all active users, their data queue status, QoS requirements and priorities. The scheduler broadcasts its allocation decisions on a downlink control channel. The downlink transmission for the allocated time-slots can then begin. In each bin, the modulation level used is that which was suggested by the appointed user.

4 Potential of Proposed Downlink

We here estimate the spectral efficiency for transmission to K active users within one base station sector, under the following simplifying assumptions:

- Flat fading AWGN channels that are time-invariant within each bin of 200kHz $\times$0.667ms and independent Rayleigh fading between bins,
- all users are assigned equal average channel power by a slow power control scheme and the channels to different users fade independently,
- accurate channel predictions,
- the target service is reliable packet data transmission, while fairness between users with different rates, QoS requirements, and delay constraints are neglected,
- all users do always have data to transmit, and the payload symbols within the allocated slots are fully utilized by the designated user.

The assumed scheduler thus works as a selection diversity scheme, where the user with the best predicted SNR in each time-frequency bin out of all K users will transmit in that bin. In the receiver we assume MRC with L antennas.

The resulting pdf of the received SNR (γ) after MRC and multiuser selection diversity can then be shown to be

[2] Algorithms for channel estimation that are located in the mobile terminals will use also the received payload symbols in their channel estimators, to obtain channel estimates with high time-frequency resolution which are then used as input data to the channel predictors.

(see e.g. pp. 186–194 in [11] for guidance in performing this calculation)

$$p(\gamma) = \frac{Ke^{-\gamma/\bar{\gamma}}(\frac{\gamma}{\bar{\gamma}})^L}{\gamma\Gamma^K(L)}\left(\Gamma(L) - \Gamma(L, \frac{\gamma}{\bar{\gamma}})\right)^{K-1}, \quad (1)$$

where $\bar{\gamma}$ is the average SNR per receiver antenna, and

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt \quad (2)$$

$$\Gamma(a, z) = \int_z^\infty t^{a-1}e^{-t}\,dt \quad (3)$$

are the gamma and incomplete gamma functions, respectively.

We utilize an adaptive modulation system that uses $N = 8$ modulation formats: BPSK, 4-QAM, 8-QAM, 16-QAM, 32-QAM 64-QAM, 128-QAM, and 256-QAM. Equal transmitter power (equal energy per symbol) is used in all bins. The target symbol error rate that determines the choice of modulation level must now be determined.

Since the target service is reliable packet transmission we need an ARQ scheme. For simplicity we assume simple ARQ without coding. For this ARQ to be useful, we select a target packet error rate of $P_E = 0.1$.

A packet error rate of 0.1 for the link level packets containing 115 payload symbols corresponds to a symbol error rate of 9.16×10^{-4}.

The raw spectral efficiency must for our proposed system be multiplied by a factor $1 - P_E = 0.9$ due to packet retransmissions, by a factor $G_c = 100/111$ due to cyclical prefixes and a factor $G_p = 115/120$ since five of the 120 symbols in a bin are used as pilots. The spectral efficiency using AMS and simple ARQ will then be obtained by a weighted sum over the modulation levels, weighted by the probabilities that those particular levels will be utilized

$$\eta \geq (1 - P_E)G_cG_p \sum_{i=0}^{N-1} k_i \int_{\gamma_i}^{\gamma_{i+1}} p(\gamma)\,d\gamma \text{ [bits/s/Hz]} \quad (4)$$

Here, $k_i = \log_2(M)$ is the number of bits transmitted using uncoded M-ary QAM modulation. The values γ_i, $i = 0, \ldots, 7$, define the symbol SNR regions in which to use the different modulation levels: 2^{k_i}-ary QAM is used in the SNR interval $[\gamma_i, \gamma_{i+1})$, with $\gamma_N = \infty$. No transmission is performed below SNR $= \gamma_0$. These γ_i values are given in Table 1 for the target symbol error rate 9.16×10^{-4}. The inequality in (4) comes from the fact that once the SNR is above the γ_i level, the symbol error rate (and thus the packet error rate) is lower than the target.

The spectral efficiency in (4) is evaluated numerically for a symbol energy to noise ratio of $\bar{\gamma} = 16$ dB per receiver antenna. The result is presented in Figure 2, which shows a dramatic improvement with an increasing multiuser selection diversity and with the number of receiving antennas. Already for two antennas and five users per sector, the spectral efficiency is considerable.

i	Modulation	k_i	γ_i (dB)
0	BPSK	1	6.86
1	4-QAM	2	10.41
2	8-QAM	3	14.57
3	16-QAM	4	17.69
4	32-QAM	5	21.03
5	64-QAM	6	24.02
6	128-QAM	7	27.16
7	256-QAM	8	30.14

Table 1: The limits γ_i for the SNR per receiver antenna in dB that determine the use of different modulation levels. These limits are calculated as the SNR required to achieve the target symbol error rate 9.16×10^{-4} using $M = 2^{k_i}$ QAM. Here k_i is the number of bits per symbol.

The spectral efficiency saturates for a high number of users, where most bins are occupied by users who can utilize a high modulation degree in that bin. The addition of more receiver diversity branches tends to decrease the multiuser diversity effect.

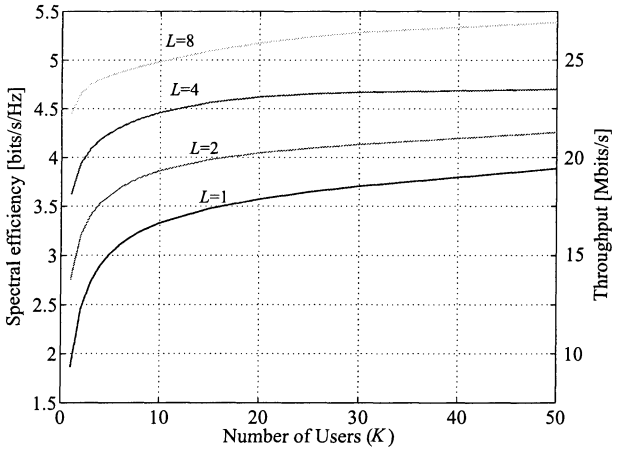


Figure 2: Spectral efficiency using AMS and simple ARQ with Lth order MRC diversity in the mobile and Kth order of selection diversity between the users. The symbol energy to noise ratio per receiver antenna is $\bar{\gamma} = 16$ dB. Right-hand scale: The corresponding capacity within one sector in a 5MHz downlink.

In theory, the spectral efficiency within one sector, η, should be multiplied by the number of sectors, S, used per cell. However, even if the interference is reduced by using many sectors, we cannot use all time-frequency bins in all sectors. The spectral efficiency of each cell would therefore be larger than is indicated by Fig. 2, but smaller than ηS. Interference between sectors has to be studied carefully for a more precise estimate.

By multiplying the spectral efficiency in bits/s/Hz by 5, it is converted into an equivalent capacity in Mbit/s of a base station sector in a system with 5MHz bandwidth. This is illustrated by the right-hand scale of Figure 2. Although the test situation is not exactly comparable, it can

be noted that the maximum capacity of the HSDPA mode of WCDMA is claimed to be 8-10Mbit/s.

4.1 Discussion

The above estimate is conservative, in that the packet error rate 10% is attained only at the limits γ_i of the SNR regions. Inside the regions, the error rate is less, with a corresponding lower need for retransmissions. On the other hand, we have neglected prediction errors, which will tend to increase the error rate and decrease the attained throughput.[3] The impact of these two effects is a topic for our near-term research.

Another important aspect is the requirement for control bandwidth by the channel quality feedback and the transmission scheduling. The channel quality feedback in the uplink is in the form of a suggested modulation format. With 8 allowed formats, this requires 3 bits for each of the 25 frequency bins in a 0.667ms time slot. With K active users, the required uplink control bandwidth is $B_{uc} = K \times 112$ kbits/s. For $K = 10$ active users per sector, this corresponds to 1.1Mbit/s. This is considerable (75 times more than for power control in WCDMA) but not overwhelming; The total uplink capacity is under the assumptions of this section expected to be on the order of 15Mbit/s per base station sector [6]. Furthermore, there are many conceivable ways of modifying the signaling to reduce the bandwidth demand.

The downlink control information for broadcasting the scheduling decisions consists of a user number per slot. For 2^n simultaneous users and a code rate of R used for the control information, the required downlink control bandwidth is $B_{dc} = (n/R) \times 37.5$ kbits/s. It is small as compared to the capacity per sector.

5 Concluding Remarks

Packet data systems that use feedback information may be constructed in a way that promises to attain extreme spectral efficiency. The type of system outlined in this paper is promising and it also raises many fascinating research questions. Some of them concern the interplay between channel prediction and adaptive modulation [7], others the interplay between Hybrid ARQ retransmission and predictive scheduling [4]. The efficient use of smart antennas to improve the resource allocation is yet another topic where many questions await answers.

[3] For predictive scheduling combined with link level retransmission, our preliminary results indicate that the use of bad power prediction (standard deviation 5dB) reduces the throughput by about 20% as compared to the use of perfect prediction. The attainable Quality of Service parameters such as delays do suffer much worse degradation [4, 5].

In this conference, [7] provides an estimate of the spectral efficiency when the levels γ_i are adjusted by taking the prediction error variance into account. For flat Rayleigh fading channels, a normalized prediction error variance level of 0.1 there results in only a negligible reduction of the spectral efficiency as compared to perfect prediction. This prediction error variance would be reasonable in our proposed downlink [10].

The interplay with higher layers is also crucial for obtaining a system with overall good performance. The flow control of TCP is here of particular interest. At present we investigate the performance of different variants of TCP over wireless links with the optimized link layers outlined above. We also investigate how the use of information from lower layers may improve the performance of higher layers. We are making a major effort to create an appropriate simulation environment for studying such issues [12]. Using real TCP and UDP stacks, it enables us to generate packets that propagate in real time over realistically simulated wireless links, so that important aspects and timescales of lower layers, and inter-layer interactions can be studied.

References

[1] The Wireless IP Project within the Swedish SSF PCC (Personal Computing and Communication) program. See M.Sternad, "The Wireless IP Project," RVK02, June 2002. www.signal.uu.se/Research/PCCwirelessIP.html

[2] 3rd Generation Partnership Project (3GPP), Technical Specification Group Radio Access Network: "UTRA High Speed Downlink Packet Access (Release 4)," *3GPP TR 25.950 V2.0.0.*

[3] R. Knopp and P.A. Humblet, "Information capacity and power control in single-cell multiuser communications," *IEEE ICC*, Seattle, 1995.

[4] N. C. Ericsson, S. Falahati, A. Ahlén and A. Svensson, "Hybrid type-II ARQ/AMS supported by channel predictive scheduling in a multi-user scenario," *IEEE VTC2000-Fall*, Boston, MA, USA, Sept. 24-28, 2000, pp. 1804-1811.

[5] N. C. Ericsson, *On Scheduling and Adaptive Modulation in Wireless Communications*. Licentiate Thesis, Signals and Systems, Uppsala University, Sweden, June 2001.

[6] T. Ottosson, A. Ahlén, A. Brunström, M. Sternad and A. Svensson, "Toward 4G IP-based wireless systems: A proposal for the uplink." *5th Wireless World Research Forum Meeting*, Temple, AR, March 2002.

[7] S. Falahati, A. Svensson, T. Ekman and M. Sternad, "Effect of prediction errors on adaptive modulation systems for wireless channels," *RVK02*, Stockholm, June 2002.

[8] D.N.Tse, "Optimal power allocation over parallel Gaussian channels," *ISIT*, Ulm, Germany, June 1997.

[9] M. Sternad, T. Ekman and A. Ahlén, "Power prediction on broadband channels," *IEEE Vehicular Technology Conference*, Spring 2001, Rhodes, Greece, May 6-9 2001.

[10] T. Ekman, M. Sternad and A. Ahlén, "Unbiased power prediction on broadband channels," *RVK02*, Stockholm, June 2002.

[11] M. D. Yacoub, *Foundations of Mobile Radio Engineering*, CRC Press, 1993.

[12] A. Ewerlid and N.C. Ericsson "Real time simulations of wireless IP communications," *RVK02*, Stockholm, June 2002.

An Algorithmic Technique for Construction of Multi Dimensional Signal Constellations

Jan-Erik Porath[1], Tor Aulin[2]

[1]Department of Technology, University of Trollhättan/Uddevalla, Box 957, SE-461 29 Trollhättan, Sweden. E-mail: jan-erik.porath@htu.se.
[2]Telecommunication Theory, Department of Computer Engineering, Chalmers University of Technology, SE-412 96 Göteborg, Sweden. E-mail: tor@ce.chalmers.se.

Abstract

An algorithmic technique is developed intended for signal constellation design for the N-dimensional Euclidean signal space. Such signals are suitable for reliable and efficient digital communications on the Gaussian channel. The minimum Euclidean distance is used as performance measure.

The basic idea is to spread out the signal points (vectors corresponding to the signals) using an iterative method modeling the behavior of equally charged particles in the free space. Initially, the vectors are randomly distributed. After a few iterations, the vectors have moved into equilibrium.

Numerical calculations for $N = 3$, 4 and 5 dimensions and $M = N+1 \ldots 20$ signals show that the generated codes are nearly as good as previously known codes and in some cases even better.

The conclusion is that the proposed algorithm may be used for construction of multidimensional modulation schemes.

1 Introduction

The concern of this presentation is to design multidimensional signal sets suitable for the additive white Gaussian noise (AWGN) channel. The AWGN channel motivates the use of maximum likelihood detection as well as representation of signals as vectors in the Euclidean space. Using multidimensional modulation may increase reliability when transmitting digitized information. During the last decades some effort has been made on this subject. In the papers [1]-[5] good signal constellations were designed from well known two-dimensional modulation schemes such as M-PSK and M-QAM. Schemes based on lattices can be found in [6] and [7]. Optimal two-dimensional signal constellations, in terms of performance on the Gaussian channel in the presence of jitter, were present in [8]. Trellis coded modulation is one approach to generate multidimensional signals [9].

Whatever the scheme selected, the receiver observes a sequence of symbols which have been transmitted over the AWGN channel. The set of possible sequences is represented by vectors in the multidimensional Euclidean space $\mathbb{R}^N$, where N is the number of dimensions and $\mathbb{R}$ is the set of real numbers. A distinct sequence of symbols, each occupying a few dimensions, may be viewed as a single signal in signal space with a larger number of dimensions. The specific signal corresponds to a unique set of information bits. In most practical cases, k information bits are mapped onto one of 2^k signal points in Euclidean space. From this point of view, the use of a specific encoder is only one out of many possible ways to generate signal points well separated in the Euclidean space. Generally, a look-up table may be used to perform the mapping from data to channel symbols. This point of view is according to Shannon [10].

The scope of the paper is to present an algorithmic approach aimed for design of signal constellations useful for the AWGN channel. The constellations are generated by a heuristic technique based on the behavior of charged electrons in free space. Other criteria are also possible, see for instance [8] where an error gradient approach was proposed. The problem of concern is to spread M signal vectors in an N-dimensional Euclidean space such that the normalized minimum Euclidean distance between all distinct pair of signals is maximized. In literature M is often 2^k where k is an integer number of information bits. In the more general case, M can take any integer value larger than N, i.e. $M > N$. Selecting $M < N$ is not fruitful because one can always transform an original signal set into a N-dimensional space where $N < M$ by applying a Gram-Schmidt orthogonalization procedure to the set [11].

Conway and Sloane et. al. have constructed codes with different properties, which are presented in electronic form on Sloane's homepage [12]. The numerical results in this paper are compared with the spherical codes presented there. Two methods to construct spherical codes with good properties are given in their book [13]. The first method is to wrap good lattices over a sphere. The other approach is to take a lattice code with good packing and then selecting vectors

with fix distance from origin. Neither method does necessarily give optimal codes.

2 Definitions

When nothing else is stated, the definitions and statements in this section are according to Wozencraft and Jacobs [11]. Particularly, the geometric representation of signals is adopted from there.

The AWGN channel is a zero-mean white Gaussian process $n_w(t)$ with power density $S_w(f) = N_0/2$. The receiver observes $r(t) = s(t) + n_w(t)$ where $s(t)$ is one of M possible transmitted signals. The time signal $s^i(t)$ can be described by the vector $\mathbf{s}^i = [s_0^i, s_1^i, ..., s_{N-1}^i]$; i=0, 1, ..., M-1. The relevant noise process is described by a random vector $\mathbf{n} = (n_0, n_1, ..., n_{N-1})$ where n_k has variance $\sigma^2 = N_0/2$.

A look-up table is a one-to-one mapping from M possible messages to M signals in the N-dimensional Euclidean space $\mathbb{R}^N$ where $M > N$. We may see the table as a mapping, a code or an N-dimensional modulator. The specific input to the modulator is denoted m^i ; $i \in [0,1, ..., M\text{-}1]$.

Each signal is carrying $k = \log_2(M)$ bits where k is not necessarily an integer. The set of M signal vectors $\{\mathbf{s}^0, \mathbf{s}^1, ..., \mathbf{s}^{M-1}\}$ is denoted S. The squared Euclidean distance D^2 between two distinct signals is given by

$$D^{ij\,2} = \sum_{n=0}^{N-1} \left(s_n^i - s_n^j\right)^2 = \left|\mathbf{s}^i - \mathbf{s}^j\right|^2 \qquad (1)$$

The symbol energy E_s^i of a specific signal $\mathbf{s}^i$ is evaluated from eqn. 1 by simply setting $\mathbf{s}^j = \mathbf{0}$. Each of the M symbols represents $\log_2(M)$ information bits. The average symbol energy used by the signal set is denoted E_s. The average bit energy is denoted E_b.

A code is spherical if all symbol elements in the set S have equal energy. That means that the signal points are located at the surface of a N-dimensional sphere.

The minimum normalized squared Euclidean distance between the signals in S is used in order to compare different schemes with each other. It is defined by

$$d_{\min}^2 = \frac{1}{2 \cdot E_b} \cdot \min_{i \neq j} \left\{ D^{ij\,2} \right\}; \quad i, j \in \{0,1,...,M-1\} \qquad (2)$$

It is assumed that all signals appear with equal probability. Moreover, all signals carry the same number of bits.

3 The algorithm

In this section we shall design constellations with good properties in terms of large minimum distance.

The idea is to copy the behavior of charged particles in free space. Imagine four equally charged particles in a two-dimensional plane. If they are tied to the origin with four strings of equal length, then in equilibrium, they are located in the corner of a square, exactly as the signal vectors in the well-known QPSK modulation scheme. It is possible to determine the forces that act between every distinct pair of particles. The total force on one particle is the superposition of forces acting on that particle, i.e. the vector sum of all forces. Related problems, in three dimensions, were treated in [14], [15] and [16].

Below, we shall outline a technique for construction of M-ary signal constellations in $\mathbb{R}^N$. In analogy with the behavior of charged particles, a 'force' acting on each of the M signals is introduced. The strength of the mutual force between two signals should decrease with the distance between the signals. Moreover, it should be directed along with the straight line passing through the two signal points. The following definition of the force $\mathbf{f}$ between two signals $\mathbf{s}^i$ and $\mathbf{s}^j$ satisfies these criteria:

$$\mathbf{f}^{ij} = \alpha \cdot \frac{\mathbf{s}^i - \mathbf{s}^j}{\left|\mathbf{s}^i - \mathbf{s}^j\right|^\beta} = \alpha \cdot \frac{\mathbf{s}^i - \mathbf{s}^j}{D^{ij\,\beta}} \qquad (3)$$

where α and β are constants. The entire force acting on signal i is

$$\mathbf{F}^i = \sum_{\substack{j=0 \\ j \neq i}}^{M-1} \mathbf{f}^{ij} \qquad (4)$$

Now, starting with a set of M randomly located signal vectors in the N-dimensional space, the force $\mathbf{F}^i$ might be calculated for each signal $\mathbf{s}^i$. The signal points are then moved an incremental distance with direction and length given by their respective force. An infinitesimal move will result in a new set of forces, so it is necessary not to move the signal points too far away at each iteration. In order to avoid the signal points tending to infinity, a constraint is applied to the total energy of the system. Simply scaling the vectors once every iteration step, such that the mean energy E_S is conserved, performs this.

In the numerical applications below, the maximum incremental step is limited to half of the minimum distance of the current constellation in order to avoid large movements when signal points are located near each other. Each $\mathbf{F}^i$ is scaled with $1/(2M) \cdot D^{ij}{}_{\min}$. Doing so, the length of $\mathbf{F}^i$ will be maximized to $(M\text{-}1) \cdot D^{ij}{}_{\min}/(2M) < D^{ij}{}_{\min}/2$.

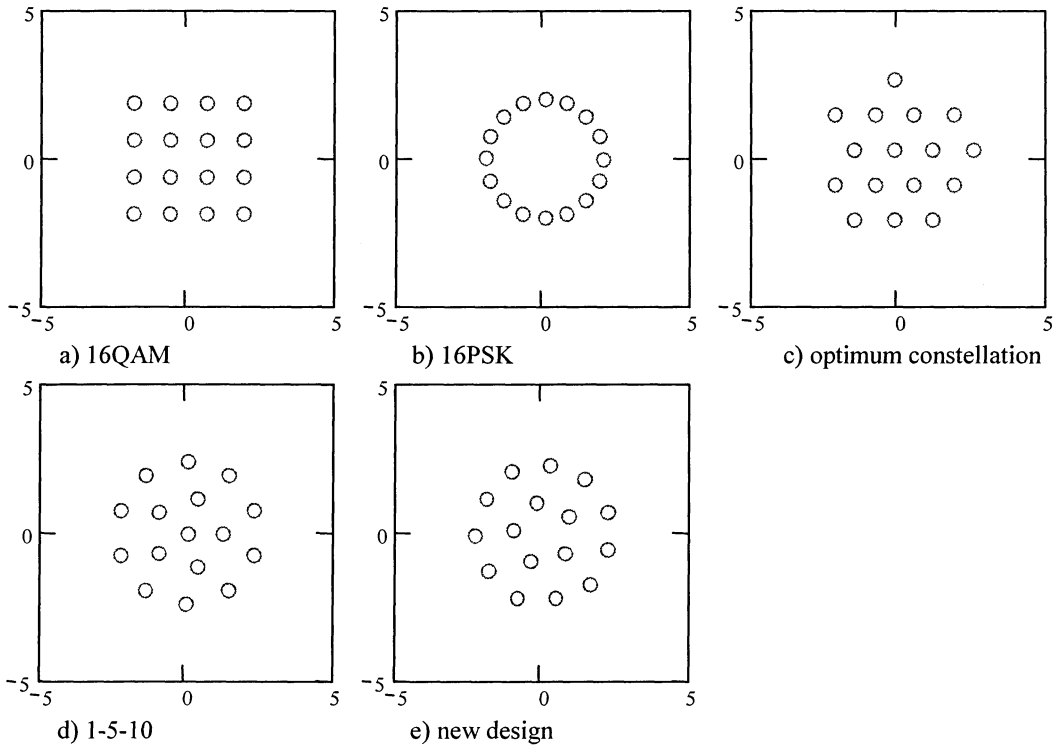


Figure 1: *Different constellations for N = 2, M = 16*

Spherical codes may be constructed with a slightly modification of the previously described method. This may be performed by scaling each signal individually such that $E_s^i = E_s^j \ \forall i, j$. Hence, the vectors are moved radially from origin onto the surface of a sphere. This can be performed at every iteration step or, alternatively, when the signals are in equilibrium.

The choice of the constants α and β in eqn. 14 determines if and how fast the algorithm will converge. Here, convergence means that no significant changes are present in the position of the signal vectors when new iterations are performed. One way to determine if equilibrium has been reached is to trace the minimum distance d_{min} as the number of iterations increase. However, a more reliable approach is to trace the maximum length l_{max} that any signal has been moved during the most previously performed iteration. This length may be used as a stop criterion. E.g. when $l_{max}/d_{min} < 0.1$ % we may stop the iteration. By experience, proper values of α and β is 1 and 5 respectively. With these values, the algorithm converge within 10-100 iteration steps for N = 2, 3, 4 and 5 dimensions with $M \leq 20$ signals.

A signal constellation evolved from running the algorithm is a set of randomly positioned points in the sense that the final positions depend on where the initial randomly chosen positions where located. Also the resulting d_{min} is varying with different runs. Therefore, it is a good idea to perform more than one simulation in order to find a constellation with good properties.

4 Numerical results

The performance of the algorithm is investigated in this section. Numerical results are given for dimensions N = 2, 3, 4 and 5 with different number of signals, M.

In order to illustrate the capability of the algorithm a constellation is designed for M = 16 signals in N = 2 dimensions. The generated constellation (Fig. 1 e) is compared to the well known 16QAM system (Fig. 1 a) and 16PSK (Fig. 1 b), the optimum design with respect to the minimum Euclidean distance due to [8] (Fig. 1 c) and the 1-5-10 constellation due to [8] (Fig. 1 d). In the latter constellation, the angles between signals is large. This is a good criterion, in terms of error rate performance, when phase jitter and Gaussian noise is present. The mean bit energy E_b = 1 for all the constellations.

The normalized minimum Euclidean distance for the five constellations in Fig. 1 is shown in Table 1. The table also includes the asymptotic gain G = $20\log(d_{min}/d_{min}(16QAM))$ [dB].

The new constellation has a very good performance, but if not optimal, compared to previously known schemes. From Table 1 we can see that the d_{min} is -0.33 [dB] from 16QAM. On the other hand, it has larger d_{min} then the optimal code for the channel characterized by phase jitter. At the same time the new constellation looks very similar to this code. We may therefore expect that the new code behaves good when phase jitter is present. The asymptotic gain G yields when SNR approaches infinity. At intermediate and low SNR

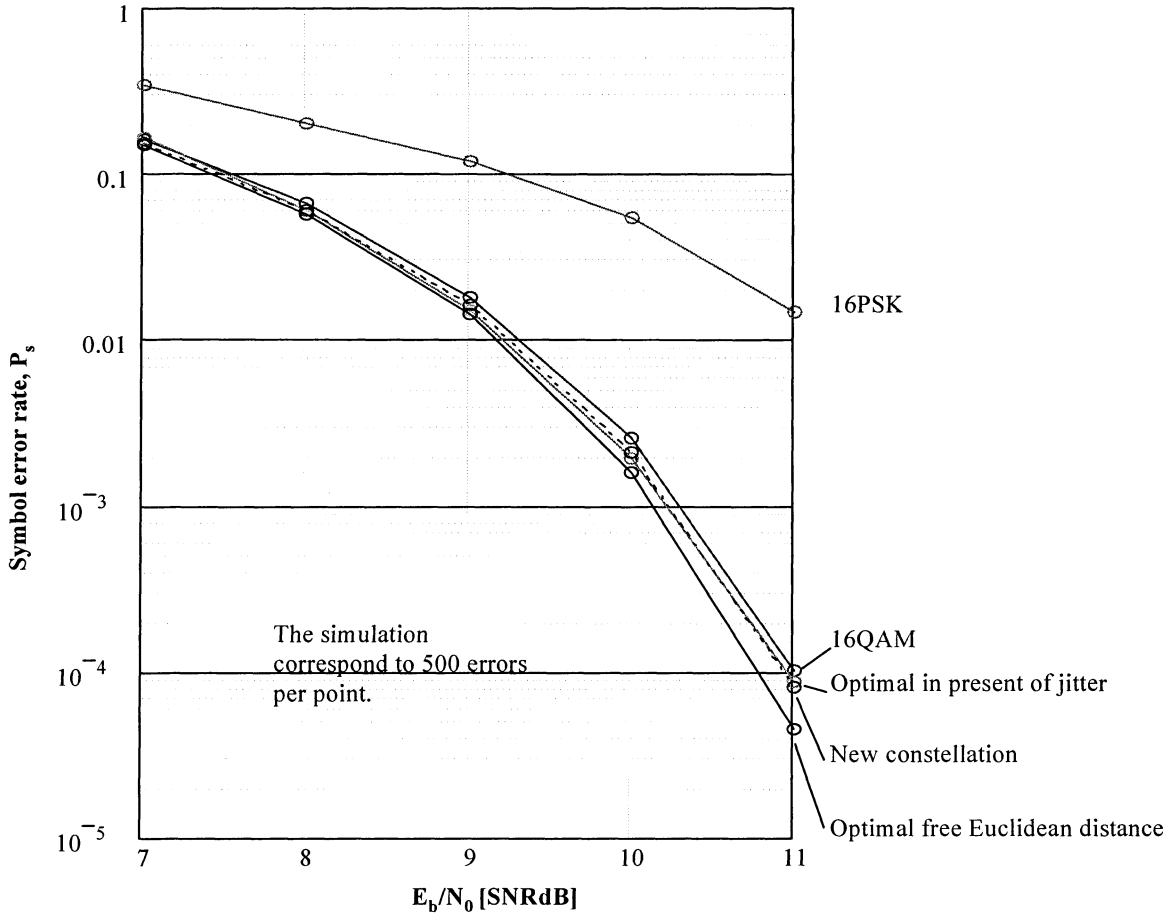


Figure 2: *Simulated symbol error rate for the constellations in Fig. 1*

the relative performance may be different as can be observed by inspecting Figure 2. In the range from SNR = 9 to 10 [dB] the new constellation is beaten only by the optimal d_{min} constellation.

The algorithm was also applied to constellations with $N = 3$, 4 and 5 with $M = N+1, N+2, \ldots, 20$ symbols. The squared minimum distances for the generated constellations are tabulated in Table 2. They are compared with the codes due to Sloane et.al. [12]. The table also gives the ratio between the minimum and the maximum symbol energy in each constellation. A code is spherical when this ratio is '1'. The ratio is zero if one signal is located in origin. New codes with larger distances than for the corresponding codes due to Sloane are indicated with bold typeface.

5 Conclusions

A new approach is proposed for construction of spherical and non-spherical codes. Many of the non-spherical codes were quasi-spherical in the sense of having low energy deviation from the mean energy. Some codes were near spherical except for one single point located near origin.

The numerical results for $N = 2$, 3, 4 and 5 indicates that the design technique generate constellations that performs well compared to other schemes. It should be noticed that the new design for $N = 2$ is very like the optimal design for the channel characterized by Gaussian noise and phase jitter. Furthermore, some codes are as good as or better than the codes presented by Sloane [12].

The algorithm may be used as a complement to other methods because of its generality. It also opens possibilities for design of constellations with special properties. E.g. when designing codes, points with opposite polarization can be placed at locations where no signal points are desired. Even lines and other geometric objects could be used for this purpose.

code	1 a	1 b	1 c	1 d	1 e
d^2_{min}	0.8	0.304	0.914	0.711	0.741
G [dB]	0	-4.194	0.579	-0.511	-0.332

Table 1: *Minimum free Euclidean distances for the constellations in Fig. 1*

	$N=3$			$N=4$			$N=5$		
M	d^2_{min}	$\min E_s/\max E_s$	Sloane et. al.	d^2_{min}	$\min E_s/\max E_s$	Sloane et. al.	d^2_{min}	$\min E_s/\max E_s$	Sloane et. al.
4	2.666	1.000	2.667	-	-	-	-	-	-
5	**2.407**	0.663	2.322	2.900	1.000	2.902	-	-	-
6	2.581	0.998	2.585	**2.702**	0.726	2.585	3.082	0.999	3.102
7	2.192	0.592	2.217	2.787	0.815	2.807	**2.902**	0.841	2.807
8	2.136	0.664	2.216	2.961	0.998	3.000	2.968	0.879	3.000
9	**2.143**	0.686	2.113	2.565	0.790	2.656	3.122	0.936	3.170
10	1.935	0.433	1.979	2.741	0.999	2.768	3.316	0.999	3.322
11	1.880	0.000	1.912	2.516	0.741	2.662	2.851	0.732	3.000
12	1.979	0.999	1.982	2.639	0.992	2.689	2.964	0.875	3.033
13	**2.003**	0.000	1.692	2.471	0.798	2.563	2.775	0.762	3.007
14	**1.782**	0.000	1.660	2.572	0.874	2.591	2.846	0.877	3.046
15	**1.685**	0.000	1.592	2.442	0.958	2.536	2.831	0.879	3.126
16	**1.605**	0.000	1.551	2.326	0.830	2.449	3.172	0.999	3.200
17	**1.581**	0.000	1.520	2.284	0.824	2.402	2.795	0.912	2.982
18	**1.595**	0.000	1.465	2.230	0.532	2.407	2.757	0.767	3.021
19	**1.438**	0.000	1.389	2.248	0.853	2.403	2.776	0.948	3.008
20	1.319	0.084	1.398	2.196	0.000	2.445	2.685	0.740	3.028

Table 2: *Largest distances found for different number of dimensions*

References

[1] E.J. Leonardo, Z. Lin and B.S Vucetic. Multidimensional M-PSK trellis codes for fading channels. *IEEE Trans. Inform. Theory*, 42(4):1093–1108, July 1996.

[2] S.S Pietrobon, R.H. Deng, A. Lafanechere, G. Ungerboeck and D.J. JR. Costello. Trellis-coded multidimensional phase modulation. *IEEE Trans. Inform. Theory*, 36(1):63 –89, Jan. 1990.

[3] S. S. Pietrobon and D. J. JR. Costello. Trellis coding with multidimensional QAM signal sets. *IEEE Trans. Inform. Theory*, 39(1):325-336, Jan. 1993.

[4] S. Benedetto, R. Garello , M. Mondin and G. Montorsi. Geometrically uniform partitions of L \times MPSK constellations and related binary trellis codes. *IEEE Trans. Inform. Theor*, 39:1773-1798, Nov. 1993.

[5] E. Biglieri and M. Elia. Multidimensional modulation and coding. *IEEE Trans. Inform. Theory*, 34:803-809, July 1988.

[6] J. Boutos, E. Viterbo, C. Rastello and J. C. Belfiore. Good lattice constellations for both Rayleigh and Gaussian channels. *IEEE Trans. Inform. Theory*, 42:502-518, Mar. 1996.

[7] X. Giraud and E. Boutillon. Algebraic tools to build modulation schemes for fading channels. *IEEE Trans. Inform. Theory*, 43:938-952, May 1997.

[8] G.J. Foschini, R. D. Gitlin and S.B. Weinstein. Optimization of twodimensional signal constellations in presence of Gaussian noise. *IEEE Trans. On Commun.*, 22:28-38, Jan. 1974.

[9] G. Ungerboeck. Channel coding with multilevel/phase signals. *IEEE Trans. Inform. Theory*, 28:55-67, Jan. 1982.

[10] C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical System Journal*, 27:379-423, 1948.

[11] Wozencraft and Jacobs. *Principles of Communication Engineering*. Wiley & Sons Inc., 1965.

[12] N.J. Sloane. www.research.att.com/~njas/packings. *Published electronically*.

[13] Conway and Sloane. *Sphere Packings, Lattices and Groupes*. Springer- Verlag, NY, 3d edn. 1998.

[14] L.L. Whyte. Unique arrangements of points on a sphere. *Amer. Math. Monthly*, 59:606-611, 1952.

[15] T. Erbert and G.M. Hockney. Equilibrium configurations of N equal charges on a sphere. *J. Phys*, A 24, 1991.

[16] L. Glasser and A.G. Every. Energies and spacings of point charges on a sphere. *J. Phys.*, A 25, 1992.

Extrinsic Information Transfer Charts for Narrow-Band Multiple Access*

Fredrik Brännström[1], Tor Aulin[1], Lars Rasmussen[1,2] and Alex Grant[2]

[1]Chalmers University of Technology, Department of Computer Engineering, SE-412 96 Göteborg, Sweden.
[2]Institute for Telecommunications Research, University of South Australia, Mawson Lakes SA 5095, Australia.
E-mail: fredrikb@ce.chalmers.se, aulin@ce.chalmers.se, larsr@ce.chalmers.se, Alex.Grant@unisa.edu.au

Abstract

Convergence analysis of iterative detectors for narrow-band multiple access is performed by using extrinsic information transfer charts. The system has no bandwidth expansion, so K users are using the same bandwidth as one single user. The load (the number of bits per channel use) of the system is therefore higher than the load in, for example, conventional CDMA systems. The analysis provide useful guidelines how to combine outer codes with the chosen mapper for different loads of the system. Both memoryless mappers and differential mappers are evaluated. The limitations of the system measured in the maximum number of users and the required SNR for different scenarios are presented.

1 Introduction

In [1], [2] constellation-constrained capacity calculations were used to find the maximum theoretically achievable information rate together with the minimum required signal-to-noise ratio (SNR) in a multiple access (MA) system known as trellis code multiple access (TCMA) [1], [3], [4]. These calculations were based on the superimposed constellations of all the users, disregarding the outer code of each user. The calculations could therefore not predict what outer code for the single user (SU) that could be combined with the chosen mapper to achieve reliable communication. Unfortunately, the minimum required SNR in terms of E_b/N_0 given from these calculations were several dB from the actual simulated threshold for the chosen combination of code and mapper [1]. E_b is the energy used to transmit one information bit and $N_0/2$ is the double sided power spectral density of the noise.

Extrinsic information transfer (EXIT) charts were introduced in [5], [6] as a tool to design mappers to be used together with demappers and outer codes. Demappers use a priori information on the input bits to the mapper together with the received observable to update the extrinsic information on the same bits. EXIT charts have also been used to design both serially and parallel concatenated codes [7], [8]. Similar methods used for convergence analysis are based on density evolution, [9].

In this paper EXIT charts are applied to narrow-band MA systems, extending the ideas in [5] to include a MA demapper. In contrast to the constellation-constrained capacity calculations [1], [2], the outer code for the SU can now be included in the convergence analysis.

2 System Model

The MA system in Figure 1 consists of K users transmitting statistically independent and identically distributed binary data x_k [1], [3], [4]. Each user's data is passed through an outer encoder $\mathcal{C}_k$ of rate R_c. The coded bits y_k are then fed to a user unique random bit interleaver Π_k, permuting blocks of L bits. The interleaved bits z_k are mapped with $\mathcal{M}_k$ onto an M_k-ary two-dimensional symbol s_k, referred to as the *user symbol*. In contrast to [1], [3], [4], $\mathcal{M}_k$ can here be a memoryless mapper or a mapper with memory.

All users synchronously modulate their user sym-

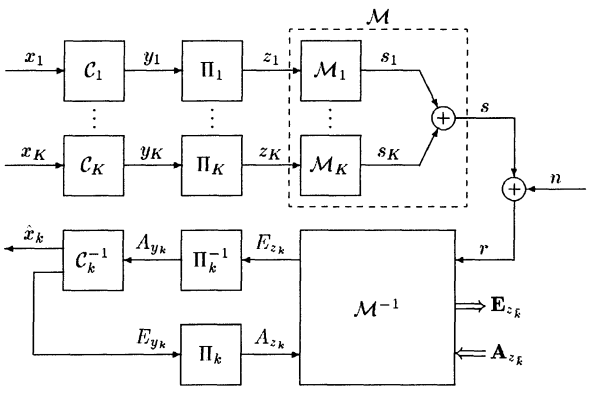


Figure 1: Narrow-band MA system model for K users and the iterative detector of user k.

*This research was conducted while Fredrik Brännström visited Dr. Alex Grant at the Institute for Telecommunications Research, University of South Australia.

bol, s_k, onto a continuous-time waveform $s(t)$. Note that there is no spreading in this system besides the outer code. Each waveform (or channel use) will therefore carry $b_k = R_c \log_2(M_k)$ information bits. These waveforms are superimposed on the AWGN channel, where s is a symbol from the joint constellation of all the user symbols [1], [3], [4].

The dashed box in Figure 1 indicates that the K mappers can be identified as a singe mapper $\mathcal{M}$, mapping the K bitstreams z_k in parallel onto a symbol s. A Gaussian noise vector n is then added to s in the vector channel model, Figure 1. The variance of each component in this vector is σ^2 which is also the double sided noise power spectral density of the continuous time system, $N_0/2$.

To be able to tell what block of information bits that corresponds to which user, each user must have a unique combination of the outer code $\mathcal{C}_k$, the interleaver Π_k and the mapper $\mathcal{M}_k$. They can be combined in different ways as long as the combination is unique among the users [1], [3], [4]. In this narrow-band MA system some of the signal points in the joint constellation can be ambiguous, [1].

If the mappers $\mathcal{M}_k$ are memoryless, the MA demapper works in principle as described in [10]. If the individual mappers have memory the MA demapper will work on a trellis, where the number of states and edges grows exponentially with the number of users. That is usually not a problem here, since only a few users are considered.

The iterative detector in Figure 1 consists of a MA demapper followed by K SU decodes. A_{y_k}, E_{y_k}, A_{z_k} and E_{z_k} denote the a priori and the extrinsic information on y_k and z_k respectively. The MA demapper $\mathcal{M}^{-1}$ uses the received observable r together with the a priori values A_{y_k} delivered from all SU decoders to calculate the extrinsic information on z_k for all users. $\mathbf{A}_{y_{\bar{k}}}$ in Figure 1 denotes the set of a priori information from all users except user k, $\mathbf{A}_{y_{\bar{k}}} = \{A_{y_1}, \ldots, A_{y_{k-1}}, A_{y_{k+1}}, \ldots, A_{y_K}\}$. The same notation is used for the extrinsic information, $\mathbf{E}_{z_{\bar{k}}}$.

The extrinsic information E_{z_k} is passed through a deinterleaver Π_k^{-1} and becomes the a priori information A_{y_k} fed to the a posteriori probability (APP) calculator of user k, $\mathcal{C}_k^{-1}$. This module makes decisions on the information bits $\hat{x}_k$ and calculates the extrinsic information of the outer coded bits y_k. The APP calculator is an extension of the BCJR algorithm [11] and a detailed description can be found in [12]. After the interleaver, the extrinsic information E_{y_k} becomes the a priori information A_{z_k} fed to the MA demapper. The next iteration of the MA demapper can start when all users have delivered their updated A_{z_k}.

3 Convergence Analysis and Performance

The mutual information between the coded bits from the outer code and the log-likelihood ratio of the a priori knowledge $I_{A_y} = I(y_k; A_{y_k})$ is calculated according to [7]. Mutual information is also used to measure the information content of the extrinsic output from the APP calculators $I_{E_y} = I(y_k; E_{y_k})$. I_{E_y} depends only on the a priori knowledge and can therefore directly be plotted versus I_{A_y} in an EXIT chart [5], [6].

Three $R_c = 1/2$ outer codes are chosen for analysis, a repetition code and two feed forward convolutional codes (CC). CC(2,3) denotes the CC with code polynomials (2,3) given in octal. CC(2,3) has two states and CC(5,7) has four states. Any half rate code will have $I_{E_y} = 0.5$ when $I_{A_y} = 0.5$ [5].

The extrinsic characteristic of the input bits to the mapper I_{E_z} depends on both the a priori knowledge I_{A_z} and the SNR. For brevity, the analysis of the inner code is here restricted to QPSK with Gray mapping, where all users have the same phase-offset. The Gray mapping will give a higher initial value of I_{E_z} than, for example, QPSK with natural mapping [5]. This scenario, where all users have the same phase-offset, is the worst case scenario in terms of achievable information rate, since some of the signal points in the superimposed constellation can be ambiguous [1]. The results can easily be extended to include higher order constellations $M_k > 4$, random phase-offset and other mappings than Gray mapping.

For a SU with a memoryless QPSK Gray mapper, I_{E_z} remains constant when increasing I_{A_z}, i.e., the level only depends on the SNR [5]. Therefore, nothing can in this case be gained by iterating between the outer code and the demapper.

The EXIT chart for two users with memoryless QPSK is shown in Figure 2a together with the SU case. Thick lines corresponds to the outer code and thin lines to the inner MA demapper. In the two-user case, I_{E_z} has a slope greater than zero and something can be gained by iterating between the outer SU code and the inner MA demapper. When there is more than one user in the system, I_{E_z} will initially start at a lower level than in the SU case, but it will for all K finally end up at the same level as in the SU case when $I_{A_z} = 1$.

For turbo codes, there is a waterfall at a critical SNR, [7]. This effect will only appear if the curve for the inner code combined with the curve for the outer code will open a tunnel at a certain SNR such that the tunnel ends in $I_{E_z} = I_{A_y} = 1$. When that happens, the a priori knowledge of the coded bits y_k will be perfectly known and the APP calculator of the outer code will have no problem to make reliable decisions on the information bits $\hat{x}_k$. The performance turns from high error rate to zero error rate within fractions

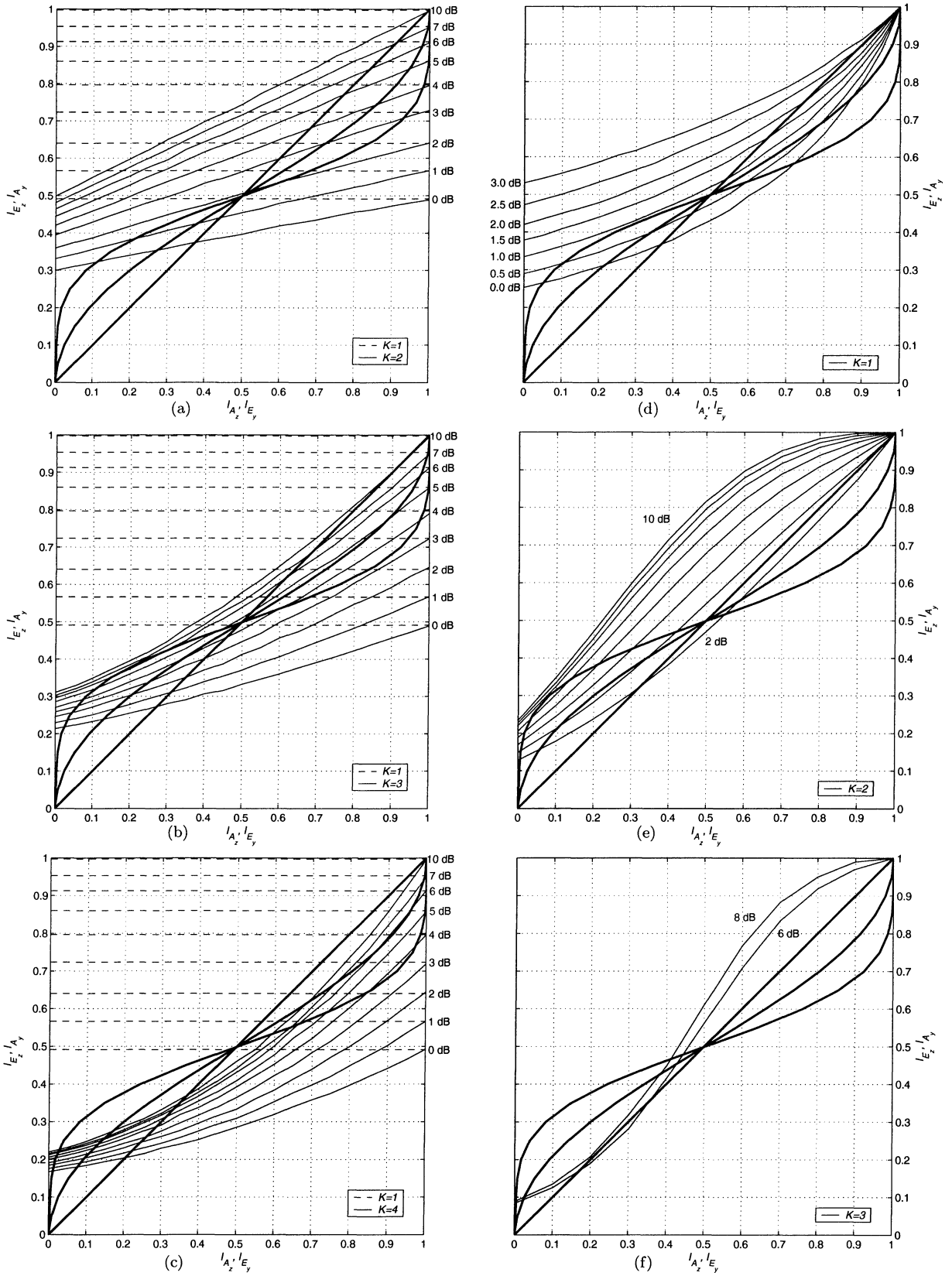


Figure 2: EXIT charts for MA demappers together with CC(5,7) (outer thick line), CC(2,3) (middle thick line) and a repetition code (straight thick line). (a)–(c) Memoryless QPSK. (d)–(f) Differential QPSK.

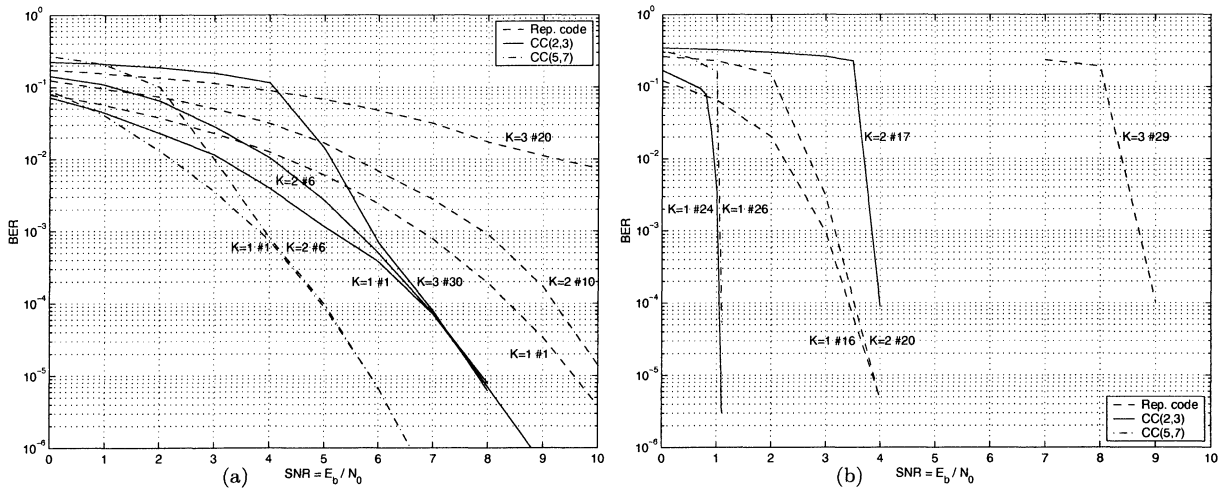


Figure 3: The performance of the repetition code, CC(2,3) and CC(5,7) for different number of users, $L = 100\,000$. (a) Memoryless QPSK. (b) Differential QPSK.

of a dB. In other words, the performance is related to how much a priori knowledge the outer code has on the coded bits [8], i.e., how high $I_{A_y} = I_{E_z}$ is.

Figures 2a–c show that I_{E_z} for the multi user case is never higher than I_{E_z} for the SU case. This means that the SU performance is the lower limit on the performance of the individual users in the multi-user case. SU performance can therefore only be achieved in the point where I_{E_z} in both the SU case and the multi-user case intersect at the same SNR and this will only happen when $I_{A_z} = I_{E_y} = 1$. For two users together with CC(5,7) this seems to be around 4–5 dB and that corresponds well with the simulation of this scenario, Figure 3a. In Figures 3a–b, 'K=2 #6' denotes the performance of $K = 2$ users after 6 iterations. For CC(2,3) SU performance is achieved at 7 dB, Figure 2a and Figure 3a.

In the two-user case with CC(5,7), the tunnel in Figure 2a is opened somewhere between 2 and 3 dB, but the curves will not end up in $I_{E_z} = 1$ and therefore no sharp waterfall in the performance will appear as shown in Figure 3a. The minimum required SNR according to the constellation-constrained capacity for this scenario with $R_c = 1/2$ and memoryless QPSK is close to 2 dB [1]. That is why I_{E_z} at 2 dB is around 0.5 when $I_{A_z} = 0.5$ in Figure 2a.

Based on the interleaver sizes used in [1], it was not possible for three users with QPSK Gray mappers together with CC(5,7) to achieve reliable communication. Figure 2b shows that this is indeed possible if large bit-interleavers and many iterations are applied. The threshold is observed to be close to 10 dB and many iterations are necessary. At this SNR the tunnel will be opened and $I_{E_z} \approx 1$, i.e., a waterfall region will appear. If the SNR is increased even more, the curve for the inner code will still not differ from the 10 dB curve, i.e., is not going to open up a tunnel for stronger outer codes. With three users and CC(2,3) SU performance is, as in the two-user case, expected at around 7 dB according to Figure 2b. This is also indicated in the performance shown in Figure 3a. In addition, Figure 3a shows that using the simple repetition code as the outer code does not provide as good performance as the other codes. Figure 2a shows that two users with the repetition code is not expected to have SU performance until the SNR is higher than 10 dB and after many iterations. For three users SU performance is probably never going to occur even for very high SNR as indicated in Figure 2b and Figure 3a.

The maximum number of users with QPSK mappers together with any half rate code is four according to the constellation-constrained capacity [1], [2]. Even for high SNR, I_{E_z} will not exceed the crucial 0.5 point when $I_{A_z} = 0.5$ as illustrated in Figure 2c. This is necessary for the inner code to be combined with any half rate outer code. According to the constellation-constrained capacity it will eventually just exceed this point when the SNR and the block size go towards infinity [1], [2].

To make the inner code recursive, a differential QPSK mapper [13] is suggested as the mapper for each user. This recursive code is a four state, rate one, code where the transmitted symbol at time index l depends on the symbol at time index $l - 1$. The trellis in the MA demapper for this implementation is a fully connected trellis with 4^K states and 16^K edges per trellis section, where K is the number of users.

The inner code will now be recursive and $I_{E_z} = 1$ when $I_{A_z} = 1$. This means that if a tunnel is opened, the performance will have a very steep slope at this SNR value, at least after many iterations. Figure 2d shows that a tunnel is opened around 1.0 dB in the SU case together with CC(5,7) and just below 1 dB with CC(2,3). That is also confirmed in Figure 3b, where the performance is shown for both cases.

The EXIT chart for two users with differential QPSK is shown in Figure 2e. The tunnel is not opened before approximately 10 dB with CC(5,7). When the tunnel is passed, only a few more iterations are required for convergence. If a weaker outer code is used, like the repetition code or CC(2,3), the EXIT chart indicates that the performance is expected to be much better at a lower SNR. Figure 3b shows that two users with repetition codes have SU performance at 4 dB. With CC(2,3) the tunnel is opened at around 3.5 dB, Figure 2e and Figure 3b. These two figures also show that two users with the repetition code have better performance than two users with CC(2,3) for SNR less than 4 dB.

The EXIT chart for the three-user case is shown in Figure 2f. The tunnel is only opened if a repetition code is used as the outer code and it will then require an SNR higher than 8 dB, Figure 3b.

EXIT charts can also be used to estimate the BER in MA system after an arbitrary number of iterations using methods similar to [7], [8]. It can also predict what the dB loss is compared to the SU case. For example, Figure 2a shows that around six iterations is required at 2 dB in the two-user case with CC(5,7) to get up to the SU 0 dB level. This means that after six iterations in the two-user case at 2 dB, the dB loss compared to the SU case is 2 dB. That can be confirmed by inspection of Figure 3a.

4 Conclusions

The narrow-band system described here has no bandwidth expansion compared to a SU with no spreading. The number of users can therefore not be as high as in MA schemes with bandwidth expansion. This fact is shown here using convergence analysis of the MA channel.

EXIT charts are used as a tool to choose the SU codes to be combined with the chosen mappers. The EXIT charts for the MA system can predict when and also at what SNR SU performance can be accomplished. For example, three users with a two-state CC and memoryless QPSK mappers can have SU performance at 7 dB. This is also confirmed by simulations.

Instead of using memoryless mappers, differential mappers can be used, making the inner code recursive. The EXIT charts show that the performance in this case is only limited by the outer code and perfect knowledge of the coded bits is achievable. Unfortunately, three users is the maximum number of users in the worst case scenario described here. Convergence is then only accomplished if the users have a repetition code as the outer code together with high SNR.

The EXIT chart predicts how many users that can have reliable communication for a given scenario. It can also predict if a threshold is expected in the performance and in that case at what SNR. These fundamental system limits provide useful guidelines when designing MA systems with high spectral efficiency.

References

[1] F. Brännström, *Trellis Code Multiple Access (TCMA) - Detectors and Capacity Considerations.* Lic. Eng. thesis, Chalmers Univ. of Techn., Göteborg, Sweden, Nov. 2000. Available: http://www.ce.chalmers.se/staff/fredrikb.

[2] F. N. Brännström, T. M. Aulin, and L. K. Rasmussen, "Constellation-constrained capacity for trellis code multiple access systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM'2001)*, San Antonio, TX, Nov. 2001, vol. 2, pp. 791–795.

[3] F. N. Brännström, T. M. Aulin, and L. K. Rasmussen, "Iterative detectors for trellis code multiple access," to appear in *IEEE Trans. Commun.*, 2002.

[4] F. N. Brännström, T. M. Aulin, and L. K. Rasmussen, "Iterative multi-user detection of trellis code multiple access using a posteriori probabilities," in *Proc. IEEE Int. Conf. Commun. (ICC'2001)*, Helsinki, Finland, June 2001, vol. 1, pp. 11–15.

[5] S. ten Brink, "Convergence of iterative decoding," *IEE Electron. Lett.*, vol. 35, pp. 806–808, May 1999.

[6] S. ten Brink, "Iterative decoding for multicode CDMA," in *Proc. IEEE Vehicular Technol. Conf. (VTC'99)*, May 1999, pp. 1876–1880.

[7] S. ten Brink, "Design of serially concatenated codes based on iterative decoding convergence," in *Proc. Int. Symp. on Turbo Codes and Rel. Topics*, Brest, France, Sept. 2000, pp. 319–322.

[8] S. ten Brink, "Iterative decoding trajectories of parallel concatenated codes," in *Proc. 3rd IEEE/ITG Conference on Source and Channel Coding*, Munich, Germany, Jan. 2000, pp. 75–80.

[9] D. Divsalar, S. Dolinar, and F.Pollara, "Iterative turbo decoder analysis based on density evolution," *IEEE J. Selected Areas Commun.*, vol. 19, pp. 891–907, May 2001.

[10] S. ten Brink, J. Speidel, and R.-H. Yan, "Iterative demapping and decoding for multilevel modulation," in *Proc. IEEE Global Commun. Conf. (GLOBECOM'98)*, Sydney, Australia, Nov. 1998, pp. 579–584.

[11] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.

[12] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Serial concatenation of interleaved codes: performance analysis, design, and iterative decoding," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 909–926, May 1998.

[13] P. Hoeher and J. Lodge, ""turbo DPSK": iterative differential PSK demodulation and channel decoding," *IEEE Trans. Commun.*, vol. COM-47, pp. 837–843, June 1999.

Multi-User Spatial Multiplexing — A Pragmatic Approach

Mats Bengtsson

Signal Processing, S3, Royal Institute of Technology, KTH, SE-100 44 Stockholm, SWEDEN,
Tel: +46 8 790 84 63, fax: +46 8 790 72 60, Email: matsb@s3.kth.se

Abstract

We suggest a new approach for spatial multiplexing in a narrowband system with several access points serving a number of mobile terminals. In contrast to previously suggested algorithms, this approach is formulated based on a systems perspective, generalizing earlier algorithms for jointly optimal beamforming and power control. Since global channel knowledge is required, the algorithm is mainly intended for system simulations where it provides a benchmark for the performance of any spatial multiplexing scheme with equal rate on the spatial channels.

1 Introduction

The concepts of MIMO (Multiple Input Multiple Output) processing for wireless communication has attracted large attention during the last years. Both information theoretic considerations and practical algorithms show that the performance, in terms of data rate and quality, can be improved significantly using multiple antennas both the receivers and transmitters [1–3].

However, most results that have appeared so far concentrate on the performance of each single link, not on the system level performance. Using more spatial degrees of freedom for each individual link increases the overall interference level in the system and may even reduce the total system throughput. There is no complete information theory for a general multiuser system, but a few studies have been published where the Shannon channel capacities of each link are added together, viewing the interference from other users as spatially colored noise [4, 5]. The interesting conclusion is that it often is better, in terms of system capacity, not to use all available spatial degrees of freedom for each link. Also when it comes to practical algorithms, only a few system level studies have been published, such as [6].

Basically, MIMO systems can be divided into two major categories, *space-time coding* systems, where the transmitter has no channel knowledge but adds redundancy to the signals to enable error correction at the receiver and *spatial multiplexing* systems, where the receiver exploits channel knowledge to form parallel spatial channels. In this contribution, we will consider spatial multiplexing systems, where full knowledge of all channels is available on a system level. On the link level, it is well-known [3] that optimal channel capacity can be reached from the singular value decomposition (SVD) of the channel matrix, using the the left and right singular vectors at the receiver and transmitter, respectively, combined with water filling to determine the data rate of each of the resulting orthogonal spatial channels. In a practical implementation it may be more convenient to use the same rate (or a small selection of different rates) for all the spatial channels. This observation is used in [7, 8] to find alternative spatial multiplexing schemes. For a single link using spatial multiplexing combined with OFDM over a frequency selective channel, the optimal solutions corresponding to a number of different optimality conditions is found using a unified approach in [9].

Here, we follow the same general idea but try to find a solution optimized at the system level, not at the link level, using the following pragmatic approach.

- Use signal to interference plus noise ratio (SINR) of the received signal as the quality of service (QoS) criterion.
- Allow for some controlled self interference between the spatial channels of each user.
- Minimize the overall interference level in the system for a given QoS level for each multiplexed data stream.

The system performance of this solution is evaluated numerically.

2 System Model

Assume a system with one or more transmitters each communicating with one or several receivers. Both

the transmitters and receivers are equipped with array antennas with N and M antenna elements each, respectively. For simplicity, we will assume a narrowband system and only study users that share the same carrier frequency. The discrete time equivalent complex valued baseband channel from transmitter t to receiver r is described by the $M \times N$ channel matrix $\mathbf{H}_{t,r}$. Thus, the signal at receiver r is given by the $M \times 1$ vector

$$\mathbf{y}_r(n) = \sum_t \mathbf{H}_{t,r}\mathbf{x}_t(n) + \mathbf{n}_r(n) \tag{1}$$

where $\mathbf{x}_t(n)$ denotes the $N \times 1$ vector of signals emitted from the antennas of transmitter t. We will assume that the additive noise $\mathbf{n}_r(n)$ is spatially and temporally white, $\mathrm{E}[\mathbf{n}_r(n_1)\mathbf{n}_r^*(n_2)] = \sigma_r^2\mathbf{I}\delta_{n_1,n_2}$.

The communication between each transmitter receiver pair is done by multiplexing the data over one or more spatial channels, formed by linear transformations at the receiver and transmitter. The lth spatial channel for receiver r is determined by a transmit beamformer $\mathbf{v}_{r,l}$ and a receive beamformer $\mathbf{u}_{r,l}$. To emphasize that self interference between spatial channels is treated exactly the same as multiuser interference and in order to keep the notation as simple as possible, a single index i is used below instead of the pair r, l to identify each spatial channel. The receiver using spatial channel i is denoted by $\rho(i)$ and the corresponding transmitter is denoted by $\tau(i)$. The signal transmitted at transmitter t is given by

$$\mathbf{x}_t(n) = \sum_{i;\tau(i)=t} \mathbf{v}_i s_i(n) \tag{2}$$

where s_i denotes the data stream transmitted over the spatial channel i and $\mathbf{v}_i$ is the corresponding transmit beamforming vector. At the corresponding receiver, the data stream s_i is estimated using the receive beamforming vector $\mathbf{u}_i$,

$$\hat{s}_i(n) = \mathbf{u}_i^*\mathbf{y}_{\rho(i)}(n)\ . \tag{3}$$

Thus, the resulting spatial channel i including transmit and receive beamformers, is a scalar channel with gain $\mathbf{u}_i^* H_{\tau(i),\rho(i)}\mathbf{v}_i$. Note that this formulation is general enough to describe not only that each access point communicates simultaneously with several mobile stations but also situations like soft hand-over, where more than one access point communicates with the same mobile.

3 Algorithm

The main strategy is a straightforward generalization of optimal downlink beamforming [10, 11], minimizing the total transmitted power under the constraint that the received signal of each spatial channel has a sufficient signal to interference plus noise ratio (SINR). Note that minimizing the total transmit power will also help reducing the overall interference level in the system.

Assume that the signals multiplexed over the spatial channels are mutually uncorrelated and have equal power normalized to one. If γ_i denotes the SINR threshold for spatial channel i, the resulting optimization problem is

$$\begin{aligned} &\min \sum_j \|\mathbf{v}_j\|^2 \\ &\text{s.t.} \ \frac{|\mathbf{u}_i^* H_{\tau(i),\rho(i)}\mathbf{v}_i|^2}{\sum_{j\neq i}|\mathbf{u}_i^* H_{\tau(j),\rho(i)}\mathbf{v}_j|^2 + \sigma_{\rho(i)}^2\|\mathbf{u}_i\|^2} \geq \gamma_i \\ &\quad \|\mathbf{u}_i\|^2 = 1, \ \ i = 1, \ldots, I \end{aligned} \tag{4}$$

This is a non-linear and non-convex optimization problem in the variables $\mathbf{v}_i$ and $\mathbf{u}_i$. For the case of a single transmitter and a single receiver, it is easy to show that the left and right singular vectors of $\mathbf{H}$ provide a local but not a global optimum. Thus, one would have to resort to global optimization methods, such as cutting plane, branch and bound or stochastic optimization methods such as genetic algorithms or simulated annealing to find the global optimum. However, an alternative heuristic approach is presented in [12], based on the observation that for fixed transmit beamformers, the optimum receive vectors are given by the standard linear MMSE multiuser detector and for fixed receive beamformers, the problem reduces to the problem of jointly optimal transmit beamforming which can be solved using the algorithms in [10, 11, 13]. Iterating between these receive and transmit solutions will decrease the cost function in each step and is thus guaranteed to converge. Numerical experiments have shown that even if the algorithm may find different local optima with different starting points, the resulting transmitted power is almost identical, indicating that the solution is very close to the global optimum. We have tried several variations of this idea, based both on the convex reformulation in [13] and the extended power control loop in [10]. The best convergence results have been obtained using a variation of the "Virtual Downlink Method" of [12], but solving explicitly for the uplink powers in each iteration instead of just applying a power control iteration. The main steps of the resulting algorithm are given by:

- Find initial receive vectors $\mathbf{u}_i$.
- Repeat until convergence:
  - Update normalized transmit vectors $\mathbf{v}_{i,\mathrm{norm}}$ and virtual power levels $\tilde{p}_i$ using the "virtual uplink" formulation.

- Find MMSE receive vectors $\mathbf{u}_i$.
- Find transmit power levels p_i to reach the target SINR levels.
- Replace negative p_i by a small positive constant.

- After convergence: set $\mathbf{v}_i = \sqrt{p_i}\,\tilde{\mathbf{v}}_{i,\mathrm{norm}}$.

The details of each step may be found in [10–12]. The last step of the main loop is necessary since the transmit beamforming and power control subproblem may be infeasible during the first iterations before the receive beamformers have converged, resulting in some negative power levels. The strategy to replace these with a positive number is similar in spirit to some modified Newton methods where an indefinite Hessian matrix is replaced by a positive definite approximation [14]. Using the semidefinite formulation of [13] to determine the transmit beamformers in each iteration has been found to require more iterations before convergence.

Unfortunately, this algorithm may fail completely to find a feasible solution even when the problem is feasible. In order to find a feasible initial point, a version of the algorithm in [9] was used to iteratively find the optimal set of beamformers for each transmitter receiver pair, viewing the signals from all other transmitters as colored noise. Even though this iteration will not converge to an optimum of (4), it is still an interesting heuristic alternative since it can be used in a decentralized implementation, where the beamformers for each link are determined using information available locally at the corresponding receive transmit pair.

4 Numerical Examples

The performance of the proposed strategy is first evaluated for a single link, i.e. point-to-point communication between two nodes. We study an idealized multipath scattering simulated scenario where the channel matrix contains i.i.d. complex Gaussian fading variables. Both the transmitter and the receiver are equipped with 4-element antenna arrays that are used to transmit 1,2,3 or 4 independent data streams. The same SINR threshold is used for all the spatial channels. Figure 1 shows the total transmit power required to reach different SINR thresholds. The channel gain and noise variance is normalized such that the required transmit power to obtain 0dB SINR using one spatial channel is 0dB. The results are averaged over 2000 independent channel realizations. For comparison, the the channel was diagonalized completely using the singular value decomposition. The dashed curves show the total transmitted power needed to obtain an SNR (note that there is no self interference

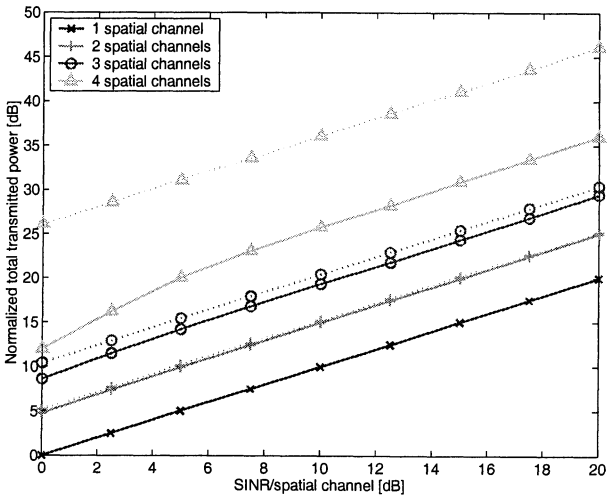


Figure 1: Spatial multiplexing over a single link, suggested algorithm (solid lines) and SVD (dashed lines).

in the SVD solution) above the threshold for all the spatial channels. This illustrates the advantage of allowing some self interference, especially when using many spatial degrees of freedom.

Even though the total data rate increases with the number of spatial channels, there is a high price paid in terms of extra transmitted power, i.e. extra background interference for possible other users.

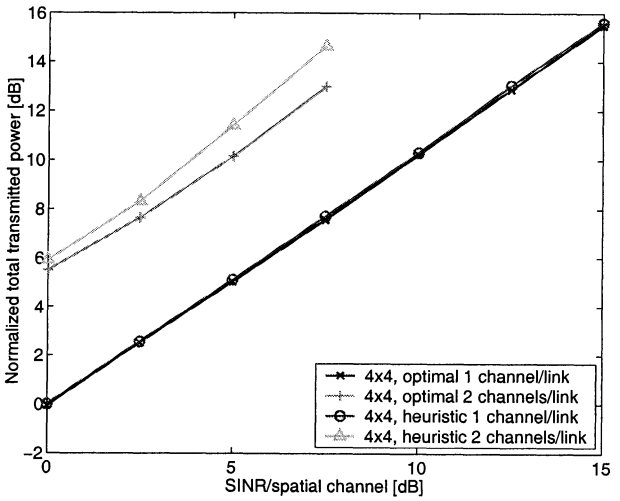


Figure 2: Spatial multiplexing in a random multiuser scenario

Next, the strategy described above was evaluated in a simulated narrowband system with four fixed transmitting access points serving four randomly placed receiving terminals. Again, the channel matrices were random with i.i.d. Gaussian elements, corresponding to an idealized rich scattering propagation. However, the gain of each channel matrix was determined

based on a simple geometrical model with distance dependent path loss and additive lognormal fading. Each terminal was allocated to the access point with highest average path gain. Note that this is not necessary the best strategy when using smart antennas [15]. Figure 2 shows the total transmitted power necessary to exceed a certain SINR threshold at all receivers. The curves for two spatial channels/link are interrupted at $\gamma_i = 7.5$dB since for higher values, a feasible solution was only found in a small fraction of the scenarios. Trying to use 3 or 4 spatial channels per user failed completely in almost all scenarios. This confirms the conclusions in [5] that it often is better to use only one or two spatial degrees of freedom for each link in a highly populated system.

For comparison, we have also included the performance of the algorithm used to find an initial feasible point. Interestingly enough, this algorithm performs almost as good and has the additional advantage that it may be used in a decentralized implementation. Note that the small difference between the two algorithms may also be an indication that the optimization procedure gets stuck in a local optimum.

5 Summary

We have proposed a pragmatic approach to the use of multiple antennas at both transmitters and receivers in a multiuser system. The main focus is to minimize the overall interference level in the system, maintaining a certain QoS level expressed in terms of the received signal to interference plus noise ratio. An advantage of this formulation is the possibility to reuse existing ideas and methods from the area of downlink beamforming, such as adding robustness to channel estimation errors or handling constrains on the dynamic range of each antenna element, see [11, 16].

In contrast to so-called space-time coding methods, we assume that the channel variations are slow enough that it is feasible to maintain accurate estimates of all the channels in the system, for example in an indoor application [17]. Much work remains to find robust optimization algorithms to solve the problem. So far, we have used an algorithm similar to the one proposed in [12] to search for the optimal solution and an iterative application of an algorithm from [9] to find a feasible solution to use as a starting point. This heuristic second algorithm has been found to produce near optimal performance and may therefore be an interesting practical alternative, especially since it may be implemented in a decentralized fashion.

Note also the difference to channel capacity achieving spatial multiplexing for a single link, which is obtained using the singular value decomposition combined with bit loading (water filling) using different data rates on each spatial channel in an optimal way. Here we mainly consider a simpler setup that uses the same modulation and coding for all spatial channels, even though it is possible to assign different QoS requirements (γ_i) for different spatial channels.

The numerical experiments confirm that it often is better not to use all spatial degrees of freedom for each link since it would cause a too high interference level in the system.

References

[1] I. Emre Telatar. Capacity of multi-antenna gaussian channels. Technical report, Bell Laboratories, Lucent Technologies, `http://mars.bell-labs.com`, 1995.

[2] Gerard J. Foschini. Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. *Bell Labs Techn. Journ.*, 1(2):41–59, autumn 1996.

[3] Gregory G. Raleigh and John M. Cioffi. Spatio-temporal coding for wireless communication. *IEEE Trans. Commun.*, 46(3):357–366, March 1998.

[4] Daniel W. Bliss, Keith W. Forsythe, Alfred O. Hero, and A. Lee Swindlehurst. MIMO environmental capacity sensitivity. In *Proc. 34rd Asilomar Conf. Sig., Syst., Comput.*, volume 1, pages 764–768. IEEE, 2000.

[5] Rick S. Blum, Jack H. Winters, and Nelson R. Sollenberger. On the capacity of cellular systems with MIMO. In *Proc. VTC2001 Fall*, volume 2, pages 1220–1224. IEEE, 2001.

[6] Magnus Karlsson Stefan Parkvall, Maria Samuelsson, Leo Hedlund, and Bo Göransson. Transmit diversity in WCDMA: link and system level results. In *Proc. VTC 2000 Spring*, volume 2, pages 864–868, 2000.

[7] Hemanth Sampath, Petre Stoica, and Arogyaswami Paulraj. Generalized linear precoder and decoder design for MIMO channels using the weighted MMSE criterion. *IEEE Trans. Commun.*, 49(12):2198–2206, December 2001.

[8] Timothy A. Thomas and Frederick W. Vook. MIMO strategies for equal-rate data streams. In *Proc. VTC 2001 Fall*, volume 2, pages 548–552, October 2001.

[9] Daniel Pèrez Palomar, John M. Cioffi, and Miguel Angel Lagunas. Joint Tx/Rx beamforming for multicarrier MIMO channels: a con-

vex optimization approach. submitted to IEEE Trans. SP, February 2002.

[10] Farrokh Rashid-Farrokhi, K.J. Ray Liu, and Leandros Tassiulas. Transmit beamforming and power control for cellular wireless systems. *IEEE J. Select. Areas Commun.*, 16(8):1437–1450, October 1998.

[11] Mats Bengtsson and Björn Ottersten. Optimal and suboptimal transmit beamforming. In Lal C. Godara, editor, *Handbook of Antennas in Wireless Communications*, chapter 18. CRC Press, August 2001.

[12] Jae-Hwan Chang, Leandros Tassiulas, and Farrokh Rashid-Farrokhi. Joint transmitter receiver diversity for efficient space division multiaccess. *IEEE Trans. Wireless Comm.*, 1(1):16–27, January 2002.

[13] Mats Bengtsson and Björn Ottersten. Optimal downlink beamforming using semidefinite optimization. In *Proc. 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 987–996, September 1999.

[14] Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts, 1995.

[15] Mats Bengtsson. Jointly optimal downlink beamforming and base station assignment. In *Proc. IEEE ICASSP*, volume V, pages 2961–2964, Salt Lake City, Utah, May 2001.

[16] Mats Bengtsson. Robust and constrained downlink beamforming. In *Proc. EUSIPCO*, volume III, pages 1433–1436. EURASIP, September 2000.

[17] D.P. McNamara, M.A. Beach, P.N. Fletcher, and P. Karlsson. Temporal variation of multiple-input multiple-output (MIMO) channels in indoor environments. In *Proc. IEE AP Conference*, volume 2, pages 578–582, 2001.

Multiuser Diversity and Maximum Entropy Scheduling

Mathias Johansson[1],

[1]Uppsala Universitet, Signals & Systems, Box 528, SE-751 20 Uppsala, Sweden. E-mail: mj@signal.uu.se.

Abstract

Multiuser diversity is exploited by tracking rapidly time-varying radio channels and allocating more transmission resources to users with good instantaneous channel conditions.

We explain how to simultaneously exploit multiuser diversity and variable and highly uncertain source bit rates to maximize system capacity. A framework for scheduling under uncertainty is derived based on the maximum entropy principle.

By using the channel and the source rates as independent sources of diversity the sum capacity is substantially improved. The performance of maximum entropy scheduling is compared to another recent proposal for exploiting multiuser diversity.

1 Introduction

In wireless communication networks channel fluctuations can be combatted by the use of diversity. Time, frequency, and space are the most commonly recognized forms of diversity. Recently, however, *multiuser diversity* has been receiving an increasing amount of attention [1], [2], [3], [4]. The term multiuser diversity was introduced by Knopp in [5] where he notes that multiuser diversity is *selection diversity performed at the transmission end.* Multiuser diversity is exploited by tracking the rapidly time-varying channels and allocating more transmission resources to users with good instantaneous channel conditions. Hence, scheduling based on channel predictions and link adaptation is recognized as a method to increase the system capacity and improve individual bit rates. As the channel fluctuations and the number of users increase so does the sum-of-rates capacity. Interestingly, the sum capacity is substantially higher even for a moderate number of users with independent Rayleigh fading than for a fixed AWGN channel [4].

Assuming perfect channel knowledge Knopp and Humblet [1] showed that in order to maximize the sum-of-rates capacity the optimal scheduling strategy for a single-cell uplink scenario is to time-divide the channel. At each time instant the transmitter should only send data to the user having the best channel. This conclusion however implicitly requires that there is always data to send. In reality, data streams from computer networks are bursty, and consequently it is important to take the instantaneous source rates into account. Otherwise, channel capacity may be allocated to users with less bandwidth need than is actually available. The result is that other users with higher instantaneous source rates needlessly experience lower throughput, potentially resulting in buffer overflows. Capacity is thus wasted. In summary, although multiuser diversity as used up to date utilizes the variability of the channels, *it neglects the variability of the source rates.*

In this paper we explain how to utilize variable and highly uncertain source rates to further increase capacity and present a scheduling framework based on [6] which explains how to optimally utilize multiuser diversity in various settings. We also extend the subject to multiple access points to obtain a methodology for optimal resource allocation in rather general multiuser communication systems. The suggested framework was introduced in [7]. Here we extend the study with comparisons to another recently proposed scheduler [4] which utilizes multiuser diversity.

2 Problem Formulation

We investigate the problem of allocating transmission resources over time. A scheduler is situated at the interface between a fixed and a wireless network. For each user there is a buffer in which source data are received from the fixed network. The aim of the scheduler is to distribute the buffer contents over the available transmitters so as to maximize the total throughput.

In Figure 1 an overview of the system is given. There are U users, and equally many buffers. We will schedule the use of the transmitters for T time slots. During the scheduling horizon T each buffer is filled with $n_u = \sum_{t=1}^{T} n_{ut}$ bits, u denoting the user index, and t time slot. A buffer may also have a number, S_u, of bits remaining in stock from previous scheduling rounds. Maximizing system throughput is equivalent to minimizing the buffer contents at the end of the scheduled time horizon. This amounts to minimizing

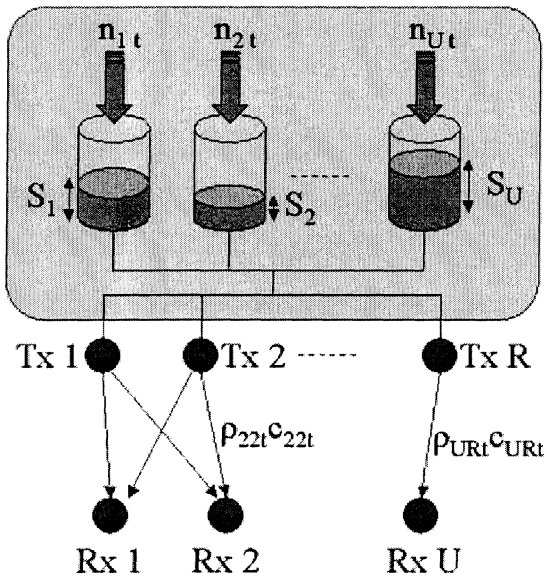


Figure 1: The system consists of U buffers, one for each receiver. R transmission resources are available and user u receives $\rho_{urt}c_{urt}$ bits at time t from transmitter r.

the loss function

$$L = \sum_{u=1}^{U} g\Big(S_u + n_u - \sum_{t=1}^{T}\sum_{r=1}^{R} c_{urt}\rho_{urt}\Big) \ , \quad (1)$$

where $g(x) = x$ if $x > 0$, else $g(x) = 0$. The time-varying effective capacity, i.e. the number of bits that can be transmitted over a specific link at a prescribed error rate, is denoted by the integer c_{urt}, with r being the transmitter index. The minimization is achieved by adjusting ρ_{urt}, the fraction of transmission resource r the scheduler allocates to user u at time t.

This would be a straightforward minimization if the actual time-varying source rates n_{ut} were known. However, as the nature of data traffic is highly unpredictable we have to resort to reasoning under uncertainty, and thus use probability theory [8], [9]. By assigning prior probabilities for the source rates and then minimizing the expectation of our loss function (1) we obtain a solution free from inconsistencies.

3 Solution

We now assume knowledge of the *average* number of incoming bits into each buffer over the scheduling horizon T and denote this average by $< n_u >$. In order to specify the prior distribution for the actual amount of incoming data, n_u, we will use the maximum entropy principle. The maximum entropy principle was introduced by Jaynes [10] and gives the least biased probability distribution possible on given testable information. The principle states that the probability distribution that corresponds to our knowledge while making the fewest additional assumptions is the one with maximum entropy subject to the constraints given by our prior knowledge. The maximum entropy distribution is simply the distribution which can arise in the greatest number of ways when the outcomes are constrained to agree with the given information. Thus it is a predictive distribution, not to be confused with assuming a certain sampling distribution.

In our case, with knowledge of average source rates, the probability distribution with maximum entropy is

$$P(n_u|I) = \frac{1}{< n_u > +1}\Big(\frac{< n_u >}{< n_u > +1}\Big)^{n_u} \ . \quad (2)$$

This is a negative exponential distribution and the maximum entropy derivation was first given in [6]. Using this distribution, we then calculate the expectation of the loss function (1) and obtain:

$$\begin{aligned} < L > \ &= \sum_{u=1}^{U}\sum_{n_u=0}^{\infty} P(n_u|I) g(S_u + n_u - x_u) \\ &= \sum_{u=1}^{U}\Big(\sum_{n_u=0}^{\infty} P(n_u|I)(S_u + n_u - x_u) \\ &\quad - \sum_{n_u=0}^{x_u - S_u} P(n_u|I)(S_u + n_u - x_u)\Big) \\ &= \sum_{u=1}^{U} < n_u > \Big(\frac{< n_u >}{< n_u > +1}\Big)^{x_u - S_u}, \quad (3) \end{aligned}$$

where

$$x_u = \sum_{t=1}^{T}\sum_{r=1}^{R} c_{urt}\rho_{urt} \ . \quad (4)$$

The full derivation of (3) is omitted here but can be found elsewhere [11]. Further, notice that we here assume knowledge of the effective capacity. In [11], the expected loss is derived for channel predictions of varying accuracy.

Finding the minimum of (3) is in general an exercise in non-linear programming. The basic constraints on ρ_{urt} are:

$$\sum_u \rho_{urt} \leq 1 \quad \forall r, t \quad (5)$$

$$0 \leq \rho_{urt} \leq 1 \quad \forall u, r, t, \quad (6)$$

but we may also have an additional number of matrix equalities and inequalities representing constraints on different resources. For instance some transmitters may only be shared by a subset of users, some may be shared by all, and yet others may be dedicated to a

certain user. Such constraints are included by restricting the set of values that ρ_{urt} can take. These types of constraints as well as many others are easily handled by available software for solving non-linear programming problems and present no conceptual difficulties. The flexibility of simple linear constraints can transform the general problem to a variety of interesting special cases, all represented by the same average loss but with different optima due to the restrictions on ρ_{urt}. A common requirement is that certain users may need a minimum average bit rate φ_u, which can be represented as additional linear constraints

$$\sum_{t=1}^{T}\sum_{r=1}^{R} c_{urt}\rho_{urt} \geq \varphi_u \ , \tag{7}$$

fitting well into the non-linear programming framework. Similarly, other forms of requirements may be appropriate in other problems (maximum allowed rates for instance) and they are handled similarly.

Fairness requirements and priorities are achieved by other means. Priorities can be introduced as multiplicative factors for each user's contribution to the total loss. Regarding the difficult subject of fairness, note that as the loss function increases with buffer sizes, fairness can be expected simply because buffer sizes are known and users with large buffers therefore good candidates for allocations. Thus, *ad hoc* compensations in later scheduling rounds are not required. In the simulations section of the full paper, one of the major results is the level of fairness the scheduler achieves.

4 Simulations

To assess the performance of the maximum entropy scheduler derived in Section 3, we compare the scheduler with another recent proposal, proportional fair scheduling [4], which exploits multiuser diversity while also including a mechanism for obtaining fairness between users.

The proportional fair scheduler works as follows. The data rates that the users' channels can support at each time slot t (the effective capacity) is known to the scheduler. The scheduler then keeps track of the average throughput $T_u(r,t)$ of each user u in a past window of length t_c. At each access point r and time slot t, the scheduler transmits exclusively to the user with the largest

$$\frac{c_{urt}}{T_u(r,t)} \tag{8}$$

The parameter t_c is used as a forgetting factor in the calculation of the windowed average throughput. It is used as a means of obtaining fairness, by giving a user access to a channel when its effective capacity is high relative to its own average capacity over the time scale t_c. In [4], a single base station is considered. Here, we adapt the proportional fair scheduler to multiple access points simply by treating an additional base station as more time slots. In other words, if we are to assign two access points and three time slots, the scheduler works exactly as if it were to schedule one base station and six time slots. After each single assignment, the average throughput $T_u(i)$ (where i indexes assignments regardless of whether it describes time slot or access point) is recalculated according to [3]:

$$T_u(i) = (1-\frac{1}{t_c})T_u(i-1) + \frac{1}{t_c}c_{u,i-1}\delta(u-u^*) \ , \tag{9}$$

where $\delta(u-u^*) = 1$ if user u was the transmitting user u^* in the most recent assignment, otherwise, $\delta(u-u^*) = 0$.

Table 1: Parameters for the comparison of proportional fair scheduling with the maximum entropy scheduler. Average inflows per time slot, $\frac{<n_i>}{T}$, Rayleigh parameter, γ, for PDF for effective channel capacity (number of bits per time slot), and the corresponding expected effective capacity, $< c_{urt} >$

	$\frac{<n_i>}{T}$	γ	$< c_{urt} >$
User 1	100	100	125
User 2	140	100	125
User 3	60	80	100

The schedulers were run on the same data sets, with source rates n_{ut} drawn from a Poisson random number generator, and effective capacities from a Rayleigh random number generator. The parameters used are listed in Table 1. The choices of distributions for source rates[1] and effective capacities are somewhat arbitrary and neither choice would in general be an accurate model of reality. Choosing a Rayleigh distribution for the effective capacities would be appropriate for a system supporting a continuous range of bit rates mapped directly from the channel gain predictions. In systems with a small set of supported bit rates, the simulated effective capacities would need to be drawn from some other system-specific distribution.

The simulated scenario consisted of three users and two access points. The scheduling horizon was $T = 3$ time slots, and the schedulers were run for a total of 60 time slots. The results listed in Table 2 are averages from 100 realizations. The table reports average throughput and average buffer levels after the 60 time slots.

These simulations show the importance of taking source rates and actual buffer levels into account. It

[1]The simulation of an individual Internet data source is a notoriously difficult problem. For a discussion see [12].

Table 2: Results for the comparison of proportional fair scheduling with the maximum entropy scheduler. The average number of bits remaining in the buffers after 60 time slots are listed in columns 1 and 2 for the proportional fair scheduler (pf) and the maximum entropy scheduler. The last two columns display average total throughput in bits.

	S_{60}(pf)	$S_{60}(c_{urt})$	Tp(pf)	Tp(c_{urt})
User 1	81	31	5922	5972
User 2	1547	77	6859	8329
User 3	22	31	3582	3573
Total	1650 bits	139 bits	16363 bits	17874 bits

Table 3: Test 2. Average inflows per time slot, $\frac{<n_i>}{T}$, Rayleigh parameter, γ for PDF for effective channel capacity (number of bits per time slot), and the corresponding expected effective capacity, $< c_{urt} >$

	$\frac{<n_i>}{T}$	γ	$< c_{urt} >$
User 1	110	100	125
User 2	110	80	100
User 3	110	110	138

also displays the level of fairness obtained by using the maximum entropy scheduler. It should be evident that fairness must be defined according to each user's need for capacity in relation to the available capacity. For instance, compare in Table 2 the precise correspondence between the average source rates of user 1 and 2, and the obtained bit rates. (Both users have the same average effective capacity.)

Comparing the results for maximum entropy scheduling and proportional fair scheduling, the difference in total throughput may not be overwhelming but the individual throughput of user 2 is severely degraded when buffer contents are neglected. In terms of buffer levels it is clear that the second user's buffer would overflow, causing further throughput degradation and increased delays due to the invoking of higher-layer mechanisms such as decreasing transmission rates or retransmissions.

It can be noted that the proportional fair scheduler would approach the performance of the maximum entropy scheduler were the transmission buffers constantly flooded with data. A more important observation is that this case is prevented from ever occurring in a real system due to the omnipresent rate-control mechanisms such as provided by TCP. It can be concluded that schedulers must take buffer contents into account. The additional use of *source rate diversity* further increases the performance of the maximum entropy scheduler.

A similar test was performed with slightly different parameters, letting all users have the same average source rates as described by Table 3. The results

Table 4: Test 2, results. The average number of bits remaining in the buffers after 60 time slots are listed in columns 1 and 2 for the proportional fair scheduler and the maximum entropy scheduler. The last two columns display average total throughput in bits.

	S_{60}(pf)	$S_{60}(c_{urt})$	Tp(pf)	Tp(c_{urt})
User 1	182	184	6424	6421
User 2	1160	292	5435	6303
User 3	116	167	6486	6435
Total	1458 bits	643 bits	18345 bits	19159 bits

are listed in Table 4. Although the setting was designed to better suit the proportional fair scheduler, the second user still obtains significantly lower bitrates when using the proportional fair scheduler.

Another important result from the first simulation can be observed by studying the throughput obtained for the second user, 8329 bits. Instead of trying to use the multiuser diversity to our advantage one could just split the available bandwidth into three equal parts and always transmit to all users. Instead of 8329 bits, user 2 would then obtain a total throughput of $\frac{125}{3} \times 2 \times 60 = 5000$ bits. Thus, the individual throughput increased by 67% when using the fluctuating channel as a source of diversity. The reason that the proportional fair scheduler only achieves an increase of 37% is that it does not take the varying source rates into account. Evidently, there are substantial benefits associated with taking advantage of the fact that, on average, the other users' source rates are lower than their effective capacities. Neglecting this source of diversity results in decreased individual and total throughput.

5 Comments

We formulated a problem of scheduling radio resources in the presence of large uncertainties concerning incoming data rates. By using the maximum entropy principle we obtained the least biased probability distribution possible given average inflows. Put differently, the maximum entropy principle yields predictive distributions representing a certain state of knowledge, irrespective of the long-run sampling distributions. Using the maximum entropy distribution for inferences should not be confused with assuming more or less arbitrarily a specific sampling distribution. Indeed, it can be shown that the frequency connections are irrelevant for making inferences. It is the demonstrable information content only, i.e. the specified mean values $< n_u >$, which is used when making inferences based on the maximum entropy principle. This is in sharp contrast to assuming a certain arrival distribution without paying attention to the actual

information content.

The maximum entropy scheduler results in solving a non-linear program, which facilitates different constraints on resource sharing and quality of service. A drawback is the computational complexity of finding the constrained minimum of a non-linear function. It would however be possible to use an approximate scheduling policy based on tabulated decision regions and thus obtain high performance at low complexity.

It was shown that the maximum entropy scheduler outperformed the proportional fair scheduler in terms of throughput and fairness. This was seen to be a consequence of not only utilizing multiuser diversity, but also acknowledging the fact that the need for bandwidth is highly variable.

References

[1] R. Knopp, P.A. Humblet, "Information Capacity and Power Control in Single-Cell Multiuser Communications", Proc. IEEE ICC 95, June 1995.

[2] D. N. Tse, "Optimal power allocation over parallel Gaussian channels", Proc. ISIT, June 1997.

[3] David N.C. Tse, "Multiuser Diversity in Wireless Networks", Presentation at Stanford University, April 16, 2001.

[4] Pramod Viswanath, David N.C. Tse and Rajiv Laroia, "Opportunistic Beamforming using Dumb Antennas", Submitted, Sep 11 2001.

[5] Raymond Knopp, *Coding and Multiple-Access over Fading Channels*, D.Sc. Thesis, Swiss Federal Institute of Technology (Lausanne), Dept. of Electrical Engineering, 1997.

[6] E. T. Jaynes, "New engineering applications of information theory", *Engineering Uses of Random Function Theory and Probability*, Bogdanoff and Kozin (eds.), Wiley, New York, pp 163-203 1963.

[7] Mathias Johansson, "Probabilistic Resource Allocation and Scheduling for Multiuser Communication Systems", Proc. IEEE Globecom'01, Nov 25-29 2001.

[8] R. T. Cox, "Probability, Frequency and Reasonable Expectation", American Journal of Physics, Vol. 14, No. 1, pp. 1-13, Jan-Feb 1946.

[9] E. T. Jaynes, *Probability Theory - The Logic of Science*, Unfinished manuscript (http://bayes.wustl.edu/etj/prob.html), 1996.

[10] E. T. Jaynes, "Information Theory and Statistical Mechanics", The Physical Review, Vol. 106, No. 4, pp. 620-630, May 15 1957.

[11] Mathias Johansson, "Resource Allocation Under Uncertainty Using the Maximum Entorpy Principle", in preparation.

[12] Sally Floyd, Vern Paxson "Difficulties in simulating the Internet", IEEE/ACM Transactions on Networking, vol. 9, issue 4, pp. 392-403, Aug 2001.

Performance Analysis of Impulse Radio Based on an Ultra Wideband Channel Model for Indoor Communications

Andres Alayon Glazunov

Telia Research AB, 123 86 Farsta, Vitsandsgatan 9, Hus B, andres.a.alayon@telia.se

Abstract

A performance analysis of a RAKE receiver for impulse radio (IR) ultra wideband (UWB) indoors communications is presented for single link pulse position demodulation scheme. The channel model, which has the broadly accepted delayed tap filter form, models the response of the environment to signals of sub-nanosecond duration. Due to the high resolution in time, the birth-death process in the distance-time plane as well as the fading variation are confined to the paths of constant distance to time ratio. Finally, it is also shown that in order to obtain a low symbol error rate or at least a performance closer to the one in AWGN channels the number of fingers in an ideal conventional RAKE receiver would be too high.

1 Introduction

The Impulse Radio (IR) technique has been known since the late 1940s. Because of the grate resolution in time obtained with the utilized wideband signals, the major application has been within the radar-based technology. But it was not before the last decade that extensive research works in the field of IR communications systems had been initiated [1&2], including both theoretical studies and measurement campaigns.

In spite of that, there are no reliable IR-UWB channel models that could be used in order to make a realistic assessment of such a technique for communications purpose. The goals of this study are first to develop an IR-UWB channel model that comprises most features that are relevant for performance assessment of a communications systems design such as time dispersion and fading variation. Then we make use of this model in order to assess some figures of merit relevant to the RAKE receiver performance in indoor multipath channels with AWGN for time shift keying (TSK) or pulse position modulation, as it is also known. The purpose is to assess the suitability of the classical RAKE receiver for IR. This paper is organized as follows: first a brief introduction to the UWB basic concepts and the applied to it time shift keying. In section 3 the UWB channel is presented followed by the numerical results and the summary.

2 UWB basics

Ultra-wide band signals are by definition signals with a fractional bandwidth of at least 25 % or spanning over a frequency band of 1.5 GHz. The fractional bandwidth is defined as twice the quotient of the difference and the sum of the largest and the lowest points at the 10 dB level of the signal spectrum.

Impulse radio makes use of pulse trains in which the information may, for instance, be transmitted by a modulation technique known as pulse position modulation (PPM) or time shift keying (TSK). As the name indicates the "0" and "1" bits are send separated in time. The base band pulses that are commonly of sub-nanosecond duration are sent with a very low duty cycle and high pulse repetition rate (PRR). The IR is a spread spectrum technique in which, like to direct sequence spreading systems (DSSS) techniques, different users are separated by different pseudo-noise random sequences of good correlation characteristics.

The basic signal used in IR is often the Gaussian monocycle. Both this function and its derivatives [3] are relevant to the modeling of the propagation channel and the analysis of the IR receiver performance. Namely, the antennas have a differentiating effect on the UWB pulse. This is a direct consequence of Maxwell's equations and how the radiated electric field is excited in the antenna. It may be shown that the radiated electric field strength is proportional to the time derivative of the drive current regardless the waveform. Distortion due to penetration into the different materials that compose the surrounding environments is also an effect to take into account as well as superposition of signals due to multipath propagation. The former phenomenon is a result of the different attenuation experienced by the UWB signal over the entire frequency spectrum occupied. The later phenomenon is the one we are going to treat in this report, in order to assess the IR receiver performance.

2.1 IR Receiver Model and Signal Processing

The analysis will be restricted to the single link case. The focus will be on binary time shift keying modulation (BTSK) performance.

The signal received at position u is a filtered version of the transmitted one [1]. The sources of filtering are the antennas and the propagation impairments imposed by the radio channel, and then the received signal for a single pulse bit at a single link is expressed as the convolution of the transmitted pulse and the channel impulse response (ChIR).

In order to detect a given transmitted bit, the matched filter concept is used by means of RAKE receiver architecture. Then the decision variable is constructed by the multiplication with the signal template corresponding to the waveform of the binary modulation scheme and integrated over the bit duration frame.

Assume that the RAKE receiver consists of L correlators. Then the received signal may basically be represented as the sum of three components comprising the L-strongest paths, the complementary set that includes the remaining L_{tot}-L (L_{tot} denotes the total available signal paths) paths and noise. The second term can be considered as a signal dependent Gaussian distributed self-noise that is statistically independent of the receiver noise.

In order to obtain an approximated performance analysis, the union bound on the symbol error probability can be used [4]. As is well known the error probability is proportional to the SNR value obtained from the decision variable at each position u. In turn the SNR is directly proportional to the channel multipath power gain obtained combining the output of L RAKE-fingers and the normalized correlation function of the L-RAKE reference signal. Further, the SNR is inversely proportional to the variance due to the presence of the AWGN noise times the same term as in the denominator plus the corresponding value that accounts for the signal-dependent self-noise, which is proportional to signal cross-correlation function between the L-RAKE reference signal and its complement.

Finally, taking the expected value with respect to u, or which is equivalent computing the algebraic average over all positions an approximated measure of the performance of the IR RAKE receiver is obtained as function of average symbol signal to noise ratio at the receiver. A more correct assessment is to derive the distribution for the received SNR and integrate the symbol error rate over all SNR.

3 UWB Channel Model

Extensive work on channel modeling for indoor communications has been performed using sinus-like signals with large signal bandwidths [5&6]. The grate part of this knowledge may be re-used in order to develop reliable IR-UWB channel models. A statistical model for UWB signals is presented in [7], tough it do not consider the channel taps evolution over time.

In this section we describe the basic features of the IR-UWB channel proposed here, such as delay spread, power delay profile (PDP) and fading behavior of multipaths.

The channel impulse response (ChIR) is modeled as a discrete tapped delay line filter at position u and time t:

$$h(u,t) = \sum_{k=1}^{L} a_k(u,t)\delta(t-\tau_k(u)), \qquad (1)$$

where $\{a_k(u,t)\}$ are the time-variant tap weight filter coefficients, $\{\tau_k(u)\}$ is the corresponding time of arrival and δ is the Dirac delta function. The weight coefficients are random parameters representing the multipath variation of the channel together with the time of arrival of each impulse response due to reflections and scattering from different obstacles in the propagation environment.

3.1 Birth-Death Process in the Distance-Time Plane

Due to the broadband nature of IR signals and the inherent to them high resolution it is possible to track each multipath component and follow it evolution over time. In IR-UWB systems there is no fading in the sense we have been used to until now. As before, the fading variations are governed by the surrounding environment and the propagation scenario, but the behavior presents more deterministic features. Consequently, ray-tracing techniques seems more suitable for ChIR modeling than before.

In lack of published measurement results for IR ChIR that we make use of the conjectures above and the only paper [8], where indoor measurements at the 1.8 GHz band with a bandwidth of 600 MHz are presented. The fractional bandwidth of this experiment equals approximately 33%. Results presented there indicate that the channel taps follows almost linear paths in the (u=position, t=delay) plane. This is the result of reflections from surrounding surfaces along the mobile paths that could be separated in time due to the high bandwidth used in the measurements. In a similar way IR signals would behave in indoor environments.

Keeping this in mind it is easy to see that the delay paths are then a linear function of the distance along the mobile path:

$$u = \alpha_k(\tau_k - \hat{\tau}_k), \qquad (2)$$

where $\{\hat{\tau}_k\}$ are the random times of arrival at position u=0 along the mobile path. The slopes $\{\alpha_k\}$ are also random variables modeled through the slope angle ϕ_k, $\alpha_k = \tan(\phi_k)$, which is modeled with a Gaussian distribution:

$$p_\phi(\phi) = \frac{1}{\sqrt{2\pi}\sigma_\phi} e^{\frac{(\phi-\pi/2)^2}{2\sigma_\phi^2}} \qquad (3)$$

where the expected value is $\pi/2$ and the standard deviation σ_ϕ. For the delay times the exponential distribution is assumed,

$$p_\tau(\tau) = \frac{e^{\left(\frac{\tau}{2\sigma_\tau}\right)}}{2\sigma_\tau}, \qquad (4)$$

where σ_τ is the modeled rms delay spread.

3.2 Power Delay Profile

In order to evaluate the short-term variation of the radio channel the average power ChIR has to be considered. It has been largely accepted to assume the exponential power decay function as the basic model for radio propagation as well indoors as in outdoor environments. The model proposed below also assumes the exponential decay of the power in the ChIR.

The power delay profile is obtained integrating over consecutive power ChIR in space (or time if the mobile is moving) along the mobile path:

$$P(t) = \frac{1}{u_{\max}} \int E\left\{|h(u,t)|^2\right\} du \qquad (5)$$

In the equation above both averaging over space, that is, along the mobile path, and statistical average are used due to the nature of the signals. In the model we present the tap weights $\{a_k(u,t)\}$ are modeled as exponentially decaying functions weighted by a random variable $\{\beta_k(u)\}$ that controls the change of polarity of the transmitted signals due to reflection and scattering phenomena and the fading process in general:

$$a_k(u,t) = \beta_k(u) e^{-\frac{\tau_k(u)}{2\sigma_\tau}}, \qquad (6)$$

The coefficients $\{\beta_k(u)\}$, are modeled by a Gaussian distribution:

$$p_\beta(\beta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\beta^2}{2\sigma^2}}, \qquad (7)$$

The first two moments of the distribution are $E\{\beta\}=0$ and $E\{\beta^2\}=0$. The autocorrelation function for the fading variation of the UWB signals have still not been published in the scientific literature. Though a first attempt to model such a process could be the exponential decay function.

3.3 Multipath Fading

In UWB IR it seems meaningless to model multipath fading at a tap level within a given delay bin. The birth-death process in the (u,t) plane suggests that the modeling should focus on the paths of constant du/dt or equivalently dt/du.

An important feature of the channel variation is the multipath gain of the channel or the variation of coherently added channel taps:

$$\hat{P}(u) = \int |h(u,t)|^2 dt, \qquad (8)$$

The most common model, found in literature, when describing the multipath envelope fading is the Rice distribution, which also comprises the Rayleigh distribution as a special case. This model is from the physical point of view suitable for CW and wideband signals, where the sounding signals are sinus-like. On the other hand in IR, as the name indicates the sounding signals are of impulsive nature, commonly with Gaussian envelopes. Further, as shown above the variation of the received signal in IR, as well as the channel variation itself, is much more dependent of the birth-death of multipaths components.

It is worthwhile to note that in virtue of the large number of variable summed and the central limit theorem this magnitude can also be considered normally distributed regardless the original distribution of the channel taps.

3.4 Channel Model Simulation Results

Five ChIR were generated using the model described in section 3 above. The parameters of the model are given in the Table.1. The only parameter changed during the simulations was the delay spread. All the remaining parameters were kept constant.

		I	II	III	IV	V
Delay (Equation 4)	σ_τ, [ns]	4	9	23	30	40
Total Number of MP	L_{tot}	20	30	60	100	140
Total number of u samples	M	200	200	200	200	200
Multipath Fading (Equation 7)	μ_β	0	0	0	0	0
	σ_β	1	1	1	1	1
du/dt=tan(ϕ) (Equation 3)	μ_ϕ	π/2	π/2	π/2	π/2	π/2
	σ_ϕ	1	1	1	1	1

Table. 1 ChIR parameters used in the simulations.

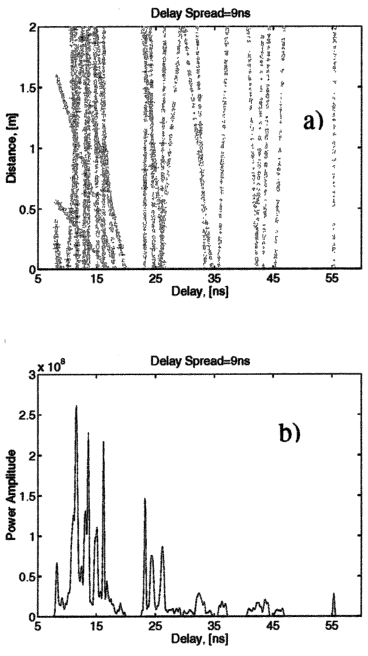


Figure 1. Evolution of multipaths in the distance-time (u,t) plane a) and average power delay profile b).

The number L_{tot}, of the random times of arrival at position u=0, $\{\hat{\tau}_k\}$, were generated according to equation (4) above. A further assumption was done in order to control it number. In our case only taps with power amplitudes larger than 30 dB relative the strongest taps were considered, which resulted in the values given in Table 1.

No second order statistics have been considered in the actual simulations. The fading variation was assumed to follow the Gaussian distribution along a line with constant du/dt accordingly to equation (7) and Table.1. An example of a simulated ChIR is given in Fig. The upper plot shows the evolution of different paths in the distance-time plane. The lower plot shows the average power delay profile.

4 Numerical Results of the Performance Evaluation of an IR-RAKE Receiver

The wide-band nature of IR-UWB signals makes it possible to implement RAKE type receivers in order to gain from the multipath behavior of the radio channel. This type of receivers if well known from the code division multiple access (CDMA) techniques used in IS-95 systems and UMTS.

The following assumptions have been made in order to assess the performance of the IR-RAKE receiver presented here,

- Perfect synchronism between transmitter and receiver.
- Infinitely fast RAKE finger tracking algorithm.
- Infinite searching delay window or at least equal with the maximum delay.
- The shape of the transmitted signal is not distorted when propagating through the radio channel or perfect estimation of the channel is assumed
- Only the effects of time dispersion are studied

4.1 Mean RAKE Power Gain

Due to the structure of the RAKE receiver the correlation properties of the received signals and of course of the basic pulse transmitted are crucial to the performance analysis of the IR-RAKE receiver. Therefore, in order to compute the receiver SNR and the error rate the autocorrelation function of the received signal containing the set of L-strongest taps as well as the cross correlation between this signal and the signal containing the complementary set of taps have been computed.

Only binary TSK was considered and only one distance in time (delay) between the "0" and the "1" bits have been considered that equals 0.3 ns. The pulse repetition rate used was 5 MHz and the pulse parameter that controls the pulse width was set to 0.8 ns.

An example of the total multipath power gain is shown in Fig.2, which corresponds to the ChIR shown in Fig.1. The results are depicted for L=1,5,7,9,11,21 and L_{tot}=30. For comparison the thicker line shows the Rayleigh cumulative distribution function.

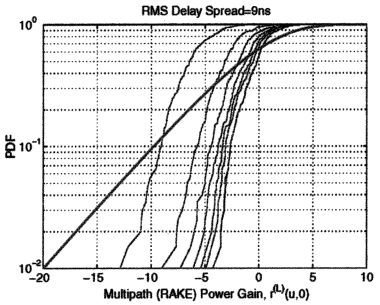


Figure 2. Cumulative probability distributions of Multipath Power Gain for L=1,5,7,9,11,21 and L_{tot}=30.

As can be seen the taps, as provided, fades much lesser than the Rayleigh signals, which results in a smaller fading margin. The mean effect collected by L taps is shown in Fig.3 in a decibel scale. From this curve the number of RAKE fingers needed to track a given percent of the total power may be derived.

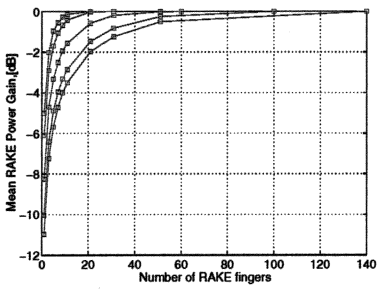


Figure 3. Mean RAKE power versus the number of RAKE "fingers" at the receiver

From Fig.3 the number of correlators needed to track a given percent of the total power available at the receiver as function of the estimated rms delay spread may be obtained, which is shown in Fig.4.

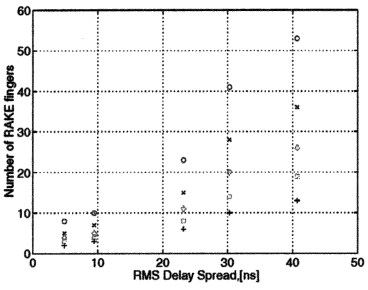


Figure 4. Number of RAKE receivers as function of the rms delay spread for different percent of the tracked power, o-90%, x-80%, ◊-70%, □- 60%, +-50%.

4.2 Numerical Results for Symbol Error Rate

The average upper bound symbol error rates as a function of the mean symbol SNR have been computed accordingly section 2.1. Results are presented in Fig.5 below. The straight line represents the inverse to the SNR, the lowest curve corresponds to performance in AWGN and the curves in between corresponds to the computed average symbol error rate in the presence of time dispersion.

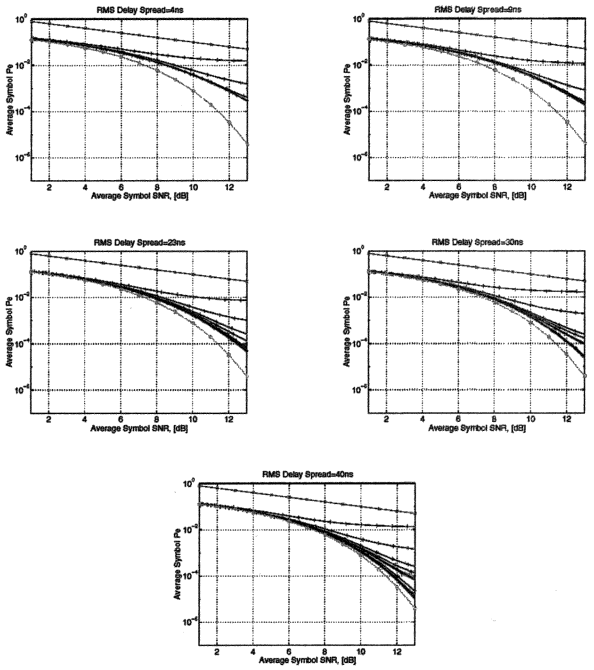


Figure 5. Upper bound average symbol error rate as function of the average received symbol SNR for channels I to V (see Table.1).

As can be seen the order of diversity (more RAKE fingers) will increase the performance, which was expected. This means that the performance of the infinite RAKE receiver will give the best performance for each channel type. When a selective RAKE is used with few fingers (L=1, 3, 5, 7), the signal-dependent self-noise is too large and the RAKE receiver performance degrades severely.

5 Summary and Conclusions

A channel model for IR indoor communications has been presented. Due to the high time resolution it is assumed that multipaths components would follow paths of constant distance-to-time ratio that in general are not parallel to each other and have a more deterministic behavior.

Further, the performance of IR-RAKE receivers at demodulating BTSK signals has been done by the theoretical formalism proposed in [4]. The obtained results show that a large number of RAKE fingers should be use to at least approach the performance in AWGN channels due to the dominance of the signal-dependent self-noise when just few RAKE correlators are used.

6 References

[1] Scholtz R.A., "Multiple Access with time hopping impulse modulation", MILCOM'93

[2] http://ultra.usc.edu/ulab/library.html

[3] Win Moe Z., et al, "Ultra wide bandwidth (UWB) signal propagation for outdoor wireless communications", IEEE, VTC'97 Vol.1, pp: 251-255

[4] Ramirez-Mireles F., Win M.Z., Scholtz R.A., "Performance analysis of ultra-wideband time-shift-modulated signals in the indoor wireless impulse radio channel", http://ultra.usc.edu/ulab/library.html

[5] Saleh A.A.M., Valenzuela R.A., "A statistical model for indoor multipath propagation", IEEE, JSAC, Vol. SAC-5, No. 2, February, 1987

[6] Hasemi H., "Impulse Response Modelling off Indoor Radio Propagation Channels", IEEE JSAC, Vol. 11, No.7, Sept. 1993

[7] Cassioli D., Win M.Z., Molisch A.F., " A statistical model for the UWB Indoor channel", IEEE, VTC'2001, Greece

[8] Kattenbach R., Englert T., Investigation of short term statistical distributions for path amplitudes and phases in indoor environment", IEEE, VTC'98

Synchronization of Code-Spread CDMA Systems

Fredrik Malmsten, Tony Ottosson and Erik G. Ström

Communication Systems Group, Dept. of Signals and Systems, Chalmers University of Technology,
SE-412 96 Göteborg, Sweden, Fax: +46 31 772 17 48
E-mail: {fredrik.malmsten, tony.ottosson, erik.strom}@s2.chalmers.se

Abstract

Delay estimation of a code-spread code-division multiple-access (CS-CDMA) system using low-rate maximum free distance (MFD) convolutional codes for bandwidth expansion is studied. The MFD codes used, have a repetitive structure, which is utilized for non-data-aided synchronization. The repetition pattern of the code is optimized for good delay estimation properties, using an optimization criterion based on the Cramér-Rao bound for the delay estimation problem. An approximate maximum-likelihood estimator that jointly estimates the channel gain, the codewords and the delay is proposed. A corresponding estimator for direct-sequence CDMA (DS-CDMA) is also derived. The performance, in terms of estimation error, outlier rate and bit-error rate, for both the systems are compared.

1 Introduction

Direct-sequence code-division multiple-access (DS-CDMA) signals are often regarded as a result of the multiplication of the data bits by a spreading sequence, clocked at a higher rate than the data sequence (where the ratio of the rates is the spreading factor). A conceptually different, but equivalent, description is that the symbol rate is raised by upsampling (repetition encoding) of the data bits by a factor equal to the spreading factor, followed by a multiplication of a scrambling sequence. The scrambling sequence is actually the same sequence as the spreading sequence, but the different names indicates that the purpose is different in the two descriptions. The spreading sequence is used both for increasing the bandwidth of the signal as well as for making the interfering signal more random, whereas the purpose of the scrambling code only is the latter. When looking at CDMA in this way, it seems, that using a forward error correcting code (FEC) that is better than the repetition code, the performance of the system could be increased. Systems where the spreading is accomplished solely by the use of a powerful FEC are referred to as code-spread CDMA (CS-CDMA) systems. In [1, 2] it is shown that maximum throughput is achieved by employing low-rate channel codes alone for bandwidth expansion.

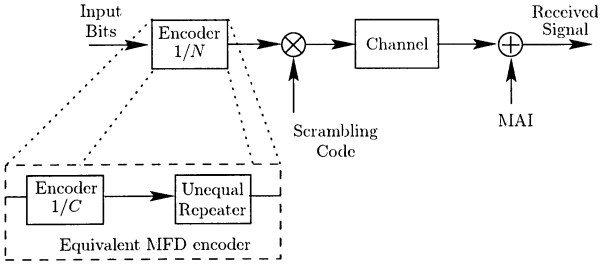


Figure 1: A code-spread CDMA system.

In [3] *maximum free distance* (MFD) low-rate convolutional codes are presented for different constraint lengths and code rates between $R = 1/4$ and $R = 1/512$. It is shown that the performance of this code-spread multiple-access system is superior to that of conventionally spread systems, assuming perfect synchronization. Furthermore, the MFD codes have the interesting property that for low rates there are significantly fewer generator polynomials than encoded bits. The output from a generator polynomial is generally used for several encoded bits. For instance, the $R = 1/64$ code used in the simulation example described later has only $C = 14$ different polynomials, which are repeated 15, 8, 7, 4, 4, 4, 4, 4, 4, 4, 2, 2, 1 and 1 times respectively. In Fig. 1 it is demonstrated that this also can be regarded as using an intermediate channel code with a higher rate ($R = 1/C = 1/14$ in this case) followed by an unequal repeater. This structure in the channel code leads to simple encoder and decoder implementations.

Synchronization algorithms for a DS-CDMA system, using short codes, can utilize the cyclostationarity of the signal. In a code-spread CDMA signal there is, in general, no such structure to exploit. The signal to noise and interference ratio is also very low, complicating the task even more. The unequal repetition in the MFD codes introduces a structure in the signal, that can be utilized by a synchronization algorithm. In this paper a non-data-aided delay estimator taking advantage of this structure is derived, using a

maximum-likelihood approach.

The input to the estimator is the asynchronous samples of the baseband signal during an observation interval of several symbols, disturbed by additive white Gaussian noise (AWGN). A single-user system is assumed, or equivalently, the multiple access interference (MAI) is regarded to be AWGN. The resulting estimator can basically be described as follows; a delay is assumed and samples belonging to the same code polynomial of the unrepeated high-rate code, are summed. In order to remove the data dependency, these sums are combined non-coherently. The result is then defined to be the value of the criterion-function, evaluated at the assumed delay. Finally, this criterion-function is searched for an extremum, to find the most likely delay, which then is chosen to be the delay estimate.

The output order of the unequal repeater can be rearranged in any way without changing the properties of the channel code. The problem of finding the best repetition pattern is a huge combinatorial problem. A good (but not guaranteed optimum) solution is found using a simulated annealing algorithm.

2 System Model

A code-spread CDMA system is schematically shown in Fig. 1. All bandwidth expansion is achieved by a low rate $1/N$ convolutional code, producing the N coded symbols. These symbols are then randomized by a non-spreading user-specific long pseudo-random scrambling sequence. The signal is transmitted over the channel. At the receiver side, similar signals from other users (the MAI) are added producing the composite signal.

We will assume that the synchronization of the scrambling sequence is already accomplished. This is not a very restrictive assumption, if the starting phase of the sequence is known. In practical systems the delay is known to be restricted to a *search window*. It will be necessary to try all possible phases of the scrambling code within the search window, and for each of the phases, use the delay estimation algorithm treated in this paper. The complexity for such a scheme will grow linearly with the number of possible delays. In practical systems the search window is kept rather small due to frame structure, access methods etc. For convenience we drop the scrambling sequence from here on. We furthermore assume that we have a single-user system or equivalently that the MAI can be regarded as AWGN.

Fig. 2 shows the system which is fed by the input bits, denoted $d(m)$, at a data rate of $1/T$. The bits are encoded using one of the MFD codes proposed in [3], with the exception that we will permute the repetitions of the code polynomials to improve

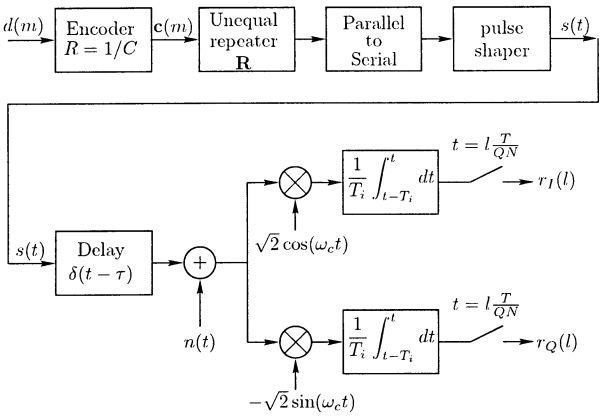


Figure 2: The equivalent baseband system model.

the delay estimation performance. For a given code rate $1/N$ these codes only have C unique polynomials, where C typically is significantly smaller than N. The input bits are fed to a convolutional encoder with the code rate $1/C$ and then mapped to BPSK-symbols. The codeword from this intermediate encoder forms the vector $\mathbf{c}(m) \in \{\pm 1\}^C$. To get the final codeword we perform an unequal repetition (and permutation) by multiplying with the *repetition matrix* $\mathbf{R} \in \{0,1\}^{N\times C}$. The i^{th} row in this matrix is all zeros except for exactly one 1, the position of which determines which code symbol of the intermediate code should be copied to the i^{th} chip. The total number of ones in each column of $\mathbf{R}$ are given by the used code, but we are free to choose the order of the chips.

Assuming that $g(t)$ denotes the chip waveform and

$$\mathbf{g}(t) = [\, g(t) \quad g(t-T_c) \quad \ldots \quad g\big(t-(N-1)T_c\big)\,],$$

we can write the baseband signal as

$$s(t) = \sum_{m=-\infty}^{\infty} \mathbf{g}\big(t-mT\big)\mathbf{R}\mathbf{c}(m). \tag{1}$$

The baseband signal is multiplied with the carrier $A\cos(\omega_c t + \theta')$ and then transmitted over an AWGN channel with the delay τ. Hence, the received signal can be written as $r(t) = \gamma s(t-\tau)\cos(\omega_c t + \theta)$, where γ is the resulting signal strength after the path loss and $\theta = \theta' - \omega_c \tau$. After the signal has been passed through the IQ-stage and the integrator it is sampled asynchronously at a rate of $\frac{NQ}{T}$, where Q is the oversampling factor, to give

$$\begin{aligned} r(l) &= \gamma \frac{1}{T_i} \int_{(l-1)T_i}^{lT_i} s(t-\tau)\,dt + n(l) \\ &= r_I(l) + j r_Q(l), \end{aligned} \tag{2}$$

where $T = NT_c = NQT_i$.

Now, gather the samples $r(l)$ into vectors of length QN which corresponds to the length of one codeword

$$\mathbf{r}(m) = \begin{bmatrix} r(mQN+1) & r(mQN+2) \\ \dots & r(mQN+QN) \end{bmatrix}^T \in \mathbb{C}^{QN}. \quad (3)$$

Furthermore the thermal noise and interference is gathered in the noise vector $\mathbf{n}(m)$ in the same manner as for $\mathbf{r}(m)$, and modeled as AWGN. Thus, the noise vector has the properties: $\mathrm{E}[\mathbf{n}(m)] = \mathbf{0}$ and $\mathrm{E}[\mathbf{n}(m)\mathbf{n}^*(p)] = \sigma^2\mathbf{I}_{QN}\delta_K(m-p)$. For the sake of formulating an algorithm we assume rectangular chip waveforms. After the evaluation of the integral in (2) and some straightforward calculations we can write the sample vector as

$$\mathbf{r}(m) = \beta\mathbf{H}\mathbf{z}(m) + \mathbf{n}(m), \quad (4)$$

where $\beta = \gamma e^{j\theta}$ is the complex path-gain multiplied with the transmitted amplitude. Moreover

$$\mathbf{H} = [\,\mathbf{H}_1\mathbf{H}_2\,] \in \{0,1\}^{QN\times 2C}. \quad (5)$$

All the τ dependence in $\mathbf{r}(m)$ is described by $\mathbf{H}$ since it contains a shifted version of the repetition matrix. More specifically

$$\mathbf{H}_1(\tau) = \left(1-\frac{\delta}{T_i}\right)\mathrm{us}(\mathbf{R},k) + \frac{\delta}{T_i}\mathrm{us}(\mathbf{R},k+1) \quad (6)$$

$$\mathbf{H}_2(\tau) = \left(1-\frac{\delta}{T_i}\right)\mathrm{ds}(\mathbf{R},N-k) + \frac{\delta}{T_i}\mathrm{ds}(\mathbf{R},N-k+1), \quad (7)$$

where $\tau = kT_i + \delta$ such that k is an integer and $\delta \in [0,T_i)$; $\mathrm{us}(\cdot,\cdot)$ is the up-shift operator; and $\mathrm{ds}(\cdot,\cdot)$ is the down-shift operator, defined as

$$\mathrm{us}([\,\mathbf{a}_1\cdots\mathbf{a}_N\,]^T, q) = [\,\mathbf{a}_{1-q}\cdots\mathbf{a}_N\mathbf{0}\cdots\mathbf{0}\,]^T$$
$$\mathrm{ds}([\,\mathbf{a}_1\cdots\mathbf{a}_N\,]^T, q) = [\,\mathbf{0}\cdots\mathbf{0}\mathbf{a}_1\cdots\mathbf{a}_{N-q}\,]^T.$$

Furthermore

$$\mathbf{z}(m) = [\,\mathbf{c}(m-1)^T\mathbf{c}(m)^T\,]^T \in \{\pm 1\}^{2C}, \quad (8)$$

contains two consecutive codewords from the rate $1/C$ encoder. Finally we assume that the channel is constant during an interval of length MT and thus we can form an observation matrix

$$\Omega = [\,\mathbf{r}(1) \quad \mathbf{r}(2) \quad \cdots \quad \mathbf{r}(M)\,]. \quad (9)$$

3 The repetition matrix

The repetition matrix $\mathbf{R}$ introduced in the previous section should be optimized for delay estimation performance. Since this optimization is a huge combinatorial optimization problem we take a simplifying approach based on the simulated annealing algorithm (see e.g. [4,5]). However, for the simulated annealing algorithm to work well it is important to find a good criterion (energy function). In our case we derived a criterion using the Cramér-Rao Bound (CRB) for the problem of jointly estimating the delay τ and complex path gain β but with known encoded bits $\mathbf{c}$. The CRB shows that good timing estimation is achieved by making the autocorrelation of each column of the repetition matrix at correlation time shift 1 as low as possible. This criterion results in codes that would allow accurate delay estimation once a good initial guess is available, but it would not necessarily be easy to find that coarse estimate. In other words, that criterion will lead us to codes that are easy to track but not necessarily easy to acquire. The criterion for improving acquisition performance is, somewhat arbitrarily, chosen to minimize the sum of the autocorrelation for all of the columns of the repetition matrix $\mathbf{R}$ for all correlation time shifts other than 0. The energy function was chosen as a weighted combination of the CRB criterion and the criterion for acquisition.

4 Proposed algorithm

The standard sliding correlator can not be applied in a code-spread CDMA system due to the lack of known spreading sequences. We can however exploit the code structure, captured by the repetition matrix $\mathbf{R}$. Since we are not primarily interested in estimating the path gain or the codewords, we can put them together to codewords with arbitrary phase and amplitude, which we denote $\zeta(m) = \beta\mathbf{c}(m)$, which implies $\xi(m) = \beta\mathbf{z}(m)$. A maximum likelihood approach is taken to derive an algorithm that jointly estimates the desired parameter τ and all the scaled codewords $\zeta(m)$, which in this case are nuisance parameters, for the observation Ω. The likelihood function for $\mathbf{r}(m)$ is given by

$$L(\mathbf{r}(m);\tau,\zeta(m)) \propto \exp\left[-\frac{\big(\mathbf{r}(m)-\mathbf{H}\xi(m)\big)^*\big(\mathbf{r}(m)-\mathbf{H}\xi(m)\big)}{\sigma^2}\right]. \quad (10)$$

Since $\mathbf{r}(p)$ and $\mathbf{r}(q)$ are uncorrelated Gaussian vectors for $p \neq q$ we can write the likelihood function of Ω as

$$L(\Omega) = \prod_{m=1}^{M} L(\mathbf{r}(m))$$

and with additive and multiplicative constants neglected we can write the negative log-likelihood function as

$$\ell(\tau, \zeta) = \sum_{m=1}^{M} \Big(2\,\mathrm{Re}\Big\{ \mathbf{r}^*(m)\mathbf{H}\xi(m) \Big\} - \xi^*(m)\mathbf{H}^*\mathbf{H}\xi(m) \Big). \quad (11)$$

The estimator for $\zeta(m)$ is found by setting the derivative of the log-likelihood function with respect to $\zeta(m)$ to 0. To avoid a coupling between $\hat{\zeta}(m)$ for different m a few terms containing $\mathbf{H}_1^*\mathbf{H}_2$ are neglected. The resulting estimator is found to be

$$\hat{\zeta}(m) = \mathbf{H}_t^{-1}\Big(\mathbf{H}_2^*\mathbf{r}(m) + \mathbf{H}_1^*\mathbf{r}(m+1) \Big), \quad (12)$$

with $\mathbf{H}_t^{-1} = \mathbf{H}_1^*\mathbf{H}_1 + \mathbf{H}_2^*\mathbf{H}_2$.

Once $\zeta(m)$ is estimated it can be inserted into (11) to yield the cost function for estimating τ

$$J(\tau) = -\sum_{m=1}^{M} \Big(2\,\mathrm{Re}\Big\{ \mathbf{r}^*(m)\mathbf{H}\hat{\xi}(m) \Big\} - \hat{\xi}^*(m)\mathbf{H}^*\mathbf{H}\hat{\xi}(m) \Big). \quad (13)$$

Given the cost function we can now search for a minimum, and let

$$\hat{\tau} = \arg\min_{\tau}\big(J(\tau)\big) \quad (14)$$

5 Numerical results

We will here present some results for delay estimation using the algorithm presented in the previous section. For comparison we also present results for a direct-sequence spread CDMA system, where the estimation algorithm is a simple modification of that for the code-spread system. For both the conventional and the code-spread systems the bandwidth expansion is N. The conventional system achieves all expansion using an orthogonal Gold code of length 32 or 128, while the code-spread system uses a MFD code of constraint length 7 and rate $1/N$, $N = 32, 64$, or 128. The generator polynomials for this convolutional code and the unequal repetition pattern can be found in [3,6,7]. The unequal repetition pattern was used as input to design the repetition matrix using the simulated annealing approach described in section 3.

The channel was in all cases an AWGN channel with constant gain ($\beta = 1$). The numerical search needed for the code-spread system to search the minimum of the cost function is performed with a tolerance of $10^{-4}T_c$. No oversampling was used, i.e. $Q = 1$. At least 2500 Monte-Carlo Simulations was done for each scenario, and independent data, channel delay and noise realizations were drawn for each Monte-Carlo run.

Fig. 3 shows the root mean-square-error (RMSE) values for the conventional and code-spread systems for 100 received code vectors, as a function of the E_b/N_0, where E_b is defined as the information bit energy. The RMSE is computed without outliers and is normalized to the chip duration T_c. An estimate is considered as an outlier if it is more than $T_c/4$ away from the true delay. As seen in Fig. 3, the RMSE for the conventional system is better than for the code-spread system at the same E_b/N_0 and code rate. This is also expected since for each symbol we get NQ complex samples from which we have to estimate the scaled code-vector (i.e. C complex parameters), whereas for the direct-sequence system we only have to estimate one scaled data symbol. We can also see that for the direct-sequence system there is no significant difference in the estimator performance between different spreading factors, since the autocorrelation properties of the spreading codes is almost the same. Remember that we compare the RMSE normalized to the chip time; in absolute numbers we will get a big difference. The code-spread system on the other hand gains from higher spreading factor, at least for high SNR. This is also intuitive since for a lower rate MFD code we have fewer parameters/sample to estimate (C is nearly constant for $N > 30$) . For low SNR the estimation error is significantly higher for the code-spread system, because the code vectors are likely to be in error.

Fig. 4 shows the probability of outliers. The code-spread system suffers form high outlier probability but it has an advantage, in that it is easy to detect an outlier. To detect an outlier, once the delay estimate is found, the data sequence could be decoded using a sequential decoding algorithm. If the metric growth is lower than a threshold, we may suspect that the delay estimate was bad. This is a way of exploiting the full structure of the signal, and not only the structure introduced by the unequal repeater.

Since the two systems put very different requirements on the estimators, the most interesting comparison is not in estimator performance. Instead, the bit-error rate (BER) of the two systems, considering estimation error effects, could be studied. In figure 5, the two systems are compared at the spreading factor $N = 32$. The timing errors have been modeled as

Gaussian with standard deviation as given by figure 3. A large number of blocks of 50000 symbols, within which the timing error was constant, were transmitted throgh the modeled system. The code was terminated with a tail and the receiver used a Viterbi decoder. The comparison in fiure 5 shows that the coding gain of the codespread system more than compensates for the worse synchronization performance.

6 Conclusions

The delay estimation error is smaller for the conventionally spread system than for the code-spread system. For high SNR this difference is small, especially for high spreading factors. When it comes to bit error probability the code-spread system's high coding gain is more than enough to compensate the loss due to imperfect synchronization. The high rate of outliers is a problem that needs to be addressed.

References

[1] J. Y. N. Hui. Throughput analysis for code division multiple accessing of the spread spectrum channel. *IEEE Journal on Selected Areas in Communications*, SAC-2(4):482–486, July 1984.

[2] A. J. Viterbi. Very low rate convolutional codes for maximum theoretical performance of spread-spectrum multiple-access channels. *IEEE Journal on Selected Areas in Communications*, 8(4):641–649, May 1990.

[3] P. Frenger, P. Orten, T. Ottosson, and A. Svensson. Code-spread CDMA using maximum free distance low-rate convolutional codes. *IEEE Transactions on Communications*, 48(1), January 2000.

[4] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[5] N. Farvardin. A study of vector quantization for noisy channels. *IEEE Transactions on Information Theory*, 36(4):799–809, July 1990.

[6] P. Frenger, P. Orten, and T. Ottosson. Combined coding and spreading in CDMA systems using maximum free distance convolutional codes. In *Proc. IEEE Vehicular Technology Conference*, pages 2497–2501, Ottawa, Canada, May 1998.

[7] P. Frenger, P. Orten, T. Ottosson, and A. Svensson. Multirate convolutional codes. Technical Report 21, Dpt. of Signals and Systems, Communication Systems Group, Chalmers University of Technology, Göteborg, Sweden, April 1998.

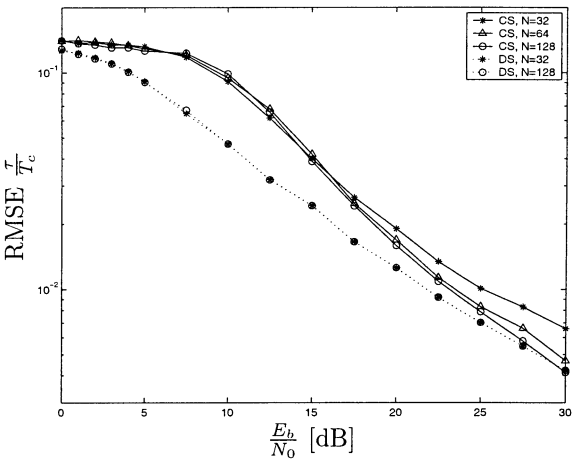


Figure 3: RMSE as a function of SNR for observation interval length, $M = 100$.

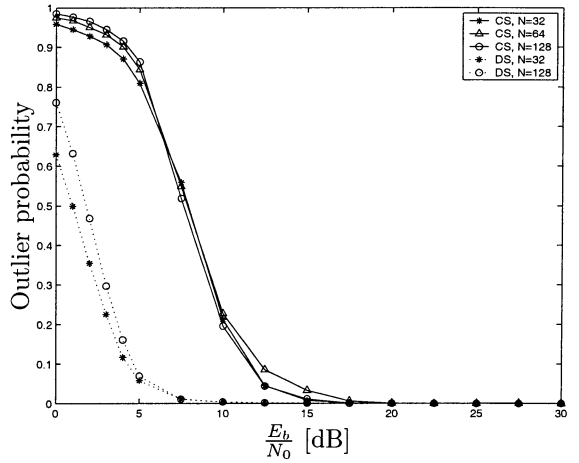


Figure 4: Probability of outliers as a function of SNR for observation interval length, $M = 100$.

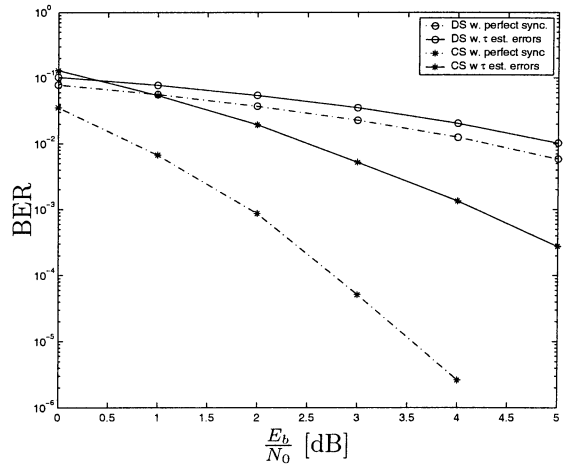


Figure 5: BER as a function of SNR for observation interval length, $M = 100$.

Tabu Search Multiuser Detection in CDMA*

Peng Hui Tan[1] and Lars Rasmussen[1,2]

[1]Chalmers University of Technology, Department of Computer Engineering, SE-412 96 Gothenburg, Sweden.
E-mail: phtan@ce.chalmers.se.

[2]University of South Australia, Institute for Telecommunications Research, Mawson Lakes SA 5095, Australia.
E-mail: larsr@ce.chalmers.se.

Abstract

In this paper, we compare several optimization methods using tabu search heuristics for solving the maximum likelihood (ML) multiuser detection problem. The purpose of using these algorithms is to provide complexity constraint alternatives to solving this NP-hard problem. It is now possible to show analytically that these methods exactly solve the ML problem, however, numerical examples show that applying these approaches, the resulting bit error rate performance is indistinguishable from the true ML performance.

1 Introduction

Direct-sequence code-division multiple access is a multiple access communication scheme in which each user is assigned a unique signature waveform for modulating the transmitted symbols. In [1], Verdú proposed and analyzed the optimal multiuser detector, which is a maximum likelihood (ML) sequence detector. This receiver minimizes, with respect to all possible transmitted bit sequences, the squared Euclidean distance[1] between the actual and the hypothesized received signal vectors. The complexity of ML detection grows exponentially with the number of users [2].

In this paper, we consider multiuser detection from a combinatorial optimization viewpoint. An obvious approach to solving a combinatorial problem is simply to list all feasible solutions of a given problem, evaluate the corresponding objective functions, and pick the best one. However, it is immediately obvious that this approach of complete enumeration illustrates the exponentially growing complexity and thus is bound to be grossly inefficient. Although it is possible in principle to solve any problem in this way, in practice it is not, because of the vast number of possible solutions to any problem of a reasonable size.

*This work was supported by the Swedish Research Council for Engineering Sciences under grant no. 27-1999-390.

[1]Assuming an additive white Gaussian noise channel.

For the sake of lower computational complexity, we can make use of heuristic methods. A heuristic technique [3] is a method which seeks local optimal solutions at a reasonable computational cost without being able to guarantee optimality. Unfortunately, it may not even be possible to state how close to optimality a particular heuristic solution is. In practice, it should be emphasized that many modern heuristic techniques do give high-quality solutions [3].

The purpose of this paper is to consider algorithms for approximating ML detection but with significantly lower computational complexity than the true optimum algorithm. Therefore, we propose a reactive tabu search (RTS) algorithm, a heuristic approach based on the tabu search (TS) algorithm [4], to seek the optimal solution in an efficient manner.

Consider a K user CDMA system with processing gain N, transmitting a sequence of L symbols chosen from an alphabet $\mathcal{D}$. Using vector notation, the LK matched filter outputs can be expressed as

$$\mathbf{y} = \mathbf{RCd} + \mathbf{z},$$

where $\mathbf{C}$ is a $KL \times KL$ matrix that contains the channel coefficients of the users, $\mathbf{R}$ is a $KL \times KL$ matrix, describing the correlation between the signature waveforms of the K users, and $\mathbf{d}$ is a KL column vector, containing the data symbols transmitted by all the users. For a more complete description of the channel model, see [5]. The Gaussian noise vector $\mathbf{z}$ has zero mean and autocorrelation matrix

$$E[\mathbf{z}\mathbf{z}^{\mathrm{H}}] = \sigma^2\mathbf{R} = \frac{N_0}{2}\mathbf{R}.$$

The vector of matched filter outputs $\mathbf{y}$, represents a sufficient statistic for ML detection of the transmitted symbols [6].

To highlight the mathematical details of the application of TS, we will only consider a synchronous CDMA system. In synchronous transmission, each interfering user produces exactly one symbol which interferes with the desired symbol. Therefore, in single-path additive white Gaussian noise channels, it is sufficient to consider the signal received in only one

signal interval for determining the optimum ML receiver. The ML detector selects the most likely hypothesis $\hat{\mathbf{d}}$ given the matched filter output, and perfect knowledge of $\mathbf{C}$ and $\mathbf{R}$. When the channel coefficients are real valued and the symbols $\mathbf{d}$ are binary, i.e, $\mathcal{D} \in \{-1, 1\}^K$, the ML detection problem can be expressed as follows

$$\hat{\mathbf{d}} = \arg \min_{\mathbf{d} \in \{\pm 1\}^K} \mathbf{d}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{R}\mathbf{C}\mathbf{d} - 2\mathbf{y}^{\mathrm{T}}\mathbf{C}\mathbf{d}. \quad (1)$$

The solution to the above problem (1) requires a search over all the 2^K possible combinations of the components of the vector $\mathbf{d}$. It is thus clear that the computational complexity increases exponentially with the number of users.

2 Tabu Search Multiuser Detection

This section considers the use of TS algorithms to seek the optimum solution of the multiuser detection problem. These techniques have the potential to reduce computational complexity without sacrificing performance. The present form of tabu search was originally developed by Glover [7], and a comprehensive account of the concepts, and recent developments can be found in [4].

Tabu Search can be considered to some extent as a local search strategy. Essentially, a local search consists of moving from one feasible solution to another in its neighborhood according to well-defined rules. Let us consider the problem of minimizing a function $f(\mathbf{x})$ on a finite set of points $\mathcal{X}$. A local search strategy starts from an arbitrary solution $\mathbf{x}_1 \in \mathcal{X}$. At each step m, a new solution $\mathbf{x}_{m+1}$ is chosen in the neighborhood $\mathcal{N}(\mathbf{x}_m)$ of the current solution $\mathbf{x}_m$. The neighborhood of $\mathbf{x}_m$ is a subset of $\mathcal{X}$ which is associated to each $\mathbf{x}_m \in \mathcal{X}$. Each solution $\mathbf{x}_{m+1} \in \mathcal{N}(\mathbf{x}_m)$ can be reached directly from $\mathbf{x}_m$ by an operation called a move, and $\mathbf{x}_m$ is said to move to $\mathbf{x}_{m+1}$ when such an operation is performed.

The most common criterion for selecting the next solution $\mathbf{x}_{m+1}$ is to pick the best one, $\mathbf{x}^*$, in the neighborhood of $\mathbf{x}_m$,

$$f(\mathbf{x}^*) \leq f(\mathbf{x}), \qquad \forall\ \mathbf{x} \in \mathcal{N}(\mathbf{x}_m). \quad (2)$$

Then, $\mathbf{x}_{m+1} = \mathbf{x}^*$ becomes the next current solution provided it is better than $\mathbf{x}_m$, i.e., $f(\mathbf{x}_{m+1}) \leq f(\mathbf{x}_m)$. Otherwise, the search is stopped. This strategy is usually called a descent strategy [3].

The main weakness of the descent algorithm is its inability to escape from local minima. For instance, it may be that all solutions in the neighborhood $\mathcal{N}(\mathbf{x}_m)$ are worse than $\mathbf{x}_m$ although, further away, there exists a global minimum of $f(\mathbf{x})$ which cannot be reached by the descent algorithm. Tabu Search [4] is a local search strategy which is explicitly designed for avoiding such a situation.

Tabu search is guided by the use of a memory structure which is implemented by the implicit recording of previously visited solutions. It is usually implemented by a tabu list of moves which have been made in the recent past of the search. These moves will be tabu or forbidden for a certain number of iterations.

2.1 A basic short term Tabu Search

An important distinction in tabu search arises by differentiating between short term memory and longer term memory defined over the span of the search history. Each type of memory is accompanied by its own special strategies. However, the effect of both types of memory may be viewed as modifying the neighborhood $\mathcal{N}(\mathbf{x})$ of the current solution $\mathbf{x}$. In this section, we focus on the short term aspects. The use of long term frequency memory will be described in the next section. In the basic short term scheme, the strategy for escaping from local minima is to move to the best possible solution $\mathbf{x}_{m+1}$ in the neighborhood of $\mathcal{N}(\mathbf{x}_m)$ even if there is no better solution than the current one, $\mathbf{x}_m \in \mathcal{N}(\mathbf{x}_m)$.

If $\mathbf{x}_m$ belongs to the neighborhood $\mathcal{N}(\mathbf{x}_{m+1})$ of $\mathbf{x}_{m+1}$ whenever $\mathbf{x}_{m+1} \in \mathcal{N}(\mathbf{x}_m)$, there is a danger of cycling when we explore $\mathcal{N}(\mathbf{x}_{m+1})$ during the next step. There is indeed a chance that $\mathbf{x}_m$ could be the best solution in $\mathcal{N}(\mathbf{x}_{m+1})$ in which case we would come back to $\mathbf{x}_m$ and from then on, oscillate between $\mathbf{x}_{m+1}$ and $\mathbf{x}_m$. To avoid this situation, we can store in a tabu list $(\mathbf{x}_{m-1}, \ldots, \mathbf{x}_{m-M'})$, a certain number M' of the last solutions encountered. If $\mathbf{x}$ is in the list, the move $\mathbf{x}_m \rightarrow \mathbf{x}$ is forbidden. An alternative is to record a characteristic or an attribute of the moves. A move with such a characteristic is then prohibited for a certain period (number of moves) P.

A tabu list of characteristics which restricts moves for a certain period, does not exactly prevent cycling. The main role of the tabu list is to diversify the search, i.e., to send the current solution to a region of the search space which was not previously explored and in particular to escape from local minima. The prohibition of solutions with a given attribute is likely to be restrictive in excluding significantly more solutions than the ones just visited. To correct the bad consequences of this (not all consequences are undesirable), it is allowed to overwrite the tabu status of a move when it leads to a sufficiently good solution. This is termed as an aspiration move and the basics for making such a move is known as the aspiration criterion. For example, a tabu move is accepted if it leads to a solution which has a better objective function value than the best available so far.

In the multiuser detection problem in (1), the so-

lution vector $\mathbf{d}$ is binary. We therefore consider the following neighborhood structure. For $\mathbf{d} \in \{-1,1\}^K$, the neighborhood with size j is defined as the set

$$\mathcal{N}_j(\mathbf{d}) = \{\mathbf{x} \in \{-1,1\}^K \mid \Delta(\mathbf{d},\mathbf{x}) \leq j\}, \qquad (3)$$

where $\Delta(\mathbf{d},\mathbf{x})$ denotes the Hamming distance between $\mathbf{d}$ and $\mathbf{x}$. For all $\mathbf{x} \in \mathcal{N}_j(\mathbf{d})$, $\mathbf{d}$ and $\mathbf{x}$ differ by at most j components. The total number of vectors in $\mathcal{N}_j(\mathbf{d})$ is $|\mathcal{N}_j(\mathbf{d})| = \sum_{i=0}^{j} \binom{K}{j}$. To allow for a fast evaluation of moves in the neighborhood, we set $j = 1$ so that only a low number of possible moves has to be inspected. In addition, the numerical examples in Section 3 show that the choice of $j = 1$ is sufficient for the TS to obtain ML performance. Hence, a move corresponds to changing the value of a single variable (i.e., setting $d_i = -d_i$ for some i).

For the choice of move attributes, we record the transformation performed on the current solution, i.e., flipping the i-th coordinate value of a binary vector from -1 to 1 or vice versa. The move is then prohibited for P iterations. In addition, we also offer the possibility of overwriting the tabu status of a move when it leads to a better solution. In our algorithm, a tabu move is accepted if it leads to a solution which has a better objective function value than the best available so far. Finally, the algorithm is terminated when the maximum number of iterations M is reached. It is not possible to determine analytically the maximum number of iterations within which the TS algorithm will converge. A qualified guess is therefore made and tested to assure a suitable M has been selected.

2.2 Reactive Tabu Search Algorithm

As previously noted, the tabu lists are typically implementing short term memory effects. It cannot guarantee the absence of cycles. In addition, choosing the prohibition period P without prior knowledge about the structure of the search space that can be generated in a given problem is difficult. For example, the prohibition period can be insufficient to avoid cycles if it is too small. If it is too large, only a small fraction of movements are admissible and the search is inefficient. Even if a constant prohibition period is suitable, its value is problem specific.

Long term frequency memory is used to monitor the alternation of intensification and diversification phases of the search [4]. It is used as a complement to short term memory and it is often frequency based, that is, based on counts of solution or transition attributes.

Intensification refers to focusing the search on certain promising types of solutions. This can be done, for example, by identifying characteristics of elite solutions met in the past and defining the neighborhood of the current solution accordingly. It is clear however that identifying these characteristics may not always be an easy task. Hence, for intensification, we choose to apply a simple local search procedure each time an improved solution was found. This procedure attempts to improve that solution further by examining all possible moves (irrespective of their tabu status) that improve the solution. The local search procedure used is the descent strategy (2) with the neighborhood structure defined in (3) with $j = 1$.

Diversification is required to enable the solution to escape from local minima so that cycling does not arise. A simple diversification procedure consists in modifying the objective function to penalizing solutions which are too close to a set of previously visited solutions. To encourage diversification and to prevent cycling, we apply the reactive tabu search (RTS) algorithm proposed in [8]. It has an additional mechanism to deal with cycles that are not avoided by using the TS and a way to adaptively change P during the search so that the value $P(m)$ is appropriate to the structure of the current search space for the problem. The prohibition period is initialized with a small value $P(0)$, and then adapted by reacting to the occurrence of repetitions. If a repetition is found, we discourage cycles or encourage diversification by increasing $P(m)$ in steps of 2. If $P(m)$ is not allowed to decrease, there is a danger that its value will remain large after a search phase with many repetitions. Therefore, we reduce $P(m)$ to $\min\{1, P(m)-2\}$ when $|f_m - f^*| > \beta f^*$, where f_m is the current objective function value, f^* is the best objective function value found so far, and $0 < \beta < 1$ is a controlled weighting factor. In the experiments, $\beta = 0.3$ and it is found to be very robust with respect to this choice.

3 Simulation Results

In this section, we present several conducted experiments to evaluate the performance of multiuser detectors based on the short term TS (STTS) and RTS. The objectives of these experiments are to investigate the bit error rate (BER) performance, near-far performance and complexity of the detectors. Synchronous CDMA systems over an AWGN channel with randomly selected long spreading codes are used. All the users are transmitting with the same power. Two different processing gains are used, $N = 16$ and $N = 32$. The ML performance of these systems are determined by the A* algorithm in [9]. The matched filter output is used as the initial solution for the STTS and RTS.

Figure 1 shows the performance gain offered by the RTS over the STTS approach in terms of BER performance of two systems with load less than one. The parameter values in round brackets in the legend

are the average CPU time[2] and the average number of times the objective function was evaluated for detecting one symbol. For the lightly loaded case ($K = 10$ and $N = 32$), the STTS based detector with $M = 20$ can approximate the ML detector BER performance closely while the RTS approach requires 15 iterations to reach ML performance. When the load increases in the system ($K = 24$ and $N = 32$), the RTS based detector requires $M = 150$ in order to obtain a good approximation. The STTS approach however does not reach ML performance, given approximately the same amount of complexity.

To investigate the effect of the load on the algorithms, we increase the load (K/N) to 1 and 1.25 in Figure 2. The RTS based detector now requires more iterations to approximate ML performance. For $K = 32$ and $N = 32$, it requires 400 iterations. For $K = 20$ and $N = 16$, it requires 350 iterations. As for the STTS based detector with 400 iterations, it has not achieved ML performance. Moreover, the best BER performance of the STTS approach can only be achieved by decreasing P.

Next, to illustrate near-far resistance, we considered the case where all but user 1 has the same SNR. User 1 stays at a fixed SNR. In Figure 3, the BER of the first user with SNR(1) = 6 dB is shown against the ratio of the strength of the interfering users' signals to the first user's signal strength (SNR(i)-SNR(1) in dB). The simulation results show no appreciable performance difference between the RTS approach and ML detection for $K = 10$ and $K = 24$, with $N = 32$.

Finally, we study the required CPU time for the RTS based detector in more detail. The required CPU time for the RTS approach with various iterations M and for the semidefinite relaxation (SDP) approach in [10, 11] are shown in Figure 4. The complexity can be approximated through curve fitting techniques. It follows that the complexity (in CPU time) is approximated as:

$$\begin{aligned} C_{SDP} &\approx 9.4 \times 10^{-6} K^3 \quad \text{(SDP)} \\ C_{RTS1} &\approx 2.3 \times 10^{-3} K \quad \text{(RTS with } M = 50\text{)} \\ C_{RTS2} &\approx 4.5 \times 10^{-3} K \quad \text{(RTS with } M = 100\text{)} \\ C_{RTS3} &\approx 6.6 \times 10^{-3} K \quad \text{(RTS with } M = 150\text{)} \\ C_{ML1} &\approx 4.0 \times 10^{-4} 2^K \quad \text{(exhaustive)} \\ C_{ML2} &\approx 5.3 \times 10^{-6} K^4 \quad \text{(SDP with BB)} \end{aligned} \tag{4}$$

Included here are also complexity estimates for the ML detector based on exhaustive search (ML1), and a true ML detector based on SDP and the branch-and-bound (BB) algorithm (ML2). The CPU time of the RTS based detectors increases linearly as the number of users increases, while the SDP approach has a complexity of K^3. Although the SDP approach requires less CPU time for small systems, the RTS approach will be competitive for large systems. The true ML detectors have significantly higher complexity requirements for the investigated examples.

[2]The CPU time is obtained in the MATLAB environment on a 600 MHz Pentium III personal computer with 512 Mb of RAM.

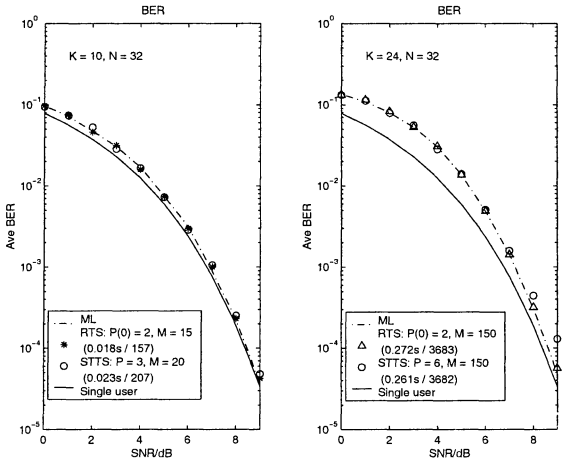


Figure 1: Comparison of average BER of the STTS and RTS based detectors for systems with load less than one.

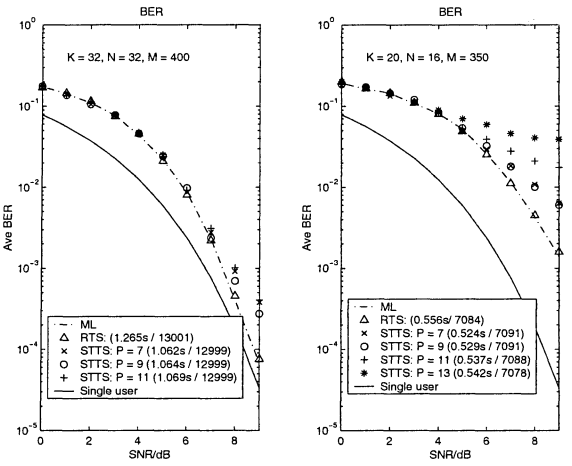


Figure 2: Comparison of average BER of the STTS and RTS based detectors for systems with load equal to or greater than one.

4 Conclusion

In this paper, we have introduced a reactive tabu search algorithm as an complexity-limiting alternative to optimal ML detection. It adheres to the symbol constraints and performs local moves in the space

of symbol solutions. It is implemented by a tabu list of moves which have been made in the recent past of the search. These moves are now forbidden according to specified rules, in order to diversify the search and avoid oscillation between local optima. Although Tabu search algorithms are designed to search for global optima, they cannot guarantee that the best solution found is indeed the global optimal solution to the problem. However, it was demonstrated that performance indistinguishable from ML performance can be achieved by RTS, for the chosen numerical examples.

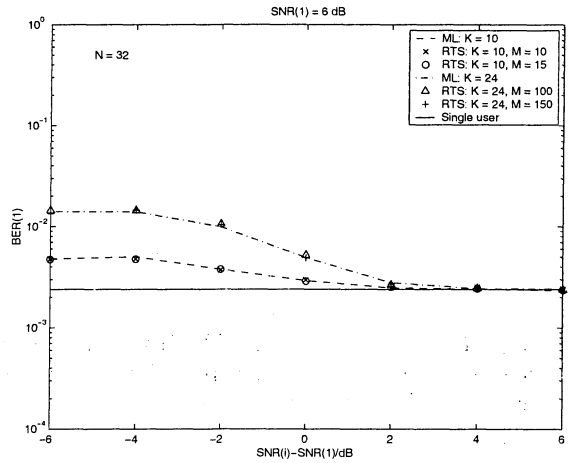


Figure 3: Average BER of RTS based detectors to illustrate near-far performance.

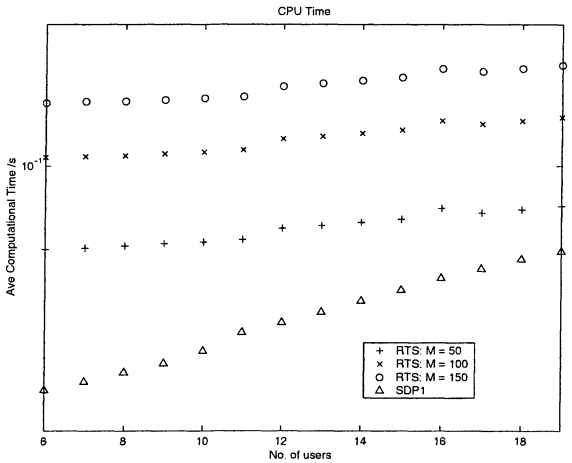


Figure 4: Average CPU time of the RTS and SDP based detectors with various number of users.

References

[1] S. Verdú, "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 32, pp. 85–96, Jan. 1986.

[2] S. Verdú, "Computational complexity of optimum multiuser detection," *Algorithmica*, vol. 4, pp. 303–312, 1989.

[3] V. J. Rayward-Smith, I. H. Osman, C. R. Reeves, and G. D. Smith, *editors, Modern Heuristic Search Methods.* John Wiley and Sons, 1996.

[4] F. Glover and M. Laguna, *Tabu Search.* Kluwer Academic Publishers, Boston MA, 1997.

[5] P. H. Tan, *Multiuser Detection in CDMA - Combinatorial Optimization Methods.* Tech. Rep. 408L, Department of Comp. Eng., Chalmers University of Technology, Sweden, 2001.

[6] S. Verdú, *Multiuser Detection.* Cambridge Univ. Press, 1998.

[7] F. Glover, "Future paths for integer programming and links to artifical intelligence," *Computers and Operations Research*, no. 5, pp. 533–549, 1986.

[8] R. Battiti and G. Tecchiolli, "The reactive tabu search," *ORSA Journal on Computing*, no. 2, pp. 126–140, 1994.

[9] P. Németh, L. K. Rasmussen, and T. M. Aulin, "Maximum-likelihood detection of block coded CDMA using A* algorithm," *in Proc. Int. Symposium Inf. Theory '01 (Washington, USA)*, p. 88, June 2001.

[10] W. K. Ma, T. N. Davidson, K. M. Wong, Z. Q. Luo, and P. C. Ching, "Efficient quasi-maximum-likelihood multiuser detection by semi-definite relaxation," *in Proc. Int. Conf. Commun. '01 (Helsinki, Finland)*, pp. 333–337, June 2001.

[11] P. H. Tan and L. K. Rasmussen, "The application of semidefinite programming for detection in CDMA," *IEEE J. Selected Areas Commun.*, pp. 1442–1449, Aug. 2001.

Utilizing the channel correlation for MAI reduction in downlink multi-carrier CDMA systems

Anders Persson, Tony Ottosson, and Erik Ström
Dept. of Signals and Systems, Chalmers University of Technology
E-mail: {Anders.Persson, Tony.Ottosson, Erik.Strom}@s2.chalmers.se

Abstract

An OFDM-CDMA system that utilizes the channel-correlation in both time and frequency to reduce multiple access interference is proposed. Orthogonal Walsh-Hadamard codes are used for spreading and in order for the codewords to stay orthogonal at the receiver, the channel response must be flat in both time and frequency. A frequency selective fading channel will, however, distort the codewords and thus introduce correlation. The proposed system, named Time-Frequency Localized CDMA (TFL-CDMA), places the codeword-chips in rectangular blocks on the time-frequency-grid in a such a way that the difference in scaling of the codeword chips is reduced. The proposed system is found to outperform competing schemes, such as Multi-Carrier (MC) and Multi-Carrier Direct-Sequence CDMA (MC-DS), in terms of bit error rate. The MC- and MC-DS-CDMA systems are special cases of the proposed scheme.

1 Introduction

In a traditional Direct-Sequence CDMA system signals are separated by user specific spreading codes, designed to have low correlation. The spreading is performed in the time domain, and the resulting signal is transmitted on a single carrier, see figure 1a. For downlink transmission, where signals can be transmitted synchronously, the orthogonal Walsh-Hadamard codes are often used. The orthogonality is however destroyed by a frequency selective channel. This effect, known as inter-path-interference, can severely degrade the performance. An OFDM system transmits with a lower rate on several parallel narrowband sub carriers so that each carrier is subject to approximately flat fading [1]. By using a cyclic prefix the sub carriers can be made orthogonal and the inter-symbol interference negligible. Placing the CDMA-chips on different carriers therefore ensures low inter-chip interference. Another advantage is that, as opposed to a single carrier DS-CDMA system, there is no need for a RAKE-receiver to resolve the time disperse channel. As a consequence there is also no need to find the exact positions of individual channel taps. Despite big implementation challenges, like designing linear amplifiers to cope with the large Peak-to-Average-Power ratio of an OFDM signal, multi carrier schemes are promising and has been subject to intensive research over the years, [2–4]. The earlier proposed Multi-Carrier CDMA (MC-CDMA) [5–7] and Multi-Carrier Direct-Sequence CDMA (MC-DS-CDMA) [8] systems spread the symbols either in the frequency or in the time direction [3]. Theses schemes are illustrated in figures 1c and d. A frequency selective fading channel will scale the chips differently and cause correlation between the codewords. As will be seen in this paper, these schemes therefore easily become interference limited.

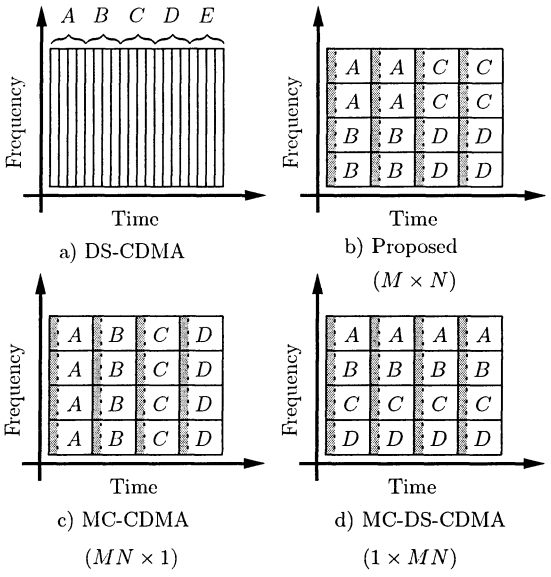


Figure 1: Spreading in different CDMA systems.

Figure 1a illustrates a wide-band single carrier system. The schemes in figure 1b, c and d are multi-carrier systems, were each column of the time-frequency grid corresponds to one OFDM-symbol. Each rectangle denotes one CDMA-chip and the letters $A - E$ represent different codewords. M and N denotes the amount of spreading in the frequency and time direction respectively of the proposed system. Due to the cyclic prefix (gray regions) the multi-carrier schemes occupy more time-frequency resources compared to the DS-CDMA scheme.

2 The proposed scheme

The proposed system utilizes the fact that the channel is correlated in both time and frequency, see figure 5. The codewords are localized to rectangular areas in order to reduce the difference in scaling among the chips. Let M and N denote the spreading in the frequency and time direction respectively of the proposed system. The MC- and MC-DS-CDMA systems can now be viewed as the $(MN\times 1)$ and $(1\times MN)$ special cases of the proposed scheme. A block-scheme of the transmitter of the proposed system is shown in figure 2. The kth users information bits, b_k, are convolutionally encoded by the encoder labeled CC, interleaved by the pseudo-random-interleaver Π and thereafter QPSK-modulated. Thereafter traditional traditional direct-sequence spreading is applied, where c_k denotes the spreading code of user k. The different users' signals, s_k, are summed up prior to the block-shaping. The block-shaping operation places the codeword-chips in $(M \times N)$ sized blocks on the time-frequency grid as illustrated in figure 1 b. The first M chips of the codeword are placed in the first column of the size $(M \times N)$ matrix, the next M in the second column, and so on. The interleaver provides interleaving in both time and frequency, i.e., it makes the position of each $(M \times N)$ block (encoded bit) on the grid pseudo-random. Several blocks per user are transmitted in parallel so that, as opposed to in a frequency hopped CDMA system, each user constantly occupies the entire bandwidth. The

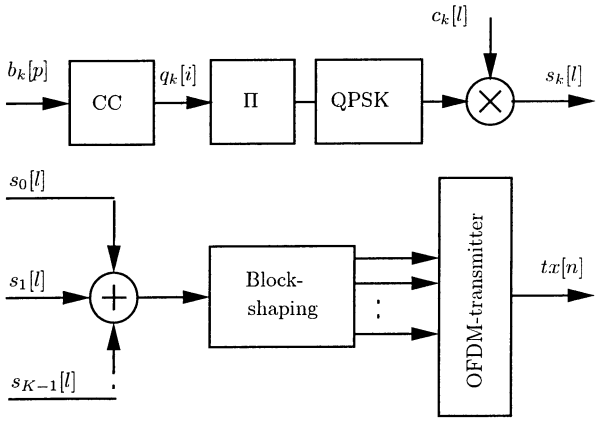


Figure 2: Transmitter block-scheme of the proposed system.

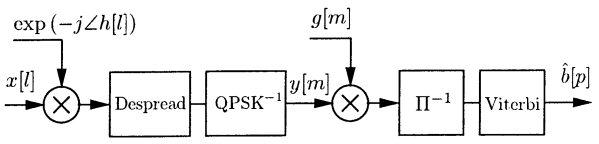


Figure 3: Correlator, demodulator, deinterleaver and decoder.

transmitter can be divided into a traditional DS-CDMA scheme with interleaving and channel coding, followed by the proposed block-shaping and a traditional OFDM transmitter. A rate$\frac{1}{2}$ convolutional code with a memory of 6 bits and generator polynomials $\{133, 171\}$ given in octal format is used. Walsh-Hadamard codes of length $MN = 64$ are used for the direct-sequence spreading. The total spreading of the DS-CDMA part of the scheme now becomes $T_{b_k}/T_{s_k} = 2\cdot\frac{1}{2}\cdot 64 = 64$, where T_{b_k} and T_{s_k} denotes the bit and chip time respectively. The OFDM transmitter uses a 256-point IFFT, a main carrier frequency of 2 GHz and a sub carrier spacing of 15.625 kHz. A 4 μs guard interval (cyclic prefix) is inserted. The resulting OFDM symbol time is 68μs and the system bandwidth approximately 4 MHz. The main carrier frequency and system bandwidth are thus similar to those of the UMTS system. One can argue that the cyclic prefix also contributes to the total spreading. Taking it into account yields a total spreading of $T_{b_k}/T_{tx} = 2\cdot\frac{1}{2}\cdot 64\cdot\frac{68}{68-4} = 68$, where T_{tx} is the sample output time (no oversampling is assumed).

3 Channel model

A frequency selective fading channel modeled as a time varying FIR-filter is used in the performance evaluation of the proposed system [9, p. 797]. The channel taps are modeled as independent complex Gaussian processes (Rayleigh fading), having the power spectrum described in [10, p. 20]. The power-delay profile, shown in table 1, is chosen according to the standardized test channel *Vehicular Reference Environment A*, [11, p. 37].

delay [ns]	$\{0, 310, 710, 1090, 1730, 2510\}$
mean power [dB]	$\{0, -1, -9, -10, -15, -20\}$

Table 1: Power-delay profile for the Vehicular A channel. Delay and mean power relative the strongest tap.

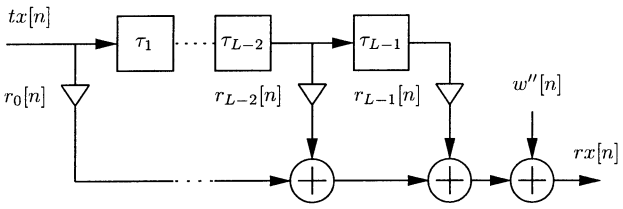


Figure 4: Frequency selective channel.

4 Performance evaluation

The performance is evaluated in a downlink scenario, where users are likely to request different transmitted powers in order to compensate for multiple access interference as well as their individual path losses. To

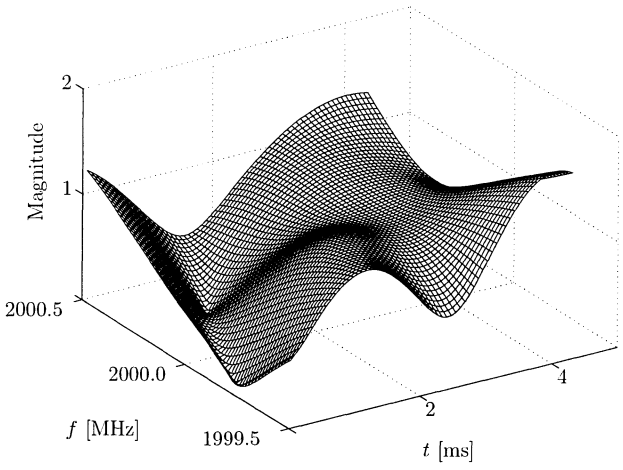


Figure 5: Envelope of the complex valued time-frequency channel response H. A realization of the "Vehicular A" channel at a carrier frequency of 2 GHz and a mobile speed of 50 km/h.

illustrate the robustness of the proposed TFL-CDMA system to this so called near-far situation, the user that is decoded is given a received power that is either 5, 7.5 or 10 dB below each of the remaining active users within the cell. The average received E_b/N_0 is, before the removal of the cyclic prefix, in all cases set to 8 dB. All figures show the performance at full system load, i.e., $K = 64$ active users. Except for in figure 9, the mobile speed is always set to 50 km/h. The optimal detector is the maximum likelihood sequence detector that jointly detects the different users signals. However, the complexity of this detector grows exponentially with the number of users and is, especially in the downlink, unrealistic with todays hardware. Therefore, sub optimal single user detectors are considered in this paper. A traditional OFDM-receiver followed by an operation that performs the inverse to the earlier described block-shaping, yields the signal x that enters the block diagram of figure 3. Thereafter channel phase compensation, despreading and demodulation is performed. The output from the QPSK-demodulator for user 0 can be written

$$\begin{aligned} y_0[m] &= \sqrt{\xi_0} q_0[m] \underbrace{\sum_{l=mS}^{(m+1)S-1} c_0^2[l]|h[l]|}_{\alpha[m]} + \\ &+ \underbrace{\sum_{k=1}^{K-1} \sqrt{\xi_k} q_k[m] \sum_{l=mS}^{(m+1)S-1} c_k[l] c_0[l] |h[l]|}_{z[m]} + \\ &+ \underbrace{\sum_{l=mS}^{(m+1)S-1} w'[l] c_0[l]}_{w[m]} \\ &= \sqrt{\xi_0} q_0[m]\alpha[m] + z[m] + w[m]. \end{aligned} \tag{1}$$

Here ξ_k, q_k and c_k denotes the chip energy, the encoded bits and the normalized spreading sequence of user k respectively. $S = MN$ is the length of the direct-sequence spreading code. The two dimensional time-frequency channel response H, see figure 5, is after the inverse block shaping denoted h. Assume that the intra-cell-interference z as well as the additive receiver noise w in equation (1) is Gaussian and independent from coded bit to coded bit (perfect interleaving). The single user maximum likelihood sequence detector, given $\{y_0\}$ in equation (1), chooses the sequence of coded bits $\{\hat{q}_0\}$ with elements $q_0[i] \in \{+1, -1\}$ as

$$\begin{aligned} \{\hat{q}_0\} &= \arg\max_{\{q_0\}} \exp\left(-\frac{1}{2} \sum_i \frac{\left(y_0[i] - \sqrt{\xi_0} q_0[i]\alpha[i]\right)^2}{\sigma_z^2[i] + \sigma_w^2} \right) \\ &= \arg\max_{\{q_0\}} \sum_i \underbrace{\frac{\alpha[i]}{\sigma_z^2[i] + \sigma_w^2}}_{g[i]} \cdot y_0[i] \cdot q_0[i] \end{aligned} \tag{2}$$

The mapping between the information bits $\{\hat{b}_0\}$ and the coded bits $\{\hat{q}_0\}$ is determined by the convolutional code. The receiver is implemented by a standard Viterbi decoder that maximizes the correlation between the pre-weighted input $\{g \cdot y_0\}$ and the hypothesis sequence $\{q_0\}$, see figure 3. The relation between the time indexes m and i is determined by the interleaver Π. From equation (2) it is seen that the sequence of weights, $\{g\}$, should be chosen as

$$g[i] = \frac{\alpha[i]}{\sigma_z^2[i] + \sigma_w^2}. \tag{3}$$

Choosing $\{g\}$ as in equation (3) requires knowledge of the channel envelope, the power of each interfering user and the variance of the receiver noise. A more practical approach is to disregard the interference part σ_z^2 and to choose $\{g\}$ as

$$g[i] = \alpha[i] \tag{4}$$

were the thermal noise now also can be disregarded since σ_w^2 can be considered constant in time.

4.1 Semi-analytical BER evaluation

The BER performance of a convolutional code can be upper bounded by a union bound [9, p. 488]

$$P_b \leq \sum_{d=d_\mathrm{f}}^{\infty} c_d P_2(d) \tag{5}$$

where d_f is the free distance of the code, $P_2(d)$ is the pair-wise error probability of two sequences separated by a Hamming distance d, and c_d is the information error weights of the code. An approximate value of the BER may be obtained by truncating the sum in equation (5) at some suitable d. To find an expression for $P_2(d)$ in a fading scenario the error probability given the instantaneous signal to interference plus

noise ratio, γ, should be averaged over the probability density function $f_d(\gamma)$. Assuming Gaussian interference and noise, yields the pair-wise error probability

$$P_2(d) = \int_0^\infty f_d(\gamma) Q\left(\sqrt{2\gamma}\right) d\gamma. \quad (6)$$

Using the soft metric weights in equation (3) results in maximum ratio combining. A closed formed expression for $f_d(\gamma)$ has not been found in this case. Histogram estimates of the pdf can however be used, see figure 6. These are obtained from computer realizations of the sum of d independent variables γ_c generated according to equation (7).

$$\gamma_c = \frac{\frac{1}{2}\alpha^2[i]}{\sigma_z^2[i] + \sigma_w^2} \quad (7)$$

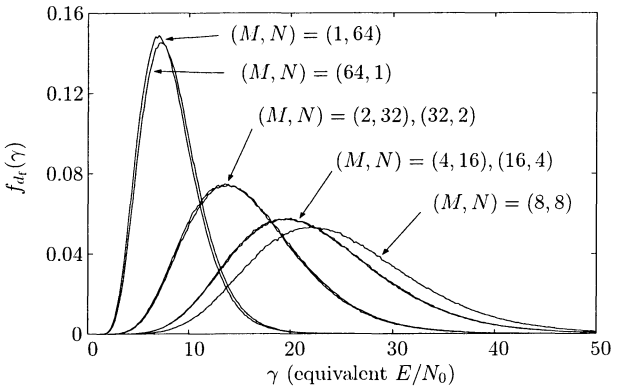


Figure 6: Histogram based estimate of the pdf of the signal to interference plus noise ratio of 10 combined *coded* bits, (the free distance of the code is 10). The horizontal axis shows the equivalent E/N_0 in linear scale. The soft metric weights are chosen as in equation (3), i.e. maximum ratio combining. The near-far ratio is 10 dB and the average received Eb/N_0 is 8 dB.

Although the analysis now becomes partly based on computer simulations, it serves as a method of validating the full Monte-Carlo simulation.

4.2 Computer simulations

Monte-Carlo simulations of the proposed system using different spreading block shapes (M, N) were performed. The main goal of this performance evaluation is to illustrate the potential gains that result from reduction of MAI, due to the block shaping operation. In order to isolate this phenomena as much as possible imperfections, such as diversity limitations due to insufficient interleaving, has to be avoided. For this reason, the simulations are performed with an interleaver size of approximately 20000 coded bits (≈ 0.15s). A system with real time demands would have to use a significantly smaller interleaver. Perfect synchronization and perfect channel estimates were also used.

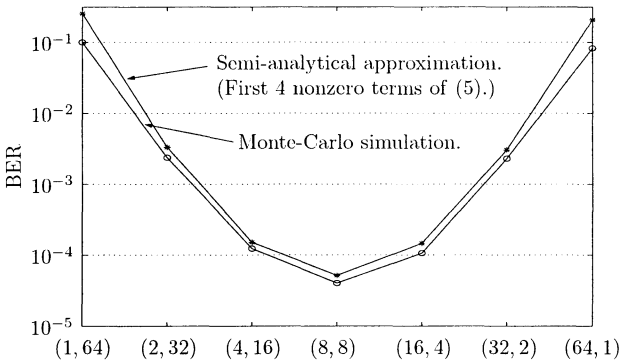


Figure 7: Bit error rate at a near-far ratio of 10 dB for different block shapes (M, N) when using the soft metric weights in equation (3). The average received E_b/N_0 equals 8 dB. The earlier proposed MC- and MC-DS-CDMA systems correspond to the $(64, 1)$ and $(1, 64)$ shapes respectively.

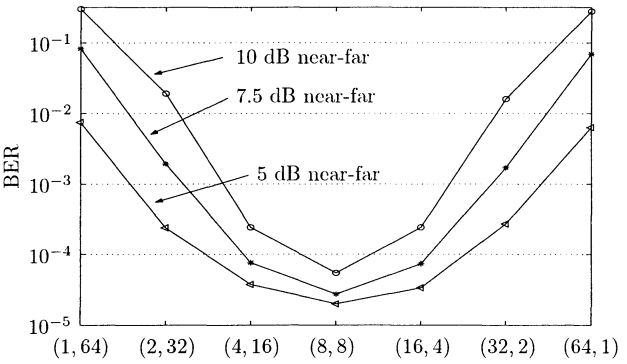


Figure 8: Computer simulation of the bit error rate for different block shapes, (M, N), at different near-far ratios and an E_b/N_0 of 8 dB. Soft metric weighting according to equation (4).

5 Discussion and conclusions

The earlier proposed MC- and MC-DS-CDMA systems correspond to the $(MN, 1)$ and $(1, MN)$ special cases of the proposed TFL-CDMA scheme. Localizing the codewords on the time-frequency grid suppresses the intra-cell-interference that results from codeword distortion. Placing the codewords non-localized, like in the MC- and MC-DS-CDMA systems, yields severe interference problems, but on the other hand ensures lower probability of placing the entire codeword in a deep fade. The choice of block shape is thus a trade-off between either obtaining interference suppression or diversity gains. However, in a downlink CDMA scenario, the overall system performance (number of simultaneously supported users at a certain bit rate) is in most situations limited by intra- and inter-cell-interference. Increased near-far robustness (lower correlation between codewords) reduces

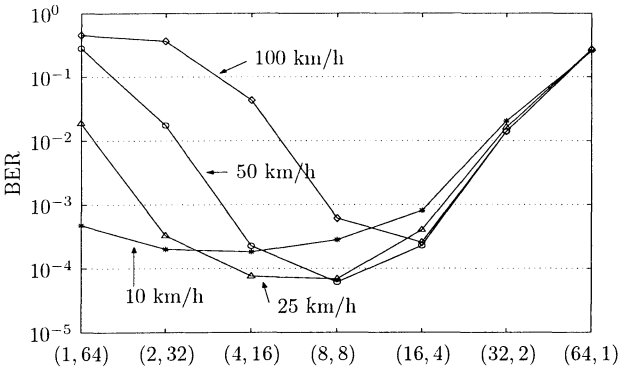


Figure 9: Computer simulations of the bit error rate at a near far ratio of 10 dB for different block shapes, (M, N), and different mobile velocities. Soft metric weighting according to equation (4) and an E_b/N_0 of 8 dB.

the amount of power necessary to maintain a certain bit error rate. As a consequence, also the amount of interference power experienced by other users in the system is reduced, and a significant increase in overall system performance can be expected. From figure 7-9 it is concluded that both the MC- and MC-DS systems are outperformed by the proposed scheme in a near-far scenario. From a comparison of figure 7 and 8 it is concluded that the loss in BER performance when using the soft metric weights in equation (4) rather than the maximum ratio weights in equation (3) is small. The preferred shape of the spreading block varies with the statistical properties of the channel. The proposed scheme is however found to be robust to different Doppler frequencies (mobile speeds), see figure 9. The increase in bit error rate in the 10 km/h case is likely caused by the diversity loss due to the slowly time varying channel (insufficient interleaving). No power control scheme, that partly can compensate for slowly time varying channels, is used in the simulations. A thorough study involving different power-delay profiles and mobile velocities is needed, as well as a fully analytical performance evaluation. Also, the use of the proposed scheme in a quasi-synchronous uplink, where comparable performance gains can be expected, is a topic for future research [4].

References

[1] R. Van Nee and R. Prasad, *OFDM For Wireless Multimedia Communications*, Artech House, 2000.

[2] D. Lee and L. B. Milstein, "Comparison of multicarrier DS-CDMA broadcast systems in a multipath fading channel," *IEEE Transactions on Communications*, vol. 47, no. 12, pp. 1897–1904, Dec. 1999.

[3] S. Hara and R. Prasad, "Design and performance of multicarrier CDMA systems in frequency-selective rayleigh fading channels," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 5, pp. 1584–1999, 1999.

[4] G. B. Giannakis, "AMOUR-generalized multicarrier transceivers for blind CDMA regardless of multipath," *IEEE Transactions on Communications*, vol. 48, no. 12, pp. 2064–2076, Dec. 2000.

[5] A. Chouly and A. Brajal S. Jourdan, "Orthogonal multicarrier techniques applied to direct sequence spread spectrum CDMA systems," in *Proc. IEEE Global Telecommunications Conference*, Huston USA, Nov. 1993, pp. 1723–1728.

[6] K. Fazel and L. Papke, "On the performance of convolutionally coded CDMA/OFDM for mobile communication systems," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Yokohama, Japan, Sept. 1993, pp. 468–472.

[7] N. Yee, J-P. Linnartz, and G. Fettweis, "Multicarrier CDMA in indoor wireless radio networks," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, Yokohama, Japan, Sept. 1993, pp. 109–113.

[8] V. M. DaSilva and E. S .Sousa, "Performance of orthogonal CDMA codes for quasi-synchronous communication systems," in *Proc. IEEE International Conference on Universal Personal Communications*, Ottawa, Canada, Oct. 1993, pp. 995–999.

[9] J. G. Proakis, *Digital Communications, 3rd edition*, McGraw-Hill, 1995.

[10] W. C. Jakes, *Microwave Mobile Communications*, IEEE Press, 1993, reissue.

[11] J. P. Castro, *The UMTS Network and Radio Access Technology*, John Wiley & Sons, 2001.

Nonlinear electromagnetic wave propagation

D. Anderson

[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden.
-mail: elfda@elmagn.chalmers.se.

Abstract

A general discussion is given of the basic differences between linear and nonlinear wave propagation. In particular, it is emphasized that in contrast to the the linear case, nonlinear wave propagation implies that the wave propagation properties depends on the intensity of the wave. Two explicit examples of electromagnetic wave propagation are discussed in more detail. One example concerns the propagation properties of wave pulses in optical fibre communication systems where the refractive index depends on the wave intensity in addition to the classical dependence on the wave frequency (so called Kerr media). It is shown that the nonlinear effect, under certain conditions, may counteract the linear phenomenon of fibre dispersion and may lead to dispersion-less pulse propagation. The second example discusses some of the properties of the microwave breakdown phenomenon, which may occur in a number of microwave applications ranging from mobile telephone communication to space communication. In this case the nonlinear properties have a threshold character in the sense that below a certain threshold power, the wave propagation behaves linearly and in accordance with classical theory for wave propagation, whereas above this threshold, the propagation behaves completely different, being strongly modified by self-induced plasma formation.

1 Introduction

The classical theory of wave propagation, which has been developed over a period of almost four hundred years, is a linear theory. The crucial feature in this theory is the superposition principle, i.e. a wave packet is considered to be a superposition of a number of fundamental wave modes. The properties of these modes are determined by the medium through which the wave propagates and are summarized in the dispersion relation, which relates the frequency, ω, and wave number, k, of the fundamental wave modes i.e. $k = k(\omega)$. The solution procedure of a linear wave propagation problem typically proceeds as follows. The initial amplitude, $\psi(0,t)$, of the wave pulse, is Fourier transformed and decomposed into its spectral components. Each frequency component is then propagated over the distance x with its corresponding wave number i.e. it is multiplied with the phase factor $\exp[-ik(\omega)x]$. This yields the Fourier transform at the propagation distance x. This transform is then inverted to give the wave amplitude, $\psi(x,t)$, at the distance x. An important consequence of this approach is the well known feature of wave interference. When two waves pass through each other, they interact and give rise to different forms of interference phenomena. However, once they have passed out of the interaction region, no trace of the interaction is left on any of the waves. In other words, it is impossible to detect afterwards that the waves have indeed interacted.

With a few notable exceptions, it was realized approximately only 40 years ago that this picture of wave propagation was incomplete and, in fact, inconsistent. It was found that in certain situations, the wave may affect its own propagation properties and furthermore when two waves collide, i.e. pass through each other, the interaction may lead to completely new phenomena, which leave enduring imprints on the waves even after they have separated.

One of the most fascinating aspect of these new phenomena is that the famous superposition principle, the basis of the theory of linear wave propagation, turned out to be invalid. This also meant that most of the theoretical framework for analyzing wave propagation suddenly became obsolete, and it became necessary to develop new analytical tools. The subsequent development of the theory of nonlinear wave propagation is one of the major achievements in mathematical physics during the last 40 years. The purpose of the present work is to give a short overview of the basic features of nonlinear wave propagation with special emphasis on electro-magnetic wave phenomena. Two particular examples will be given, one taken from the area of high speed optical fibre communication and the second from microwave communication in terrestrial as well as space environments.

2 Nonlinear Kerr media

For a certain (wide) class of nonlinear media (the so called Kerr media), the refractive index of the medium, n, depends, not only on the frequency of the wave, but also on its intensity, I, i.e. $n = n(\omega, I)$. In most materials the intensity dependence is very weak and one can expand $n(\omega, I)$ around $n(\omega, 0) = n_0(\omega)$ as follows:

$$n = n(\omega, I) \approx n_0(\omega) + \frac{dn(\omega, 0)}{dI} I \qquad (1)$$

Note that $n_0(\omega)$ is the (usual) linear part of the refractive index and $dn(\omega, 0)/dI = n_2$ is the nonlinear Kerr coefficient. This coefficient is usually very small, e.g. in silica fibres, $n_2 \approx 10^{-16} cm^2/W$. The power, P , of optical pulses in fibre communication systems is typically $n_2 \approx$ 100mW, which when distributed over the cross section of the fibre with a diameter approximately $10\mu m$, gives rise to an intensity $I \approx 10^5 W/cm^2$. Consequently, the relative change in refractive index due to the nonlinear contribution, $\Delta n/n_0$, is then

$$\frac{\Delta n}{n_0} = \frac{n_2 I}{n_0} \approx \frac{10^{-16} 10^5}{1} = 10^{-11} \qquad (2)$$

A natural question arises directly in connection with this result: If the relative change of the refractive index caused by the nonlinearity as so extremely small, how can possibly the nonlinear term be important? In order to investigate this problem let us consider the change in wave number by expanding it around the carrier frequency ω_0, keeping terms to second order. This yields

$$k(\omega, I) \approx k_0 + \frac{dk_0}{d\omega}(\omega - \omega_0) + + \frac{1}{2}\frac{d^2 k_0}{d\omega^2}(\omega - \omega_0)^2 + \frac{\omega_0}{c} n_2 I \qquad (3)$$

where $k_0 = k(\omega, 0)$ and c denotes the velocity of light. The first term on the right hand side of this expansion is the largest, but only causes a phase shift, common for all parts of the pulse. The second term is smaller and gives rise to the well known group velocity of the pulse. The two last terms are the smallest in the expansion. The third term is responsible for dispersive pulse broadening. In optical fibre communication one makes strong efforts to design the fibre in such a way as to make this term as small as possible since dispersion is one of the factors which tends to limit the achievable transmission rate of the communication system. On the other hand, the nonlinear term is an inherent property of the fibre. This implies that the importance of the nonlinear term should be inferred, not by comparing it with k_0, but rather by comparing it with the dispersion term, i.e. with the very small third term on the RHS of eq.(3). It turns out that the "internal" dynamics of the wave pulse is in fact determined by the interplay between the two last very small terms.

The magnitude of the dispersive term can be estimated as

$$\frac{1}{2}\frac{d^2 k_0}{d\omega^2}(\omega - \omega_0)^2 \approx \frac{k_0''}{\tau_p^2} \qquad (4)$$

where τ_p is the pulse length. The dispersion term and the nonlinear term are of the same order when $k_0''/\tau_p^2 \approx \omega_0 n_2/c$. For 10-ps pulses in an optical fibre with low dispersion (5-10 ps/(nm km)), the terms are of the same magnitude for pulse powers of the order of a few 100's of mW. This clearly shows that the nonlinear effects can indeed be expected to play an important role for the pulse dynamics in an optical fibre.

Another fascinating and important consequence of this is the conclusion that if $k_0'' < 0$ (i.e. the fibre has anomalous dispersion), the two last terms in the expansion have different signs (n_2 is always positive) and tend to counteract each other. In fact, if they also have the same magnitude, the nonlinear term should cancel the dispersive term and the pulse should propagate without dispersion i.e. with unchanged form. This is the famous soliton pulses predicted theoretically in 1976 by Hasegawa and Tappert, and later (1980) verified in a very beautiful experiment by Mollenauer, Stolen and Gordon at Bell Laboratories. They investigated the influence of pulse power on dispersive pulse propagation in optical fibres and the agreement with theoretical predictions was complete.

3 The nonlinear Scrödinger equation

The characteristic equation determining the evolution of the slowly varying envelope, $\psi(x, \tau)$, of the pulse is the nonlinear Schrödinger equation given by

$$i\frac{\partial \psi}{\partial x} = \frac{k_0''}{2}\frac{\partial^2 \psi}{\partial \tau^2} + \kappa |\psi|^2 \psi \qquad (5)$$

where $\tau = t - k_0' x$ denotes time measured in the coordinate system moving along with the pulse i.e. moving with the corresponding group velocity, $v_g = 1/k_0'$. It is clear that for low pulse intensities the last term should be negligible and the pulse propagates according to the linear Schrödinger equation, which describes the classical dispersive broadening of the pulse. The solution of the corresponding equation can be obtained using the full arsenal of linear methods, in particular Fourier transform techniques. However, when the intensity becomes so high that the dispersive and nonlinear terms are comparable, the corresponding nonlinear equation is not amenable to such linear solution techniques. Instead a completely new theory

has been developed, the inverse scattering theory, to solve the Nonlinear Schrödinger equation. One of the most striking features of the equation is the soliton solutions, wave pulses which are sech-shaped in time and propagate with unchanged form. The possibility of using such pulses as information bits in optical fibre communication for high bit rates and/or long propagation distances was directly realized and during the 20 years since the experiment of Mollenauer et al, there has been a strong effort directed at investigating the possibilities provided and limitations set by nonlinear effects in optical fibres. However, there has also been an intensive research directed at the possibility of using the nonlinear Kerr effect for manipulating light by light e.g. for optical switching purposes or for optical pulse compression.

4 Microwave-induced breakdown

In the linear picture of microwave propagation e.g. in a wave guide, the propagation properties of the wave are given by the geometry of the device, which directly determines the dispersion relation for the wave. However, in the gas filling the wave guide, there are always a small, but finite, amount of free electrons, being created e.g. by cosmic radiation. These electrons will oscillate in the microwave field and if the RF field is strong enough, they will acquire energies that are so high that they will cause ionization either when they collide with the neutral gas molecules (corona discharges) or, if the mean free path between collisions is longer than the geometrical dimensions of the device, when they strike the walls of the device (multipactor). This tends to increase the free electron density and clearly this process may lead to an avalanche-like growth of the electron density, if the electron losses are not strong enough to stop this growth. In the case of corona discharges, which occur for not too small pressures, the ionization rate does exceed the loss rate of the electrons for sufficiently high microwave fields, and the electron density rapidly increases. This evolution have profound effects on the wave propagation since the wave guide medium, the gas, changes from having very low conductivity to becoming a good conductor. It is well known that an electromagnetic wave propagating in a weakly ionized medium (electron density n_e) has a refractive index given by

$$n^2 = 1 - \frac{\omega_p^2}{\omega(\omega + i\nu)} \tag{6}$$

where ν denotes the collision frequency between electrons and neutrals and $\omega_p \sim n_e^{1/2}$ is the characteristic plasma frequency determined by the electron density. It is clear that the electron density will increase until the refractive index has changed significantly, in fact, if we assume that the collision frequency is much smaller than the wave frequency, the increase will continue until $n \approx 0$ and the wave will become reflected by its self-generated plasma. This phenomenon is called microwave breakdown and is another dramatic example of nonlinear wave propagation, where the wave affects its own propagation properties and the analysis cannot be carried out within the framework of linear theory. It is interesting to note that, whereas in the case of the nonlinear Kerr effect, the nonlinearly induced change in refractive index is very small, in the case of microwave breakdown, it is generally very large, in fact $\Delta n/n \approx 1$, although certain microwave devices e.g. resonators may be much more sensitive to an increase of the electron density and may be affected already at much lower changes of the refractive index. We emphasize the difference in the nonlinear behaviour between the two cases considered: in the first case there is a continuous increase of the nonlinear effects for increasing intensity, whereas in the second case, the dependence has a pronounced threshold behaviour; for intensities below the threshold intensity, propagation behaves linearly, whereas once the intensity exceeds the threshold, dramatic changes in the propagation properties take place.

In many situations, it is important primarily to calculate the threshold microwave field at which breakdown occur. In this case a perturbation approach can be used where the microwave field is calculated using linear theory and for this field, the evolution of the electron density is then obtained from the continuity equation viz.

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n) + \nu_i n - \nu_a n \tag{7}$$

where D is the diffusion coefficient, ν_i is the ionization frequency and ν_a is the attachment frequency. All these characteristic frequencies depend on physical parameters, in particular gas pressure, however, in addition, the ionization frequency depends strongly on the electric field strength. It is clear that attachment and diffusion represent electron loss mechanisms which counteract the increase of the electron density caused by ionization.

A heuristic way of deriving the breakdown threshold condition is to introduce the diffusion loss frequency ν_D by writing

$$\nabla \cdot (D\nabla n) \approx -\frac{D}{L_{eff}^2} n \equiv -\nu_D n \tag{8}$$

where L_{eff} is the effective diffusion length. This simplified analysis makes it possible to write the continuity equation as

$$\frac{\partial n}{\partial t} = \nu_{\text{net}} n \tag{9}$$

where $\nu_{net} = \nu_i - \nu_D - \nu_a$. Clearly exponential growth of the electron density occurs when $\nu_{net} \geq 0$, which

defines the breakdown threshold for the microwave field by the condition $\nu_{net} = 0$.

The physical situation discussed above is relevant for gases of not too small pressures. At close-to-vacuum conditions, another electron multiplication process occurs, which is driven by secondary electron emission when electrons accelerated by the microwave field hit the walls of the device. This process, multipactor, is of a resonant character, since the transit times of the electrons between the walls of the device must coincide with half a period of the RF field in order for the electron avalanche to build up and furthermore requires the secondary emission coefficient of the walls to be larger than unity. Thus, multipactor breakdown tends to have both a lower and an upper threshold and to occur in certain bands of RF power.

Microwave breakdown in gases under various physical and technical conditions is a familiar and much studied problem. However, the technical development of microwave devices continues to give rise to new situations and parameter regimes where microwave breakdown plays an important role and where previous results and predictions concerning breakdown are not directly applicable. One example illustrating the situation is provided by the trend to design microwave devices as small and compact as possible. This development leads to concern about the concomitant breakdown strength of the design, in particular in situations involving more complicated geometries as e.g. in microwave wave guides, resonators and filters. The physical and technical situations where microwave breakdown may occur varies over a wide range from e.g. mobile telephone communication to space borne RF communication.

5 Summary

Two general examples of nonlinear effects in connection with propagation of electromagnetic waves have been discussed; the nonlinear Kerr effect in optical fibre communication systems and electrical breakdown in wave guides and other devices for microwave communication. Several more detailed examples of such nonlinear wave propagation phenomena will be presented in the poster session. The applications include three optical problems; "Incoherent light propagation in nonlinear media", "Pulse dynamics in optical amplifiers in the presence of bandwidth limitation and gain saturation", "Suppression of nonlinear effects by phase alternation in dispersion-managed optical fibre transmission" and three microwave problems; "Importance of geometry on microwave breakdown thresholds in resonators and filters; the concept of effective diffusion length", Microwave breakdown in slots" and Suppression of multipactor breakdown in RF equipment".

Nonlinear self focusing of partially coherent light beams

L. Helczynski [1], M. Lisak[1], D. Anderson[1]
B. Hall[1], R. Fedele[2], V. E. Semenov[3]

[1]Department of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden. E-mail: elfda@elmagn.chalmers.se.
[2]Dipartimento di Scienze Fisiche, Universita Federico II di Napoli and INFN Sezione di Napoli, Complesso Universitario di MS Angelo, Via Cintia, I–80 126 Napoli, Italy
[3]Institute of Applied Physics, Russian Academy of Sciences, 46 Ulyanov Street, Nizhny Novgorod, 603 600, Russia

Abstract

Propagation of partially incoherent light in nonlinear media is analyzed using the Wigner transform method. The power and versatility of this approach is illustrated by analyzing the effect of partial incoherence on the coherent self-focusing collapse instability. It is clearly demonstrated how partial incoherence tends to suppress the instability. The suppression can be described in terms of a Landau-like damping effect, which counteracts the coherent growth rate of the instability. In particular, it is shown how the nonlinear focusing effect becomes successively smaller as the coherence length of the light decreases and eventually no collapse phenomenon occurs.

1 Introduction

Propagation of electromagnetic waves and beams in dispersive nonlinear media (for instance, propagation of light pulses in optical fibers or bright beams in photorefractive crystals) has attracted strong interest during the last 20 years. The primary drive for this interest is the possibility to balance spatial diffraction or temporal dispersion by nonlinear effects and to create wave structures propagating with unchanged form—solitons. Applications range from high speed optical fibre communication over long distances to the possibility of creating light bullets. The conventional treatment of these phenomena assumes coherent waves. However, a series of experimental and theoretical works during the last five years, see e.g. Ref.[1] has demonstrated the existence of similar wave structures also when the wave is partially coherent. In particular, spatially incoherent/partially coherent optical solitons have attracted much attention. This research field is still at a very early stage of development, but the prospect for future applications is very promising.

From a theoretical standpoint, partially coherent light propagation presents new challenges. In particular, it becomes necessary to use a statistical rather than a deterministic approach. The main progress so far has been made for situations where the material response time is much longer than the time scale of the stochastic variations i.e. the material responds to the statistical average of the stochastic field variations rather than to the variations themselves.

In order to analyze the dynamics of partially coherent optical waves and beams in dispersive and nonlinear media, we have developed a novel, general, statistical theory based on the Wigner transform method and including a Klimontovich statistical averaging of the stochastic fluctuations. Starting from a generalized nonlinear Schrödinger equation (NLSE) describing evolution of the slowly varying wave amplitude in a dispersive medium with an arbitrary nonlinear response, we derive the Wigner-Moyal equation for generalized Wigner function. This equation is a propagation equation for Wigner transform of the wave field, in the same sense that NLSE is the propagation equation for the wave field. It reduces, in the geometrical optics approximation, to the classical Liouville or Vlasov-like equation describing the conservation of optical quasi-particles (photons) in phase-space.

This approach makes it possible to investigate a number of dynamical problems. In particular, the recent discovery that modulational instability, as well as spatial collapse of cylindrical light beams, tend to be suppressed by the incoherence is conveniently explained by our new theory in terms of a damping effect, analogous to the well known Landau damping of plasma waves, see Ref.[2]. However, in this presentation we will concentrate on the effects of partial incoherence on the self-focusing collapse phenomenon. For more details see Ref.[3].

2 The Wigner formalism and the Wigner-Moyal equation

The analysis is based on the nonlinear Schrödinger equation for the slowly varying envelope field of the partially coherent optical wave $\psi(t,\mathbf{r})$, which reads

$$i\frac{\partial\psi}{\partial t}+\frac{\beta}{2}\nabla^2\psi+\kappa\langle\psi^*\psi\rangle\psi=0 \qquad (1)$$

In writing the equation in this form we have, without loss of generality, taken t and $\mathbf{r}$ as the evolution and spatial dispersive/diffractive variables, respectively. Furthermore, β is the diffraction (or second order dispersion) coefficient, κ is the nonlinear coefficient and the bracket $\langle\ldots\rangle$ denotes statistical ensemble average. For this description to be valid it is required that the relaxation time of the medium response is much longer than the characteristic time of the statistical wave intensity fluctuations.

In order to analyze the dynamics of the partially coherent optical field, it is convenient to introduce the Wigner transform of the envelope function together with the Klimontovich statistical average, which is defined by (s denotes the dimension of $\mathbf{r}$)

$$\rho(\mathbf{r},p,t)= \\ =\frac{1}{(2\pi)^s}\int_{-\infty}^{+\infty}d^s\xi e^{i\mathbf{p}\cdot\xi}\langle\psi^*(\mathbf{r}+\frac{\boldsymbol{\xi}}{2},t)\psi(\mathbf{r}-\frac{\boldsymbol{\xi}}{2},t)\rangle \qquad (2)$$

It can then be shown [2] that if $\psi(t,\mathbf{r})$ satisfies the NLSE (1), then the Wigner transform $\rho(\mathbf{r},p,t)$ satisfies the following Wigner-Moyal (WM) equation, viz

$$\frac{\partial\rho}{\partial t}+\beta\mathbf{p}\cdot\frac{\partial\rho}{\partial\mathbf{r}}+2\kappa\langle|\psi|^2\rangle\sin\left(\frac{1}{2}\frac{\overleftarrow{\partial}}{\partial\mathbf{r}}\cdot\frac{\overrightarrow{\partial}}{\partial\mathbf{p}}\right)\rho=0. \qquad (3)$$

The sine operator is defined by its series expansion and the arrows indicate that the derivatives act to the left and right, respectively. Note that Eq. (2) is to be solved subject to the integral condition

$$\langle\psi^*(\mathbf{r},t)\,\psi(\mathbf{r},t)\rangle=\int_{-\infty}^{+\infty}\rho(\mathbf{p},t,\mathbf{r})\,d^s\mathbf{p} \qquad (4)$$

Thus, the WM-equation is an integro-differential equation and this characteristic equation for $\rho=\rho(\mathbf{r},p,t)$ forms the basis of the subsequent analysis.

3 Collapse of partially incoherent optical beams

A number of investigations, both theoretical and experimental, has been made to study the propagation of partially incoherent 1D and 2D light beams in nonlinear media [1, 2, 3, 4]. In particular, the 2D self-focusing collapse phenomenon of partially incoherent light beams was recently analyzed using the coherent density function formalism [4]. Based on a virial theorem for the beam width defined in the rms sense, it was shown that the presence of partial incoherence leads to increase of the threshold for collapse.
We will here re-investigate the phenomenon of self focusing and collapse using the Wigner formalism and point out the similarities, but also the differences between the methods. In this application we denote the evolution variable as z being the distance of propagation, let $\mathbf{r}=(x,y)$ be the transverse coordinates of the beam and set $I=\langle|\psi|^2\rangle$. Thus, the WM equation for the Wigner distribution is written as

$$\frac{\partial\rho}{\partial z}+\beta\left(p_x\frac{\partial\rho}{\partial x}+p_y\frac{\partial\rho}{\partial y}\right) \\ +2\kappa I\sin\left(\frac{1}{2}(\frac{\overleftarrow{\partial}}{\partial x}\frac{\overrightarrow{\partial}}{\partial p_x}+\frac{\overleftarrow{\partial}}{\partial y}\frac{\overrightarrow{\partial}}{\partial p_y})\right)\rho=0 \qquad (5)$$

We will investigate the behaviour of some characteristic moments defined as

$$\langle\langle w(\boldsymbol{r},\boldsymbol{p})\rangle\rangle\equiv\frac{\iint w(\boldsymbol{r},\boldsymbol{p})\rho(\boldsymbol{r},\boldsymbol{p},z)d\boldsymbol{r}d\boldsymbol{p}}{\iint\rho(\boldsymbol{r},\boldsymbol{p},z)d\boldsymbol{r}d\boldsymbol{p}} \qquad (6)$$

Of particular importance are the lowest order moments $\langle\langle\boldsymbol{r}\rangle\rangle,\langle\langle\boldsymbol{r}^2\rangle\rangle,\langle\langle\boldsymbol{p}\rangle\rangle$ and $\langle\langle\boldsymbol{p}^2\rangle\rangle$. Using Eq. (5) it is possible to prove the following relations

$$\frac{d\langle\langle\boldsymbol{r}\rangle\rangle}{dz}=\beta\langle\langle\boldsymbol{p}\rangle\rangle, \qquad \frac{d\langle\langle\boldsymbol{p}\rangle\rangle}{dz}=0, \\ \frac{d\langle\langle\boldsymbol{r}^2\rangle\rangle}{dz}=2\beta\langle\langle\boldsymbol{r}\cdot\boldsymbol{p}\rangle\rangle, \quad \frac{d\langle\langle\boldsymbol{p}^2\rangle\rangle}{dz}=2\kappa\langle\langle\boldsymbol{p}\cdot\nabla_\perp I\rangle\rangle. \qquad (7)$$

Going one step further by taking the second derivative of $\langle\langle\boldsymbol{r}^2\rangle\rangle$, one can prove a virial theorem for the evolution of the beam width, σ, defined as

$$\sigma^2\equiv\langle\langle\boldsymbol{r}^2\rangle\rangle-\langle\langle\boldsymbol{r}\rangle\rangle^2 \qquad (8)$$

The theorem states that

$$\frac{d^2\sigma^2}{dz^2}=2\beta(H-\beta\boldsymbol{P}^2) \qquad (9)$$

where H and $\boldsymbol{P}$ are the Hamiltonian and momentum, respectively, defined as

$$H=\beta\langle\langle\boldsymbol{p}^2\rangle\rangle-\kappa\langle\langle I\rangle\rangle, \qquad \boldsymbol{P}=\langle\langle\boldsymbol{p}\rangle\rangle \qquad (10)$$

We emphasize that both H and $\boldsymbol{P}$ are constants of motion, which clearly implies that σ^2 is a polynomial of second order in z i.e.

$$\sigma^2(z)=\sigma^2(0)(1+c_1z+c_2z^2) \qquad (11)$$

where $\sigma(0)$ is the initial rms width of the beam and the constants c_1 and c_2 are constants determined by the initial properties of the beam. The result given

by Eqs. (9, 10) can conveniently be used to study the effect of the partial incoherence on the beam collapse. In particular, for a beam which is symmetrical around $\mathbf{r} = \mathbf{0}$ and $\mathbf{p} = \mathbf{0}$, we obtain $c_1 = 0$ and

$$c_2 = \frac{\beta}{\sigma^2(0)} H = \frac{\beta}{\sigma^2(0)} [\beta \langle\langle r^2 \rangle\rangle - \kappa \langle\langle I \rangle\rangle]_{z=0} \quad (12)$$

From this result we note that in the linear case ($\kappa = 0$) the coefficient c_2 is always positive implying that the beam always has positive curvature and diffracts monotonously. However, in the nonlinear case [$c_2 \neq 0$] and if κ and β have the same signs, the coefficient c_2 will become negative provided the intensity is high enough. This implies that the beam will focus and the beam width will in fact go to zero at a finite distance–the famous nonlinear beam collapse phenomenon.

In order to be explicit, we assume the initial profile to be a cylindrically symmetric Gaussian of the form $\Psi(x, y, 0) = \sqrt{I_0} \exp[(x^2 + y^2)/2a^2 + i\theta(x, y)]$. The function $\theta(x, y)$ characterizes the partial incoherence of the beam and is assumed to have such properties that

$$\langle \exp\{i[\theta(x + \eta/2, y + \zeta/2) - \theta(x - \eta/2, y - \zeta/2]\} \rangle = \\ = R(\xi) = \exp(-\xi^2/b^2), \quad (13)$$

where $R(\xi)$ denotes the autocorrelation function and $\xi = \sqrt{\eta^2 + \zeta^2}$. This implies that the initial Wigner distribution is given by

$$\rho(\mathbf{r}, \mathbf{p}, 0) = \frac{1}{4\pi} I_0 D^2 \exp\{-[(x^2 + y^2)/a^2 + \\ + D^2(p_x^2 + p_y^2)/4]\} \quad (14)$$

where $D^{-2} = 1/4a^2 + 1/b^2$. In the spatially diffractive 2D case, β and κ can be taken as positive constants. In this case the constant c_2 becomes

$$c_2 = \frac{\beta}{a^2} [2\beta(\frac{1}{a^2} + \frac{4}{b^2}) - \kappa I_0] \quad (15)$$

From this result it is clearly seen that the partial incoherence tends to cause an enhanced diffraction of the beam. This counteracts the nonlinear focusing effect and enhances the beam power required to cause collapse in a way which can be expressed explicitly as

$$I_0 > \frac{2\beta}{\kappa} (\frac{1}{a^2} + \frac{4}{b^2}) \quad (16)$$

A similar result has been obtained by Bang et al [4] based on the coherent density function formalism, where the governing equation is a Vlasov type equation consistent with the lowest order approximation of the WM equation only. In fact, the structure of the sine operator in the WM equation is such that the dynamics of simple moments like those in Eqs.(7) is—without any simplification or approximation—the same as for the Vlasov-like approximation. This is due to the fact that the contributions from all the higher order terms of the sine operator vanish. However, differences will indeed appear for the higher moments, expressing finer details of the beam dynamics.

4 Conclusions

The Wigner transform method has been used to analyze the self-focusing collapse instability of 2D partially incoherent light beams propagating in nonlinear Kerr media. It is shown that the partial incoherence of the beam leads to an enhanced diffraction effect which opposes the nonlinear self-focusing effect. In particular, the threshold condition for collapse of the beam is found to be increased if the coherence length of the phase variations occurring over the cross section of the beam becomes smaller than the beam width itself. The latter result is identical to that observed in a recent study [4]. However, it is important to realize that the approach used in that study, as well as most other treatments, is based on a Vlasov-type equation which in fact is equivalent to the geometrical optics approximation. The WM-approach used in the present paper is more general, although for coarse beam characteristics like the *rms* beam width, both approaches give identical results.

References

[1] D. N. Christodoulides, E. D. Eugenieva, T. H. Coskun, M. Segev, and M. Mitchell Phys. Rev. E. **63**, 035601 (2001).

[2] B. Hall, M. Lisak, D. Anderson, R. Fedele and V. E. Semenov, Phys. Rev. E. (to be published)

[3] L. Helczynski, D. Anderson, R. Fedele, B. Hall, M. Lisak, IEEE Journal of Quantum Electronics (to be published)

[4] O. Bang, D. Edmundson and W. Krolikowski, Phys. Rev. Lett. **83**, 5479 (1999).

Pulse dynamics in optical amplifiers in the presence of bandwidth limitation and nonlinear gain saturation

D. Anderson[1], P. Johannisson[1], M. Lisak[1], M. Marklund[1],
J. Dudley[2], J. Harvey[3], R. Kruhlak[3], A. Peacock[3] and E. Vanin[4]

[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden.
E-mail: elfda@elmagn.chalmers.se.
[2]Laboratoire d'Optique P. M. Duffieux, Université de Franche-Comté, Besançon, France.
[3]University of Auckland, Auckland, New Zealand.
[4]Optical Networks Research Laboratory, Ericsson Telecom AB, Stockholm, Sweden.

Abstract

Recent progress in the modelling of the amplification dynamics of short and/or powerful optical pulses is discussed. Particular emphasis is given to situations where effects due to bandwidth limited gain and/or nonlinear gain saturation become important. Simulation results and possible analytical approaches appropriate for such situations are reviewed.

1 Introduction

Modelling of pulse propagation in nonlinear, dispersive and amplifying optical media has been the subject of intensive investigations during recent years with potential applications to optical fibre amplifiers and high-speed optical fibre communication. The universal model equation describing the pulse dynamics in such a medium was for a long time taken as the Nonlinear Schrödinger (NLS) equation, augmented to include gain and fibre loss, parameters which were taken as constants. However, for amplification of very short pulses, the limited bandwidth of the amplifier becomes important and significantly affects the pulse dynamics. Furthermore for high pulse powers, gain saturation also plays a crucial role in the pulse amplification dynamics. Several efforts have been made to analyse the importance of these effects by generalising the NLS equation to include gain saturation and bandwidth limited gain as small perturbation terms. However, little attention has so far been given to situations where the pulses are so short and/or powerful that neither the bandwidth limitation nor the gain saturation effect can be considered as small.

In a collaboration between Chalmers University of Technology, Université de Franche-Comté in Besançon, France, University of Auckland, Auckland, New Zealand and Ericsson Telecom, Stockholm, we have started to analyse a generalised model equation, which allows for a more accurate description of the pulse dynamics. This new model takes into account the full behaviour of the (assumed Lorentzian) spectral gain profile as well as the nonlinear saturation of the gain. A detailed investigation of the properties of the concomitant pulse dynamics is currently being pursued using approximate analytical techniques as well as numerical simulations. The analytical and numerical results show that the solitary pulse character of the Ginzburg-Landau solutions is present even when the approximations of parabolic-gain and negligible-gain-saturation are not made, but the properties of the solutions are changed significantly.

2 Basic Theory

The fundamental equation, which is obtained by generalising the NLS equation to include (perturbatively) the effects of the bandwidth limited gain and gain saturation is the Ginzburg-Landau equation, which reads

$$i\frac{\partial\psi}{\partial x}+\beta\frac{\partial^2\psi}{\partial\tau^2}+\kappa|\psi|^2\psi=i(\gamma-\alpha)\psi+ \\ +i\mu\frac{\partial^2\psi}{\partial\tau^2}-i\nu|\psi|^2\psi\equiv R[\psi], \qquad (1)$$

where ψ is the slowly varying envelope of the optical field, x is the distance of propagation, τ is the reduced time, β characterises the group velocity dispersion and κ the nonlinear Kerr coefficient of the fibre, and α and γ denote fibre loss and linear amplifier gain respectively. Finally, the terms proportional to ν and μ model nonlinear saturation and bandwidth limitation of the gain respectively. The modelling of the latter effect corresponds to expanding the spectral gain characteristics parabolically around the point of

maximum gain (γ). We emphasise that the left hand side of Eq. (1) corresponds to the classical nonlinear Schrödinger and the right hand side introduces a simple modelling of the loss and gain characteristics of the amplifier.

It is interesting to note that Eq. (1) allows exact stationary solutions of the form

$$\psi = A\,\mathrm{sech}^{1+ib}(a\tau)\exp(icx), \tag{2}$$

where the parameters A, a, b and c can be expressed in terms of the parameters of the Ginzburg-Landau equation. For simplicity we give these relations for the case when the bandwidth limitation is the dominating mechanism:

$$\begin{aligned} A^2 &= \frac{3b}{\kappa\mu}a^2(\beta^2+\mu^2) \\ a^2 &= \frac{\gamma-\alpha}{\mu-b\beta} \\ c &= a^2\left[-\beta+(3\beta^2+2\mu^2)\frac{b}{\mu}\right], \end{aligned} \tag{3}$$

where the remaining parameter b is given by

$$b = -\frac{3\beta}{2\mu}+\sqrt{\left(\frac{3\beta}{2\mu}\right)^2+2}. \tag{4}$$

We note that since κ and μ are positive, the condition on the amplitude A^2 implies that also b must be positive. Obviously, in the case when also β is positive (anomalous dispersion), and in the limit when $\gamma = \alpha$ and μ and $\nu \to 0$, the solution reduces to the ordinary soliton solution $\psi = A\,\mathrm{sech}(a\tau)\exp(icx)$, where $A^2 = 2\beta a^2/\kappa$ and $c = \kappa A^2/2$. The Ginzburg-Landau solution involves two important differences as compared to the ordinary soliton solution; it contains no free parameters, the pulse characteristics is completely determined by the parameters of the equation, and the solution contains a time dependent frequency or chirp, ω_c, defined as $\omega_c \equiv -\partial\arg(\psi)/\partial\tau$, where $\arg(\psi)$ denotes the phase of the complex amplitude ψ. In fact, for the Ginzburg-Landau soliton we have

$$\omega_c = \mu a\,\tanh(a\tau). \tag{5}$$

This result implies that the instantaneous frequency varies through the pulse, a new feature which plays an important role in creating a stationary solution.

3 Parabolic approximation of bandwidth limitation and fully nonlinear gain saturation

A generalisation of the previous situation is to consider cases where the parabolic approximation of the spectral gain curve is sufficient, but where the nonlinear gain saturation cannot be described in the perturbative way used in Eq. (1). Such a model equation has been suggested in Ref. [1] and is given by

$$\begin{aligned} i\frac{\partial\psi}{\partial x}+\beta\frac{\partial^2\psi}{\partial\tau^2}+\kappa|\psi|^2\psi = \\ = i\left(\frac{\gamma_0}{1+W/W_{\mathrm{sat}}}-\alpha\right)\psi+i\mu\frac{\partial^2\psi}{\partial\tau^2}, \end{aligned} \tag{6}$$

where

$$W = \int_{-\infty}^{+\infty}|\psi|^2\,d\tau \tag{7}$$

and W_s denotes a characteristic pulse saturation energy. It is interesting to note that an exact solution in the form of the Ginzburg-Landau soliton still exist, i.e., Eqs. (2) and (3) still hold, with only a small change; the gain parameter γ is replaced by

$$\gamma \to \frac{\gamma_0}{1+W/W_{\mathrm{sat}}}, \tag{8}$$

where γ_0 is the linear unsaturated gain. Consider e.g. the situation when the parabolic gain coefficient μ is much smaller than the dispersion coefficient β. In this case the soliton relations can be simplified to read

$$\begin{aligned} b &\approx \frac{2\mu}{3\beta} \ll 1 \\ A^2 &= \frac{2\beta}{\kappa}a^2 \\ a^2 &\approx \frac{3(\gamma-\alpha)}{2\mu} \end{aligned} \tag{9}$$

This means that the stationary solution has the same amplitude-width relation as the ordinary soliton, but the pulse width is determined as a balance between the nonconservative effects of bandwidth limitation and gain only. Furthermore, using the fact that the soliton energy is given by $W = 2A^2/a$, we infer that the net gain, γ_{net} can be expressed as

$$\gamma_{\mathrm{net}} \equiv \gamma-\alpha \approx \frac{\gamma_0}{1+\frac{4\beta}{\kappa W_s}\sqrt{\frac{3\gamma_{\mathrm{net}}}{2\mu}}}. \tag{10}$$

This implies that for small γ_0, the net gain increases linearly with γ_0, whereas for large γ_0, it increases more slowly, in fact, proportionally to $\gamma_0^{2/3}$. Finally we emphasise the fact that for increasing gain, the pulse width decreases, the pulse spectrum becomes broader and ultimately the basis for the Ginzburg-Landau equation becomes violated.

4 Approximate solutions by direct variational methods

A powerful approach for approximate analysis of nonlinear evolution equations like Eq. (1) is direct variational methods based on trial functions and subsequent Rayleigh-Ritz optimisation, c.f. e.g. Ref. [2].

The variational approach is based on the fact that the NLS equation can be expressed as the variational problem $\delta \iint L\, dx d\tau = 0$, where

$$L = \frac{i}{2}\left(\psi^* \frac{\partial \psi}{\partial x} - \psi \frac{\partial \psi^*}{\partial x}\right) + \beta \left|\frac{\partial \psi}{\partial \tau}\right|^2 - \frac{\kappa}{2}|\psi|^4. \quad (11)$$

Approximate solutions of the NLS equation can be found by expressing the expected solution for $\psi(\tau, x)$ in terms of a flexible and physically reasonable trial function, $\psi_T(\tau; p_1(x), p_2(x), \ldots, p_N(x))$, with a prescribed dependence on τ, but involving N unknown parameter functions p_j, $j = 1, 2, \ldots, N$, which depend on the distance of propagation, x. The trial function is inserted into the variational integral, which can then be integrated over τ to give a reduced variational problem; $\delta < L >= 0$, where $< L > \equiv \int_{-\infty}^{\infty} L(\psi_T)\, d\tau$. This variational problem only involves the ordinary functions $p_j(x)$ and is a much simpler problem than the original one. This method has been used successfully (both analytically and numerically) in a number of investigations of nonlinear pulse propagation, see [2]. It can also be applied to generalisations of the NLS equation, written in the form

$$i\frac{\partial \psi}{\partial x} + \beta \frac{\partial^2 \psi}{\partial \tau^2} + \kappa |\psi|^2 \psi = R[\psi], \quad (12)$$

where in the present case the operator $R[\psi]$ is given by

$$R[\psi] \equiv i(\gamma - \alpha)\psi + i\mu \frac{\partial^2 \psi}{\partial \tau^2} - i\nu |\psi|^2 \psi. \quad (13)$$

In this case the variational equations for the parameter functions become

$$\frac{\delta < L >}{\delta p_j} = 2\,\mathrm{Re} \int_{-\infty}^{\infty} R[\psi_T] \frac{\partial \psi_T^*}{\partial p_j}\, dx. \quad (14)$$

The integration of the functionals and the subsequent optimisation may in principle be done analytically or semi-analytically. However, it is obvious that a completely analytical analysis requires a careful choice of trial functions and a not too complicated geometry, whereas a semi-analytical or a completely numerical analysis makes it possible to investigate more complicated problems. We emphasise that if a trial function of the form

$$\psi_T = A\,\mathrm{sech}^{1+ib}(a\tau)\exp(icx) \quad (15)$$

is chosen for the variational procedure, the Ginzburg-Landau soliton solution is recovered.

5 Lorentzian spectral gain profile

A few attempts have been made to use an approach, which avoids the parabolic gain approximation and keeps the full information about the spectral characteristics of the linear gain, see e.g. [3]. If the spectral gain profile is assumed to be Lorentzian, the gain term in the NLS equation is replaced by

$$\gamma\psi + i\mu \frac{\partial^2 \psi}{\partial \tau^2} \rightarrow \gamma \int_{-\infty}^{t} e^{-\frac{t-t'}{T_r}} \psi(z, t') \frac{dt'}{T_r}, \quad (16)$$

where T_r is the dipole relaxation time of the two-level atom amplifier system. If T_r is small as compared to the pulse duration, the integral expression for the gain can be partially integrated twice to give

$$\gamma \int_{-\infty}^{t} e^{-\frac{t-t'}{T_r}} \psi(z, t') \frac{dt'}{T_r} \approx \gamma \left(\psi - T_r \frac{\partial \psi}{\partial \tau} + T_r^2 \frac{\partial^2 \psi}{\partial \tau^2}\right). \quad (17)$$

The term proportional to $\partial\psi/\partial\tau$ implies a change of the group velocity and can be transformed away by changing to a new coordinate system with an appropriately chosen group velocity. The remaining equation is then simply the parabolic gain approximation studied previously. Thus, in the more detailed analysis required when investigating short pulse amplification, the full nonlocal response function for the gain process must be considered. Numerical simulations of the complete equation have demonstrated that stationary solutions reminiscent of the Ginzburg-Landau solitons still exist. An example of such a simulation, taken from Ref. [3], is shown in Fig. 1.

In relation to Fig. 1 we note that the chirp variation remains essentially tanh-shaped, but that the amplitude of the chirp increases significantly. The amplitude, which is sech-shaped for small gain, progressively grows more cusp-like when the gain is increased, but the pulse width does not increase significantly.

We are currently trying to develop an approximate analytical description of the properties of the solutions. This approach is based on the direct variational approach outlined in the previous section. Guided by the simulation results, the trial functions are taken to be of the same form as the classical Ginzburg-Landau solutions, but the relations between the parameters will be different.

6 Conclusion

The current situation with respect to the modelling of optical amplifiers for generation of short pulses with high powers have been outlined. The limitations of previous modelling approaches have been pointed out as primarily due to an inadequate description of the effects of bandwidth limitation of the gain (for very short pulses) and nonlinear saturation effects (for high power pulses). Recent numerical simulations give a good picture of the behaviour of the solutions for increasing gain, which ultimately leads to very short

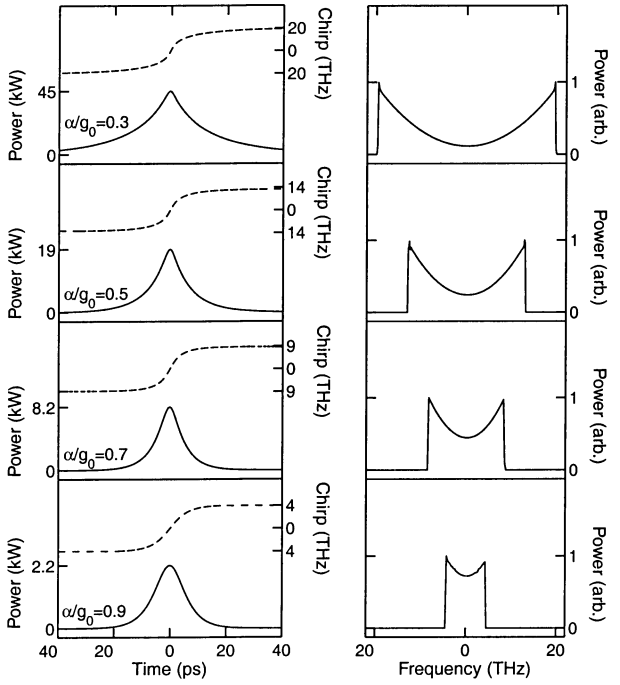


Figure 1: Simulation results from Ref. [3] showing the evolution of the pulse characteristics in the case of normal dispersion and for increasing gain and concomitant broader pulse spectrum. The right hand side figures show the broadening of the pulse spectrum, which is primarily due to the increased chirp.

pulses and a violation of the standard model based on a parabolic approximation of the spectral gain profile. A promising way to obtain approximate analytical or quasi-analytical solutions of the problem has also been discussed. This approach is based on direct variational methods involving trial functions incorporating the main features of the solution and subsequent Ritz optimisation.

References

[1] E. V. Vanin and S. Helmfrid, Proc. of Nonlinear Guided Waves and Their Applications, Dijon, France, USA 1–3 September 1999, ThD1-1, p. 262.

[2] D. Anderson, M. Lisak, A. Berntson, Pramana Vol. 57, 917 (2002)

[3] A. C. Peacock, R. J. Kruhlak, J. D. Harvey, and J. M. Dudley, Optics Communication (to be published).

Suppression of Multipactor Breakdown in RF Equipment

D. Anderson[1], P. Ingvarson[3], U. Jordan[1], U. Jostell[3], A. Kryazhev[2]
G. Li[1], M. Lisak[1], V.E. Semenov[2], R. Udiljak[1]

[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden.
E-mail: elfda@elmagn.chalmers.se.
[2]Institute of Applied Physics, Nizhny Novogorod, Russia
[3]Saab Ericsson Space AB, Göteborg, Sweden

Abstract

The possibility to suppress multipactor by means of amplitude modulation is looked into analytically, numerically as well as experimentally. It is shown that in contrast to the monochromatic case, by using two carriers with appropriate frequency separation and amplitudes, the growth of free electrons can be inhibited. Furthermore, this combination of frequency separation and amplitudes is independent of gap size and frequency range. The results open up new possibilities of avoiding multipactor in devices of complicated geometry in multi-carrier operation.

1 Introduction

Multipactor, or microwave breakdown in vacuum, is a well known problem in modern microwave systems such as e.g. high power microwave generators, rf accelerators, and space-borne rf equipment. The phenomenon was discovered by Farnsworth in the early 1930's, but during the last five decades it has mainly been studied as a problem for microwave systems.

Multipactor breakdown will occur in a microwave device under certain circumstances. The mean free path between gas molecules inside the device must be much greater than the wall spacing, otherwise ordinary gas breakdown may occur. This implicates close to perfect vacuum, which is the normal environment for many satellites in orbit around the earth. The multipactor process starts when a free electron inside the device is accelerated by an electric field of such a strength that the electron will gain enough energy to be able to liberate new electrons upon impact with one of the device walls. If the yield of secondary electrons is greater than unity and the field reverses when the first electron strikes the wall or slightly before, the newly released electrons will be accelerated towards the other wall. If this process is repeated at the other device wall, an avalanche-like growth of electrons will commence, which ultimately leads to multipactor breakdown. The multipactor generates noise and both absorbs and reflects the incident power degrading the system performance. It can also cause physical damage to the part and in some cases the multipactor will give rise to outgassing from the device walls, which could lead to a very harmful gas discharge.

Most studies in the past decades have been focused on monochromatic rf fields and design methods have been established, which successfully deal with single carrier multipactor. However, many modern communication systems are based on multi-carrier operation, i.e. several signals at different frequencies are transmitted simultaneously. In such situations the interference between different channels gives rise to a strong time variation of the total rf signal. This will result in occasional high power peaks when the signals interfere constructively and these peaks are potentially dangerous from a multipactor point of view. On the other hand, multipactor is a resonant process, as described above, and not only does it need high power to initiate but it also requires the power to remain high for a sufficiently long time during which the resonant state must be maintained. Thus a time varying signal power may also destroy the resonance condition which calls for a more detailed analysis of the multipactor breakdown in multi-carrier operation.

A special case between multi-carrier and single carrier operation is the use of two carriers. In a collaboration between Chalmers University of Technology, Institute of Applied Physics and Saab Ericsson Space (SES) a study has been made aimed at assessing the possibility to suppress multipactor in multi-carrier operation. Analytical work has been made on two carriers based on the classical multipactor model supported by extensive numerical simulations. Experiments have then been performed at SES with results that are in line with the theoretical work. The results will serve as a foundation for continued work on the more complicated case of several carriers.

2 Basic multipactor theory

Multipactor breakdown is caused by resonantly driven electrons which collide with the walls inside the rf device and liberate more electrons than prior to the impact. The growth of the electron density, N_e, depends on the secondary emission yield, δ, which is material dependent and a function of the impact energy of the electron. In the case of a single carrier with constant amplitude, the secondary emission yield is constant and provided that δ is greater than unity the growth will be exponential according to

$$N_e = N_0 \delta^q \tag{1}$$

where N_0 is the initial density and q is the number of impacts.

In the simplest case of multipactor, two plane-parallel surfaces are considered. The multipactor starts when a free electron between the surfaces is accelerated by an electric field. If the field is sinusoidal with frequency f, the force felt by the electron is governed by Newton's second law and can be written

$$m\frac{d^2x}{dt^2} = eEsin(\omega t) \tag{2}$$

where m and e are the mass and charge of the electron respectively, E is the amplitude of the electric field, $\omega = 2\pi f$, and t represents time. Integrating (2) once will yield the velocity of the electron and a second integration will result in its position. By setting the initial velocity of the electron to v_0 and the electron's phase in relation to the electric field to ϕ, an expression for the electron velocity as a function of time is obtained:

$$\frac{dx}{dt} = \frac{e}{m\omega}E(cos(\phi) - cos(\omega t)) + v_0 \tag{3}$$

The position can now be determined by setting $x = 0$ at the time of electron emission:

$$x = \frac{eE}{m\omega^2}[(\omega t - \phi)cos(\phi) - sin(\omega t) + sin(\phi)] + \frac{v_0}{\omega}(\omega - \phi) \tag{4}$$

The resonant behaviour of multipactor requires the electron transit time between the walls to be an odd number of half cycles of the rf field, which means the boundary condition can be stated

$$x(t = \frac{n\pi + \phi}{\omega}) = d \tag{5}$$

where d is the gap size and n is an odd positive integer representing the number of half cycles required for the electron to traverse the gap. Using (5) in the expression for the position of the electron (4) yields the required value of the amplitude of the electric field as follows.

$$E = \frac{m\omega(\omega d - n\pi v_0)}{e[n\pi cos(\phi) + 2sin(\phi)]} \tag{6}$$

Normally, the electric field amplitude is set by the device specification and a multipactor will initiate when a free electron has the right phase and velocity. Considering the resonance condition it may appear that such an electron would be very unlikely. However, due to a phenomenon called *phase focusing*, an electron which is slightly out of phase during the beginning of a half cycle, will by the end of the half cycle have reduced its phase error, ϵ, to a smaller value designated ϵ' according to the following relation.[1]

$$\epsilon' = \epsilon\frac{n\pi}{2}tan(\phi) \tag{7}$$

As mentioned before, the secondary emission factor is an important parameter when it comes to the multipactor evolution and it depends on the impact velocity of the electron. Thus it is of interest to know this velocity, which is denoted v_i. The electron will collide with the wall when $x = d$ and from the boundary condition (5) it can be seen that this will occur when $t = \frac{n\pi+\phi}{\omega}$. Using this in the velocity relation (3) the following expression for the impact velocity is obtained.

$$v_i = \frac{2eE}{m\omega}cos(\phi) + v_0 \tag{8}$$

The impact velocity is a parameter of importance as the secondary emission factor is a function of this quantity. Multipactor can only occur as long as δ is greater than unity. A graph showing the typical behaviour of the secondary emission factor as a function of the impact energy is shown in figure 1. The secondary emission is above unity between the first and second cross-over points, E_1 and E_2.

However, multipactor is a resonant phenomenon and other parameters than just the impact velocity will affect whether it can exist or not. By making an assumption on the ratio between the initial velocity and the impact velocity, equations (6) and (8) can be combined into an expression which determines a number of different areas within which multipactor is likely to occur. These are plotted for different materials in so called susceptibility charts and are used by engineers when designing multipactor free microwave devices.[2]

The exponential growth of the multipactor can not go on indefinitely and there must be some kind of

[1] A detailed discussion and derivation of this effect can be found in [2].

[2] In *Diagnostic Investigations into the Multipactor Effect, Susceptibility Zone Measurements and Parameters Affecting a Discharge* by A. Woode and J. Petit (Estec Working Paper No.1556, Nordwijk, Nov. 1989) susceptibility charts for many different materials commonly used in space applications can be found.

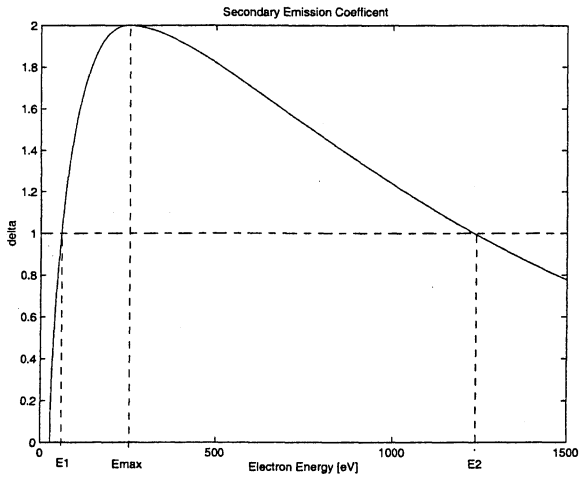


Figure 1: Example showing the dependence of the secondary emission factor on the impact energy of the electron. The horizontal dashed line represents the emission factor above which multipactor is possible.

mechanism of saturation. In most cases the multipactor will reach saturation within a few nanoseconds and the predominant saturation mechanism is often the space charge effect, which causes the electrons to spread to such an extent that the "outermost" electrons will no longer fulfil the resonance condition.

3 Suppression of multipactor breakdown

As multipactor is an unwanted phenomenon, work has been done to find ways of not only avoiding the problem, but also suppressing it. Two main methods can be recognised. The first deals with the dimensions of the device. By increasing the gap size of narrow passages within the device a greater electric field will be required to obtain multipactor, which will increase the power handling capability of the component. The drawback with this approach is that the device may become larger and heavier than wanted, something which is especially disadvantageous in space applications. The other method relies on surface treatments and choice of material. By choosing a material which has a high first cross-over point, the multipactor threshold will increase. In multi-carrier operation it is of value to prolong the time it takes for a multipactor to grow to a level where it significantly affects the input signal and this can be achieved with a material that has a low secondary emission yield.

In multi-carrier operation there are more ways of suppressing multipactor than what is available for single carrier systems. One is mentioned in the previous paragraph and another promising method is to utilise the carriers themselves in an intelligent way. In the following, a method of suppressing multipactor in the case of two carriers by choosing the frequencies and amplitudes properly, will be presented (for a detailed discussion see [1]).

The method is based on destroying the resonance condition of multipactor by modulating the first signal with a second one. The envelope of the total signal will be a slowly varying sine signal, which periodically will drop below the multipactor threshold or in some cases the upper threshold will be exceeded (see fig. 2). Even though the multipactor grows fairly quickly when the resonance condition is fulfilled and the secondary emission ratio is greater than unity, the free electrons will disappear in a much shorter time in the non-resonant state.

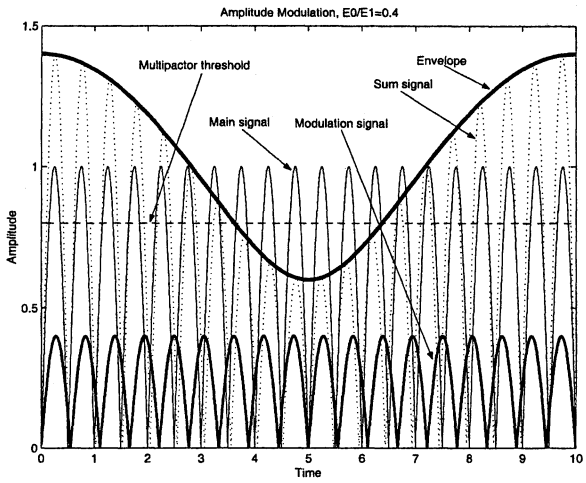


Figure 2: Amplitude modulation using two signals with amplitude ratio $\frac{E_1}{E_0} = 0.4$ and frequency ratio $\frac{\omega_1}{\omega_0} = 0.9$. Note that the main carrier is well above the multipactor threshold, but the envelope, due to interference between the two signals, drops below the threshold for a period of several half cycles.

To suppress multipactor with this method a number of conditions must be fulfilled. First of all, the factor of modulation should be large enough to vary the control parameter $\frac{\omega d}{V_\omega}$, where $V_\omega = \frac{eE}{m\omega}$, to such an extent that it exceeds the resonant bandwidth. Secondly, the modulation should be fast enough not to allow the multipactor to grow disastrously during the time the envelope exceeds the multipactor threshold. Finally, the method relies on the applicability of classical multipactor theory as governed by equations (6) - (8) above.

Classical theory has a shortcoming, namely that it does not take initial velocity spread into account. The usual assumption in classical theory is either $v_0 = 0$ or $v_i/v_0 = constant$. In reality the initial velocities will follow some kind of distribution. The spread will broaden the resonance bands, making multipactor

possible for a wider range of the control parameter. The spread will also result in some of the electrons arriving at the second electrode at a moment when the electric field is such that it will not allow them to return to the first electrode. This will result in a loss of free electrons. If secondary yield is low it will not be able to make up for these losses and the multipactor will be suppressed. The higher the order, n , of resonance and velocity spread, the greater the rate of these losses. For low secondary yield materials, this means multipactor may not be possible at all for higher orders of resonance. It becomes clear that the value of secondary emission yield and the spread of initial velocities will have a significant impact on the occurrence of multipactor.

In simulations based on the XPDP1 code with $\delta_{max} = 1.2$ suppression of multipactor by means of amplitude modulation has been studied. In the simulations, two signals, ω_1 and ω_0, were used and results from one of the sets of runs will be presented below. The ratio between the signal frequencies is $\frac{\omega_1}{\omega_0} = 0.95$. The velocity spread was taken into account by assuming Maxwellian distribution,

$$f(v) \propto e^{-\frac{(v-v_0)^2}{2V_T^2}} \qquad (9)$$

where $v_0 = 0$ and $V_T \approx 0.1$. The ratio between the amplitudes of the signals, E_1/E_0, was varied and the behaviour of the growth of the electron density, N_e, was monitored.

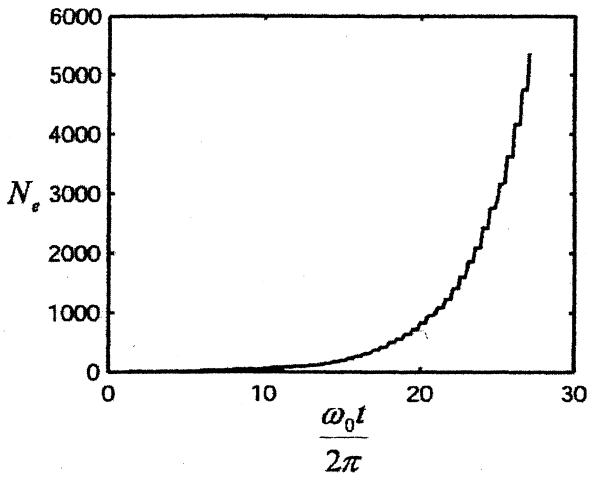


Figure 3: Development of the electron density, N_e, in a monochromatic field. Parameters used in the simulation are $v_0 = 0$, $V_T \approx 0.1$, $\omega_0 = 2\pi \cdot 10^9 s^{-1}$, $\omega_0 d/V_\omega = 3.8$.

When the modulation signal was turned off, the total number of free electrons grew basically exponentially as shown in figure 3. This behaviour is typical for multipactor in a monochromatic field.

Figure 4 shows that the multipactor is regularly suppressed as the envelope of the two signals drops below the multipactor threshold. However, the modulation is not sufficient and eventually the multipactor will grow to the level where saturation sets in and limits further growth as in the monochromatic case.

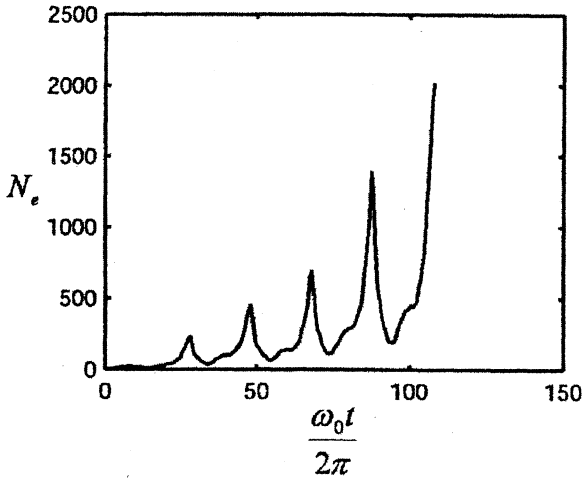


Figure 4: The electron density decreases regularly but the modulation ratio, $E_1/E_0 = 0.3$, is not sufficient to completely suppress the multipactor. The frequency ratio is $\omega_1/\omega_0 = 0.95$. The rest of the parameters as in figure 3.

Finally, when the amplitude of the modulation signal is strong enough it will completely suppress the multipactor (see fig.5). The electron density grows temporarily as the resonance condition is fulfilled, but as soon as it drops below the threshold, the free electrons are lost in just a few half cycles. The variation in the peak levels is due to the velocity spread.

An important fact to keep in mind is that the energy of the electrons at impact is at all times greater than E_1, which means the secondary yield is greater than unity, but due to the electron losses as the resonance condition is violated, the multipactor is suppressed.

The ratio between the signal frequencies in the example is 0.95, but if this is close to unity it is likely that the multipactor will be able to build up an electron density during one period of the envelope, which is great enough to significantly affect the signal to noise ratio. Thus it is of importance to choose a frequency separation, which is sufficiently wide for the material in question. The greater the secondary yield, the wider frequency separation is needed.

One interesting conclusion of this study is that the result is valid regardless of gap width and rf frequency. This is a major advantage, since it suggests that the suppression method will work even if the geometry of the component is very complicated.

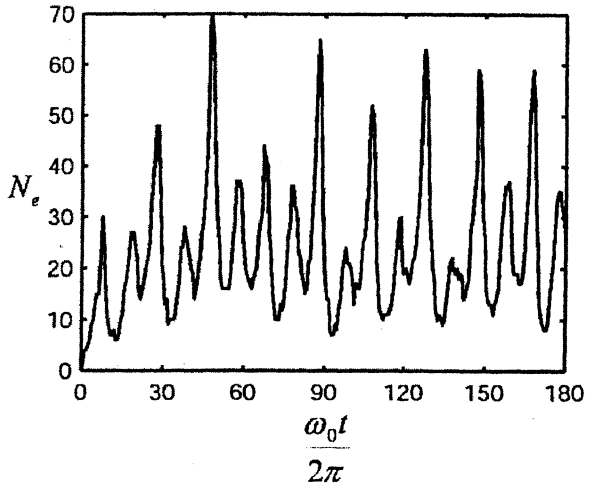


Figure 5: The amplitude ratio is increased to $E_1/E_0 = 0.4$ and now the multipactor is prevented from growing, keeping the electron density low at all times. Parameters as in figure 4

4 Multipactor experiment

In an attempt to experimentally confirm the theoretical results in the previous section, a multipactor test was performed at Saab Ericsson Space in Göteborg [2]. The test sample was a resonant cavity with three ports and it was designed to be resonant at three specific frequencies, namely the frequency of the main signal, the modulation signal, and the close-to-carrier noise (in this case 100 MHz from the main frequency). The material was aluminium, which has a maximum secondary emission ratio around 3.

As mentioned in the previous section, materials with a high secondary emission yield may be able to compensate for the losses incurred during the non-resonant period of the total envelope period. As aluminium has such a high secondary emission it is probable that suppression of multipactor will not work or at least will require a high amplitude ratio. In the tests in question, the maximum achieved amplitude ratio was 0.46 (limited by the power handling capability of available filters) and as it turned out, that was not sufficient to suppress the multipactor[3]. Furthermore, since the total power is increased when using two signals instead of one, the multipactor generated noise will be higher than in the single carrier case. Figure 6 shows a plot from one of the runs.

[3]In recent simulations using a material with δ_{max} around 3 and the same gap size and frequencies as in [2], it was found that suppression of multipactor could be possible with an AM ratio around 40%. However, the exact material parameters of the material used in the experiments are not known and that can explain the unsuccessful suppression. In another simulation with a gap size, which allowed the electron velocity to reach the value necessary for maximum secondary emission, suppression was not possible even for an AM ratio of 90%.

In the beginning only one signal is applied and the upper graph, which shows the close-to-carrier noise, indicates that multipactor is occurring (the noise level when there is no multipactor is around 80 dBm). When the second signal is applied at t = 22 seconds, the noise goes up and the impedance matching goes down, which is another indication of a more severe multipactor.

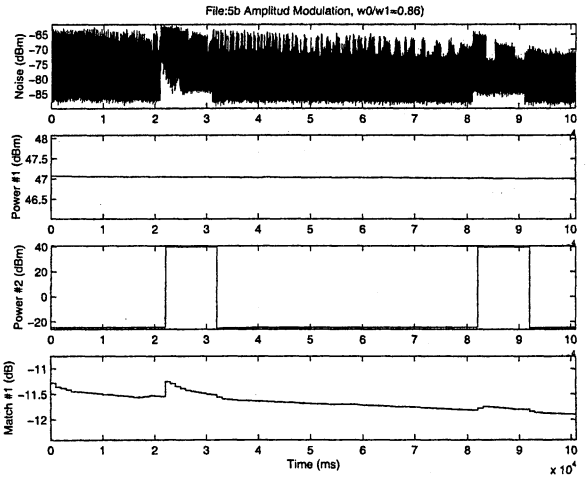


Figure 6: Multipactor experiment with $E_1/E_0 = 0.36$ and $\omega_1/\omega_0 = 1.16$. Due to a high secondary emission yield, multipactor suppression is not possible.

5 Conclusion

Suppression of multipactor by means of amplitude modulation of the main signal is possible provided the secondary emission yield is not too high. The effect on the breakdown conditions by the properties of the secondary emission must be studied more in detail before making any final conclusions regarding the possibility to suppress multipactor in multi-carrier operation.

References

[1] V. Semenov, A. Kryazhev, D. Anderson and M. Lisak, *Multipactor suppression in amplitude modulated radio frequency fields*, Phys. Plasmas, Vol 8, No. 11, November 2001

[2] R. Udiljak and G. Li, *Threshold Level Determination of Multicarrier Multipaction and AM Suppression of Multipaction in Resonant Cavity*, Master's Thesis, Chalmers University of Technology, January 2002

Suppression of nonlinear effects by phase alternation in dispersion-managed optical fibre transmission

P. Johannisson[1], D. Anderson[1], M. Marklund[1]
A. Berntson[2], M. Forzati[2], J. Mårtensson[2]

[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden.
E-mail: pontus.johannisson@elmagn.chalmers.se.
[2]Optical Networks Research Laboratory, Ericsson Telecom AB, SE–126 25 Stockholm, Sweden.

Abstract

The nonlinear effects amplitude jitter and ghost pulse generation, which decrease the signal quality in optical communication systems implementing dispersion management can be suppressed by alternating the phase of the bits. An explanation for this effect is given which shows that by choosing the phase modulations suitably the processes that give rise to the nonlinear effects can be made to counteract each other.

1 Introduction

Dispersion management is considered to be a key technology for upgrading existing communication systems to higher bit rates. Using this technique dispersion-compensating fibre (DCF) is inserted into already installed fibre links consisting of standard fibre (SMF) giving a total system with zero net dispersion. Due to the practical importance of this technique the properties of dispersion-managed (DM) communication systems are by now very thoroughly examined. The large pulse overlaps occurring in long systems with high local dispersion give rise to nonlinear effects via the Kerr nonlinearity in the fibre, a process which perturbatively can be viewed as if every combination of two or three pulses give rise to timing jitter (the signal pulses attract or repel each other), amplitude jitter (the amplitude of the signal pulses changes) or ghost pulses (smaller pulses appearing in bit slots supposed to be empty) [1, 2]. These processes decrease the signal quality making the nonlinear effects a limiting factor. All bit slots will be affected by many such nonlinear processes driven by different combinations of signal pulses. In the present work we show that the nonlinear effects can be reduced by applying a phase shift, $\Delta\phi$, to every second bit slot, see Fig. 1. This affects the phase in the nonlinear processes, and the total nonlinear effect can be minimised by a suitable choice of $\Delta\phi$. It is also found that the dependence on $\Delta\phi$ is periodic with a period π and symmetric about any multiple of $\pi/2$. A detailed investigation, based on a perturbation analysis, is given of the physical background to this and the analysis is supported by numerical simulations.

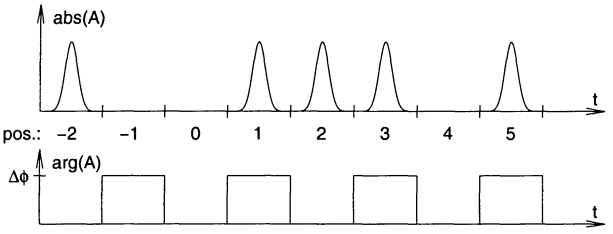


Figure 1: Initial amplitude and phase modulation of a short bit stream using the suggested phase modulation. The signal pulses are chirp-free Gaussian pulses.

2 Theory

2.1 The perturbation analysis

The nonlinear effects in strongly DM systems have been examined by the means of a perturbation analysis [3, 4, 5]. The analysis is based on the nonlinear Schrödinger equation

$$i\frac{\partial A}{\partial z} = \frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2} - \gamma|A|^2 A - i\frac{\alpha}{2}A, \qquad (1)$$

which describes the propagation of the envelope to the amplitude, $A(z,t)$, and includes the effects from the group-velocity dispersion, β_2, the Kerr nonlinearity, γ and the attenuation in the fibre, α. In a strongly DM system (i.e. a long system with high local dispersion—the typical case for a practical system) the nonlinear effects are weak and the envelope, A, can be split $A = A_l + A_p$, where A_l is the linear solution corresponding to $\gamma = 0$ (which appear perfectly compensated at the receiver) and A_p is the nonlinear perturbation. Inserting this into Eq. (1) we obtain to the

lowest order for the perturbation

$$i\frac{\partial A_p}{\partial z} - \frac{\beta_2}{2}\frac{\partial^2 A_p}{\partial t^2} + i\frac{\alpha}{2}A_p = -\gamma|A_l|^2 A_l. \qquad (2)$$

This is a linear differential equation for A_p where the analytically known A_l acts as a source term. For a pulse train $A_l = \sum_k A_k(z,t) = \sum_k A(z, t - kT_0)$, where $A(z,t)$ is a linearly propagating Gaussian pulse, T_0 is the bit slot and k runs over the bit slots containing signal pulses. Expansion of $|A_l|^2 A_l$ gives source terms of the type $A_r A_s A_t^*$, where $A_r = A(z, t - rT_0)$ etc., and the linearity of Eq. (2) guarantees that A_p is the sum of the individual contributions from these source terms. There are four different types of source terms. The first type, $|A_r|^2 A_r$, gives rise to the familiar effect of self-phase modulation of the signal pulse A_r. Terms of the type $|A_s|^2 A_r$, $r \neq s$, give cross-phase modulation of A_r and cause timing jitter [6]. In the present analysis we consider terms of the types $A_r^2 A_s^*$, $r \neq s$, and $A_r A_s A_t^*$, r, s and t all different. Depending on whether their contribution overlap with a signal pulse or not they cause amplitude jitter and ghost pulses respectively, effects known to degrade DM systems.

2.2 Destructive interference

The contribution from $A_r A_s A_t^*$ will asymptotically end up in bit slot number $r + s - t$ [4]. Thus it can be predicted which source terms that contribute to the perturbation in a specified bit slot. We choose to study bit slot number zero which, for the time being, we suppose to be without phase shift. The condition that makes a contribution appear there is $r+s-t = 0$. Phase shifting of the signal pulses is equivalent to multiplying the different source terms by complex constants. By assuming that all odd bit slots are phase shifted $\Delta\phi$ the contributions acquire the phase shifts in Tab. 1 relative to their phases when $\Delta\phi = 0$. All contributions in the zeroth bit slot are shifted either 0 or $2\Delta\phi$, which explains why the dependence on $\Delta\phi$ is periodic with a period π. It also means that we can sum the contributions into two phasors. The first is not affected by a phase shift and the second is rotated $2\Delta\phi$, see Fig. 2a and 2b. By optimising $\Delta\phi$ destructive interference can be obtained and the nonlinear effects can be minimised.

2.3 The symmetry

The amounts of amplitude jitter and ghost pulse generation are symmetric about any multiple of $\pi/2$, i.e., the phase shifts $n\pi/2 \pm \Delta\phi$, where n is an arbitrary integer, give the same results. To explain this the phase dependence of the contributions in a phase shifted bit slot is examined. By assuming that all even bit slots are phase shifted the results listed in Tab. 2 are found.

r	s	$t = r + s$	φ
even	even	even	0
even	odd	odd	0
odd	even	odd	0
odd	odd	even	$2\Delta\phi$

Table 1: The phase shift, φ, of the contribution to the perturbation from a specific source term depends on whether r, s and t are odd or even. It is assumed that all odd bit slots are phase shifted.

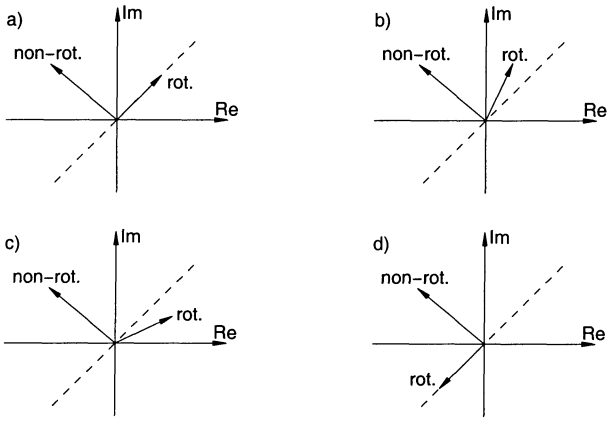


Figure 2: a) The rotating and non-rotating phasors when $\Delta\phi = 0$. b) A phase shift $\Delta\phi$ gives a rotation $2\Delta\phi$. c) The situation in a phase shifted bit slot. The direction of rotation is changed. d) $\Delta\phi = \pi/2$. All bit slots are identical.

The contributions unaffected by a phase shift in Tab. 1 are now shifted $\Delta\phi$. However, the phase shift relative to the signal in the even bit slot is still zero. The rotating phasor is shifted $-\Delta\phi$, or $-2\Delta\phi$ relative to the signal, Fig. 2c. Thus the rotating phasor rotates in the opposite direction as compared to a non-phase shifted bit slot. When $\Delta\phi$ is a multiple of $\pi/2$ the rotating phasors in phase shifted and non-phase shifted bit slots are identical. The two cases $\Delta\phi = 0$ and $\Delta\phi = \pi/2$ are seen in Fig. 2a and Fig. 2d. Starting from a multiple of $\pi/2$ and adding a phase shift means that the rotating phasors will rotate in different directions in the two bit slots. However, when comparing adding a phase shift to subtracting the same phase shift the only difference is that the situation in the phase shifted bit slot has changed places with that in the non-phase shifted bit slot, and this will not change the average amplitude jitter or ghost power.

It should however be noticed that the rotating and non-rotating phasors are not identical in all bit slots. They depend on which contributions that give rise to them, and this in turn depends on where the signal pulses are located around the bit slot. The above symmetry argument can be applied to all pairs of bit

r	s	$t = r + s$	φ
even	even	even	$\Delta\phi$
even	odd	odd	$\Delta\phi$
odd	even	odd	$\Delta\phi$
odd	odd	even	$-\Delta\phi$

Table 2: The phase shift, φ, of the contributions in bit slot zero if all even bit slots are phase shifted.

slots, one phase shifted and one non-phase shifted, interacting with identical surrounding bit patterns. The range of the interaction is approximately equal to the maximum pulse width and this determines how many bit slots we have to account for. It will always be possible to group the bit slots into pairs provided that the sequence is long enough. With large pulse breathing the interaction range is longer and a longer random bit sequence is needed to find matching pairs.

3 Numerical results

In order to illustrate these results simulations have been carried out for a system consisting of 80 km SMF followed by 16 km of DCF. The SMF has dispersion $D = 16$ ps/nm/km, nonlinear parameter $\gamma = 2$ W^{-1}km^{-1} and attenuation $\alpha = 0.25$ dB/km, and the DCF has $D = -80$ ps/nm/km, $\gamma = 8$ W^{-1}km^{-1} and $\alpha = 0.6$ dB/km. The power input to the first fibre is 1 mW and after the fibres noise free amplifiers are inserted, which compensate the attenuation in the two fibres respectively. The bit slot is 25 ps (40 Gbit/s), and the pulse width is 7 ps FWHM. Third-order dispersion has been neglected.

A common measure of the transmission performance is the Q-value defined as

$$Q = \frac{I_1 - I_0}{\sigma_1 + \sigma_0}, \tag{3}$$

where $I_{1,0}$ are the mean powers and $\sigma_{1,0}$ are the power standard deviations for the ones and zeros respectively. With small nonlinear effects the $\Delta\phi$ dependence of Q is determined by the denominator, $\sigma_1 + \sigma_0$, since $I_1 \gg I_0$ and $I_1(\Delta\phi)$ is essentially constant. However these standard deviations are of different order of magnitude. In the zeros the power of the ghost pulses $I_0 = |A_l + A_p|^2 = |A_p|^2$ and in the ones $I_1 = |A_l + A_p|^2 \approx |A_l|^2 + 2\,\mathrm{Re}(A_l A_p^*)$. Since A_p is a small perturbation it can be expected that $\sigma_1 \gg \sigma_0$.

Plots of I_0 and σ_1 can be seen in Fig. 3 and 4. Although insignificant for Q in this case I_0 is still interesting for two reasons. The mean power of the ghost pulses, I_0, can be suppressed by approximately 40% by choosing $\Delta\phi \approx \pi/2$ and the shape of $I_0(\Delta\phi)$ can easily be predicted. By assuming that we can speak about an averaged identical rotating and non-rotating phasor in every bit slot, A_r and A_n respectively, and that the numbers of phase shifted and non-phase shifted bit slots are equal the mean power of the ghost pulses is expected to be

$$\begin{aligned} I_0(\Delta\phi) &= \frac{1}{2}\left(|A_r e^{2i\Delta\phi} + A_n|^2 + |A_r e^{-2i\Delta\phi} + A_n|^2\right) \\ &= |A_r|^2 + |A_n|^2 + 2\,\mathrm{Re}(A_r A_n^*)\cos 2\Delta\phi. \end{aligned} \tag{4}$$

The calculated $I_0(\Delta\phi)$ has been plotted together with a fitted curve $k_1 + k_2 \cos 2\Delta\phi$ in Fig. 3.

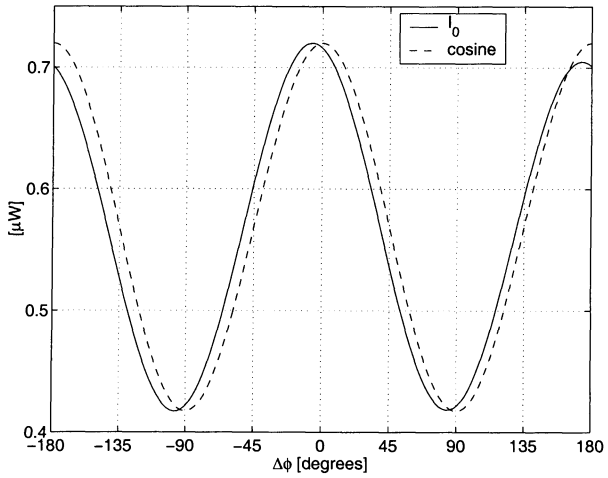


Figure 3: The solid line is the mean power of the zeros, I_0, as a function of the phase shift, $\Delta\phi$. The dashed line is a fitted curve $k_1 + k_2 \cos 2\Delta\phi$.

A plot of σ_1 can be seen in Fig. 4, and like Fig. 3 it shows the predicted symmetry around $\Delta\phi = 0$ and $\Delta\phi = \pm\pi/2$. The curve $\sigma_1(\Delta\phi)$ looks similar to a cosine, but it is significantly harder to say something about the curve shape in this case. It has been found that it depends strongly on the system parameters and the bit sequence used, although the symmetries are preserved for long random sequences.

It should be noticed that all the symmetries require the random pulse train to be long enough. The slight shift to the left of $I_0(\Delta\phi)$ in Fig. 3 is due to this. It is easy to construct a pulse train which do not show these symmetries. A PRBS with 256 bits has been used for these plots, but in general a significantly longer PRBS has to be used in order to see a convergent behaviour.

We have also carried out system simulations checking the dependence on different system parameters. This has confirmed the results stated here and shown that an increase of the Q-value is possible [7].

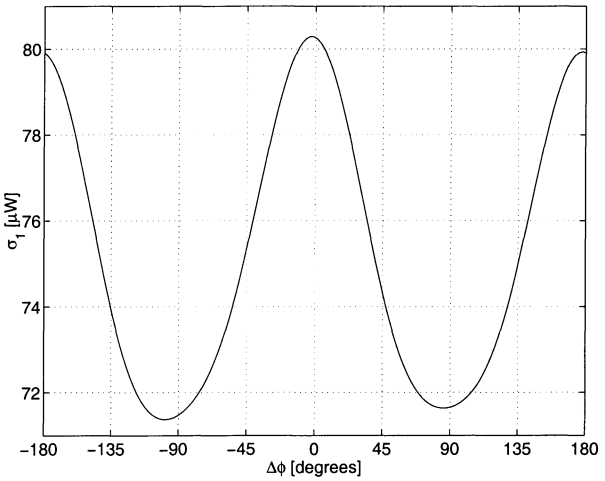


Figure 4: The standard deviation of the ones, σ_1, as a function of the phase shift, $\Delta\phi$.

4 Conclusions

In conclusion the phase alternation technique suggested has been proved to be able to significantly reduce the nonlinear effects in transmission links. This makes it interesting for practical purposes. The physical reason for the improvement of the signal quality is that different contributions to the nonlinear effects can be made to counteract each other. It has been clarified why $Q(\Delta\phi)$ is symmetric around points where the phase modulation is a multiple of $\pi/2$ and why the period is π. It would be interesting to follow up the analysis and the numerics with an experimental investigation.

References

[1] P. V. Mamyshev and N. A. Mamysheva, *Pulse-overlapped dispersion-managed data transmission and intrachannel four-wave mixing*, Opt. Lett. **24**, 1454 (1999).

[2] R.-J. Essiambre, B. Mikkelsen, and G. Raybon, *Intra-channel cross-phase modulation and four-wave mixing in high-speed TDM system*, Electron. Lett. **35**, 1576 (1999).

[3] A. Mecozzi, C. B. Clausen, and M. Shtaif, *Analysis of Intrachannel Nonlinear Effects in Highly Dispersed Optical Pulse Transmission*, IEEE Photon. Technol. Lett. **12**, 392 (2000).

[4] M. J. Ablowitz, and T. Hirooka, *Resonant nonlinear intrachannel interactions in strongly dispersion-managed transmission systems*, Opt. Lett. **25**, 1750 (2000).

[5] P. Johannisson et. al., *Generation and dynamics of ghost pulses in strongly dispersion-managed fiber-optic communication systems*, Opt. Lett. **26**, 1227 (2001).

[6] J. Mårtensson et. al., *Timing jitter owing to intra-channel pulse interactions in dispersion-managed transmission systems*, Opt. Lett. **26**, 55 (2001).

[7] M. Forzati et. al., *Reduction of intra-channel four-wave mixing using the alternate-phase RZ modulation format*, submitted to Photon. Technol. Lett.

Microwave breakdown thresholds in resonators and filters; the concept of effective diffusion length

D. Anderson[1], U. Jordan[1], L. Lapierre[2], M. Lisak[1], T. Olsson[3], J. Puech[2], V. Semenov[4], and J. Sombrin[2]

[1]Dept. of Electromagnetics, Chalmers University of Technology, SE–412 96 Göteborg, Sweden. E-mail: elfda@elmagn.chalmers.se.

[2]Centre Nationale d'Études Spatiales, Toulouse, France.

[3]Allgon Systems AB, Täby, Sweden.

[4]Institute of Applied Physics, Nizhny Novgorod, Russia.

Abstract

An investigation is made of microwave breakdown in gas-filled waveguide structures. Special attention is given to determining the breakdown threshold in physical conditions where the threshold is set by electron losses due to diffusion to the walls of the device. It is shown that the diffusion losses can be characterized by an effective diffusion length, which is a characteristic of the geometry of the device. Several different methods for determining the effective diffusion length are discussed including approximate analytical techniques like direct variational methods, fully numerical methods and direct experimental measurements. Some examples of the agreement between analytical and numerical predictions for the breakdown threshold and experimental results are shown.

1 Introduction

Microwave breakdown in gases under various physical and technical conditions is a familiar and much studied problem. However, the technical development of microwave devices continues to give rise to new situations and parameter regimes where microwave breakdown plays an important role and where previous results and predictions concerning breakdown are not directly applicable. One example illustrating the situation is provided by the trend to design microwave devices as small and compact as possible. This development leads to concern about the concomitant breakdown strength of the design, in particular in situations involving more complicated geometries as e.g. in microwave waveguides, resonators and filters.

Microwave breakdown in an RF device is due to an avalanche-like growth in time of the free electron density in the gas filling the device. This increase is caused by impact ionization of neutral gas molecules by free electrons that are being accelerated to high energies by the microwave field. In competition with the ionization are several processes leading to loss of free electrons. At small and moderate gas pressures, the dominating loss mechanism is diffusion of free electrons to the walls of the device. Microwave breakdown occurs when the ionization frequency is larger than the loss frequency due to diffusion. Whereas the ionization frequency depends on the microwave electric field strength, the loss frequency due to diffusion depends strongly on the geometry of the device. To characterize the loss frequency, the concept of an effective diffusion length, L_{eff}, is often introduced. Although qualitatively L_{eff} can be estimated to be of the order of half the smallest dimension of the waveguide, resonator or filter, such simple estimates are not enough in more complicated geometries. Resort must then be taken to approximate analytical methods or extensive numerical calculations in order to determine the microwave breakdown threshold.

A further complication in many microwave devices is that the geometrical configuration also may give rise to strongly inhomogeneous microwave fields. This inhomogeneity affects the effective diffusion length (and the diffusion losses), which consequently becomes a function also of the characteristic scale length of the field inhomogeneity.

In a collaboration between Chalmers University of Technology, Allgon Systems AB, Institute of Applied Physics and Centre National d'Études Spatiales, work is currently carried out in order to find simplified approximate expressions for the effective diffusion length in a number of geometrical situations, characteristic of microwave waveguides, resonators and filters. The analysis is semi-analytical, based on direct variational methods, but is bench marked against full numerical solutions of the breakdown problem and also compared with experimental results.

The results provide explicit predictions for microwave breakdown thresholds for several different device geometries. They can also be used for simple testing of the breakdown strength of a device by extrapolating data obtained for the moderate pressures around the minimum of the Paschen curve (which are easily obtained experimentally) to lower or higher pressures (which are cumbersome and consequently more costly to perform).

2 Basic theory of microwave breakdown

Microwave breakdown is caused by impact ionization of the neutral gas molecules by free electrons which are being accelerated to high energies by the microwave field. The evolution of the free electron density, $n(\mathbf{x},t)$, is governed by the continuity equation, which can be taken as, cf Ref. [1]

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n) + \nu_i n - \nu_a n \qquad (1)$$

where D is the diffusion coefficient, ν_i is the ionization frequency and ν_a is the attachment frequency.

A heuristic way of deriving the breakdown threshold condition is to introduce the effective diffusion length, L_{eff} , as

$$\nabla \cdot (D\nabla n) \approx -\frac{D}{L_{\text{eff}}^2} n \equiv -\nu_D n \qquad (2)$$

where ν_D is the loss frequency due to diffusion. This simplified analysis makes it possible to rewrite eq. (1) as

$$\frac{\partial n}{\partial t} = \nu_{\text{net}} n, \qquad (3)$$

where $\nu_{\text{net}} = \nu_i - \nu_D - \nu_a$. Clearly exponential growth of the electron density occurs when $\nu_{\text{net}} \geq 0$, which defines the CW breakdown threshold as $\nu_{\text{net}} = 0$.

All three characteristic frequencies contributing to ν_{net} depend on pressure, p, as well as on microwave electric field strength, E. In fact, ν_a is directly proportional to pressure, whereas ν_D varies inversely with p, due to the inverse dependence of the diffusion constant, D on p. Furthermore, they depend only weakly on the electric field strength and is usually taken as constants with respect to the electric field. The ionization frequency, ν_i, has a more complicated dependence on p and its dependence on electric field is often modeled as a power dependence on the effective electric field, E_e, defined as $E_e^2 = E_{rms}^2/(1+\omega^2/\nu_c^2)$, where ν_c is the electron-neutral collision frequency, ω is the angular frequency of the electric field, and E_{rms} is the rms electric field strength. Summing up, the characteristic frequencies can be modelled as $\nu_a = \alpha_a p$, $\nu_D = \alpha_D/(pL_{\text{eff}}^2)$, $\nu_c = \alpha_c p$, $\nu_i = \alpha_i p(E_e/p)^\beta$,

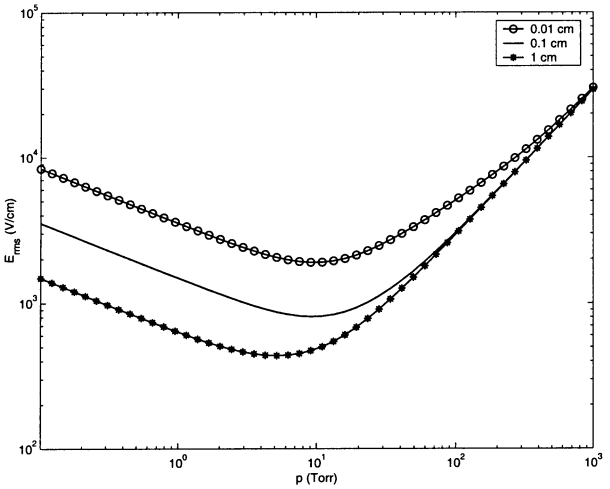


Figure 1: Breakdown threshold electric field vs pressure at $f = 10$ GHz for different effective diffusion lengths.

where the parameters α_j and β are numerical coefficients which depend on the gas. For breakdown in air, good approximations of the coefficients are $\alpha_a \simeq 6 \cdot 10^4$ ($\text{s}^{-1}\text{Torr}^{-1}$), $\alpha_c \simeq 5 \cdot 10^9$ ($\text{s}^{-1}\text{Torr}^{-1}$), $\alpha_D \simeq 10^6$ (cm^2s^{-1} Torr), $\alpha_i \simeq 9 \cdot 10^{-4}$ where p is in Torr and E_{rms} is in V/cm. Finally $\beta \simeq 5.33$. It is clear from the scaling of the characteristic frequencies with pressure that for large pressures, the breakdown threshold is set by attachment loss whereas for low pressures, diffusion loss is the dominant loss mechanism. We also emphasize the pressure dependence of the effective field, which is another important pressure scaling. Together, these dependencies give the characteristic Paschen curve describing the breakdown electric field as a function of pressure, cf Fig. 1.

3 The effective diffusion length in homogeneous electric fields

The weak point in the simplified picture presented in section 2 is that there is no easy way to predict the crucial parameter, L_{eff}, which determines the diffusion losses. Qualitatively we expect L_{eff} to be of the order of half the shortest diffusion distance in the system, i.e. essentially by half the geometrical dimension of the device. This means that L_{eff} can be estimated in geometrically simple situations e.g. in (i) a one-dimensional wave guide of height a, $L_{\text{eff}} \simeq a/2$, in (ii) a two dimensional wave guide of dimensions a and b, $L_{\text{eff}} \simeq \min(a/2, b/2)$, in (iii) a cylindrical waveguide of radius R, $L_{\text{eff}} \simeq R/2$, and in (iv) a coaxial cylindrical wave guide with radii a and R, $L_{\text{eff}} \simeq (R-a)/2$, provided that the electric field is considered as homogeneous.

However, such crude estimates are generally not

enough in order to determine L_{eff} and an accurate determination involves solving the eigenvalue problem obtained from eq. (1) in the limit case when $\partial n/\partial t = 0$, which reads

$$\nabla^2 n + \frac{\nu_i - \nu_a}{D} n = 0, \qquad n(\partial V) = 0 \tag{4}$$

where ∂V denotes the boundary of the region V occupied by the microwave electric field. Obviously introducing $(\nu_i - \nu_a)/D = \lambda$, eq. (4) becomes a classical eigenvalue problem, which can easily be solved exactly for the cases discussed qualitatively above. If we now instead define L_{eff} rigorously from the breakdown condition $\nu_{\text{net}} = 0$ as

$$L_{\text{eff}}^{-2} \equiv \frac{\nu_i - \nu_a}{D} = \lambda \tag{5}$$

the effective diffusion length will be determined by the spatial mode structure of the electron density in the volume. For the different cases considered above this yields the exact results: (i) $L_{\text{eff}} = a/\pi$, (ii) $L_{\text{eff}}^{-2} = \pi^2(a^{-2} + b^{-2})$, (iii) $L_{\text{eff}} = R/j_0$ where $j_0 \approx 2.4$ is the first zero of the Bessel function $J_0(x)$, and (iv) L_{eff} is here determined by a transcendental equation involving Bessel functions. The solution in this case is again qualitatively in agreement with the physical estimate given above. In particular, in the limit when $(R-a)/R \ll 1$ we regain the one dimensional solution with $L_{\text{eff}} = (R-a)/\pi$. Thus, the agreement between the heuristic approach and the exact solutions is qualitatively good, but obviously not enough in order to make quantitative predictions.

4 Effective diffusion length in inhomogeneous electric fields

A further complication in determining the effective diffusion length is due to the fact that the electric field has an inhomogeneous mode structure determined separately by the electrodynamic equations for the microwave field. Consequently, the ionization frequency appearing in eq. (1) is, in realistic situations, inhomogeneous i.e. $\nu_i = \nu_i(\mathbf{x})$. This generalization complicates the detailed analysis of the breakdown problem, however in many geometries it does not imply a conceptually different situation as compared to the homogeneous case. The full diffusion equation can now be written

$$\frac{\partial n}{\partial t} = \nabla \cdot (D \nabla n) + (\nu_i(\mathbf{x}) - \nu_a) n \tag{6}$$

We consider a physical situation corresponding to a symmetric geometry involving a microwave field with a maximum in the centre of the device (at $\mathbf{x} = \mathbf{0}$) and vanishes together with the electron density on the walls. With a slight generalization of the previous heuristic approach we can again introduce an effective diffusion length by approximating eq. (6) as

$$\frac{\partial n}{\partial t} \approx -\frac{Dn}{L_{\text{eff}}^2} + (\max \nu_i - \nu_a) n \tag{7}$$

where $\max \nu_i$ denotes the maximum of the ionization frequency over the volume occupied by the microwave field. The breakdown threshold condition clearly becomes

$$\nu_{\text{net}} \equiv \max \nu_i - \nu_D - \nu_a = 0 \tag{8}$$

where ν_D still is defined by eq. (2).

However, the characteristic diffusion length, L_{eff}, is now even more difficult to determine accurately than in the homogeneous case, since it depends on the spatial mode structure of the electric field (which is given) in addition to the dependence on the spatial mode structure of the electron density. The rigorous analysis for determining the breakdown threshold involves solving the following new eigenvalue problem

$$\nabla^2 n + \lambda s(\mathbf{x}) n - q n = 0, \qquad n(\partial V) = 0 \tag{9}$$

where $\lambda = \max \nu_i(\mathbf{x})/D$, $s(\mathbf{x}) = \nu_i(x)/\max \nu_i(\mathbf{x})$, $q = \nu_a/D$. It is obvious from eq. (9) that the eigenvalue will now also be affected by the field inhomogeneity through the normalized ionization form factor $s(\mathbf{x})$.

Nevertheless, having found the eigenvalue, λ, by solving eq. (9), the breakdown condition, eq. (8) can be used, a posteriori, to infer the proper effective diffusion length through the relation

$$L_{\text{eff}}^{-2} \equiv -\frac{\nabla^2 n(0)}{n(0)} = \lambda - q \tag{10}$$

Since in most inhomogeneous situations $\max \nu_i \gg \nu_a$, we obtain $L_{\text{eff}}^{-2} \approx \lambda$. However, the inhomogeneous eigenvalue problem given by eq. (9) is difficult to solve analytically for geometries of practical interest and resort must then be taken to approximate analytical techniques or to numerical means. This is particularly so in situations where the electric field is strong close to the boundaries of the device where the electron density goes to zero. Examples of this will be given in section 7.

5 Approximate solutions by direct variational methods

A powerful approach for approximate analysis of the eigenvalue problem given by eq. (9) is direct variational methods based on trial functions and subsequent Raleigh-Ritz optimization. The variational approach is based on the fact that the characteristic

eigenvalue λ can be expressed as the variational problem $\delta\lambda[n] = 0$, where

$$\lambda = \frac{\int_V [(\nabla n)^2 + q(\mathbf{x})n^2]dV}{\int_V s(\mathbf{x})n^2 dV} \tag{11}$$

and V is the volume occupied by the field.

Approximate solutions of the eigenvalue problem are found by expressing the expected solution for $n(\mathbf{x})$ in terms of a flexible and physically reasonable trial function, $n_T(\mathbf{x}; p_1, p_2 ... p_N)$ with a prescribed dependence on $\mathbf{x}$, but involving N unknown parameters p_j, j=1,2...N. The trial function is inserted into the variational integral, which can then be integrated over $\mathbf{x}$ to give a reduced variational problem, which is simply an optimization problem involving an algebraic function of the parameters p_j. This method has been used successfully (both analytically and numerically) in many previous studies of breakdown problems, cf [2] and references therein.

The integration of the functionals and the subsequent optimization may in principle be done analytically, semi-analytically or numerically. However, it is obvious that a completely analytical analysis requires a careful choice of trial functions and a not too complicated geometry, whereas a semi-analytical or a completely numerical analysis makes it possible to investigate more complicated problems.

6 Variational approach to breakdown in cylindrical cavities

As an illustration of the variational approach, we consider diffusion-controlled ($q = 0$) breakdown in a circular cylindrical resonator of radius R with the microwave field excited in the TM_{010} mode where the electric field can be written as $E = E_0 J_0(j_0 r)$, where $J_0(x)$ is the Bessel function of order 0, j_0 denotes the first zero of $J_0(x)$, and r has been normalized with respect to the radius R of the resonator. This seemingly simple problem cannot be solved exactly analytically. In order to apply the variational approach it is convenient to use a simpler approximation for the ionization profile. It can be shown, cf Ref.[2] that a good approximation for the Bessel function $J_0(j_0 r)$ is $J_0(j_0 r) \simeq (1 - r^2)^m$ where $m = (1 + \sqrt{2})/2 \simeq 1.21$. With this simplification it is obvious that a natural trial function to use in the variational procedure is $n(r) = (1 - r^2)^\alpha$ where the parameter to be optimized is the exponent α. The reduced eigenvalue functional becomes simply

$$\lambda = \frac{2\alpha(2\alpha + m\beta + 1)}{2\alpha - 1} \tag{12}$$

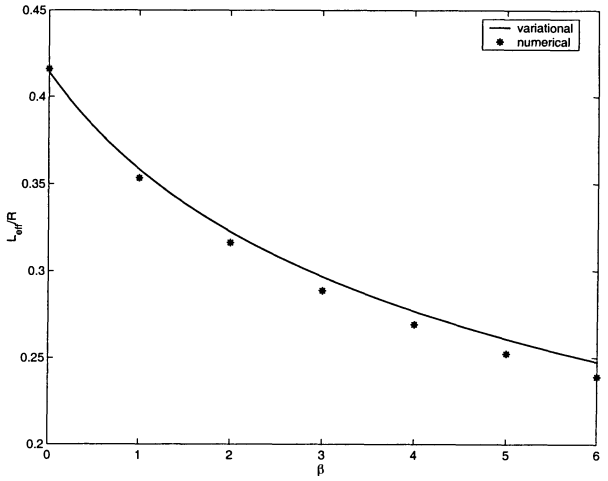


Figure 2: Normalized effective diffusion length L_{eff}/R vs ionization exponent β for the TM_{010} mode of a long cylindrical resonator at low pressure.

Optimizing this expression with respect to the free parameter α, we find $\alpha = (1 + \sqrt{2 + m\beta})/2$ and $\lambda = (1 + \sqrt{2 + m\beta})^2$. Finally we note that this implies that the effective diffusion length, L_{eff} , can be written as

$$L_{\text{eff}} = \sqrt{\frac{D}{\nu_i(0)}} = \frac{R}{\sqrt{\lambda}} = \frac{R}{1 + \sqrt{2 + m\beta}} \tag{13}$$

As expected, the effective diffusion length depends on the geometrical size of the wave guide (R), but also on the inhomogeneous mode structure of the microwave field (m) and on the degree of nonlinearity in the ionization frequency (β). In Fig. 2 we demonstrate the excellent agreement between the variational predictions and the numerical results for the effective diffusion length.

7 Numerical and experimental approaches to breakdown calculations and the effective diffusion length

It is clear that analytical and semi-analytical approaches, based e.g. on the variational approach, provides good physical insight into the characteristic physics involved in the breakdown process and gives accurate explicit results in several important situations. Nevertheless, they cannot provide accurate predictive answers in situations involving complicated geometries, in which case resort must be taken to numerical computations. For determination of breakdown thresholds, the numerical procedures typically proceeds in two steps. In the first step, the microwave field structure is calculated in the device, without taking into account the breakdown plasma. This microwave field is

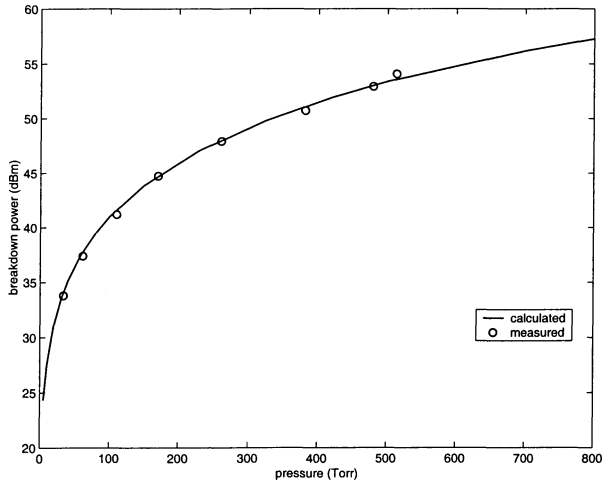


Figure 3: CW-breakdown power threshold vs pressure for a coaxial $\lambda/4$-resonator. The solid line indicates the numerically calculated threshold and the circles indicate measurement points.

then used to determine the ionization function $s(\mathbf{x})$ in the electron density equation. In the second numerical step, the corresponding eigenvalue problem is solved numerically. This procedure provides very accurate predictions for the breakdown thresholds and can be used to determine the effective diffusion length, a posteriori, by matching the results to a Paschen curve with appropriate effective diffusion length. An example of a comparison between numerically obtained predictions and experimental results for the breakdown threshold is shown in Fig. 3. For more details see Ref. [2].

Correspondingly, the effective diffusion length can also be obtained by experimentally determining the Paschen curve and using this result to find the proper value of L_{eff}, as suggested in Refs. [3, 4].

8 Conclusion

The present investigation has emphasized the importance of the effective diffusion length as a convenient measure of the diffusion losses in microwave breakdown studies. This quantity is a characteristic of the geometry of the RF device and can be found by solving the breakdown equation, being the stationary form of the continuity equation for the electron density. In simple geometries involving the constant electric field assumption, the effective diffusion length can be determined exactly. When the inhomogeneity of the microwave electric field is taken into account, few exact solutions are available, but approximate solutions can be found for a number of technically important situations. In more complicated geometries, numerical computations must be used to establish the breakdown threshold and the concomitant effective diffusion length. Finally, it can also be inferred from experiments by finding the Paschen curve that closest matches the experimental results. The effective diffusion constant makes it possible to accurately describe the breakdown characteristics of an RF device and is a convenient parameter to use e.g. for reliable extrapolations of low pressure tests to high pressure regions, cf [2].

References

[1] A. D. MacDonald Microwave Breakdown in Gases (New York; Wiley) 1966.

[2] D. Anderson, U. Jordan, M. Lisak, T. Olsson, M. Åhlander IEEE MTT **47**, 2547 (1999).

[3] J. Puech, U. Gutierrez, Proc. "Workshop on Multipactor, RF and DC Corona and Passive Intermodulation in Space RF Hardware", WPP-178, 25, ESTEC, Noordwijk, The Netherlands, 4–6 September 2000.

[4] A. D. Rawlins Proc. "Workshop on Multipactor, RF and DC Corona and Passive Intermodulation in Space RF Hardware", WPP-178, 31, ESTEC, Noordwijk, The Netherlands, 4–6 September 2000.

Ahlen, Anders 36 709
Alfredsson, Mattias 123
Aljasmi, Rahman 249
Anderson, D 291 754 758 761 765 770 774
Anderson, John B 46 47 52
Andersson, Henrik 16
Andersson, Michael 392
Andersson, Sören 6
Andersson, Stefan 133
Andersson, Tomas 487
Asplund, Henrik 273
Astély, David 553 592
Aulin, Tor 338 517 606 714 719
Auvert, Marine 657
Axelsson, Sune 254 269 278 506 640
Backenius, Erik 197
Barta, József 553
Beach, Mark 311 548 635
Belitsky, Victor 82
Bengtsson, Mats 311 436 724
Berg, Jan-Erik 273
Bergljung, Christian 1 11 635
Berntson, A 770
Biro, Thomas 396 400
Bodin, Ulf 167
Bondeson, Anders 410
Booth, Roy 87
Boyd, Stephen 533
Bria, Aurelian 415 426
Browne, David 273
Brun Madsen, Ole 652
Brunström, Anna 709
Brändström, Urban 109
Brännström, Fredrik 719
Bäckström, M 291 295 300
Börjesson, Per Ola 501 630
Caire, Giuseppe 57
Caputa, Peter 133
Carlsson Jan 452
Carlsson, Charlie 452
Carlsson, Jonas 128
Carlsson, Patrik 357
Carr-Motycková, Lenka 511
Christoffersson, Jan 362

Cook, Greg 558
Dahlén, Anders 538
Dam, Hai Huyen 601 625
Dam, Mads 644
Dán, György 676
Danielsson, Andreas 700
Derneryd, Anders 577
Dhaou, Imed Ben 72
Dudley, J 761
Edfors, Ove 57 141 153
Edman, Fredrik 324
Edvardsson, Olle 467
Eklund, Jan-Erik 491
Ekman, Jonas 457
Ekman, Torbjörn 36 234
Elbornsson, Jonas 491
Eneroth, Peter 635
Engquist, Björn 405
Ericsson, Nilo Casimiro 229
Eriksson, Kurt 300
Eriksson, Thomas 26
Eriksson, Tomas 52
Ernström, Per 538
Escudero-Pascual, Alberto 527
Ewerlid, Anna 229
Falahati, Sorour 234
Falk, Johan 264
Falk, Lars 172
Fedele, R 758
Ferrer, I 291
Fiedler, Markus 357
Firozabadi, Babak S 644
Flament, Maxime 415 431
Fletcher, P.N 548
Florén, Fredrik 153
Fodor, Viktória 676
Foo, Sze Ern 635
Forkman, Peter 471
Forsgren, Stefan 362
Forslund, Ola 563
Forzati, M 770
Fransson, Johan E.S. 244
Fransson, Pierre 163
Frölind, Per-Olov 244
Furuskär, Anders 31

Färnqvist, Daniel 345
Geijer Lundin, Erik 41
Gessler, Fredrik 415
Giese, Jochen 616
Glazunov, Andres Alayon 734
Grant, Alex 719
Grönkvist, Jimmi 667
Gunnarsson, Fredrik 41 352
Gunnarsson, Frida 352
Gustafsson, Jonas 380
Gustafsson, Oscar 192 473
Gustafsson, Fredrik 41 352 491
Gustafsson, Mats 392
Gustavsson, Anders 244
Gustavsson, Andreas 123
Gustavsson, Björn 106 109
Hafezi, P 11
Hagdahl, Stefan 405
Hagerman, Bo 6 553 592
Hall, B 758
Hamberg, Lennart 696
Hamnerius, Yngve 446 692
Hannu, Hans 362
Hansson Mild, Kjell 703
Hansson, Anders 338
Haridi, Seif 224
Harvey, J 761
Havelka, Dragan 224
Hedelin, Per 333
Hedin, Mikael 101 106
Helczynski, L 758
Hespanha, Pedro João 345
Himdi, Mohamed 316
Hjalmarson, Emil 182 187
Hjalmarsson, Håkan 657
Hjalmarson, Åke 77
Holm, Maria 587
Holmberg, Torgny 148
Howard, J. 700
Huo, Jiaquan 621
Hyberg, Per 596
Häggblad, Jan-Olof 305
Hägglund, Robert 182 187 197
Häggström, Ingemar 97 101 106
Händel, Peter 264 487

Höijer, Magnus 295
Ingelström, Pär 410
Ingvarson, P 765
Janches, Diego 97
Jantsch, Axel 178
Jansson, Carl Gustaf 207
Jansson, Magnus 264 596
Jenvey, Stewart 457
Johannisson, Pontus 761 770
Johansson, Albin 501
Johansson, Anders J 705
Johansson, Erika 681
Johansson, Håkan 496
Johansson, Karl Henrik 345 657
Johansson, Mathias 729
Johansson, Mikael 533
Jonsson, Andreas 163
Jordan, Ulf 291 765 774
Jostell, U 765
Jöngren, George 328
Kamenetsky, Max 420
Karlsson, Anders 705
Karlsson, Gunnar 522 543 644 657 662 672
Karlsson, Johan M 141 143 148 158
Karlsson, Peter 1 11 311 548 635
Karlsson, Robert 16
Kildal, Per-Simon, 446 452
Kim, A 291
Kjellgren, Jan 249
Kopecký, Rudolf 692
Kristensen, F 11
Kristensson, Gerhard 577
Krulahk, R 761
Kryazhev, A 765
Kurup, Dhanesh G 316
Kämpe, Andreas 118
Lahne, Lars Magnus 350
Landström, Daniel 630
Lantz, Martin 66
Lapierre, L 774
Larsson, Per 16
Larzon, Lars-Åke 167
Lassing, Johan 26
Lau, Buon Kiong 6 558
Leanderson, Carl Fredrik 57

Leung, Yee Hong 6 558
Leyser, Thomas 109
Li, Fengyi 375
Li, G 765
Li, Weidong 128 482
Lin, Zihuai 606
Lindblom, Jonas 333
Lindh, Thomas 647
Lisak, M 291 758 761 765 774
Lorén, Jörgen 295
Lundbäck, Jonas 572
Lundén, O 291
Lundgren, Urban 457
Lundin, Henrik 487
Lundqvist, Henrik 543
Lämmermann, Sven 216
Lövehagen, Nina 696
Löwenborg, Per 182
Lövstrand, Karl-Gunnar 300
Magnusson, Håkan 118
Malmqvist, Robert 123
Malmsten, Fredrik 739
Marcus, Carina 568
Marklund, M 761 770
Marsh, Ian 375
Más, Ignacio 662

Mathews, John D 97
Mattisson, Sven 66
Mattsson, Lennart 563
McNamara, Darren 311 548
Medbo, Jonas 273
Meisel, David D 97
Mielczarek, Bartosz 239
Mohammed, Abbas 572
Morogan, Matei Ciobanu 212
Mäkikaltio, Per 462
Nelander, Anders 259
Nellström, H. 692
Nilsson, Arne A 342 350 357
Nilsson, Börje 396 400
Nilsson, Olle 114
Nilsson, Peter 62
Njölstad, Tormod 128
Nordberg, Jörgen 601 625

Nordebo, Sven	572				
Nordholm, Sven	601	621	625		
Nordström, Barbro	300				
Norgren, Martin	384				
Novak, Mirek	52				
Nykvist, Johan	511				
Nylund, Mikael	137				
Ohlsson, Henrik	478				
Olsson, Björn	300				
Olsson, Bo	582				
Olsson, T	774				
Olsson, Thomas	62				
Ossipov, Evgueni	672				
Ottersten, Björn	310	311	328	436	596
Ottosson, Tony	709	739	749		
Ouacha, Aziz	123				
Palmkvist, Kent	128				
Parkvall, Stefan	31				
Paulsson, Lars-Erik	689				
Peacock, A	761				
Pellinen-Wannberg, Asta	97				
Persson, Anders	749				
Persson, Henrik	158				
Persson, Kristin	392				
Persson M	692	700			
Persson, Magnus	31				
Petersson, Niklas	501	630			
Petersson, Sven	553				
Pettersson, Lars	568				
Porath, Jan-Erik	714				
Puech, J	774				
Queseth, Olav	415				
Radasky, W.A	283				
Rantakokko, Jouni	611				
Rasmussen, Lars	517	719	744		
Risacher, Christophe	82				
Rosengren, Kent	452				
Rydberg, Anders	316				
Rylander, Thomas	410	577			
Samuelsson, Maria	31				
Sandgren M	11				
Sandström, Monica	703				
Schelén, Olov	167				
Schiöler, Henrik	652				
Semenov, VE	758	765	774		

Sergienko, Tima 109
Siegbahn, Martin 696
Skoglund, Mikael 328 487 616
Smith, Gary 244
Sombrin, J 774
Son, Than Le 652
Stadler, Rolf 644
Stake, Jan 410
Steen, Åke 109
Sternad, Mikael 36 202 234 709
Sterner, Ulf 681
Strandemar, Katrin 345
Stranne, André 153
Stridh, Rickard 415 436
Ström, Erik 739 749
Svanbro, Krister 362
Svantesson, Thomas 21
Svensson, Arne 234 239 431 709
Svensson, Christer 133
Svensson, Linnéa 496
Svensson, Tommy 26
Tan, Peng Hui 744
Tapio, Mikael 319
Tenhunen, Hannu 72
Teo, Kok Lay 621
Thibblin, Ulf 380
Thidé, Bo 78
Thorén, Karl 47
Tronarp, Otto 611
Törnevik, Christer 686 696
Törnqvist, Mattias 273
Udiljak, R 765
Uhlemann, Elisabeth 517
Ulander, Lars M.H. 244
Unbehaun, Matthias 415 420
Wahlgren, Bo 288
Wallace, Jon 21
Waller, Peter 577
Walter, Fredrik 244
Wanhammar, Lars 128 182 187 192 473 478 482
Vanin, E 761
Wannberg, Gudmund 97
Vassilev, Vessen 82
Velayos, Héctor 522
Werle, Patrik 207

Vesterbacka, Mark 197
Westman, Assar 97
Wiberg, Per-Arne 517
Widenberg, Björn 387
Wik, Manuel W 283
Wikström, Patrik 16
Wilén, Jonna 703
Winnberg, Anders 471
Wisten, Åke 462
Wosinska, Lena 367
Wu, Jiang 415 441
Wänstedt, Stefan 362
Xiao, Lin 533
Yiu, Ka Fai Cedric 621
Yu, Kai 311
Zander, Jens 415
Zander, Roland 143
Zetterberg, Rolf 176
Zepernick, Hans-Jürgen 601
Zhuang, Shengxian 128 482
Ödling, Per 501 630
Öwall, Viktor 46 47 324