lstván Frigyes János Bitó Péter Bakki *Editors*

Advances in Mobile and Wireless Communications

Views of the 16th IST Mobile and Wireless Communication Summit



Volume 16

István Frigyes · János Bitó · Péter Bakki (Eds.)

Advances in Mobile and Wireless Communications

Views of the 16th IST Mobile and Wireless Communication Summit



Frigyes István Budapest Department of Technology and Economics Dept. of Broadband Infocommunication and Goldmann Gy. tér 3 Budapest 1111 Hungary frigyes@mht.bme.hu

János Bitó Budapest Department of Technology and Economics Dept. of Broadband Infocommunication and Goldmann Gy. tér 3 Budapest 1111 Hungary bito@mht.bme.hu Péter Bakki Budapest Department of Technology and Economics Dept. of Broadband Infocommunication and Goldmann Gy. tér 3 Budapest 1111 Hungary peter.bakki@mht.bme.hu

ISBN: 978-3-540-79040-2

e-ISBN: 978-3-540-79041-9

Library of Congress Control Number: 2008926495

© 2008 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: eStudio Calamar S.L.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

All are but parts of one stupendous whole, Whose body Nature is, and God the soul; -Alexander Pope

Contents

Co	ontrib	outors	XVII
Pr	eface		XXIII
In	trodu	iction	. XXV
Pa	rt I I	Physical	1
1		erleaving Strategies for Multidimensional	2
	Con	Icatenated Zigzag Codes	
		Zigzag Codes	
	1.2	1.2.1 Encoding of Zigzag Codes	
	1.3	Multiple Interleavers	
	1.5	1.3.1 Problem	
		1.3.2 Congruential Interleavers	
		1.3.3 Cyclic Shifted Multiple Interleavers	
		1.3.4 UMTS Based Interleaver	
		1.3.5 Interleavers for Zigzag Codes	
	1.4	Simulation Results	
	1.5	Conclusions	19
	Refe	erences	19
2	Sim	plified Channel-Aware Greedy Scheduling and	
		enna Selection Algorithms for Multiuser MIMO	
		tems Employing Orthogonal Space Division Multiplexing	23
		Introduction	
	2.2	System Model	26
		2.2.1 Block Diagonalization (BD)	27
		2.2.2 Successive Optimization (SO)	28
	2.3	· · · · ·	
		2.3.1 Joint User and Antenna Selection for Block	
		Diagonalization (BD)	30

		2.3.2 Joint User and Antenna Selection for Successive	24
		Optimization (SO) 2.3.3 Subspace Correlation Based User Grouping	
	2.4	Impact of Receive Antenna Selection (RAS)	
	2.4	2.4.1 Receive Antenna Selection (RAS)	
		2.4.1 Receive Antenna Selection Algorithm 1 (RAS-I)	
	2.5	Simulation Results	
	2.3	2.5.1 Correlation Threshold	
		2.5.2 Average Sum Rate Performance	
	2.6	Conclusions	
	Ack	nowledgements	
		erences	
3		the Impact of Channel Estimation and Quantization	=0
		ors for Various Spatio-Temporal Transmission Schemes	
	3.1 3.2	Introduction Single Link MIMO Transmission	
	3.2	3.2.1 System Model	
		3.2.2 Transmission Schemes	
	3.3	Multi-User Downlink MIMO Transmission	
	5.5	3.3.1 System Model	
		3.3.2 ZF Precoding for Multi-User Downlink Transmission	
	3.4	Impact of Channel Estimation Errors at Receiver	
		3.4.1 Channel Estimation Error Model	
		3.4.2 Simulation Results	
	3.5		
		3.5.1 Random Vector Quantization	
		3.5.2 Simulation Results	64
	3.6	Concluding Remarks and Further Comments	68
	Refe	erences	69
4	On	Multi-Cell Cooperative Signal Processing	
-		Backhaul-Constrained Cellular Systems	71
		Introduction	
		Notation	
	4.3	System Model	
		4.3.1 Optimization Framework	
		4.3.2 Optimization Problem	
	4.4	Optimization Approach	
		4.4.1 User Grouping	
		4.4.2 Virtual MIMO Configuration	

		4.4.3	Beamforming Matrices	
		4.4.4	Power Allocation	
	4.5	A Clo	oser View on Subsystem Partitioning	
	4.6	Simu	lation Results	
	4.7		lusions	
	Refe	erences	s	94
D				05
Pa	rt II	Access	s	97
5			Multi-Bid Auction and Pricing	
	in D	ynam	ic Spectrum Allocation Networks	99
	5.1	Intro	duction	99
		5.1.1	Trading and Liberalization	99
		5.1.2	Rights and Obligations	100
		5.1.3	Interference	100
	5.2	Relat	ed Works	101
	5.3	Spatio	o-temporal DSA Model	102
		5.3.1	Interference and Spectrum Efficiency	102
			Feasible Allocation	
		5.3.3	Interference Tolerance	105
	5.4	Pricir	ng Scheme in the Proposed DSA Model	105
			Inputs	
			Allocation and Pricing Rules	
	5.5		ple	
	5.6	Conc	lusion	111
	Refe	erences	s	112
6	Dag		Allosoficer Studenting for SDMA (OFDMA Suctoring	115
U	6 .1		Allocation Strategies for SDMA/OFDMA Systems duction	
	6.2		m Model	
	6.3	•	e-Frequency/Time Resource Allocation	
	0.5			119
		0.5.1	SDMA Algorithm: Greedy Regularized	120
		(2)	Correlation-Based Algorithm	120
		0.3.2	Frequency/Time Assignment Algorithm:	101
		())	Group-to-Resource Assignment	
	<i>с</i> 1		Resource Allocation Strategy	
	6.4		ency/Time-Space Resource Allocation	
			SDMA Algorithm: Successive Projections Algorithm	n 124
		6.4.2	Frequency/Time Assignment Algorithm:	10-
		<i>.</i>	User-to-Resource Assignment	
	- -		Resource Allocation Strategy	
	6.5	Analy	ysis and Simulation Results	126

		Conclusions 1	
	Refe	erences1	.32
D			
Pa	rt III	Techniques and Technologies1	35
7	Mor	nent-Based Estimation of the Signal-to-Noise	
		o for Oversampled Narrowband Signals1	37
	7.1	Introduction	
	7.2	Equivalent Baseband Model1	
	7.3	Moment-Based SNR Estimation	
	7.4	Computation of the Correlation Index1	40
	7.5	Higher-Order Statistics	
	7.6	Simulation Results	
	7.7	Conclusions	47
	Refe	rences1	47
8		mation of Rain Attenuation Distribution on Terrestrial	
	Mic	rowave Links with General N-State Markov Model1	
	8.1	Introduction	
	8.2	Stationary Examination of Rain Attenuation Process 1	
	8.3	The N-state Markov Model 1	
	8.4	Model Parameterization 1	
	8.5	Applying the Proposed Model for a Designated Link 1	
	8.6	Results	
	8.7	Conclusion	
		nowledgement1	
	Refe	prences1	.61
9		Investigation of the Applicability of Fade Duration	
		kov Model in Attenuation Time Series Synthesis	
for Multipath Fading Channel			
		Introduction	
		Description of the Measured Data	
	9.3	Stationarity Investigations of the Attenuation Process	
		Event Modeling with Two-State Markov Chain	.6/
	9.5	The Attenuation Threshold Dependent Fade Duration Model	68
	9.6	From the Fade Duration Model to the Two-State	
		Fade/Non-fade Model1	71
	9.7	Simulate a Single Fading Event 1	
	9.8	Modeling the Scintillation	
	9.9	Evaluation of the Synthesized Time Series	

	9.10 Summary	178
	Acknowledgment	
	References	
10		
10	Cost-Optimised Active Receive Array Antenna	101
	for Mobile Satellite Terminals	
	10.1 Introduction	
	10.2 System Scenario	
	10.3 Simulation Assessments	
	10.3.1 Environmental Conditions	
	10.3.2 Directional Land Mobile Satellite Channel Model	
	10.3.3 Conformal Array Simulations	
	10.4 Antenna Design	
	10.5 Digital Beamforming	
	10.5.1 Implementation Aspects	
	10.6 Conclusions	
	References	199
11	Scheduling Techniques for Mobile Broadcast	• • • •
	and Multicast Services	
	11.1 Introduction	
	11.2 Problem Analysis	
	11.2.1 Introduction of MBMS	
	11.2.2 Scheduling and Congestion Control in MBMS	
	11.3 Concepts and Algorithms	
	11.3.1 Dynamic MBMS Resource Scheduler	
	11.3.2 Scheduling of Carousel Services	
	11.3.3 Scheduling of Streaming Services	
	11.4 Performance Evaluation	
	11.4.1 Evaluation of Scheduling for Carousel Services	
	11.4.2 Evaluation of Scheduling for Streaming Services	
	11.4.3 Evaluation of Dynamic MBMS Resource Scheduler	r217
	11.5 Conclusions	217
	Acknowledgment	
	References	
Pa	rt IV Networks	221
12	Body Area Network and Its Standardization at	
	IEEE 802.15.BAN	223
	12.1 Introduction	
	12.2 SG-BAN and BAN Definition	
		····· 44T

	12.3 BAN Applications and Usage Models	
	12.3.1 Medical and Healthcare Applications	
	12.3.2 Applications for Assisting Persons with Disabilit	
	12.3.3 Entertainment Applications	231
	12.4 Some Short-Range Technologies and a Prototype BAN.	
	12.4.1 BAN Requirements and Some Short-Range	
	Wireless Technologies	
	12.4.2 A Prototype BAN System	233
	12.5 Issues in Discussion and Future Work	
	12.5.1 Wearable BAN and Implant BAN	235
	12.5.2 Frequency regulations	
	12.6 Conclusion	237
	References	
13	Generic Abstraction of Access Performance and	220
	Resources for Multi-Radio Access Management	
	13.1 Introduction	
	13.2 Multi-Radio Access Architecture	
	13.3 Service Requirements	
	13.3.1 Reliability Requirements of Applications	
	13.3.2 Rate Requirements of Applications	
	13.3.3 Delay Requirements of Applications	
	13.4 Service Specification13.5 Generic Access Performance Abstraction	
	13.6 Generic Access Resource Abstraction	
	13.6.1 Generic Access Resource Metrics	
	13.6.2 Access Resource Structures and Combined	
	Access Resource Metrics	251
	13.7 Access Selection Process	
	13.7.1 Policy-based Access Selection	
	13.7.2 Dynamic Access Selection	
	13.8 Conclusion	
	Acknowledgments	
	References	
14	A Decentralized RAT Selection Algorithm	
	Enabled by IEEE P1900.4	
	14.1 Introduction	
	14.2 RAT Selection Enablers Defined by IEEE P1900.4	
	14.3 Case Study: Interference Reduction through	
	Decentralized RAT selection	
	14.4 Simulation Model	

	14.5 Results in a Single Service Scenario	
	14.6 Results in a Multi-service Scenario	
	14.7 Conclusions	
	Acknowledgement	
	References	
Pa	rt V Applications	
15	Business Models for Local Mobile Services Enabled by	• • • •
	Convergent Online Charging	
	15.1 Introduction	
	15.2 Business Models for Mobile Services	
	15.2.1 The LOMS Role Model	
	15.2.2 Categories of Mobile Services & Different	
	Charge Types	
	15.3 A News Publishing Scenario	
	15.3.1 Charging of Service Usage and Revenue Sharing .	
	15.3.2 Mobile User Outside the Event Area	
	15.3.3 Mobile user inside the event area	
	15.3.4 Requirements for Convergent Online Charging	
	15.4 Extending Charging and Billing Systems	
	15.4.1 Design of the Online Charging Interfaces	
	15.4.2 System Design and Benefits	
	15.5 Conclusion & Outlook	
	References	
16	Rights Management for User Content	
	16.1 Introduction	
	16.2 Background	
	16.3 Devices in the Home Network	
	16.4 Rights Management	
	16.5 Usage Scenarios	
	16.5.1 Commercial Content	
	16.5.2 User Content	
	16.6 Authentication and Encryption	
	16.7 Service Architecture	
	16.8 Future Work	
	16.9 Conclusion	
	References	

Pa	Part VI Systems		
17	Distributed Cross-Layer Approaches for VoIP Rate Control over DVB-S2/RCS	.315	
	17.1 Introduction		
	17.2 System Model	.316	
	17.2.1 Centralized vs. Distributed Approaches		
	17.3 QoS Model		
	17.4 RTCP-Driven Cross-Layer Distributed VoIP Rate Control	.319	
	17.4.1 RTCP Reports	.319	
	17.4.2 Bank of Narrowband Non-Adaptive Codecs	. 320	
	17.5 AMR-WB-Based Cross-Layer VoIP Distributed Rate Control		
	17.5.1 AMR-WB Codec		
	17.5.2 Cross-layer VoIP Rate Control	.321	
	17.6 Delay Budget Model and Performance Model	. 322	
	17.7 Numerical Results	. 324	
	17.8 Conclusion	. 329	
	References	. 329	
18	Optical Satellite Downlinks to Optical Ground Stations and		
	High-Altitude Platforms		
	18.1 Introduction		
	18.2 Solving the Challenge of Cloud-Blockage		
	18.3 System Comparison		
	18.3.1 Earth Observation Scenario		
	18.3.2 State of the Art RF Downlink		
	18.3.3 Proposed RF Downlink		
	18.3.4 Proposed Optical Downlink		
	18.3.5 Proposed Combined RF-Optical Downlink		
	18.3.6 Proposed GEO Relay		
	18.3.7 Proposed HAP Relay		
	18.4 Comparison of Downlink Scenarios		
	18.5 Cloud Cover Statistics and OGS-Diversity		
	18.6 Availability of OGS-Networks		
	18.6.1 OGS Network within Germany		
	18.6.2 OGS Network within Europe		
	18.6.3 World Wide OGS Network		
	18.7 Wavelength Selection and Terminal Architecture		
	18.8 Conclusion		
	References	. 349	

19	Wireless Applications in Healthcare and Welfare	351
	19.1 Introduction	351
	19.2 Wireless Hospital Concept	
	19.2.1 Network Topologies	
	19.3 Application Areas	
	19.3.1 Wireless Hospital	
	19.3.2 Wireless Sensors	
	19.3.3 Sport Training	362
	19.3.4 Enterprise Resource Planning System	
	19.4 Conclusions and Future Visions	
	Acknowledgment	
	References	
20	Analytical Analysis of the Performance Overheads of IPsec	
	in MIPv6 Scenarios	365
	20.1 Introduction	365
	20.2 Security Configurations	366
	20.3 Reference Scenario and Network Model	
	20.4 Performance Analysis	
	20.4.1 Calculation Method	
	20.4.2 Input Parameters	
	20.4.3 Results	
	20.5 Concluding Remarks and Future Work	
	Acknowledgments	
	References	
		-
Ind	lex	387

Contributors

Ramón Agüero, M.Sc.

Communications Engineering Department, University of Cantabria, Avda. los Castros s/n, 39005 – Santander, Spain

Ramon Agustí, Prof.

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), c/ Jordi Girona, 1-3, Campus Nord, 08034, Barcelona, Spain

Angeliki Alexiou, Dr.

Bell Labs Research, Alcatel-Lucent, The Quadrant, Stonehill Green, Westlea, Swindon SN5 7DJ, United Kingdom

Nigel Baker

Mobile and Ubiquitous Systems Group, CCCS Research, University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, England

Gerhard Bauch, Dr.

DoCoMo Euro-Labs, Landsberger Strasse 312, 80687 Munich, Germany

János Bitó, Dr.

Budapest University of Technology and Economics, Goldmann Gy. Tér 3., 1111, Budapest, Hungary

Frank Bormann, Dr.

Research & Development, Orga Systems GmbH, Am Hoppenhof 33, 33104 Paderborn, Germany

Anna Brunstrom, Professor.

Department of Computer Science, Karlstad University, Universitetsgatan 2, SE-651 88 Karlstad, Sweden

Robin Chiang

Mobile and Ubiquitous Systems Group, CCCS Research, University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, England Pau Closas, MSc.

Technical University of Catalonia (UPC), Jordi Girona 1-3, Campus Nord, 08034, Barcelona, Spain

László Csurgai-Horváth, MSc.

Budapest University of Technology and Economics, Goldmann Gy. Tér 3., 1111, Budapest, Hungary

Khadija Daoud, Dipl.-Ing.

Orange Labs, 38-40 avenue du général leclerc, 92794, Issy Les Moulineaux, France

Bernhard Epple, Dipl.-Inf.

Institute of Communications and Navigation, German Aerospace Center (DLR), Münchner Straße 20, 82234 Weßling, Germany

Zoltán Faigl, MSc.

Mobile Innovation Center, Budapest University of Technology and Economics, Bertalan Lajos utca 2. H-1111, Budapest, Hungary

Péter Fazekas, MSc.

Department of Telecommunications, Budapest University of Technology and Economics, Magyar Tudósok krt. 2., H-1117, Budapest, Hungary

Juan A. Fernandez-Rubio, Prof. Technical University of Catalonia (UPC), Jordi Girona 1-3, Campus Nord, 08034, Barcelona, Spain

Gerhard Fettweis, Prof. Dr.-Ing.

Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Georg-Schumann-Str. 11, 01187 Dresden, Germany

Stephan Flake, Dr.

Research & Development, Orga Systems GmbH, Am Hoppenhof 33, 33104 Paderborn, Germany

Markus Flohberger, Dipl.-Ing.

Institute of Communication Networks and Satellite Communications, Graz University of Technology, Inffeldgasse 12, 8010 Graz, Austria

István Frigyes, Prof.

Department of Broadband Infocommunications, Budapest University of Technology and Economics, 1111 Budapest, Goldmann György tér 3, Hungary Wilfried Gappmair, Dr.techn.

Institute of Communication Networks and Satellite Communications, Graz University of Technology, Inffeldgasse 12, 8010 Graz, Austria

Jens Gebert, Dipl.-Ing.

Alcatel-Lucent Deutschland AG, Holderaeckerstrasse 35, 70499 Stuttgart, Germany

Dirk Giggenbach, Dr.-Ing.

Institute of Communications and Navigation, German Aerospace Center (DLR), Münchner Straße 20, 82234 Weßling, Germany

Matti Hämäläinen, Dr.Tech.

Centre for Wireless Communications, University of Oulu, Erkki Koiso-Kanttilan katu, 90570 Oulu, Finland

Balázs Héder, MSc.

Budapest University of Technology and Economics, Goldmann Gy. Tér 3., 1111, Budapest, Hungary

Joachim Horwath, Dipl.-Ing.

Institute of Communications and Navigation, German Aerospace Center (DLR), Münchner Straße 20, 82234 Weßling, Germany

Jari Iinatti, Dr.Tech., Professor. Centre for Wireless Communications, University of Oulu, Erkki Koiso-Kanttilan katu, 90570 Oulu, Finland

György Kálmán, MSc. University of Oslo / UniK - University Graduate Center, Instituttveien 25., 2007 Kjeller, Norway

Anja Klein, Prof. Dr.-Ing.

Institute of Telecommunications, Technische Universität Darmstadt, Merckstrasse 25, 64283 Darmstadt, Germany

Michael Knappmeyer

Mobile and Ubiquitous Systems Group, CCCS Research, University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, England

Ryuji Kohno, Ph.D., Professor,

Yokohama National Univerity, 79-7 Tokiwada, Hodogaya-ku, Yokohama, 240-8501, Japan

Otto Koudelka, Univ.-Prof.

Institute of Communication Networks and Satellite Communications, Graz University of Technology, Inffeldgasse 12, 8010 Graz, Austria

Georgios P. Koudouridis, Tekn.Lic. TeliaSonera Corporate R&D, Vitsandsgatan 9, 123 86 Farsta, Sweden

László Kovács, Msc.

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar tudósok krt. 2, Hungary

Witold A. Krzymień, Dr. (Professor);
Electrical & Computer Engineering, University of Alberta, and TRLabs,
2nd Floor ECERF Building, Edmonton, Alberta T6G 2V4, Canada

Katsutoshi Kusume DoCoMo Euro-Labs, Landsberger Strasse 312, 80687 Munich, Germany

Huan-Bang Li, Dr. of Eng., Senior Researcher,National Institute of Information and Communications Technology,3-4 Hikarino-oka, Yokosuka-shi, Kanagawa, 239-0847, Japan

Stefan Lindskog, Associate Professor

Department of Computer Science, Karlstad University, Universitetsgatan 2, SE-651 88 Karlstad, Sweden

Oliver Lücke, Dr.-Ing. TriaGnoSys GmbH, Argelsrieder Feld 22, 82234 Wessling, Germany

Tarcisio F. Maciel, M.Sc.

Institute of Telecommunications, Technische Universität Darmstadt, Merckstrasse 25, 64283 Darmstadt, Germany

Patrick Marsch

Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Georg-Schumann-Str. 11, 01187 Dresden, Germany

Francesco Meago, Ing.

Nokia Siemens Networks SpA, Via Monfalcone 1, 20092 Cinisello Balsamo (MI), Italy

Florian Moll, Dipl.-Ing. (FH)

Institute of Communications and Navigation, German Aerospace Center (DLR), Münchner Straße 20, 82234 Weßling, Germany

Catherine Morlet, Dr.

ESA's Research and Technology Centre (ESTEC), European Space Agency (ESA), Keplerlaan 1, 2201AZ Noordwijk, The Netherlands

Markus Muck, Dr.

Motorola Labs, Parc Les Algorithmes, 91193, Gif-sur-Yvette, France

Jad Nasreddine, Dr.

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), c/ Jordi Girona, 1-3, Campus Nord, 08034, Barcelona, Spain

Josef Noll

University of Oslo / UniK - University Graduate Center, Instituttveien 25., 2007 Kjeller, Norway

Alberto Pellon

Space Services - Componentes Electrónicos, Lda., Taguspark - Nucleo Central, Sala 203, 2780-920 Oeiras., Portugal

Jordi Pérez-Romero, Dr.

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), c/ Jordi Girona, 1-3, Campus Nord, 08034, Barcelona, Spain

Pekka Pirinen, Dr.Tech.

Centre for Wireless Communications, University of Oulu, Erkki Koiso-Kanttilan katu, 90570 Oulu, Finland

David Pradas Fernández, Eng.

Department of Telecommunications and Systems Engineering, Universitat Autònoma de Barcelona (UAB), Campus UAB (ETSE), 08193 Bellaterra (Barcelona), Spain

Mikael Prytz, Ph.D. Ericsson Research, Torshamnsgatan 23, 164 80 Stockholm, Sweden

Teemu Rinta-aho, M.Sc.

Ericsson Research, Hirsalantie 11, 02420 Jorvas, Finland

Joachim Sachs, Dipl.-Ing.

Ericsson Research, Ericsson Allee 1, 52134 Herzogenrath, Germany

Oriol Sallent, Dr.

Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), c/ Jordi Girona, 1-3, Campus Nord, 08034, Barcelona, Spain

Zach Shelby

Sensinode Ltd. (Mr., CTO), Teknologiantie 6, 90570 Oulu, Finland

Shreeram Sigdel, MSc.

Electrical & Computer Engineering, University of Alberta, and TRLabs, 2nd Floor ECERF Building, Edmonton, Alberta T6G 2V4, Canada

Jürgen Tacken, Dr.

Research & Development, Orga Systems GmbH, Am Hoppenhof 33, 33104 Paderborn, Germany

Haitao Tang, Dr.

RTP, Nokia Siemens Networks, Linnoitustie 6, 02600 Espoo, Finland

Ralf Tönjes, Prof. Dr.-Ing.

Faculty of Engineering and Computer Science, University of Applied Sciences Osnabrück, P.O. Box 1940, 49009 Osnabrück, Germany

M. Ángeles Vázquez Castro, Dr.

Department of Telecommunications and Systems Engineering, Universitat Autònoma de Barcelona (UAB), Campus UAB (ETSE), 08193 Bellaterra (Barcelona), Spain

Attila Vidács, Dr.

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar tudosok krt. 2, Hungary

Kai Yu, Dr.

Bell Labs Research, Alcatel-Lucent, The Quadrant, Stonehill Green, Westlea, Swindon SN5 7DJ, United Kingdom

Preface

Lectori Salutem!

This is another book – among the myriads – dealing with wireless communications. The reader might be aware: this topic is really among bestsellers in technology – bestsellers in technology itself and that in technical literature. Communications is one of the leading techniques in information society and mobile/wireless communications is one among the (maybe not more than two with optics the second) leading techniques in communications.

Development of wireless communications was and is really spectacular in the last decade of the 20th and first decade of the 21st century. Such topics as MIMO, wireless networking, security in the technological field, new business models in the service providing field, various applications in the users' side, to mention a few only, were undergoing an unprecedented evolution. So it is not surprising that the number of conferences and the number of books in this field grows and grows, in a nearly unbounded way.

I strongly hope that in spite of this abundance our book yields some valid contribution to this mass. It is a sample (I feel and hope: a rather significant sample) of what was achieved in the last year/last few years in this field in the world. It is also a sample of what topics are felt as important in Information Society Technology by the European Union. And also a sample of what was felt as important by this conference, the 16th IST Mobile and Wireless communication Summit. Our effort was to put together a valid show of communications' art by these samples.

Dear Reader: enjoy this book!

Prof. István Frigyes Budapest University of Technology and Economics

Introduction

I. Frigyes

The conference-series called in recent years IST Mobile and Wireless Communication Summit grew during its 16 years of history to a major conference being very likely the most important in its subject in Europe. The original aim of the Summits was to report on the progress of European Union-sponsored R&D projects; however, they became general open call conferences, with covered subjects much wider than these projects. Its 16th edition having been held 1–5 July 2007 in Budapest Hungary was similar to its predecessors, in subject, in niveau and in the number of participants.

This book is a selection of the topics covered in the conference. Authors of about 15% of the presented papers were invited to contribute to the book, with extended versions of their papers. The result is the present book. (Interestingly enough, the final selection contains, by chance, 10 chapters strictly related to EU projects and 10 being independent from these.)

The book, divided into six parts comprises twenty chapters. The four chapters of Part I deal with problems of the physical layer. Chapter 1 written by Bauch and Kusume investigates high performance (multidimensional) zigzag codes; the main problem here is interleaver design; attempt of optimization is made. Next two chapters deal with MIMO techniques. Chapter 2 (Sigdel and Krzymien) proposes and investigates two novel algorithms for scheduling and antenna selection in Orthogonal SDMA (Space Division Multiple Access) multiuser MIMO downlink. In Chapter 3 Yu and Alexiou, after briefly describing spatial multiplexing and diversity performance of MIMO transmission investigate the effect of various errors. In Chapter 4 (by Marsch and Fettweiss) various aspects and advantages of multi-base-station cooperation are discussed.

Two chapters of Part II deal with access. Chapter 5 written by Kovács and Vidács deals with the resource allocation (RA) problem from a techno-economic point of view. They propose a novel pricing method which enhances tolerance of users toward other users of the network. Chapter 6 (by Maciel and Klein) investigates and compares two approaches of RA in the case of Orthogonal SDMA-FDMA; with

their proposed (sub-optimal) method the targeted maximizing of the weighted capacity is fairly well achieved.

Various techniques and technologies are collected in Part III. Chapter 7 (Gappmair & al.) describes a "blind" SNR estimator being unbiased at intermediate SNR magnitudes; blind means here needing neither clock nor carrier synchronization. Chapter 8 and 9 are about channel characterization. Héder and Bitó in chapter 8 propose an N-state Markov chain for modeling the non-stationary process of rain attenuation; a method for terrestrial link design is based on that model. In Chapter 9 (by Csurgai-Horváth and Bitó) multipath-fade-duration is modeled; based on their model fade attenuation time series for mobile satellite links is designed. Chapter 10 written by Lücke & al. gives a very detailed description of the design of a receiver antenna array for small satellite ground stations: of channel model, requirements, conformal array design, beamforming and implementation. Chapter 11 by Knappmeyer & al. deals with scheduling techniques in mixed broadcast/multicast services.

Three chapters of Part IV are devoted to networks. Chapter 12 written by Li and Kohno, deals with Body Area Networks. This is a somewhat exceptional topic in this book; while belonging strictly to wireless communications I have the feeling that it is rather unknown among average radio engineers. Therefore two chapters of ICT in healthcare are included, both being of rather tutorial character; this chapter is the first among them dealing with network aspects - the second is in Part VI, systems. Chapter 13 and 14 are dealing with heterogeneous networks. That of Sachs & al transfer Application requirements to Radio Resource characteristics in order to make appropriate access possible in a scenario of various applications and various resources. To do that appropriate abstraction of available network characteristics is proposed. Pérez-Romero & al in chapter 14 advocate for a distributed resource management in a heterogeneous network in which decision on the choice of radio access technology is made in the user terminal; a simulated case study shows that compared to central management with this choice overhead needed is significantly less and so throughput is higher.

Part V deals with applications. In Chapter 15 Bormann & al. discuss business models for what is called Local Mobile Services; they discuss the role of each of the more-than-usual players in this type of services and propose principles of billing/pricing. Chapter 16 by Kálmán and Noll deal with content security. While personal communication becomes more and more widespread this issue is of great importance; the chapter discusses two situations – commercial and self-generated content, respectively – and proposes criteria and solutions for sufficient security.

Part VI contains a pot-pourri of various systems. Chapter 17 by Vasquez-Castro & al is about cross-layer optimization of Voice-over-IP in satellite systems; two different systems are investigated for codec-rateadaptation in a Ka band GEO (Geostationary Earth Orbit) satellite link; the aim is to reduce delay and delay jitter; feasibility of the proposed methods is shown. Chapter 18 by Giggenbach & al. deals with a rather special system: possibilities to transmit the huge amount of information collected by optical Earth observation satellites; it is shown that classical microwave transmission is hardly applicable, optical downlink being much more suitable; two systems are investigated, a purely optical and a mixed optical/HAP system. Chapter 19 by Hämäläinen & al is the second dealing with wireless healthcare; thisone gives a descripton of system aspects of the wireless hospital system. Chapter 20 written by Faigl & al develops a queuing-theory-based analysis model of the performance overheads of IPsec, which can be applied in various mobility scenarios; based on that, decision for the best security configuration can be made by specifying the trade-off between security and performance.

As it is usually the case with similar works: there is no unified concept in this book. However, similarly to mosaics: individual building blocks being completely independent consolidate finally to a more unique picture. The picture, in our case, is: constantly advancing wireless communications. We hope that the ensemble of our mosaic building blocks yield a valid contribution to the general picture.

Part I Physical

1 Interleaving Strategies for Multidimensional Concatenated Zigzag Codes

Gerhard Bauch and Katsutoshi Kusume

1.1 Introduction

Zigzag codes have been proposed in [1] and extended in [2,3,4,5,6,7,8]. They are attractive because of their low encoding and decoding complexity and their excellent performance particularly with high code rate. Concatenated zigzag codes with iterative decoding perform only about 0.5 dB worse than parallel concatenated convolutional codes (turbo codes) while having significantly lower complexity. A single zigzag code is a very weak code due to its small minimum Hamming distance of $d_{min}=2$. Strong codes are obtained by concatenation of zigzag codes where each constituent code encodes an interleaved version of the data sequence. Since a zigzag code has usually relatively high rate, multiple zigzag codes have to be concatenated in order to obtain codes with reasonable rate and error correction capabilities.

Building low rate codes by concatenation of several constituent codes is a difference to turbo codes, where usually only two constituent codes are concatenated. This implies a new problem in interleaver design: Not only one interleaver which is optimized for iterative decoding is needed but several mutually independent interleavers are needed. Furthermore, we wish to construct the interleaver permutation rule by a simple equation or by simple permutations from a common mother interleaver in order to minimize memory requirements for storing the interleaver pattern.

We propose new interleavers which take the specific properties of zigzag codes into consideration. We compare the BER performance to straightforward interleaver design approaches such as congruential or s-random interleavers.

1.2 Zigzag Codes

1.2.1 Encoding of Zigzag Codes

The principle of regular zigzag codes is illustrated on the left hand side of Fig. 1.1, where \oplus denotes the modulo 2 sum. The data bits $d_{i,j} \in \{0,1\}$ are arranged in an $I \times J$ matrix. Each row of the matrix is called a segment of the zigzag code. The parity bits p_i are determined as the modulo 2 sum over each segment *i* including the previous parity bit p_{i-I} . The zigzag code is completely described by the two parameters *I* and *J*.

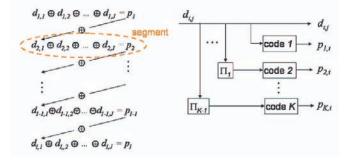


Fig. 1.1 Zigzag code and concatenated zigzag codes

A single zigzag code has weak performance since the minimum Hamming distance is $d_{min}=2$. This is easily verified when two data bits within a segment *i* are flipped. In this case, the parity bit p_i will remain unchanged and consequently no other bits in the code word are effected. A code word with minimum Hamming weight $w_H=d_{min}=2$ occurs, if the data sequence contains only two bits with value $d_{i,j}=1$ which are located within the same segment as depicted on the left hand side of Fig. 1.2. Low weight code words with Hamming weight $w_H=3$ are generated if a single bit with value 1 appears in two consecutive segments.

In order to build a powerful code, several constituent zigzag codes have to be concatenated. Each of the respective constituent encoders encodes an interleaved sequence of the data bits as shown on the right hand side of Fig. 1.1, where Π_k indicates the permutation rule of interleaver *k*.

Concatenated zigzag codes are decoded using an iterative algorithm similar to decoding of turbo codes [1]. However, the decoding complexity in terms of number of operations is about a factor 10 less than that of turbo codes. Since the code rate $R_c=J/(J+1)$ of a zigzag code is usually relatively high, we need to concatenate several zigzag codes. For the concatenated

constituent codes, only the parity bits are transmitted. With *K* constituent codes, the overall code rate becomes $R_c = J/(J+K)$.

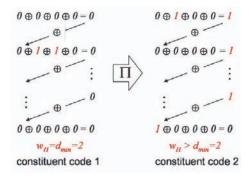


Fig. 1.2 Hamming weight for concatenated zigzag codes

The use of multiple interleavers is an essential difference to turbo codes, where usually only two constituent codes are concatenated. This implies a new problem in interleaver design: Not only one interleaver which is optimized for iterative decoding is needed but several interleavers are needed. Each of those interleavers should provide good performance of iterative decoding while the interleavers should be mutually as independent as possible. It is still an open problem what is a good criterion for mutual independency of multiple interleavers.

One intuitive condition might be that, in order to increase the minimum Hamming distance of the overall code, the interleaving should make sure that bits which are within one segment at the input of a certain constituent encoder are not mapped to the same segment or adjacent segments at the input of any other constituent encoder. The problem is illustrated for an example in Fig. 1.2. Code words with minimum Hamming weight $w_H=d_{min}=2$ at the output of a zigzag constituent code are generated by data sequences, where two bits of value $d_{i,j}=1$ are located within the same segment *i*. In order to increase the minimum Hamming distance of the overall code, the interleaver should make sure that those data sequences produce a code word with higher Hamming weight at the output of the other constituent encoders. Therefore, the two 1-valued bits should be spread as far as possible as demonstrated on the right hand side of Fig. 1.2.

1.3 Multiple Interleavers

1.3.1 Problem

For the concatenation of several zigzag constituent codes, we need K-1 different interleavers. In general, multiple interleavers can be generated randomly. However, apart from the fact that it cannot be guaranteed that interleavers with good mutual properties are generated, a significant practical problem appears in terms of memory requirements since each of the K-1 permutation patterns needs to be stored. Moreover, a system may have to support various block lengths. In this case, we have to store K interleaver patterns for each possible block length. Therefore, we wish to generate multiple interleavers from a simple equation or by simple operations on a common mother interleaver.

The problem of saving memory for multiple interleavers has been addressed in the context of zigzag codes in [7,8]. Here, it is proposed to build a multidimensional zigzag code by arranging the data bits in a cube and performing zigzag encoding in various directions through the cube. However, by doing so, the design space is limited and particularly the parameter J of the constituent zigzag codes may be fixed and differ for the various directions of the cube.

In [8], a proposal is presented which allows to use zigzag codes and the parallel concatenated convolutional codes as specified for UMTS within the same framework. As far as interleaving is concerned, the author proposes to use the interleaver specified for the UMTS turbo code and its transpose as the interleaver for a third concatenated constituent zigzag code. Even though this scheme is very simple, it is limited to two interleavers. It is further suggested in [8] to produce additional interleavers from the UMTS interleaver by swapping of addresses of the interleaver. However, no information is given on how exactly this should be done.

Only a few papers on design of multiple interleavers exist and an appropriate design criterion is still an open problem. Multiple interleavers are considered in [11,12] in the context of multiuser detection in code division multiple access (CDMA) and in [13,14] for interleave division multiple access (IDMA). Multiple interleavers for multidimensional turbo codes are proposed in [15,16,17].

In [12], the authors derive design criteria for random congruential interleavers in order to minimize the impact of multiuser interference under the condition of certain convolutional codes. A disadvantage is that if the code properties are changed, e.g. by puncturing in a rate-adaptive coding scheme, the interleavers have to be changed. Furthermore, a minimum interleaver size is required in order to meet the design criteria. Particularly, for low rate codes this minimum interleaver size becomes prohibitively large.

Another design criterion is discussed in [11], where the authors use the heuristic criterion of minimizing the intersection, i.e. the set of common code words, between the resulting codewords after interleaving. However, it turns out that interleavers which violate this criterion but instead satisfy criteria on individual interleavers for good turbo processing, show better performance. This again stresses the difficulties in finding appropriate design criteria and justifies to rely more on heuristic approaches and evaluation by simulations. Furthermore, except for the case of congruential symbol interleavers in combination with convolutional codes, no construction methods are given in [11]. The heuristic approach taken in [13,14] will be explained in more detail in Sect. 1.3.3.

Interleaver design in the context of multidimensional parallel concatenated convolutional codes is addressed in [16,17]. Here, the fact that data sequences which are divisible by the feedback polynomial of the recursive convolutional constituent codes produce low weight code words is taken into account. The interleaver should permute those divisible patterns to non divisible patterns. In order not to put too many restrictions on one interleaver, the idea in [16,17] is that each interleaver takes care of a subset of critical patterns which have to be broken. This ensures that at least one constituent code contributes weight to the codeword. However, the design criteria are limited to parallel concatenated convolutional codes and depend on the particular choice of the constituent codes. Furthermore, no simple construction method is given which would allow low cost implementation.

The proposal in [15] is limited to parallel concatenation of three convolutional codes and focuses on small block lengths. Here, the goal is to ensure that all constituent codes terminate in the same state.

In the following subsections, we explain and propose several possibilities to generate multiple interleavers. The proposals underlie different requirements. We start with congruential interleavers in Sect. 1.3.2. This is a straightforward approach which allows to construct multiple interleavers for different block lengths from a simple equation. However, congruential interleavers introduce limited randomness which results in suboptimum performance. In Sect. 1.3.3, we propose to generate interleavers from a common mother interleaver by simple operations such as cyclic shifts and self-interleaving. Only the mother interleaver or its construction rule has to be stored. Any good interleaver, e.g. the interleaver which has been specified for turbo codes in UMTS [18], can be used as mother interleaver. A different philosophy underlies the proposal in Sect. 1.3.4. Here we use intermediate steps in the construction of the interleaver specified for turbo codes in UMTS [18] in order to obtain several interleavers. The idea is that building blocks, i.e. the interleaver, which are available in a system such as UMTS can be reused if e.g. zigzag codes are introduced as an optional coding scheme. This would allow to use turbo codes and zigzag codes within the same framework. Both coding schemes could share the building blocks for interleaver construction even though zigzag codes need more interleavers than the turbo code. The price we pay for this low-cost version is that the obtained interleavers may be suboptimum.

Finally, we propose two versions of interleavers in Sect. 1.3.5, which are specifically designed for zigzag codes. The design criterion is to avoid worst case interleaver mappings. Hence, we optimize the interleavers for performance in the error floor region, i.e. for medium to high SNR. The first proposal in Sect. 1.3.5.1 only puts the necessary restrictions but apart from that the interleavers are constructed randomly. This may yield good performance but does not solve the memory problem of storing interleaver patterns. In contrast, the proposal in Sect. 1.3.5.2 gives a more structured construction method which meets the specific requirements of interleaving for zigzag codes. The interleaver is generated using several small subinterleavers which reduces memory requirements and is suitable for parallel decoder implementations.

1.3.2 Congruential Interleavers

A simple method to construct multiple interleavers is to use congruential interleavers with different seed. The permutation rule of a congruential interleaver is given by [19]

$$\Pi_k(n) = s_k + nc_k \mod N, \ n = 0, ..., N - 1, \tag{1.1}$$

where s_k is an integer starting value, N is the interleaver size and c_k is an integer which must be relatively prime to N in order to ensure an unique mapping. Multiple interleavers can be generated by using different c_k and s_k . We may choose the values of c_k such that adjacent bits in the data sequence are mapped to positions with a predetermined minimum spacing of s bits. In this case, the interleaver is called an s-random congruential interleaver.

1.3.3 Cyclic Shifted Multiple Interleavers

Generating multiple interleavers from one common mother interleaver using cyclic shifts and self-interleaving was proposed in [13,14] in the context of interleave division multiple access (IDMA) where users with low rate FEC coding are separated by different interleavers. The advantage is that only a single interleaving pattern has to be stored. Other interleavers can be constructed if needed based on very few parameters, i.e. the cyclic shifts. The use of cyclic shifts for generation of multiple interleavers is motivated by an observation for multiuser detection which showed that asynchronism between users, i.e. the user's signals arrive with different delay at the multiuser receiver, allows to separate them as well as userspecific random interleavers even if the same interleaver is used for all users [13]. It was proposed to construct the interleaving pattern Π_k for user k from a common interleaver Π by user-specific cyclic shifts $\Delta_{k,c}$ and interleaving of the permutation pattern by itself as indicated in Fig. 1.3. With about D=3 such cyclic shifts and self-interleaving operations, the same performance as with randomly chosen interleavers could be obtained in IDMA with synchronous users. We now apply the same idea to interleaving in a concatenated zigzag code.

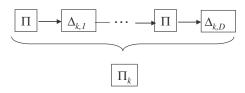


Fig. 1.3 Cyclic shifted interleaver Π_k from mother interleaver Π

1.3.4 UMTS Based Interleaver

The interleaver which is defined for the parallel concatenated convolutional code (turbo code) in UMTS [18], is optimized for performance in iterative decoding of turbo codes while allowing relatively simple construction for different interleaver sizes. A simple method to obtain a second interleaver is to use the transpose permutation matrix of the UMTS interleaver as suggested in [8]. Here, we propose a method in order to obtain more interleavers by reading out permutation rules at intermediate steps of the UMTS interleaver construction.

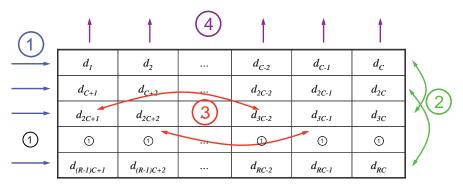


Fig. 1.4 Construction of the UMTS interleaver

The UMTS interleaver is constructed in several steps as illustrated in Fig. 1.4: First, the data bits are written row by row into a matrix of dimension $R \times C$. The values of R and C depend on the interleaver size. Next, rows are exchanged according to certain rules (for details see [18]). Finally, intra-row permutations are performed within the rows. The permuted bits are read out column by column.

In principle, we can generate multiple interleavers by reading out intermediate permutation rules at each step either column-wise or row-wise. E.g. reading out column by column after the first step, i.e. writing data into the matrix, would yield a block interleaver. After the row exchange, we can again read out column by column or row by row, which yields two additional interleavers. The same can be done after each intra-row permutation step. Naturally, not all of those interleavers will show good performance.

We propose to improve the spreading properties of those intermediate interleavers by simple operations. One possibility is to use the transpose Π_k^T of the respective permutation matrix rather than Π_k itself. Actually, Π_k^T is the deinterleaver corresponding to Π_k . Therefore, this operation imposes no additional complexity compared to using Π_k directly, since in a bidirectional communication the deinterleaving rule has to be determined anyway. We just exchange the role of interleaver and deinterleaver. An iterative decoder anyway needs both interleaver and deinterleaver. Interestingly, the original version and its transpose have almost no mappings in common. This might be a hint for good mutual independency and, hence, applicability in concatenated zigzag codes.

Another method for randomization can be obtained by modifying the order in which we read out the interleaver. We propose to read out only column wise but to change the starting value and the order in which the columns are read out. A possible implementation is, to use the permutation rule of a congruential interleaver in order to determine the order in which the columns are read out. Let C be the number of columns. Then, we can determine the order in which the columns are read out according to

$$\Pi_{c,k}(i_c) = s_{c,k} + c_{c,k}i_c \mod C, \ i_c = 0, \dots C - 1, \tag{1.2}$$

where $c_{c,k}$ is an integer which must be relatively prime to *C* in order to ensure that each column is read out once and only once. $s_{c,k}$ is an integer starting value. Both $c_{c,k}$ and $s_{c,k}$ can be randomly chosen and should be different for different interleavers Π_k . Column i_c is read out as the $\Pi_{c,k}(i_c)$ -th column. The special case $c_{c,k}=1$ yields just a cyclic shift by $s_{c,k}$ of the order in which the columns are read out.

As a further option we may choose the row index at which reading out of column i_c starts either randomly or according to a predetermined rule. We may further specify that row i_c is read out upwards, i.e. towards lower row indices, or downwards, i.e. towards higher row indices, in a cyclic manner.

In order to obtain more or improved interleavers, we may take the transpose of the permutation matrix for all or some of the interleavers which are generated by the modified read out process.

For interleavers which are constructed differently from the UMTS interleaver, it may make sense to apply an analogous row-wise read out process.

We suggest to construct interleavers Π_k in the following order plus the above proposed operations such as transpose or modified read out order: Use the complete UMTS interleaver for $\Pi_I \cdot \Pi_2 = \Pi_1^T$ may be constructed as the transpose of Π_I . For Π_3 we may use the block interleaver which results from reading out the data column by column after step 1. The next interleaver Π_4 is obtained by reading out column by column after the row exchange in step 2. By doing so, we can obtain interleavers which have very few bit mappings in common.

1.3.5 Interleavers for Zigzag Codes

In the following we take the requirements of concatenated zigzag codes explicitly into account in the construction of interleavers. The asymptotic performance of a concatenated zigzag code is determined by low weight codewords of the overall code [1]. Our objective is to avoid those low weight codewords in order to increase the minimum Hamming distance of the concatenated code while providing sufficient randomness by the interleavers. More precisely, our main goal is to avoid codewords with weight $w_H = d_{min} = 2$. Codewords with Hamming weight $w_H = 2$ occur, if the original data sequence has only two bits with value $d_{i,j} = 1$ which are located within the same segment and which are mapped to the same segment by the interleavers (see Fig. 1.2). Consequently, a restriction to the interleavers should be that data bits which are within the same segment in the original sequence are mapped to different segments by the interleaver. As indicated in Fig. 1.2, the weight of the resulting codeword will be the higher the farer the segments to which the two bits with value 1 are mapped are separated. This can be taken into account when putting the even harder interleaver restriction that bits which are within one segment in the original sequence are mapped to different segments with a minimum separation of at least *B* segments.

As a secondary criterion, we may wish to care also about code words with the second smallest possible Hamming weight $w_H=3$. Those codewords are generated if the data sequence contains two bits with value 1 which are located in adjacent segments. The two 1-bits should be spread farer apart by the interleaver. Particularly, a situation should be avoided, where both 1-bits are mapped to the same segment and, hence, a weight $w_H=d_{min}=2$ codeword results.

The interleaver design criteria may be summarized as follows:

- 1. Bits which are located in the same segment in the interleaver input sequence must be mapped to different segments which are separated by at least *B* segments, where $B \ge 1$ is a design parameter.
- 2. Bits which are located in adjacent segments in the interleaver input sequence should be mapped to different segments which are separated by at least *n* segments, where $n \ge 2$ is a design parameter.

1.3.5.1 Restricted Random Interleaver

Our first proposal is a random interleaver construction with restrictions. The approach is illustrated in Fig. 1.5, where the abscissa denotes the indices of the interleaver input sequence and the ordinate denotes the indices of the output sequence.

The input indices are successively mapped to output indices starting from index (i,j)=(1,1) up to index (i,j)=(I,J). The first index (i,j)=(1,1) is randomly mapped to an index (i',j'). The (i',j')-th row is marked as blocked area such that no further input indices are mapped to the same output index. In order to meet the above mentioned criterion 1, we further block an area within the first segment i=1 consisting of a predetermined number aJ of rows above and bJ below the J rows which belong to the assigned segment i'. In most cases, we may choose a=b=k. Next, we randomly assign the next index (i,j)=(1,2) to (i',j'), where (i',j') must not be located in the blocked area. This ensures that all bits which are located within the same segment at the interleaver input are mapped to different segments which are separated by at least $B=\min\{a,b\}$ segments. All further indices are assigned accordingly.

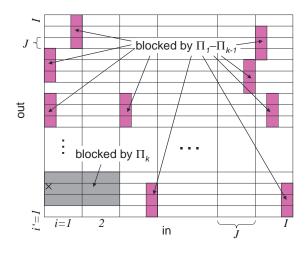


Fig. 1.5 Construction of multiple random interleavers with restrictions

In order to meet also the above mentioned criterion 2, we can block the respective rows above and below segment *i*' for the two segments *i* and i+1 of the input sequence rather than only for segment *i*. This ensures that data bits which are located in two adjacent segments of the input sequence are mapped to segments which are separated by at least $B=\min\{a,b\}$ segments. Interleaver construction is impossible if (2J-1)(a+b+1)>I. In order to enable convergence of the proposed algorithm, we should choose

$$(2J-1)(a+b+1) \ll I. \tag{1.3}$$

So far, we have described the construction for one interleaver. If multiple interleavers are required, we may wish to ensure that they are mutually independent. One criterion for *mutually independent* might be that they have no mappings in common. This can be achieved if we start construction of the *k*-th interleaver Π_k with a blocked area which consists of a part of the blocked area from previously constructed interleavers Π_I to Π_{k-I} . More precisely, we propose to block in each column the index pairs (i',j') to which the input index pair (i,j) of the respective column has been mapped by previously constructed interleavers Π_I to Π_{k-I} as well as a predetermined number of *m*, $m \ge 0$ elements above and below those index pairs

(i',j') as indicated in Fig. 1.5. If this blocking results in a situation during construction of interleaver Π_k , where no output index pair can be assigned for a particular input index pair, then the blocking which is due to previously assigned interleavers Π_I to Π_{k-I} is deleted for this column. If still no mapping can be found, then also the blockings set during construction of the current interleaver Π_k are deleted. Alternatively, we may start construction of the random generator.

For higher input index pairs (i,j), the degrees of freedom are reduced due to the already put restrictions. This can be taken into account if we change the starting index of the algorithm for each constructed interleaver Π_k . A simple approach is to start the algorithm from (i,j)=(1,1) for odd k. For even k, we can do a reverse order, i.e. start at the highest index pair (i,j)=(I,J). Even more randomness can be achieved if we choose the next index pair (i,j) randomly at each step. However, in this case we have to block the respective rows not only for input segments i and i+1 but also for segment i-1 in order to meet the abovementioned criterion 2.

1.3.5.2 AB Interleavers

A more structured method for generation of interleavers which meet the requirements for zigzag codes will be described in this section. Again, we use a square interleaver representation as depicted in Fig. 1.6 for illustration, where the abscissa denotes the indices of the interleaver input sequence and the ordinate denotes the indices of the output sequence.

In order to ensure that worst case patterns are avoided, we restrict the area of allowed mappings. We wish to make sure that bits which are located in the same segment in the original sequence are mapped to different segments. This can be achieved if we mark an allowed area of A segments for the mapping of each input bit as shown in Fig. 1.6. The allowed areas of two bits in the same input segment shall be separated by at least B segments. Therefore, the allowed area of A segments for the *j*-th bit in each input segment *i* starts with output segment i'=(j-1)(A+B)+1, j=1,...,J, for odd numbered segments *i*. For even numbered segments *i*, the allowed areas are shifted upwards by *B* segments in order to obtain a unique mapping to all output index pairs. For the sake of simplification of ensuring unique mappings, we restrict ourselves to the case A=B=I/2. This implies that I/2 is a multiple of *J* which is not a strong restriction and is met for all published regular zigzag codes.

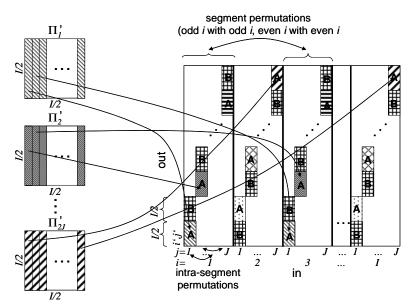


Fig. 1.6 Construction of AB interleaver

Next, all input bits which share the same allowed area are stacked to one block and interleaved by an interleaver of size I/2 as indicated in Fig. 1.6. We can use any interleaver type for those interleavers Π'_{1} to Π'_{2J} , e.g. random interleavers, congruential interleavers or interleavers as specified for the UMTS turbo code [18]. The interleavers Π'_{1} to Π'_{2J} can be identical or different. Using several smaller interleavers yields the advantage of lower required memory for the permutation pattern or lower effort for the interleaver construction, respectively, as well as a relaxation of the memory access collision problem. The disadvantage is a reduced interleaver size and, hence, less randomization effect. The permutation patterns of the interleavers Π'_{1} to Π'_{2J} are then remapped to the full interleaver pattern as indicated in Fig. 1.6.

Further randomization can be obtained by doing intra-segment permutations, i.e. permutations of columns within one segment i and segment permutations, i.e. group wise permutations of column groups of size J which belong to the same input segment i. Those permutations can be done pseudo randomly or according to any deterministic rule. However, we need to make sure for the segment permutation that odd segments i are only exchanged with odd segments and even segments are only exchanged with even segments. Multiple interleavers can easily be generated by doing different column permutations. A further but slightly more complex method for obtaining multiple interleavers is using different size I/2 interleavers Π'_{1} to Π'_{2J} .

The interleavers generated in this way guarantee that worst case patterns are avoided. The above mentioned criterion 1 is met. Also, it is guaranteed that bits of adjacent segments are not mapped to the same segment. However, the above mentioned secondary criterion 2 cannot be completely met. It is not ensured that bits of adjacent segments are spread farer apart.

1.4 Simulation Results

For evaluation of the interleaving schemes, we show performance results for zigzag codes with I=256 and J=4, i.e. an interleaver size of N=1024bits. We concatenate K=3 or K=4 constituent codes, which results in a code rate of $R_c=4/7$ and $R_c=1/2$, respectively. BPSK modulation was applied and the code bits have been transmitted over an AWGN channel with two-sided noise power spectral density $N_0/2$. The decoder performs 8 iterations. We compare the following interleavers:

- Random interleavers, where *K* interleavers Π_k are randomly generated without any restrictions.
- Random congruential and *s*-random congruential interleavers according to Sect. 1.3.2.
- Cyclic shifted random interleavers according to Sect. 1.3.3, where the mother interleaver is a randomly generated interleaver or the UMTS interleaver. We use D=1 or D=3 randomly chosen cyclic shifts $\Delta_{k,l}$.
- UMTS-based intermediate interleavers as described in Sect. 1.3.4.
- Restricted random interleavers according to Sect. 1.3.5.1.
- AB interleavers according to Sect. 1.3.5.2

The BER with K=3, i.e. two interleavers, is depicted in Fig. 1.7. It can be observed, that cyclic shifted interleavers require D=3 cyclic shifts and self interleaving operations for good performance. Cyclic shifted interleavers with a random mother interleaver or the UMTS interleaver as mother interleaver perform similar with a slight advantage of the UMTS based interleaver in the error floor region. Congruential interleavers show relatively poor performance. However, an s-random interleaver with the choice s=2J+1 performs very well. It shows almost the same performance as the restricted random interleaver as proposed in Sect 1.3.5.1. The AB interleaver proposed in Sect. 1.3.5.2 performs slightly worse in the waterfall

region. This is due to the poorer randomization effect of the small subinterleavers. However, it shows the best performance in the error floor region.

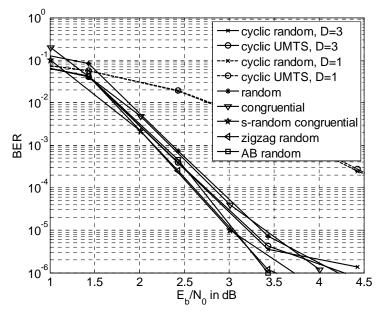


Fig. 1.7 BER of zigzag codes with different interleavers. I=256, J=4, K=3

The BER performance with a higher number of interleavers, i.e. K=4, is depicted in Fig. 1.8. In contrast to the case of K=3, s-random congruential interleavers perform significantly worse than our new proposed interleavers. Obviously, the congruential construction rule fails to provide mutual randomness between different interleavers.

Performance results for zigzag codes with interleavers which are constructed from intermediate UMTS interleavers are depicted in Fig. 1.9. For the curves in Fig. 1.9, we used the full UMTS interleaver and the block interleaver after step 1 in the interleaver construction as well as their transpose and randomly read out versions as intermediate interleavers. The worst performance is obtained, when we use the UMTS interleaver, the row-wise read UMTS interleaver and the block interleaver after step 1. This is due to the bad, only local permutation obtained when reading out the UMTS interleaver row-wise. The performance is significantly improved, if the same intermediate interleavers are used but randomly read out as suggested above.

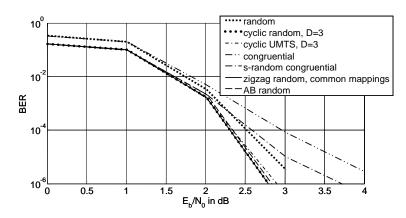


Fig. 1.8 BER of zigzag codes with different interleavers. I=256, J=4, K=4

Very good performance is obtained, when we use the UMTS interleaver, its transpose and the block interleaver after step 1. The performance of this combination can even be slightly improved by reading out randomly. Those interleaving schemes outperform s-random congruential interleavers. The performance when using the transpose of the block interleaver after step 1 instead of the transpose of the UMTS interleaver is similar to that when randomly reading out three different intermediate interleavers.

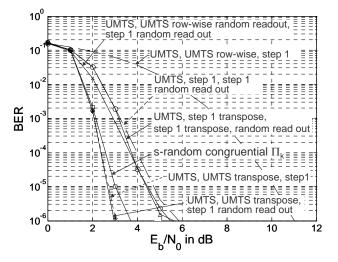


Fig. 1.9 BER with modified intermediate UMTS interleavers. I=256, J=4, K=4

1.5. Conclusions

We have proposed various interleaving strategies for concatenated zigzag codes. Cyclic shifted interleavers require low memory for storing of the interleaving patterns since multiple interleavers are generated by cyclic shifts and self interleaving from a single common mother interleaver. However, the specific properties of zigzag codes are not taken into account by cyclic shifted interleavers in order to avoid worst case patterns which result in low overall Hamming weight. Those properties are taken into consideration in our proposals of restricted random interleavers and AB interleavers. We showed that our proposed interleavers outperform straightforward interleaving approaches such as s-random congruential interleavers particularly when the number of required interleavers increases where the congruential construction rule fails to provide mutual randomness between different interleavers. Using intermediate steps in the construction of the UMTS turbo code internal interleaver with modified read out process allows to use turbo codes and zigzag codes within the same framework with minimum hardware modifications. Even with those intermediate interleavers, excellent performance can be obtained.

References

- L. Ping, X. Huang, and N. Phamdo, "Zigzag codes and concatenated zigzag codes," *IEEE Transactions on Information Theory*, vol. 47, pp. 800–807, February 2001.
- S.-N. Hong and D.-J. Shin, "Construction and analysis of rate-compatible punctured concatenated zigzag codes," in *International Symposium on Information Theory (ISIT)*, pp. 840–844, IEEE, September 2005.
- 3. S.-N. Hong and D.-J. Shin, "Optimal rate-compatible punctured concatenated zigzag codes," in *International Conference on Communications (ICC)*, pp. 627–631, IEEE, May 2005.
- 4. S.-N. Hong and D.-J. Shin, "Design of irregular concatenated zigzag codes," in *International Symposium on Information Theory (ISIT)*, pp. 1363–1366, IEEE, September 2005.
- L. Ping, S. Chan, and K. Yeung, "Iterative decoding of multi-dimensional concatenated single parity check codes," in *International Conference on Communications (ICC)*, pp. 131–135, IEEE, June 1998.

- X. Wu, Y. Xue, and H. Xiang, "On concatenated zigzag codes and their decoding schemes," *IEEE Communications Letters*, vol. 8, pp. 54–56, January 2004.
- N. Nefedov, "Multi-dimensional zigzag codes for high data rate transmission," in *International Symposium on Turbo Codes and Related Topics*, pp. 215–218, September 2003.
- 8. N. Nefedov, "Evaluation of low complexity concatenated codes for high data rate transmission," in *International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1868–1872, September 2003.
- L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory*, vol. IT-20, pp. 284–287, March 1974.
- J. Hagenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Transactions on Information Theory*, vol. IT–42, pp. 425–429, March 1996.
- A. Tarable, G. Montorsi, and S. Benedetto, "Analysis and design of interleavers for iterative multiuser receivers in coded CDMA systems," *IEEE Transactions on Information Theory*, vol. 51, pp. 1650–1666, May 2005.
- S. Brück, U. Sorger, S. Gligorevic, and N. Stolte, "Interleaving for outer convolutional codes in DS-CDMA systems," *IEEE Transactions on Communications*, vol. 48, pp. 1100–1107, July 2000.
- 13. K. Kusume and G. Bauch, "Cyclically shifted multiple interleavers," in *Globecom*, IEEE, November/December 2006.
- 14. K. Kusume and G. Bauch, "Some aspects of interleave division multiple access in ad hoc networks," in *International Symposium on Turbo Codes and Related Topics*, April 2006.
- 15. A. Barbulescu and S. Pietrobon, "Interleaver design for three dimensional turbo codes," in *IEEE International Symposium on Information Theory*, p. 37, September 1995.
- N. Ehtiati, M. Soleymani, and H. Sadjadpour, "Interleaver design for multiple turbo codes," in *Canadian Conference on Electrical and Computer Engineering*, pp. 1605–1607, May 2003.
- N. Ehtiati, M. Soleymani, and H. Sadjadpour, "Joint interleaver design for multiple turbo codes," in *IEEE Vehicular Technology Conference (VTC Fall)*, pp. 2302–2306, September 2004.
- 18. 3GPP, "Turbo code internal interleaver," in 3G TS 25.212 V3.2.0 (2000–03), Release 1999, 2000.
- S. Crozier, J. Lodge, P. Guinand, and A. Hunt, "Performance of turbo-codes with relative prime and golden interleaving strategies," in *International Mobile Satellite Conference*, pp. 268–275, 1999.
- D. Divsalar and F. Pollara, "Multiple turbo codes for deep-space communications," TDA Progress Report 42-121, May 1995.
- T. Sohdoh, N. Nishinaga, and Y. Iwadare, "The new interleaver design of turbo codes for error floor improvement," in *Symposium on Information Theory and its Applications*, pp. 13–16, December 1997.

- 22. G. Bauch, K. Kusume: "Interleaving strategies for concatenated zigzag codes." *IST Mobile and Wireless Communications Summit*, July 2007.
- 23. G. Bauch, K. Kusume: "Simple construction of multiple interleavers for concatenated zigzag codes." *International Wireless Communications and Mobile Computing Conference (IWCMC)*, August 2007.

2 Simplified Channel-Aware Greedy Scheduling and Antenna Selection Algorithms for Multiuser MIMO Systems Employing Orthogonal Space Division Multiplexing

Shreeram Sigdel and Witold A. Krzymień

2.1 Introduction

In the downlink of multiuser multiple-input multiple-output (MIMO) systems, the multiple antennas at the base station allow for spatial multiplexing of transmissions to multiple users in the same time slot and frequency band. Due to their geographical dispersion, coordination among users is difficult, which makes the downlink of a multiuser system more challenging compared to single user MIMO systems. From information theoretic point of view, the sum capacity achieving precoding or preprocessing technique is dirty paper coding (DPC) [1], which uses successive interference pre-cancellation approach employing complex encoding and decoding. This approach is very complex and its value is primarily theoretical. For this reason an intensive research effort is underway to devise suboptimal but more practical approaches to multiuser downlink signal processing. Beamforming or transmit pre-processing is a suboptimal and reduced complexity (compared to DPC) strategy, where each user stream is coded independently and multiplied by a beamforming weight vector for transmission through multiple antennas. Proper design of the beamforming weight vectors allows the interference among different streams to be minimized (or eliminated), thereby supporting multiple users simultaneously. The challenge is thus to design transmit and receive processing vectors such that space-division multiplexing is effectively achieved. Despite its suboptimality, for independently fading channels linear beamforming has been shown to achieve the best trade-off between performance and complexity [2]. In conjunction with scheduling, it has been shown to achieve a fairly large portion of DPC capacity when the base station has multiple antennas and each user has a single antenna [3,4].

In this Chapter, we consider orthogonal linear precoding techniques to achieve orthogonal space division multiplexing (OSDM) in the downlink of multiuser MIMO systems, in which both base station as well as mobile stations employ multiple antennas. The orthogonal precoding allows transmission to the mobile users to be multiuser interference free. OSDM for multiuser MIMO systems has been proposed by several researchers [5,6]. We consider two OSDM techniques that use subspace projection methods to design precoding matrices: block diagonalization and successive optimization. With block diagonalization, each user's precoding matrix is designed such that the transmitted signal of that user lies in the null space of all other remaining users' channels, and hence multiuser interference is pre-eliminated. Block diagonalization takes the sum rate maximization approach with the sum power constraint, which tends to select the strong users often causing unfairness among users. Hence, minimizing transmit power while achieving desired quality of service for users may be interesting in practice [5]. In this context, [5] proposes two solutions for the transmit power minimization problem. One is based on block diagonalization, where precoding matrix of a user is designed as in block diagonalization and waterfilling power allocation is performed with individual power constraint P_k (e.g. power required to achieve rate R_k). Another way of minimizing transmit power is to use a method called successive optimization [5], where precoding matrix of a user is designed such that it does not interfere with any of the previously precoded users and transmit power for that user is individually allocated to achieve required quality of service. The latter leads to a simpler solution.

Due to the null space dimensional requirements of block diagonalization and successive optimization techniques, the numbers of users supported in the same time/frequency slot are limited for a given number of transmit antennas. Therefore these techniques should be combined with scheduling so that a multiuser diversity gain can be achieved. Multiuser diversity arises when a large number of users with independently fading channels are present, and hence it is likely that a user or multiple users experience high channel gain in any given time/frequency slot [15]. The objective of channel-aware scheduling (scheduling with channel state information available at the transmitter) is to maximize system throughput by allocating radio resources to the user or the multiple users that experience the highest channel gains. In this context, several suboptimal but simplified algorithms to schedule a subset of users have been proposed in the literature for the case of multiuser multiple antenna downlink with single receive antenna users [7,8,9]. On the other hand, to reduce the hardware complexity at the mobile units, receive antenna selection is a promising technique, with which a smaller subset of antennas is selected for reception. Similarly, power saving modes with receive antenna selection at mobile stations have also been considered in current 802.11n wireless local area network standard proposals [10]. Additionally, receive antenna selection allows block diagonalization or successive optimization to accommodate more users in a given time slot, which may translate into the sum rate gain.

In this Chapter, we consider channel-aware joint user scheduling and receive antenna selection problem for block diagonalization and successive optimization techniques. Optimal selection of antenna and user subsets involves exhaustive search through all combinations of active users and receive antennas, which is computationally very complex for systems with a large number of users. Hence, we propose simplified user and antenna selection algorithms for block diagonalization and successive optimization. Our scheduling algorithms stem from greedy scheduling algorithms for systems with single antenna users [7,8,9]. We also propose a user grouping technique at each step of the proposed algorithms so that the search complexity is further reduced. Moreover, we propose scheduling with proportional fairness for both block diagonalization and successive optimization in order to achieve fairness in throughput among users. Introducing proportional fairness criterion is not trivial for either block diagonalization or successive optimization because the supported rate of a user is unknown before all the supported users have been selected. To overcome this problem, we propose simplified scheduling metrics. We also propose two receive antenna selection algorithms, which work in conjunction with user scheduling. They are shown to further enhance multiuser diversity gain achieved through scheduling.

The rest of the Chapter is organized as follows. Section 2.2 describes the system model of the proposed system along with brief review of block diagonalization and successive optimization techniques. The proposed channel-aware user scheduling and receive antenna selection algorithms are discussed Sect. 2.3. A subspace correlation based user grouping technique is also proposed in this Section. The impact of receive antenna selection on the multiuser MIMO system employing block diagonalization is discussed in Sect. 2.4. Simulation results are given in Sect. 2.5, and Sect. 2.6 presents some concluding remarks and possible directions of further work.

2.2 System Model

We consider the downlink of a multiuser MIMO system with M_T transmit antennas and N_k , k = 1, 2, ..., K receive antennas at kth mobile user as shown in Fig. 2.1, where K is the number of multiple antenna users.

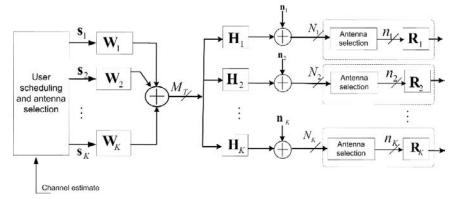


Fig. 2.1 Block diagram of the proposed multiuser MIMO system

Let $\mathbf{H}_k \in \mathcal{C}^{N_k \times M_T}$, denote the downlink channel matrix of the *k*th user. We assume a flat Rayleigh fading channel model so that the elements of \mathbf{H}_k , k = 1, 2...K can be modeled as independent and identically distributed (i.i.d) complex Gaussian random variables with variance of 0.5 per dimension. The data vector of user k, $\mathbf{s}_k \in \mathcal{C}^{n_k \times 1}$ is preprocessed at the transmitter with an $M_T \times n_k$ beamforming matrix \mathbf{W}_k to produce the transmitted signal vector $\mathbf{x}_k \in \mathcal{C}^{M_T \times 1}$. n_k denotes the number of transmitted streams to user k, which is assumed to be equal to the number of radio frequency chains at the *k*th mobile station. The $N_k \times 1$ received signal vector of the *k*th user can be expressed as

$$\mathbf{y}_k = \mathbf{H}_k \sum_{j=1}^{K} \mathbf{W}_j \mathbf{s}_j + \mathbf{n}_k$$
(2.1)

Where $\mathbf{n}_k \in \mathcal{C}^{N_k \times 1}$ denotes zero mean additive white Gaussian noise with $E\{\mathbf{n}_k \mathbf{n}_k^H\} = \sigma_n^2 \mathbf{I}_{N_k}$. Linear processing at the receiver yields,

$$\hat{\mathbf{y}}_k = \mathbf{R}_k^H (\mathbf{H}_k \sum_{j=1}^K \mathbf{W}_j \mathbf{s}_j + \mathbf{n}_k)$$
(2.2)

where, $\mathbf{R}_k \in \mathcal{C}^{N_k \times n_k}$ is the receive processing matrix of user k. For multiuser interference free transmission, \mathbf{W}_k should be designed such that

$$\mathbf{H}_k \mathbf{W}_j = \mathbf{0} \quad \text{for all } k \neq j \text{ and } 1 \le k, j \le K$$
(2.3)

The multiuser MIMO system is then decomposed into parallel single user MIMO channels and $\mathbf{y}_k = \mathbf{H}_k \mathbf{W}_k \mathbf{s}_k + \mathbf{n}_k$.

We assume that base station transmitter has the channel state information (CSI) of all active users in the system obtained via channel reciprocity for time division duplex (TDD) systems or through feedback links. The issue of inaccuracy of CSI at the transmitter as well as the emerging partial CSI feedback schemes are some of the practically important issues for multiuser MIMO downlink transmission. However, separate research is necessary on those questions and they are beyond the scope of this study.

2.2.1 Block Diagonalization (BD)

Block diagonalization (BD) [5] is a linear orthogonal precoding technique to achieve complete multiuser interference cancellation as defined in (2.3). Let us denote the aggregate channel and precoding matrices of all users, respectively, as $\mathbf{H} = (\mathbf{H}_1^T \ \mathbf{H}_2^T \dots \mathbf{H}_K^T)^T$ and $\mathbf{W} = (\mathbf{W}_1 \ \mathbf{W}_2 \dots \mathbf{W}_K)$. The precoding matrices of users are designed using null space projection method as follows. Let us define a $\sum_{j=1, j \neq k}^K N_j \times M_T$ aggregate channel interference matrix for user *k* as

$$\bar{\mathbf{H}}_{k} = (\mathbf{H}_{1}^{T} \dots \mathbf{H}_{k-1}^{T} \mathbf{H}_{k+1}^{T} \dots \mathbf{H}_{K}^{T})^{T}$$
(2.4)

Zero multiuser interference constraint requires that the precoding matrix \mathbf{W}_k of user k lies in the null space of \mathbf{H}_k , which requires the null space of $\bar{\mathbf{H}}_k$ to have a dimension greater than zero. This condition imposes a constraint on the number of base station antennas to be $M_T \ge \sum_{i=1}^K N_i$, assuming that the channel matrices are of full rank for all K users (which occurs with probability of one in i.i.d. Gaussian channels). We denote the SVD of $\bar{\mathbf{H}}_k$ as $\bar{\mathbf{H}}_k$ as $\bar{\mathbf{H}}_k = \bar{\mathbf{U}}_k (\bar{\mathbf{\Sigma}}_k \mathbf{0}) (\bar{\mathbf{V}}_k^1 \bar{\mathbf{V}}_k^0)^H$ where $\bar{\mathbf{\Sigma}}_k$ is the $\bar{r}_k \times \bar{r}_k$ diagonal matrix containing \bar{r}_k non-zero singular values of $\bar{\mathbf{H}}_k$, and $\bar{r}_k = \operatorname{rank}(\bar{\mathbf{H}}_k)$. $\bar{\mathbf{V}}_k^{\mathbb{I}}$ holds the $M_T - \bar{r}_k$ right singular vectors as its columns. These columns constitute the orthonormal basis for the null space of $\bar{\mathbf{H}}_k$. We assume that the fading among the antennas of a user as well as among the users is independent (hence the matrices are of full rank). This means at most $M_T - \bar{r}_k$ streams can be transmitted to user k. Hence, there exists $N_k = M_T - \bar{r}_k$ columns of $\bar{\mathbf{V}}_k^0$, which form the null-space basis of $\bar{\mathbf{H}}_k$. Constructing the precoding matrix with the columns of $\bar{\mathbf{V}}_k^{\mathbb{I}}$ will satisfy the zero multiuser interference constraint (2.3). With this, the multiuser channel decouples into K parallel non-interfering single-user MIMO channels (also referred to as null space projected channels), which is expressed as

$$\mathbf{H}_{k,e} = \mathbf{H}_k \bar{\mathbf{V}}_k^0 \tag{2.5}$$

With sum power constraint P, the achievable throughput of the block diagonal system is

$$R_{BD} = \max_{\mathbf{Q}_k: \mathbf{Q}_k \ge 0, (\sum_{k=1}^{K} \operatorname{trace}(\mathbf{Q}_k)) \le P} \sum_{k=1}^{H} \log_2 \det(\mathbf{I} + \frac{1}{\sigma_n^2} \mathbf{H}_{k,e} \mathbf{Q}_k \mathbf{H}_{k,e}^H)$$
(2.6)

where $\mathbf{Q}_k = E\{\mathbf{s}_k \mathbf{s}_k^H\} \in \mathcal{C}^{N_k \times N_k}$ is the *k*th user's input covariance matrix. Then, the throughput maximization problem turns into a single-user capacity maximization problem. The solution of it is to choose the precoding matrix (of equivalent single user channel) to be the right singular vectors of $\mathbf{H}_{k,\epsilon}$ and perform waterfilling power allocation [11]. Let us denote the SVD of $\mathbf{H}_{k,\epsilon}$ as

$$\mathbf{H}_{k,e} = \mathbf{U}_{k,e} \begin{bmatrix} \boldsymbol{\Sigma}_{k,e} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{k,e}^1 & \mathbf{V}_{k,e}^0 \end{bmatrix}^H$$
(2.7)

where $\Sigma_{k,\epsilon}$ is an $L_{k,\epsilon} \times L_{k,\epsilon}$ diagonal matrix of singular values; $L_{k,\epsilon}$ is the rank of $\mathbf{H}_{k,\epsilon}$ and $\mathbf{V}_{k,\epsilon}^1$ holds the first $L_{k,\epsilon}$ singular vectors. Thus, the product of $\bar{\mathbf{V}}_{k}^{0}$ and \mathbf{V}_{k}^{1} gives the precoding matrix for user k that maximizes the throughput of user k under zero multiuser interference constraint [5], i.e. $\mathbf{W}_{k} = \bar{\mathbf{V}}_{k}^{0} \mathbf{V}_{k,\epsilon}^{1}$ and $\mathbf{R}_{k} = \mathbf{U}_{k,\epsilon}$. Then, the product \mathbf{HW} block-diagonal matrix, which can be expressed a is as $\mathbf{HW} = \operatorname{diag}(\mathbf{H}_1 \bar{\mathbf{V}}_1^0 \mathbf{V}_{1\,e}^1, \mathbf{H}_2 \bar{\mathbf{V}}_2^0 \mathbf{V}_{2\,e}^1, \dots, \mathbf{H}_K \bar{\mathbf{V}}_K^0 \mathbf{V}_{K\,e}^1). \quad \text{We denote}$ the power loading coefficients corresponding to $\Sigma_{k,\epsilon}$ as $\Lambda_{k,\epsilon}$. Then, from (2.7), we define an aggregate matrix of singular values as $\Sigma = \text{diag}(\Sigma_{1,e}, \Sigma_{2,e}..., \Sigma_{K,e})$ and corresponding aggregate power loading matrix $\Lambda = \text{diag}(\Lambda_{1,e}, \Lambda_{2,e}, \dots, \Lambda_{K,e})$. Then, sum rate of BD can be [5] $R_{BD} = \max_{(\Lambda, \operatorname{trace}(\Lambda) \le P)} \log_2 \det \left(\mathbf{I} + \frac{1}{\sigma_n^2} \Sigma^2 \Lambda \right),$ expressed as where the diagonal elements of Λ are obtained from waterfilling on the diagonal elements of Σ for the sum power constraint *P*.

2.2.2 Successive Optimization (SO)

Successive optimization (SO) [5] is a technique to solve a problem of transmit power minimization while achieving a desired quality of service target (e.g. minimum or average rate) for each user. Users are successively precoded and power is allocated individually. To eliminate interference to the previously precoded users, SO employs the null-space projection method such that the transmit processing matrix of the user i lies in the null space of the aggregate channel of i = 1, 2, ..., i - 1 previously precoded

users. Power allocation to user *i* is done individually by taking interference from previous (*i*-1) users into account and fulfilling the rate requirement R_s . This is different from BD, where power is allocated to all users jointly by designing a block diagonal structure as discussed in the previous section. We denote the aggregate channel matrix of the previously selected users as $\hat{\mathbf{H}}_i = (\mathbf{H}_1^T \ \mathbf{H}_2^T \dots \mathbf{H}_{i-1}^T)^T$ and the basis for the null space of $\hat{\mathbf{H}}_s$ as $\hat{\mathbf{V}}_i^{\mathbb{C}}$. Then, precoding matrix \mathbf{W}_s can be chosen as a linear combination of the basis vectors $\mathbf{W}_i = \hat{\mathbf{V}}_i^0 \mathbf{B}_s$ for some choice of matrix \mathbf{B}_s . Hence, the objective of SO is to find \mathbf{B}_s that solves

$$R_{i} = \log \det \left(\mathbf{I} + \mathbf{H}_{i} \hat{\mathbf{V}}_{i}^{0} \mathbf{B}_{i} \mathbf{B}_{i}^{H} \hat{\mathbf{V}}_{i}^{0^{H}} \mathbf{H}_{i}^{H} \mathbf{Q}_{i,Z}^{-1} \right)$$
(2.8)

where $\mathbf{Q}_{i,Z} = (\sigma_n^2 \mathbf{I} + \sum_{j=1}^{i-1} \mathbf{H}_i \mathbf{W}_j \mathbf{W}_j^H \mathbf{H}_i^H)$ is the covariance of noise and interference from users precoded up to (*i*-1)th step. Then, the objective is to solve (2.8) such that trace($\mathbf{B}_i \mathbf{B}_i^H$) is minimized [5]. That leads to a water-filling solution as in the case of MIMO link with cochannel interference [12]. Let us define the equivalent single user channel matrix $\mathbf{H}'_{i,\epsilon}$ and the SVD as

$$\mathbf{H}_{i,e}' = \mathbf{Q}_{i,Z}^{-1/2} \mathbf{H}_i \hat{\mathbf{V}}_i^0 = \mathbf{U}_{i,e}' (\mathbf{\Sigma}_{i,e}' \ \mathbf{0}) (\mathbf{V}_{i,e}'^1 \ \mathbf{V}_{i,e}'^0)^H$$
(2.9)

Then, $\mathbf{W}_i = \hat{\mathbf{V}}_i^0 \mathbf{V}_{i,e}^{'1} \mathbf{\Lambda}_{i,e}^{'1/2}$, where $\mathbf{\Lambda}_{i,\epsilon}^{'}$ denote the diagonal power loading matrix with the elements obtained through waterfilling on the diagonal elements of $\boldsymbol{\Sigma}_{i,\epsilon}^{'}$ such that trace $(\mathbf{\Lambda}_{i,e}^{'}) = P_i$; P_i is the power required to achieve the desired rate of R_i . The transmit power is obtained by the sum of the elements of all $\mathbf{\Lambda}_{i,\epsilon}^{'}$, i = 1, 2, ...K. The transmit power minimization problem also depends on the order in which the users are processed. However, detailed study of the user ordering and its impact on sum rate is beyond the scope of this study.

2.3 Fair Scheduling and Antenna Selection Algorithms

Due to the constraint on the number of simultaneous users supported by orthogonal space-division multiplexing (OSDM) techniques discussed above, active users in the system have to be appropriately scheduled in time so that multiuser diversity gain is maximized. In this context, several low complexity user selection algorithms driven by throughput maximization have also been proposed [13,14]. Scheduling algorithms based on throughput maximization may result in some users starved due to bad channel conditions (e.g. users closer to the edge of a cell). Proportional fairness based scheduling [15] attempts to provide fairness among users, yet maintaining multiuser diversity gain. We propose simplified proportionally fair user scheduling algorithms and evaluate the effect of antenna selection on the system throughput for both block diagonalization (BD) and successive optimization (SO).

Let $\mathcal{U} = \{1, 2, ..., K\}$ denote the set of all users and $\mathcal{U}_l \subset \mathcal{U}$ denote the possible subset of users such that $|\mathcal{U}_l| \leq K_{\mathbb{C}}$, $i = 1, 2, ... {K \choose K_0}$, where $|\mathcal{U}_l$ denotes the cardinality of the set \mathcal{U}_i . $K_{\mathbb{C}}$ denotes the maximum number of supported users for given $M_{\mathcal{T}}$ and $N_k, k = 1, 2, ..., K$. Hence, $K_0 \leq K$ users have to be scheduled at a given time slot from the pool of K active users. We denote the selected subset of users as \mathcal{U}_s .

Proportionally fair and other classes of scheduling algorithms can generally be described through the weighted sum-rate maximization requirement [4]

$$\max_{\mathcal{U}_l \subset \mathcal{U}, P_k \ge 0, \sum_{k \in \mathcal{U}_l} P_k = P} \sum_{k \in \mathcal{U}_l} \mu_k(t) R_k(t)$$
(2.10)

where, $\mu_k(t) = 1/\bar{R}_k(t)$ and $R_k(t)$ are the priority weight and supported rate of kth user during tth scheduling interval. The maximization is done over achievable rate vectors with a given transmission scheme in a given time slot. $\mu_k(t) = 1$ leads to maximum throughput scheduling. $\bar{R}_k(t)$ is the average throughput achieved by user k up to time t, which is updated as

$$\bar{R}_k(t+1) = \delta \bar{R}_k(t) + (1-\delta)R_k(t), \qquad k \in \mathcal{U}_{\varepsilon}$$
(2.11)

$$\bar{R}_k(t+1) = \delta \bar{R}_k(t), \qquad k \notin \mathcal{U}_{\varepsilon}$$
(2.12)

where, $\delta = 1 - \frac{1}{T_c}$ is the forgetting factor, which is related to a sliding window with width equal to a number of time slots T_c , over which the throughput of user K is averaged, and $R_k(t)$ is the rate of user k during th transmission interval (zero if not scheduled).

2.3.1 Joint User and Antenna Selection for Block Diagonalization (BD)

Let $\Theta_k = \{1, 2, ..., N_k\}, k = 1, 2, ..., K$ denote the set of available antennas and $\theta_j \subset \Theta_k, |\theta_j| \le n_k, j = 1, 2, ... {N_k \choose n_k}$ denote the possible subset of antennas. The scheduling problem with user and antenna selection for block diagonalization (BD) can be expressed as,

$$\max_{\theta_j, \mathbf{Q}_k(\theta_j), \mathcal{U}_l} \sum_{k \in \mathcal{U}_l} \mu_k(t) R_{k, BD}(\mathcal{U}_l)$$
(2.13)

subject to $\mathbf{Q}_k(\theta_j) \ge 0$, $\sum_{k \in \mathcal{U}_l} \operatorname{trace}(\mathbf{Q}_k(\theta_j)) \le P$, where

$$R_{k,BD}(\mathcal{U}_l) = \log_2 \det(\mathbf{I} + \frac{1}{\sigma_n^2} \mathbf{H}_{k,e}(\theta_j) \mathbf{Q}_k(\theta_j) \mathbf{H}_{k,e}^H(\theta_j))$$

 $\mathbf{H}_{k,e}(\theta_j)$ and $\mathbf{Q}_k(\theta_j)$ denote the equivalent channel matrix and input covariance matrix corresponding to the subset of antennas denoted by θ_j , respectively.

Naturally, the exhaustive search based solution that maximizes (2.13) is computationally prohibitive. Hence, low complexity algorithms to select the best receive antenna and user subset are desirable. In [13], we proposed a maximum throughput simplified greedy joint scheduling and antenna selection algorithm for block diagonalization (BD). In that paper, simplified Frobenius norm based selection metrics were designed for joint selection of users and their best subset of antennas. However, due to the iterative nature of the proposed algorithm, the supported rate of a user can only be computed after all possible users have been selected. It makes the implementation of proportionally fair scheduling difficult. Hence. approximation to a user's rate is necessary at each step of the greedy scheduling algorithm. To achieve this, we propose two simplified scheduling metrics: capacity based and Frobenius norm based. The selection metrics are of heuristic nature and stem from the greedy scheduling algorithms in [7,8,9] for zero-forcing beamforming for single antenna users, but perform within a few dB from the exhaustive search algorithms.

The *i*th iteration of the algorithm in [13], however, still requires a search through K - i users. In [16] we propose an algorithm, which eliminates the users that do not meet some specified criterion at each step of the algorithm. To achieve this, the *i*th step of the proposed algorithm forms an intermediate user group \mathbb{U}_i of users less correlated with the subspace of the aggregate channel of the users selected up to (i - 1)th step based on some specified threshold, ξ . The user grouping technique will be discussed in Sect. 2.3.3. At step i + 1, the user search is restricted to the subset \mathbb{U}_i , which achieves significant complexity reduction compared to searching through K - i users for a large K. The proposed simplified proportionally fair (PF) scheduling algorithm for block diagonalization is summarized in Table 2.1. The proposed scheduling metrics as well as the operation of the algorithm (we call this algorithm Algorithm-I) are described as follows.

The first proportionally fair (PF) scheduling metric is based on the rate of the equivalent single user channel with equal power allocation at the transmitter. This is appropriate for high SNR region as the throughput gain with water-filling power allocation is minimal compared to the equal power allocation [17]. Hence, the proposed rate based scheduling metric (called Metric 1) to be used at the *i*th step of the proposed algorithm (see Table 2.1) is expressed as

$$\Omega_{1,k}(t) = \log_2 \det(\mathbf{I} + \frac{P}{M_T} \mathbf{H}_{k,e}(\hat{\theta}_k) \mathcal{P}_i^{\perp} \mathbf{H}_{k,e}^H(\hat{\theta}_k)) \text{ for } \forall k \in \mathbb{U}_i$$
 (2.14)

where $\mathcal{P}_i^{\perp} = \mathbf{I}_{M_T} - \mathbf{V}_{i-1}^H \mathbf{V}_{i-1}$ is the subspace projection matrix; \mathbf{V}_{i-1} is the row basis of $\mathbf{H}(\mathcal{U}_s)$, the aggregate channel of the users selected up to (i-1)th step. $\mathbf{H}_{k,e}(\hat{\theta}_k)$ denotes the selected channel matrix of user $k \in \mathbb{U}_i$.

Table 2.1 Simplified proportionally fair (PF) greedy scheduling algorithm for block diagonalization (BD): Algorithm-I (PF)

- i ← 1; U = {1, 2, ..., K}; U_s = {}; Θ_k = {1, 2, ...N_k}, k = 1, 2, ...K.
 Select a user u₁ such that u₁ = arg max μ_{1,k}(t) log det(**I** + P/M_T**H**_k**H**^H_k) (Metric 1); u₁ = arg max μ_{2,k}(t)||**H**_k||²_F (Metric 2);
 Set U_s = U_s ∪ {u₁}; U₁ = U \ {u₁}.
 Select n_{u1} receive antennas out of N_{u1} using proposed RAS algorithms. Let **H**[']_{u1} = **H**_{u1}(θ̂_{u1}); denote the selected channel corresponding to the best subset θ̂_{u1}
 i ← i + 1; Find the projector matrix P[⊥]_i = **I**_{MT} - **V**^H_{i-1}**V**_{i-1}, where **V**_{i-1} is the row basis of **H**(U_s); and **H**(U_s) = [**H**^{'T}_{u1}**H**^{'T}_{u2}...**H**^{'T}_{u1}]^T
 If (|U_s| < K₀); a. Find U_i = { k ∈ U_{i-1}, k ∉ U_s | ||**H**_k**W**^H_{i-1}|| < ξ }. b. If (|U_i| ≠ 0)
 - Select a user such that $u_i = \arg \max_{k \in U_i} \mu_{1,k}(t) \Omega_{1,k}(t)$ (Metric 1); $u_i = \arg \max_{k \in U_i} \mu_{2,k}(t) \Omega_{2,k}(t)$ (Metric 2);
 - Let $\mathbf{H}'_{u_i} = \mathbf{H}_{u_i}(\hat{\theta}_{u_i}), \mathcal{U}_s = \mathcal{U}_s \cup \{u_i\}$ and $\mathbb{U}_i = \mathbb{U}_i \setminus \{u_i\}$. Go to Step 2. Else end.

With Metric 1, the step 1 of the Algorithm-I (see Table 2.1) selects the first user such that

$$u_1 = \arg\max_{k \in \mathcal{U}} \mu_{1,k}(t) \log \det(\mathbf{I} + P/M_T \mathbf{H}_k \mathbf{H}_k^H)$$
(2.15)

Then, the antenna of the selected user is selected using the proposed receive antenna selection (RAS) algorithms (see Sect. 2.4). User grouping is performed at Step 3a by computing the correlation of the *k*th user's channel matrix with the signal space of $H(\mathcal{U}_{s,k})$, the aggregate channel matrix of users selected up to (i - 1)th step. User grouping is discussed in detail in Sect. 2.3.3. Then, if there are some users that meet the specified correlation criterion, the next user is selected from group U_{s} using the scheduling metric (2.14). Otherwise, the algorithm terminates with fewer than maximum number of supported users. In Step 3b a user u_{s} is selected such that

$$u_i = \arg\max_{k \in \mathbb{U}_i} \ \mu_{1,k}(t)\Omega_{1,k}(t)$$

$$(2.16)$$

Once the maximum supported numbers of users have been selected, block diagonalization algorithm discussed in Sect. 2.2.1 is used to compute actual rate of those users as

$$R_k(t+1) = \log_2 \det(\mathbf{I} + \frac{P}{M_T} \boldsymbol{\Sigma}_k^2 \boldsymbol{\Lambda}_k)$$
(2.17)

Then, (2.11) and (2.12) are used to update the average rate of all users and $\mu_{1,k}(t+1) = 1/R_k(t+1)$.

However, since the computation of Metric 1 involves logarithm and determinant operations at each step of the algorithm, it is still very complex for large K. To simplify it, we propose to employ the fairness in total power gain of each user's equivalent channel. A similar argument is also used in [18]. The rationale is to scale the kth user's channel gain (including the power allocation from waterfilling) so that the users with inferior channels are considered in later scheduling decisions. In this context in [16] the selection metric is designed such that at each iteration of the algorithm the null space of the kth user's channel is computed first and previously selected users' channels are individually projected to its subspace (we refer to this projection as forward projection). This simplification significantly reduces the complexity compared to a similar algorithm proposed in $[14]^1$ without noticeable loss in sum rate performance. Then, the sum of the Frobenius norm of forward projections along with the Frobenius norm of the projection of the kth user channel on the null space of the aggregate channel of the previously selected users (we refer to it as backward projection) constitutes the user selection metric. The objective of our selection metric design is to implicitly incorporate some orthogonality measure in it. This is because block diagonalization achieves the sum capacity for a given channel, when users' channels are orthogonal to each

¹ Reference [14] considers user selection only.

other [19]. The squared Frobenius norm is used as a selection metric as it gives the total power gains from the eigenmodes of a channel matrix, i.e. $||\mathbf{A}||_F^2 = \sum_i \lambda_i$; λ_i are the eigenvalues of $\mathbf{A}\mathbf{A}^H$. Additionally, antenna selection of a user in each step is carried out using one of the proposed receive antenna selection (RAS) algorithms discussed in Sect. 2.4.

We incorporate the gain due to forward and backward projections in our second scheduling metric (Metric 2) as follows.

$$\Omega_{2,k}(t) = \underbrace{||\mathbf{H}_k(\hat{\theta}_k)\mathcal{P}_i^{\perp}||_F^2}_{\text{Backward projection}} + \underbrace{\sum_{k'=1}^{i-1} ||\mathbf{H}_{k'}\mathcal{P}_k^{\perp}||_F^2}_{\text{Forward projections}} \text{ for } \forall \ k \in \mathbb{U}_i$$
(2.18)

where \mathbf{H}_{k^*} denotes the selected channel submatrix of user $k' \in \mathcal{U}_{k}$. With Metric 2, Step 1 of Algorithm-I selects the first user $u_{\mathbb{I}}$ such that

 $u_1 = \arg\max_{k \in \mathcal{U}} \mu_{2,k}(t) ||\mathbf{H}_k||_F^2$ (2.19)

Similarly, in Step 3b user u_* is selected such that

$$u_i = \arg\max_{k \in \mathbb{U}_i} \ \mu_{2,k}(t)\Omega_{2,k}(t)$$
(2.20)

with Metric 2 defined in (2.18). Once the maximum supported number of users have been selected, block diagonalization algorithm discussed in Sect. 2.2.1 is used to compute the actual power gain (including power allocated from waterfilling) of the *k*th user's channel as trace($\Sigma_k^2 \Lambda_k$) (also see (2.17)). Then, average value of the scheduling Metric 2 is updated as

$$\bar{\Omega}_{2,k}(t+1) = \delta \bar{\Omega}_{2,k}(t) + (1-\delta) \operatorname{trace}(\boldsymbol{\Sigma}_k^2 \boldsymbol{\Lambda}_k), \quad k \in \mathcal{U}_{\varepsilon}$$
(2.21)

$$\bar{\Omega}_{2,k}(t+1) = \delta \bar{\Omega}_{2,k}(t) \quad k \notin \mathcal{U}_{\mathfrak{s}} \tag{2.22}$$

and $\mu_{2,k}(t+1) = 1/\overline{\Omega}_{2,k}(t+1)$. Once again, Metric 2 is of a heuristic nature, which serves to explain the concept of the proportionally fair scheduling for block diagonalization. However, its actual impact on the sum rate is difficult to obtain analytically. The motivation behind it is that the capacity is closely related to power gains from the eigenmodes of a channel.

2.3.2 Joint User and Antenna Selection for Successive Optimization (SO)

The objective of successive optimization (SO) technique is to maintain desired quality of service for the users while minimizing transmit power. Hence, some level of fairness among the selected users in a given time slot is inherently present in this technique. In addition to the fairness among the selected users in a given time slot, the objective is to achieve fairness in throughput for all users through proportionally fair (PF) scheduling. Hence, the objective of the user scheduling is to acquire multiuser diversity gain while minimizing transmit power in a given time slot. Since the users are precoded successively, the users in a given subset \mathcal{U}_i have to be processed in some specified order. Let $\pi(\mathcal{U}_l) : \{u_1, u_2, ..., u_{K_0}\}$ denote a function assigning the processing orders of the users, which accounts for all $K_0!$ possible orders of $K_{\mathbb{C}}$ users in \mathcal{U}_i ; $u_k, k = 1, 2, ..., K_{\mathbb{C}}$ denote user indices. Then, the user and antenna selection problem for SO can be expressed as

$$\arg \max_{\theta_j, \mathbf{Q}_k(\theta_j), \mathcal{U}_l, \pi(\mathcal{U}_l)} \sum_{k \in \mathcal{U}_l} \mu(k) R_k(t)$$
(2.23)

such that trace $(\mathbf{Q}_k(\theta_j)) \leq P_k$, where $\mathbf{H}'_{k,e}(\theta_j) = \mathbf{Q}_{k,Z}^{-1/2} \mathbf{H}_k(\theta_j) \hat{\mathbf{V}}_i^{\mathsf{C}}$ and

$$R_{k}(t) = \log_{2} \det \left(\mathbf{I} + \mathbf{H}_{k,e}^{\prime}(\theta_{j}) \mathbf{Q}_{k}(\theta_{j}) \mathbf{H}_{k,e}^{\prime H}(\theta_{k}) \right)$$
(2.24)

and $\mathbf{Q}_{k,Z} = (\sigma_n^2 \mathbf{I} + \sum_{q=1}^{i-1} \mathbf{H}_k(\theta_j) \mathbf{W}_q \mathbf{W}_q^H \mathbf{H}_k^H(\theta_j))$ is the covariance of noise plus interference from users $u_1, u_2, ..., u_{i-1}$. $\mathbf{Q}_k(\theta_j)$ is the $n_k \times n_k$ input covariance matrix of user $k \in \mathcal{U}_i$ and $\hat{\mathbf{V}}_i^{\mathbb{C}}$ is the null space basis of the aggregate matrix of (i-1) previously selected users (see Sect. 2.2.2).

The optimal solution based on (2.23) is computationally very complex as it has to search through all possible combinations of users and antennas and each of such combinations has to be further searched through all possible user orders.² The total search size will thus be $\binom{K}{K_0}\binom{N_k}{n_k}K_0!$. Hence, simplified but suboptimal algorithms are of interest. We propose a greedy algorithm that works in a similar fashion as Algorithm-I discussed in the previous subsection. The algorithm summary is given in Table 2.2. The scheduling metrics and the explanation of each step of the algorithm follow.

Due to the nature of the proposed greedy algorithm, the power to be allocated to the u_i th user at the *i*th step of the algorithm u_i has to be decided before the user selection decision. One way to solve this problem is to search through all users and select the one that needs the least transmit power. This leads to overwhelming search complexity for the case with a large number of users, as water-filling power allocation has to be computed for each unselected user. To reduce this complexity, we propose to

² User ordering for successive zero-forcing precoding is discussed in [20] and references therein. User ordering algorithms are beyond the scope of this study. Hence, we will not take user ordering issue into account in our algorithm formulation.

allocate the power after a user has been selected. We propose two scheduling metrics as in the case of BD.

The algorithm works in a similar fashion as the Algorithm-I. The first metric is capacity based and the second is simplified Frobenius norm based. The capacity based scheduling metric (Metric 3) for successive optimization to be used in the *i*th step of the Algorithm-II (see Table 2.2) is expressed as,

$$\Omega_{1,k}(t) = \log \det(\mathbf{I} + \frac{P}{M_T} \mathbf{H}_k(\hat{\theta}_k) \mathbf{V}_i^0 \mathbf{V}_i^{0^H} \mathbf{H}_k^H(\hat{\theta}_k) \mathbf{Q}_{k,Z}^{-1}) \text{ for } \forall k \in \mathbb{U},$$
(2.25)

A user at the *i*th step of the algorithm is selected such that $u_i = \arg \max_{k \in \mathcal{U}_i} \mu_{1,k}(t) \Omega_{1,k}(t)$. Similarly, the first user u_1 in Step 1 of Algorithm-II is selected such that

$$u_1 = \arg\max_{k \in \mathcal{U}} \mu_{1,k}(t) \log_2 \det(\mathbf{I} + P/M_T \mathbf{H}_k \mathbf{H}_k^H)$$
(2.26)

The best subset of antennas of a user selected in each step of Algorithm-II is selected using receive antenna selection (RAS) algorithms proposed in Sect. 2.4. In Step 3 of the algorithm, user grouping is performed as discussed in Sect. 2.3.3. If there are some users in the group \mathbb{U}_i , the selection is carried out from that group. Otherwise, the algorithm ends with fewer users than maximum supported. In Step 3b of Algorithm-II, user u_i is selected such that $u_i = \arg \max_{k \in \mathbb{U}_i} \mu_{1,k}(t)\Omega_{1,k}(t), k \in \mathbb{U}_i$ using the metric defined in (2.25). Unlike in block diagonalization, once a user is selected, the precoding matrix and the rate of that user are computed at each step of the algorithm using the procedure described in Sect. 2.2.2. Then, average rate of that user can be updated using (2.11) and (2.12) and $\mu_{1,k}(t+1) = 1/R_k(t+1)^k$.

For large K, the logarithm, determinant and matrix inversion operations associated with Metric 3 may lead to high computational complexity. To simplify this, we propose to use fairness in the channel norm with a simplified Frobenius norm scheduling metric as follows. We call this Metric 4. Again, Metric 4 is of a heuristic nature. It is designed to reflect the ratio of the total power gains from the eigenmodes of the equivalent single user channel and equivalent interference power from the users selected up to (i - 1)th step of the algorithm. It is given as

$$\Omega_{2,k}(t) = \frac{\|\mathbf{H}_k(\hat{\theta}_k)\mathbf{V}_i^0\|_F^2}{\sum_{q=1}^{i-1} \|\mathbf{H}_k(\hat{\theta}_k)\mathbf{W}_q\|_F^2} \qquad \text{for } \forall k \in \mathbb{U},$$
(2.27)

A user in the *i*th step of Algorithm-II (see Table 2.2) is then selected such that $u_i = \arg \max_{k \in \mathbb{U}_i} \mu_{1,k}(t) \Omega_{1,k}(t)$. Similarly, at the Step 1 of Algorithm-II the first user u_1 is selected such that

$$u_1 = \arg\max_{\mathcal{U}} \mu_{2,k} ||\mathbf{H}_k||_F^2 \tag{2.28}$$

In Step 3 of the algorithm, user grouping is performed as discussed in Sect. 2.3.3. Similarly, Step 3b of Algorithm-II should be modified so that the user u_i is selected such that $u_i = \arg \max_{k \in \mathbb{U}_i} \mu_{2,k}(t) \Omega_{2,k}(t)$ with

Table 2.2 Simplified proportionally fair (PF) scheduling algorithm for successive optimization (SO): Algorithm-II (PF)

- 1. $i \leftarrow 1; \mathcal{U} = \{1, 2, ..., K\}; \mathcal{U}_s = \{\}; \Theta_k = \{1, 2, ..., N_k\}, k = 1, 2, ...K.$
 - Select a user such that $u_1 = \arg \max_{k \in \mathcal{U}} \mu_{1,k}(t) \log \det(\mathbf{I} + P/M_T \mathbf{H}_k \mathbf{H}_k^H)$ (Metric 3) $u_1 = \arg \max_{k \in \mathcal{U}} \mu_{2,k}(t) ||\mathbf{H}_k||_F^2$ (Metric 4)
 - Select n_{u1} receive antennas out of N_{u1} using proposed RAS algorithms. Let
 H'₁ = H_{u1}(θ̂_{u1}) denote the selected channel corresponding to the best
 subset θ̂_{u1} of user u₁.
 - Find H₁' = U₁Σ₁[V₁¹ V₁⁰]^H, W₁ = V₁¹Λ₁^{1/2}, where water-filling coefficients Λ₁ are chosen such that the rate requirement R_{u1} of u₁ is satisfied. Set U_s = U_s ∪ {u₁}; U₁ = U \{u₁}. Update the scheduling coefficients for user u₁.

2.
$$i \leftarrow i+1$$
; $\mathbf{H}(\mathcal{U}_s) = [\mathbf{H}_1^{'T}\mathbf{H}_2^{'T}...\mathbf{H}_{i-1}^{'T}]^T = \mathbf{U}_i \boldsymbol{\Sigma}_i [\mathbf{V}_i^1 \ \mathbf{V}_i^0]^H$.

3. If $(|\mathcal{U}_s| < K_{\mathbb{C}})$

a. Find
$$\mathbb{U}_i = \left\{ k \in \mathbb{U}_{i-1}, k \notin \mathcal{U}_s | \frac{\|\mathbf{H}_k \mathbf{v}_i^{\|H\|}}{\|\mathbf{H}_k\| \|\mathbf{v}_i^{\|}\|} < \xi \right\}.$$

- b. If $(|\mathbb{U}_i| \neq \mathbb{C})$
 - Select a user such that $u_i = \arg \max_{k \in U_i} \mu_{1,k}(t) \Omega_{1,k}(t)$ (Metric 3) $u_i = \arg \max_{k \in U_i} \mu_{2,k}(t) \Omega_{2,k}(t)$ (Metric 4)
 - Let $\mathbf{H}'_i = \mathbf{H}_{u_i}(\hat{\theta}_{u_i})$, $\mathbf{R}_i = \sigma_n^2 \mathbf{I} + \sum_{j=1}^{i-1} \mathbf{H}'_i \mathbf{W}_j \mathbf{W}_j^{II} \mathbf{H}'_i^{II}$. Compute $\mathbf{V}_i^{0^H} \mathbf{H}_i^{'II} \mathbf{R}_i^{-1} \mathbf{H}'_i \mathbf{V}_i^0 = \mathbf{M}_i \boldsymbol{\Sigma}_{w,i} \mathbf{M}_i^{II}$. With $\boldsymbol{\Sigma}_{w,i}$, compute waterfilling coefficients $\mathbf{A}_{w,i}$, such that rate requirement R_{u_i} is satisfied; Compute $\mathbf{W}_i = \mathbf{V}_i^0 \mathbf{M}_i \mathbf{A}_{w,i}^{1/2}$.
 - Set U_s = U_s ∪ {u_i}, U_i = U_i \{u_i}. Update the scheduling coefficients for user u_i.

Else end.

metric defined in (2.27). Once the power is allocated to the selected user, the actual total power gain of its channel (accounting for the power allocated) can be computed as trace $(\Sigma_{i,e}^{\prime 2} \Lambda_{w,i})$ (also see (2.9) and the discussion

that follows it), where $\Sigma'_{i,e}$ and $\Lambda_{w,i}$ are diagonal matrix of singular values of $\mathbf{H}'_{i,e}(\hat{\theta}_i)$ and diagonal matrix of power allocation coefficients obtained from waterfilling, respectively. Then, the average value of the scheduling metric (Metric 4) is updated as

$$\bar{\Omega}_{2,k}(t+1) = \delta \bar{\Omega}_{2,k}(t) + (1-\delta) \operatorname{trace}(\boldsymbol{\Sigma}_{k,e}^{\prime 2} \boldsymbol{\Lambda}_{k,w}), \quad k \in \mathcal{U}_{s}$$
(2.29)

$$\Omega_{2,k}(t+1) = \delta\Omega_{2,k}(t), \quad k \notin \mathcal{U}_s \tag{2.30}$$

and $\mu_{2,k}(t+1) = 1/\overline{\Omega}_{2,k}(t+1)$.

Since there are K_0 ! ways the users can be precoded, the order of processing the users is important. For a system with unequal user SNRs, precoding a user with the most attenuated channel first might yield power savings [5]. Hence, incorporating user ordering in the user selection process may further enhance the performance in terms of transmit power minimization while maintaining the quality of service to users. However, detailed performance evaluation with various user ordering algorithms is beyond the scope of this study.

On the other hand, the base station can adaptively allocate different numbers of spatial modes (or data streams) to users to further enhance system throughput. With M_T antennas, base station can serve up to $K_0 = M_T$ users in a given time slot by allocating one stream to each user. Simplified algorithms to do so are of great importance for downlink MIMO multiuser schedulers. Similarly, if the base station and users can coordinate their processing, further increase in throughput can be obtained. For such systems, the overhead involved in the exchange of information about transmit and receive processing matrices should also be evaluated.

2.3.3 Subspace Correlation Based User Grouping

The user grouping strategies proposed in [21,22,23] have focused on the cases where all possible space-division multiplexing groups are formed and one of such groups is selected. Even though those proposals may be effective for small K, they become computationally very complex for large K, which is of interest in our study. Instead of first forming smaller subsets of all active users and exhaustively searching through the subsets, our approach is to dynamically adjust the intermediate group size at the *i*th step of the algorithm. Hence, the user grouping needs to be done only once at each step of the algorithm. The intermediate user group size decreases as the user selection progresses. A similar approach is taken in [7] for single antenna users and zero-forcing beamforming. The main idea behind our

user grouping mechanism is that the regular greedy user selection algorithms [13] have to search through K - i users to find the next best user at the *i*th step of the algorithm, which may still be too complex for large K. However, we show in [16] that the intermediate user grouping approach significantly reduces the complexity compared to regular greedy user selection algorithms.

The proposed user grouping algorithm works as follows (please refer to the algorithm summarized in Table 2.1). The *i*th step of the proposed algorithm forms a group \mathbb{U}_i of users less correlated with the subspace of the aggregate channel of the users selected up to (i - 1), th step based on some specified threshold ξ . The reason for such approach is that the sufficient condition for the rank of the equivalent single user channel $\mathbf{H}_{k,\epsilon}$ to be greater than one (to make sure the transmission to user k is possible) is that at least one row of \mathbf{H}_k should be linearly independent of the rows of \mathbf{H}_k . To meet this requirement, we should avoid multiplexing users with highly correlated channels [5]. Similarly, reference [19] shows that if the user channels are orthogonal to each other, then block diagonalization achieves the sum capacity for a given channel. Hence, it is justified to group the users that are spatially less correlated with each user (or spatially more compatible) for throughput maximization. At the *i*th step of the algorithm the user grouping is performed such that

$$\mathbb{U}_{i} = \left\{ k \in \mathbb{U}_{i-1}, k \notin \mathcal{U}_{s} | \frac{\|\mathbf{H}_{k}\mathbf{V}_{i-1}^{H}\|}{\|\mathbf{H}_{k}\|\|\mathbf{V}_{i-1}\|} < \xi \right\}$$
(2.31)

Equation (2.31) indicates that the threshold selection is critical to avoid the loss in sum rate. Smaller thresholds result in smaller user subset at the *i*th step of the algorithm, which in turn lowers the search complexity. However, too small threshold incurs throughput loss due to the loss in multiuser diversity (due to smaller pool of users in the search size). The objective is to select thresholds that yield throughput closer to full search greedy selection (i.e. with $\xi = 1$). Yoo and Goldsmith in [7] derive bounds on the loss of the channel gain and the multiuser diversity gain due to threshold selection for single antenna users and zero-forcing beamforming. This is, however, very difficult in case of multiantenna users. Nevertheless, [7] also uses the optimal thresholds obtained through simulation in its numerical results. We study the threshold selection issue through simulation in Sect. 2.5.1.

2.4 Impact of Receive Antenna Selection (RAS)

Antenna selection is a simple but promising technique for hardware complexity reduction at the transmitter, receiver or both. There has been a considerable research effort put into antenna selection techniques for single user MIMO systems. Receive antenna selection (RAS) for multiuser MIMO system has been discussed by several researchers recently. For example, impact of RAS on the sum rate of MIMO broadcast channel with block diagonalization has been discussed in [24]. Similarly, we have studied the impact of antenna selection on the performance of multiuser MIMO orthogonal space-division multiplexing systems with block diagonalization and successive optimization in conjunction with downlink multiuser MIMO scheduling in [16]. Details of antenna selection algorithms and performance analysis will be discussed in Sect. 2.5, but first we outline main advantages of RAS in the downlink of multiuser MIMO systems employing block diagonalization. Similar arguments apply for other orthogonal linear precoding techniques as well.

Consider a case where n_k antennas out of N_k antennas are selected at the *k*th receiver. Then, the dimension of the aggregate channel $\mathbf{\bar{H}}_j$ in (2.4) for all the other users $j \neq k$ is

$$\sum_{i=1,i\neq j}^{K} n_i \times M_T \tag{2.32}$$

which means the dimension of the equivalent single user channel $\mathbf{H}_{k,\epsilon}$ is

$$n_k \times (M_T - \sum_{i=1, i \neq k}^K n_i)$$
 (2.33)

From (2.33) we see that the row dimension of the *k*th user's channel matrix has diminished but at the same time the column dimension increases by up to $\sum_{i=1,i\neq k}^{K} (N_i - n_i)$. The increase in column dimension may provide some diversity advantage and yield higher singular values of the projected channel, which helps to minimize the rate loss of kth user without receive antenna selection (RAS) (note here that the removal of receive antennas means individual rate loss for the kth user). Overall gain in sum rate has been demonstrated in [13] and [24]. On the other hand, RAS allows transmit precoding schemes to accommodate more users, which may translate into improved fairness in throughput among users. As discussed above, RAS at some of the users releases transmission resources that may be utilized to serve other more needy users in the system. For example, for BD with $M_T = 4$, $N_k = 4$, only one user can be served at a time (assuming i.i.d. Gaussian channel). However, when RAS is used such that each user selects $n_k = 2$ antennas out of $N_k = 4$ available antenna elements, two users can be served at a time. When RAS is combined with user scheduling, it has been shown that RAS improves the multiuser diversity gain that is achieved with user scheduling [16]. A similar observation is made in [25] in a situation where a random beamforming technique is combined with RAS. With opportunistic beamforming, multiuser diversity greatly improves the system performance for sufficiently large mobile user population. However, in cases where the user population is not sufficiently large for opportunistic beamforming to perform well, combining it with RAS is an effective means to enhance the system throughput. Intuitively, employing multiple receive antennas at mobiles can be viewed as increasing the number of users in the system (each antenna can be thought of as a single antenna user).

We propose two receive antenna selection (RAS) algorithms to work in conjunction with the proposed greedy scheduling algorithms. The first is a Frobenius-norm based exhaustive search selection, which selects the best subset of receive antennas of a user that maximizes the Frobenius norm of the null space projected channel. The second takes an incremental subspace projection approach, which is similar to the greedy single user selection algorithm [8].

2.4.1 Receive Antenna Selection Algorithm 1 (RAS-I)

RAS-I selects antennas of the first selected user in Step 1 of the algorithm as $\hat{\theta}_{u_1} = \arg \max_{\theta_j \subset \Theta_{u_1}} \|\mathbf{H}_{u_1}(\theta_j)\|_F^2$. At the *i*th step, the best antenna subset is selected as

$$\hat{\theta}_k = \arg \max_{\theta_j \subset \Theta_k} \|\mathbf{H}_k(\theta_j) \mathcal{P}_i^{\perp}\|_F^2$$
(2.34)

This algorithm exhaustively searches the best antenna subset that provides the maximum Frobenius-norm of the projection to the null space of the aggregate channel of the previously selected users. This may be used when users have small number of receive antennas, which is the assumption of this Chapter. Several other antenna selection algorithms proposed for single user MIMO systems may also be employed to the equivalent single user channel or the null-space projected channel $\mathbf{H}_k \mathcal{P}_i^{\perp}$ (e.g. incremental or decremental receive antenna selection algorithms in [26] for single user MIMO systems).

2.4.2 Receive Antenna Selection Algorithm 2 (RAS-II)

RAS-II takes incremental approach similar to that used in user selection, where one antenna is selected at a time until n_k antennas have been selected.

Let us denote the rows of \mathbf{H}_k as $\mathbf{h}_{k,1}, \mathbf{h}_{k,2}, ..., \mathbf{h}_{k,N_k}$. At the first step of algorithm, RAS-II starts by selecting an antenna (a row of \mathbf{H}_k) with maximum Frobenius norm. Then, r_j th antenna, $j = 2, ..., n_k$, that yields the maximum Frobenius norm of the projection to the null space of the aggregate channel $\mathbf{H}(\mathcal{U}_s, k) = [\mathbf{h}_{k,r_1}^T, ..., \mathbf{h}_{k,r_{j-1}}^T]^T$ is selected as,

$$r_{j} = \operatorname*{argmax}_{1 \le n \le N_{k} - j + 1} \|\mathbf{h}_{k,n}(\mathbf{I}_{M_{T}} - \mathbf{V}_{k,j}^{H}\mathbf{V}_{k,j})\|_{F}^{2}$$
(2.35)

where $\mathbf{V}_{k,j}$ is the orthogonal basis of $\mathbf{H}(\mathcal{U}_s, k)$. At the *i*th step of the algorithm, antenna selection follows (2.35) with $\mathbf{H}(\mathcal{U}_s, k) = [\mathbf{H}(\mathcal{U}_s)^T, \mathbf{h}_{k,r_1}^T, ..., \mathbf{h}_{k,r_{j-1}}^T]^T$. When one antenna is added, basis vectors for the null space of the aggregate channel matrix can be recursively computed using Gram-Schmidt orthogonalization as

$$\mathbf{V}_{k,j} = [\mathbf{V}_{i-1}^T, \mathbf{v}_{r_j}^T]^T$$
(2.36)

where \mathbf{v}_{r_s} is computed as

$$\mathbf{v}_{r_j} = \frac{(\mathbf{h}_{r_j} - \sum_l \mu_{jl} \mathbf{v}_l)}{(||\mathbf{h}_{r_j} - \sum_l \mu_{jl} \mathbf{v}_l||)}$$
(2.37)

 \mathbf{v}_i denotes the *l*th row of \mathbf{V}_{i-1} and $\mu_{jl} = |\mathbf{h}_{r_j} \mathbf{v}_l^H$. This algorithm may reduce complexity compared to RAS-I for the cases with large N_k . Also, since the algorithm selects the least correlated antenna at each step, it is more suitable for spatially correlated channels.

2.5 Simulation Results

First, the effect of the correlation threshold, ξ , on the average sum rate of the system is discussed. Then, sum rate versus the number of users for block diagonalization (BD) and successive optimization (SO) are presented. We assume that all users have same number of receive antenna elements and radio frequency chains.

2.5.1 Correlation Threshold

A detailed discussion of the effect of threshold selection has been presented in [16] for homogenous received SNRs (users with same average received SNR). Smaller thresholds result in small user subset at each step of the algorithm, which in turn lowers the search complexity. However, too small threshold incurs a throughput loss due to a loss in multiuser diversity. The objective is to select thresholds that yield throughput close to full search greedy user selection ($\xi = 1$). The threshold values are mainly determined by the number of transmit antennas, number of receive antennas at mobiles and the number of active users. For example, for BD, the threshold in the range of 0.55–0.47 may be selected for $M_T = 4$ and K = 10-80 at SNR = 10 dB [16]. For $M_T = 8$, it ranges from 0.45 to 0.4 for K from 10 to 80. Similarly, the range of thresholds for SO has been reported in [16] to be from 0.475 to 0.38 for K from 10 to 80. In our simulations, we use the optimal threshold³ values obtained through simulations in [16].

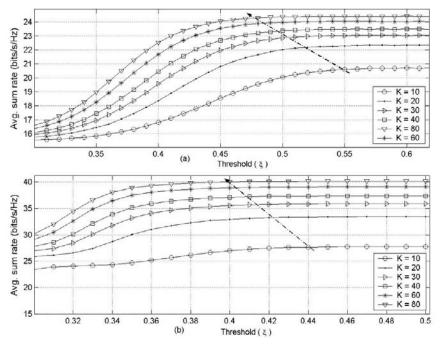


Fig. 2.2 Average sum rate with block diagonalization versus threshold ξ ; RAS-I;; asymmetric received SNRs: user SNRs range from 0 to 20 dB on a log-linear scale; (a) $M_T = 4$, $N_k = 4$, $n_k = 2$, $K_0 = 2$; (b) $M_T = 8$, $N_k = 4$, $n_k = 2$, $K_0 = 4$. K denotes the number of active users. The slanted *dash-dot line* with an arrow indicates the thresholds that yield similar throughput as a full search greedy algorithm

³ The optimal threshold is a threshold that maximizes the average sum rate. For the fading environment under consideration a look-up table, which relates the number of active users to the optimal threshold is generated through simulation. In each scheduling interval, the scheduling decision involves the selection of threshold from the look-up table. However, a different threshold for each packet is not necessary (e.g. if multiple slots are allocated to a user).

To study the threshold range for asymmetric received user SNRs (i.e. users have different average received SNRs), we plot the average sum rate with block diagonalization (BD) versus the threshold ξ in Fig. 2.2. To simulate this case, we assume that K users' SNRs range from 0 to 20 dB on a log-linear scale. For example, when K = 30 the first user's SNR = 0 dB and the 30th user's SNR = 20 dB. Figure 2.2 shows that the range of threshold (with respect to the number of users) for asymmetric user SNRs case is not significantly different from the symmetric user SNRs case in [16]. This is because the spatial correlation (2.31) is normalized by the channel norm (the channel attenuation of user k is reflected in the elements of \mathbf{H}_k). To achieve the throughput closer to that obtained with full search greedy selection algorithms [13], the threshold range similar to symmetric received SNRs case in [16] can be selected for asymmetric receive SNRs case as well. Even though the decreasing threshold trend with the increasing number of active users has been observed to be valid for both cases, our other simulation results have shown that the optimal thresholds (thresholds that yield the maximum throughput with the proposed simplified greedy scheduling algorithm) are different for symmetric and asymmetric user SNR cases. However, to achieve throughput similar to that with full search greedy algorithms, we conjecture that the range of thresholds for symmetric user SNRs case can also be used for asymmetric SNRs case.

2.5.2 Average Sum Rate Performance

In Figs. 2.3 and 2.4, we compare various downlink multiuser MIMO scheduling strategies for block diagonalization (BD) and successive optimization (SO), respectively. For these plots we assume that the average received SNR is equal for all users (the channels are symmetric) and it is subject to independent Rayleigh fading. The plots are obtained by averaging over 5000 independent channel realizations and the optimal value of ξ is used in all simulations. The averaging of present and past throughputs is done over a sliding window 100 time slots long, so that the forgetting factor δ is equal to 0.99 [4]. In Fig. 2.3 we compare the performance of the proposed scheduling metrics for proportionally fair scheduling and also present the performance with maximum sum rate scheduling (denoted as Max-rate sch. in the figures; it corresponds to $\mu_k(t) = 1, \forall k$) for $M_T = 4$ and $M_T = 8$ at SNR of 10 dB. Fig. 2.3a also presents the performance with optimal Max-rate scheduling for BD (exhaustive search user selection). The performance of RAS-I and RAS-II have been found to be very similar

for all MIMO antenna configurations considered in the simulations. Therefore, only the performance with RAS-I is presented in most of the plots.

For block diagonalization (BD), performance with proportionally fair (PF) scheduling Metric 1 is found to be marginally better in case of a lower number of transmit antennas (e.g. $M_T = 4$) and the difference between the performance of the two metrics widens for higher number of transmit antennas (e.g. $M_T = 8$, see Fig. 2.3). For the $M_T = 8$ case, PF scheduling with Metric 1 is found to yield 1 bps/Hz better sum rate compared to PF scheduling with Metric 2. However, the complexity of the PF scheduling with Metric 1 is higher than that with Metric 2. The proposed algorithms perform close to exhaustive search selection with much lower complexity. The spectral efficiency loss compared to exhaustive search selection is only about 1 bps/Hz for $M_T = 4$ but it increases for larger M_T (e.g. $M_T =$ 8). However, the proposed algorithms are much less complex than the optimal ones. Obviously, the loss in sum rate with PF scheduling compared to the max-rate scheduling comes with a gain in fairness. As seen from the figures, the loss is about 1 bps/Hz at SNR = 10 dB for both $M_T = 4$ and $M_T = 8$ cases.

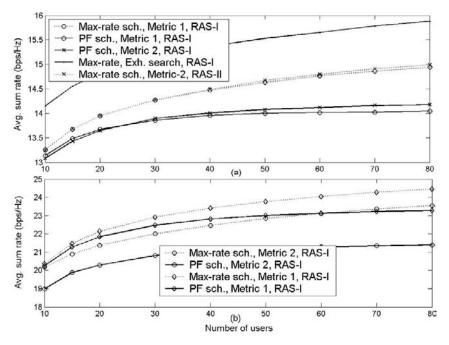


Fig. 2.3 Average sum rate versus K with symmetric channels and block diagonalization; Algorithm-I; SNR = 10 dB. (a) $M_T = 4$, $N_k = 4$, $n_k = 2$, $K_0 = 2$; (b) $M_T = 8$, $N_k = 4$, $n_k = 2$, $K_0 = 4$

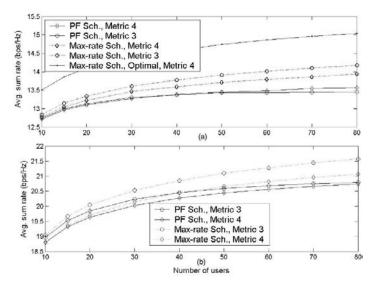


Fig. 2.4 Average sum rate versus K with symmetric channels and successive optimization; equal power allocation to the selected users; Algorithm-II, RAS-I, SNR = 10 dB. (a) $M_T = 4$, $N_k = 4$, $n_k = 2$, $K_0 = 2$; (b) $M_T = 8$, $N_k = 4$, $n_k = 2$, $K_0 = 4$

In Fig. 2.4, we show the performance of the proposed scheduling metrics for successive optimization (SO) (see Algorithm-II in Table 2.2). For this simulation we consider a simple case, where equal power is allocated to the selected users. In general, the rate requests and power allocations to users are different, which in real systems would involve other factors like adaptive modulation and coding, the number of transmit streams (or spatial modes) to be allocated to users, etc. A closer look at the plots in Fig. 2.4 reveals that the difference between the sum rate of maxrate scheduling and PF scheduling strategies has decreased (compared to the block diagonalization (BD) case; see Fig. 2.3) even for the simple case of equal power allocation to the selected users. This is not unexpected as SO is supposed to improve fairness compared to BD.

To compare fairness achieved through the proposed proportionally fair (PF) scheduling algorithms, we plot the average rates of individual users in Figs. 2.5 and 2.6 for block diagonalization (BD) and successive optimization (SO), respectively. To study the fairness for the users with asymmetric SNRs, for this particular case we consider a pool of 40 users and the average received SNRs of users ranging from 0 to 20 dB on a log-linear scale as in [7]. This means that user 1 has SNR of 0 dB and user 40 has SNR of

20 dB. As observed in Fig. 2.5, max-rate scheduling favors the users with high SNRs. For example, user 40 enjoys much higher rate compared to virtually no transmission to users with low SNRs (e.g. users 1–15). The proposed PF scheduling algorithm mitigates this unfairness.

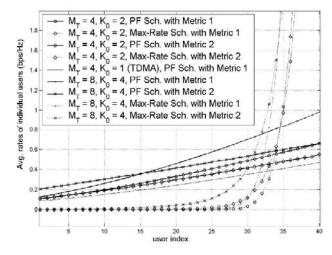


Fig. 2.5 Average rates of individual users with proposed scheduling and antenna selection algorithm with block diagonalization (Algorithm-I); K = 40. RAS-I used in all cases. *Solid lines* denote the performance with proportionally fair scheduling and the *dotted lines* correspond to the performance with max-rate scheduling

The use of Metric 2 is found to enhance average rate to low SNR users more than with Metric 1, particularly in a case of higher number of transmit antennas. For SO, on the other hand, the fairness achieved with Metric 4 has been found to be much better than that with Metric 3. Even with max-rate scheduling, Metric 4 has outperformed Metric 3 for both antenna configurations (see Fig. 2.6) for successive optimization.

Similarly, the average sum rate versus the number of active users K for block diagonalization (BD) is plotted in Fig. 2.7 for asymmetric average received SNR case. For this simulation, we use a simple model to realize the asymmetric average received SNRs. We assume that the average received SNRs of users range from 0 to 20 dB (on a log-linear scale) irrespective of the number of active users; i.e. when K = 10, user 1 has SNR of 0 dB and user 10 has the SNR of 20 dB. Similarly, when K = 80, user 1 has SNR of 0 dB and user 80 has the SNR of 20 dB. All other configurations for the proportionally fair (PF) scheduling are the same as in the previous examples. From Fig. 2.7 we see that Metric 1 performs slightly better than Metric 2. A large difference between sum rates with max-rate and

PF scheduling is due to the fact that low SNR users are virtually not getting any throughput with max-rate scheduling. Since higher SNR users are being selected, sum rate is higher with max-rate scheduling. PF scheduling improves fairness while exploiting the multiuser diversity gain.

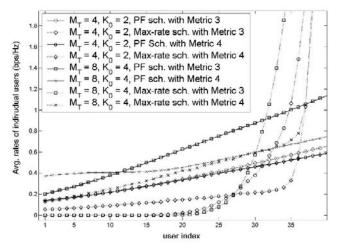


Fig. 2.6 Average rates of individual users with proposed scheduling and antenna selection algorithm with successive optimization (Algorithm-II); K = 40. RAS-I used in all cases. *Solid lines* denote the performance with proportionally fair scheduling and the *dotted lines* correspond to the performance with max-rate scheduling

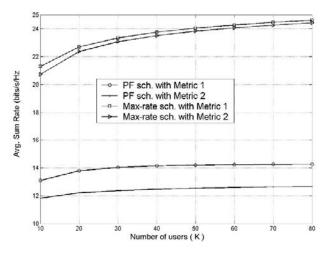


Fig. 2.7 Average sum rate versus the number of active users K for block diagonalization; The average SNRs of users range from 0 to 20 dB (on a log-linear scale)

2.6 Conclusions

We have investigated channel-aware joint multiuser scheduling and antenna selection problem for two downlink multiuser MIMO transmission strategies: block diagonalization and successive optimization. Simplified greedy algorithms for proportionally fair scheduling have been proposed and analyzed. Since the user rate at the *i*th step of the proposed greedy algorithm for block diagonalization is unknown before all users have been selected, simplified proportionally fair scheduling metrics have been proposed. Similarly, for successive optimization a scheduling metric based on approximate capacity of an equivalent single user MIMO channel with cochannel interference has been proposed. To further simplify the algorithm a channel matrix Frobenius norm based proportionally fair scheduling metric is also proposed, which is shown to perform close to the capacity based scheduling metric. To reduce the search complexity further for a large number of users K, a user grouping technique is used so that the search through K - i users at the *i*th step of the algorithm is avoided. Proposed algorithms have been demonstrated to improve fairness among users with unequal average received SNRs.

Scheduling algorithms and impact of antenna selection on the performance of other multiuser MIMO precoding algorithms, including nonlinear precoding, are interesting topics for further study. Similarly, low complexity algorithms to incorporate user ordering into scheduling for successive optimization are important. Application of the proposed algorithms to allocate subcarriers in MIMO-OFDM transmission is also of interest. Similarly, scheduling and antenna selection with partial channel state information at the transmitter is an important area of research that has gained significant attention and needs to be developed.

Acknowledgements

Funding for this work provided by TRLabs, the Rohit Sharma Professorship, and the Natural Sciences and Engineering Research Council (NSERC) of Canada is gratefully acknowledged.

References

- 1. M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.
- 2. G. Caire, "MIMO downlink joint processing and scheduling: A survey of classical and recent results," in *Proc. Workshop on Information Theory and Its Applications*, San Diego, CA, Jan 2006.
- 3. G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 7, pp. 1691–1706, July 2003.
- H. Viswanathan, S. Venkatesan, and H. Huang, "Downlink capacity evaluation of cellular networks with known-interference cancellation," *IEEE J. Select. Areas Commun.*, vol. 21, no. 5, pp. 802–811, June 2003.
- Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Trans. Signal Processing*, vol. 52, no. 2, pp. 461–471, Feb 2004.
- 6. L. Choi and R. D. Murch, "A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 20–24, Jan 2004.
- T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 528–541, March 2006.
- 8. G. Dimic and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Trans. Signal Processing*, vol. 53, no. 10, pp. 3857–3868, Oct 2005.
- 9. Z. Tu and R. S. Blum, "Multiuser diversity for a dirty paper approach," *IEEE Commun. Lett.*, vol. 7, no. 8, pp. 370–372, Aug 2003.
- 10. R. Rao and A. Eltawil, "Introduction to OFDM and MIMO communication systems with emphasis on DVB-H, 802.11n and 802.16e," Sept 2006, VTC'06-Fall Tutorial.
- G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, no. 3, pp. 311–335, Mar 1998.
- F. R. Farrokhi, G. J. Foschini, A. Lozano, and R. A. Valenzuela, "Linkoptimal space-time processing with multiple transmit and receive antennas," *IEEE Commun. Lett.*, vol. 5, no. 3, pp. 85–87, Mar 2001.
- 13. S. Sigdel and W. A. Krzymień, "Antenna and user subset selection in downlink multiuser orthogonal space-division multiplexing," in *Proc. Wireless Personal Multimedia Commun. Conf. (WPMC'06)*, Sept 2006, pp. 671–675.
- Z. Shen, R. Chen, J. G. Andrews, J. R. W. Heath, and B. L. Evans, "Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization," *IEEE Trans. Signal Processing*, vol. 54, no. 9, pp. 3658–3663, Sept 2006.

- P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.
- S. Sigdel and W. A. Krzymień, "Simplified antenna selection and user scheduling for orthogonal space-division multiplexing," in *Proc. IEEE WCNC'07*, Mar 2007, pp. 1728–1733.
- R. Chen, R. W. Heath, and J. G. Andrews, "Transmit selection diversity for unitary precoded multiuser spatial multiplexing systems with linear receivers," *IEEE Trans. Signal Processing*, vol. 55, no. 3, pp. 1159–1171, Mar 2007.
- G. Primolevo, O. Simeone, and U. Spagnolini, "Channel aware scheduling for broadcast MIMO systems with orthogonal linear precoding and fairness constraints," in *Proc. IEEE ICC'05*, May 2005, pp. 2749–2753.
- Z. Shen, R. Chen, J. G. Andrews, R. W. Heath, and B. L. Evans, "Sum capacity of multiuser MIMO broadcast channels with block diagonalization," in *Proc. IEEE ISIT*'06, July 2006, pp. 886–890.
- P. Tejera, W. Utschick, G. Bauch, and J. A. Nossek, "Subchannel allocation in multiuser multiple-input multiple-output systems," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4721–4733, Oct 2006.
- 21. F. Shad, T. D. Todd, V. Kezys, and J. Litva, "Dynamic slot allocation (DSA) in indoor SDMA/TDMA using a smart antenna base station," *IEEE/ACM Trans. Networking*, vol. 9, no. 1, pp. 69–81, Feb 2001.
- 22. D. J. Mazzarese and W. A. Krzymień, "Scheduling algorithms and throughput maximization for the downlink of packet-data cellular systems with multiple antennas at the base station," *Wireless Personal Commun.*, vol. 43, no. 2, pp. 215–260, Oct 2007.
- Q. H. Spencer and A. L. Swindlehurst, "Channel allocation in multi-user MIMO wireless communication systems," in *Proc. IEEE ICC'04*, June 2004, pp. 3035–3039.
- B. C. Lim, C. Schlegel, and W. A. Krzymień, "Sum rate maximization and transmit power minimization for multi-user orthogonal space division multiplexing," in *Proc. IEEE GLOBECOM'06*, Nov 2006. pp. 1–5.
- 25. L. Zan and S. A. Jafar, "Combined opportunistic beamforming and receive antenna selection," in *Proc. IEEE WCNC'05*, Mar 2005, pp. 1007–1011.
- A. Gorokhov, D. A. Gore, and A. J. Paulraj, "Receive antenna selection for MIMO spatial multiplexing: Theory and algorithms," *IEEE Trans. Signal Processing*, vol. 51, no. 11, pp. 2796–2807, Nov 2003.

3 On the Impact of Channel Estimation and Quantization Errors for Various Spatio-Temporal Transmission Schemes

Kai Yu and Angeliki Alexiou

3.1 Introduction

Recently, wireless broadband systems equipped with antenna arrays have attracted much attention both in academia and industry. They can provide diversity gain and/or spatial multiplexing (SM) gain comparing to the traditional systems using single transmit/receive antenna, hence increase the system capacity [4,10] and/or lower the bit error rate (BER). Some popular spatial-temporal transmission schemes proposed before include orthogonal space-time block codes (OSTBCs) [1,9], SM scheme [3,16] and hybrid transmission scheme [8,11] that balances the benefit of diversity gain and SM gain. To further improve the performance, an algorithm has been proposed in [5] to optimally switch between the SM scheme and the Alamouti scheme.

The spatio-temporal transmission schemes mentioned above are designed for single link multiple-input multiple-output (MIMO) transmissions. When the base station is equipped with multiple antennas, it can also communicate with multiple users simultaneously, so called multi-user downlink MIMO transmissions. Different approaches have been proposed including both the linear (such as zero-forcing precoding) [7,13] and nonlinear (such as MMSE THP precoding) [13] approaches.

All schemes mentioned above require channel state information (CSI) at the receive side, and some require CSI at the transmitter as well. In real systems, the channel estimation procedure can never be perfect, and therefore there always exist the channel estimation errors. The main sources of channel estimation errors include the noise in the estimation procedure, the interpolation process, the quantization, and the feedback delay. The first two sources have impact over the whole system, while the last two sources impact mainly at the transmitter. In [17], the impact of channel estimation error at the receiver has been studied together with the V-BLAST receiver. In this chapter, we further study the impact of channel estimation error and channel quantization errors on various single link MIMO transmission schemes as well as the multi-user MIMO transmission methods.

This chapter is organized as follows. Section 3.2 describes several spatio-temporal transmission schemes for single link MIMO transmission. The zero-forcing (ZF) precoding for multi-user downlink MIMO transmission is briefly described in Sect. 3.3. In Sects. 3.4 and 3.5, we study the impact of channel estimation errors and channel quantization errors at the receiver and the transmitter respectively. The simulation results are also presented in this section. Finally, we conclude and make some further comments in Sect. 3.6.

3.2 Single Link MIMO Transmission

In this section, we briefly describe several spatio-temporal transmission schemes for single link MIMO transmission, namely the spatial multiplexing, the Alamouti scheme, the hybrid transmission scheme and the dominant eigenmode transmission.

3.2.1 System Model

Assume there are N_t transmit antennas and N_r receive antennas. The input-output relationship for a narrow band system can be expressed in the baseband as

$$\mathbf{y} = \sqrt{\frac{E_s}{N_t}} \mathbf{H} \mathbf{s} + \mathbf{n} , \qquad (3.1)$$

where E_s is the total average energy transmitted over a symbol period, **H** is the $N_r \times N_t$ channel matrix, **s** is the transmitted signal vector, **y** is the received signal vector, and **n** is the additive white Gaussian noise with variance σ^2 .

3.2.2 Transmission Schemes

3.2.2.1 Spatial Multiplexing

The SM scheme transmits data symbols over parallel spatial subchannels and achieves high data rate. This has been demonstrated in [3,16] where the Bell Labs Layered Space-Time (BLAST) architectures were proposed along with a coding and decoding scheme. In this chapter, we assume the transmitter has no CSI, therefore the power is equally allocated to each transmit antennas. Furthermore, we assume the receiver only knows the estimated CSI and study the impact of channel estimation error on the SM scheme using four different types of receivers. We briefly describe these receivers below.

Zero-Forcing Receiver

The ZF receiver belongs to the linear receiver, and can be expressed as [6]

$$\mathbf{G}_{ZF} = \sqrt{\frac{E_s}{N_t}} \mathbf{H}^+, \qquad (3.2)$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudo inverse.

Minimum Mean Square Error Receiver

Another type of linear receiver is the minimum mean square error (MMSE) receiver, which can be expressed as [6]

$$\mathbf{G}_{MMSE} = \sqrt{\frac{E_s}{N_t}} (\mathbf{H}^H \mathbf{H} + \frac{N_t}{\rho} \mathbf{I}_{N_t})^{-1} \mathbf{H}^H , \qquad (3.3)$$

where ρ is the average SNR at the receive side, and $(\cdot)^{H}$ denotes complex conjugate transpose.

V-BLAST Receiver

In [16], the non-linear vertical BLAST (V-BLAST) receiver has been proposed. The main idea is to successively decode the symbols layer by layer. By using symbol cancellation, the interference from the decoded symbols is removed. This approach can be combined with either the ZF or MMSE receivers mentioned above. See [16] for more details on V-BLAST receiver.

3.2.2.2 Alamouti Scheme

The Alamouti transmission scheme [1] is designed to exploit the transmit diversity for any system with 2 transmit antennas. It is a special case of OSTBCs [9] which can exploit diversity for any number of transmit antennas. Note that the Alamouti code is the only full rate code available among the OSTBCs [9]. In this chapter, we focus on studying a 2×2 system using Alamouti scheme.

Assume the channel remains constant for two consecutive symbol periods, the input-output relationship for the Alamouti scheme can be written as

$$\mathbf{Y} = \sqrt{\frac{E_s}{2}} \mathbf{H} \begin{bmatrix} s_1 & -s_2^* \\ s_2 & s_1^* \end{bmatrix} + \mathbf{N}, \qquad (3.4)$$

where $(\cdot)^*$ denotes complex conjugate, **Y** is the received signal matrix, and **N** is the noise matrix.

The above expression can be rewritten using the effective channel matrix as

$$\mathbf{y}_{eff} = \sqrt{\frac{E_s}{2}} \mathbf{H}_{eff} \mathbf{s} + \mathbf{n}_{eff} , \qquad (3.5)$$

where $\mathbf{y}_{eff} = [\mathbf{Y}(:,1)^T, \mathbf{Y}(:,2)^T]^T$, $\mathbf{N}_{eff} = [\mathbf{N}(:,1)^T, \mathbf{N}(:,2)^T]^T$, $\mathbf{s} = [s_1, s_2]^T$, (·)^T is transpose, and the effective channel matrix \mathbf{H}_{eff} is

$$\mathbf{H}_{eff} = \begin{bmatrix} h_{1,1} & h_{1,2} \\ h_{2,1} & h_{2,2} \\ h_{1,2}^* & -h_{1,1}^* \\ h_{2,2}^* & -h_{2,1}^* \end{bmatrix}.$$
(3.6)

Using the fact that \mathbf{H}_{eff} is orthogonal irrespective of real channel matrix \mathbf{H} , the symbols can be easily decoded after multiplexing the received signal with \mathbf{H}_{eff}^{H} , i.e. $\mathbf{z} = \mathbf{H}_{eff}^{H} \mathbf{y}$.

3.2.2.3 Hybrid Transmission Scheme

For MIMO systems equipped with 4 transmit antennas, instead of transmitting using the well-known SM scheme or OSTBC scheme, a hybrid MIMO transmission scheme has been proposed [18] in order to obtain a good balance between the SM gain and the diversity gain. The main idea is to group two antennas together so that there are two sets of antenna pair. The SM scheme is used between two sets of antenna pairs. While within each antenna pair, the Alamouti scheme is deployed. This hybrid scheme has also been proposed in the 3GPP meeting [11].

Assume the MIMO channels are stationary for two consecutive symbol periods. In the first symbol period, the transmitted symbol vector is $\mathbf{s} = [s_1, s_2, s_3, s_4]^T$. The transmitter transmits $\mathbf{\tilde{s}} = [s_2^*, -s_1^*, s_4^*, -s_3^*]^T$ in the second symbol period. Using the effective MIMO channel matrix \mathbf{H}_{hyb} , the input-output relationship can be written as [18]

$$\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2^* \end{bmatrix} = \sqrt{\frac{E_s}{4}} \mathbf{H}_{hyb} \mathbf{s} + \begin{bmatrix} \mathbf{n}_1 \\ \mathbf{n}_2^* \end{bmatrix}, \qquad (3.7)$$

where the effective MIMO channel matrix \mathbf{H}_{hyb} is

$$\mathbf{H}_{hyb} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \mathbf{h}_3 & \mathbf{h}_4 \\ \mathbf{h}_2^* & -\mathbf{h}_1^* & \mathbf{h}_4^* & -\mathbf{h}_3^* \end{bmatrix},$$
(3.8)

 \mathbf{r}_i is the received vector at symbol period *i* (*i* = 1, 2), and \mathbf{n}_i is the noise vector at symbol period *i*. At the receive side, a linear ZF/MMSE or V-BLAST receiver can be used to decode the received signals [18].

The hybrid MIMO transmission scheme described above does not select specific antennas to form the antenna pairs. Hence it does not provide optimal balance between the SM gain and diversity gain. In [11], the antenna shuffling has been presented based on long-term channel statistics to further improve the system performance. Reference [8] proposed an antenna shuffling approach based on maximizing the minimum post-processing SNR using the long-term channel information.

One method to obtain the optimal antenna pairs is based on the capacity of the effective MIMO channel matrix, i.e.

$$[M_1, M_2] = \underset{M_1, M_2}{\operatorname{arg\,max}} \log_2 \det(\mathbf{I}_4 + \frac{\rho}{4} \mathbf{H}_{hyb} \mathbf{H}_{hyb}^H), \qquad (3.9)$$

where det(·) denotes determinant, ρ is the average SNR at the receive side, M_1 and M_2 represent two sets of antenna pairs.

In real system when the exact value of the effective channel capacity is not a concern, the above equation can be rewritten as below to reduce the computation complexity, i.e.

$$[M_1, M_2] = \underset{M_1, M_2}{\operatorname{arg max}} \det(\mathbf{I}_4 + \frac{\rho}{4} \mathbf{H}_{hyb} \mathbf{H}_{hyb}^H), \qquad (3.10)$$

using the fact that $\log_2(\cdot)$ is an increasing function. Note that with antenna shuffling, the receiver needs to find the best antenna pairs among 6 candidates and a 3 bits feedback overhead is required.

3.2.2.4 Dominant Eigenmode Transmission

When the CSI is available at the transmitter, the dominant eigenmode transmission is proposed to achieve the same diversity gain but higher array gain than the Alamouti scheme [6]. The main idea is to transmit via the subchannel that associated with the largest singular value σ_{max} of channel matrix **H**. Let the singular value decomposition (SVD) of **H** be written as

$$\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^H \,. \tag{3.11}$$

Let us assume **w** and **g** are the right and left singular vectors that associated with σ_{max} . The transmitter then multiplies the symbol *s* with the weighting vector **w**, and the received signal is

$$\mathbf{y} = \sqrt{\frac{E_s}{N_t}} \mathbf{H} \mathbf{w} s + \mathbf{n} \,. \tag{3.12}$$

We then multiply the received signal with another weighting vector \mathbf{g} , i.e.

$$z = \mathbf{g}^H \mathbf{y} \,. \tag{3.13}$$

This can be simplified to

$$z = \sqrt{E_s} \sigma_{\max} s + n, \qquad (3.14)$$

and the received SNR for the dominant eigenmode transmission is

$$\eta = \frac{\left\| \mathbf{g}^{H} \mathbf{H} \mathbf{w} \right\|_{F}^{2}}{N_{t} \left\| \mathbf{g} \right\|_{F}^{2}} \rho = \sigma_{\max}^{2} \rho , \qquad (3.15)$$

where $\left\| \cdot \right\|_{F}$ is the Frobenius norm.

3.3 Multi-User Downlink MIMO Transmission

Below, we study a ZF precoding method for multi-user downlink MIMO Transmission as an example. Other approaches, such as the MMSE precoding, have also been proposed, see [7, 13] for more details.

3.3.1 System Model

Let us assume one base station equipped with N_t transmit antennas is communicate simultaneously with *K* users, each equipped with one receive antenna. The system model can be expressed in the baseband as

$$\mathbf{y} = \sqrt{\frac{E_s}{N_t}} \mathbf{HFs} + \mathbf{n} , \qquad (3.16)$$

where $\mathbf{H} = [\mathbf{h}_1^T, \mathbf{h}_2^T, ..., \mathbf{h}_K^T]^T$ is a $K \times N_t$ channel matrix.

3.3.2 ZF Precoding for Multi-User Downlink Transmission

Using the ZF method, the ZF precoding matrix \mathbf{F} is designed as

$$\mathbf{F} = \mathbf{H}^{H} (\mathbf{H}\mathbf{H}^{H})^{-1}.$$
(3.17)

By using the ZF precoding matrix designed above, the interferences among different users can be completely removed. This means the multiuser downlink channel is separated into K independent parallel channels that are orthogonal with each other. Therefore the throughput for the whole system can be calculated as the sum of K independent channels with small amount of interferences from other users.

3.4 Impact of Channel Estimation Errors at Receiver

In this section, we focus on the channel estimation errors at the receiver side caused by the noise during channel estimation procedure. Note that even in the very high SNR region, the channel estimation errors at the receiver will still not disappear completely due to the interpolation error. Such errors will be discussed briefly in the last section of this chapter.

3.4.1 Channel Estimation Error Model

One popular channel estimation error model is to model the elements of the channel estimation error matrix using independent and identically distributed (iid) zero mean complex Gaussian variables [17], i.e.

$$\hat{\mathbf{H}} = \sqrt{1 - \varepsilon^2} \mathbf{H} + \varepsilon \tilde{\mathbf{H}}, \qquad (3.18)$$

where $\hat{\mathbf{H}}$ is the estimated MIMO channel matrix, $\tilde{\mathbf{H}}$ is the normalized iid zero mean complex Gaussian channel estimation error matrix, and the parameter $\mathcal{E} \in [0,1]$ measures the accuracy of channel estimation. Note that $\mathcal{E} = 0$ means no channel estimation error exists while $\mathcal{E} = 1$ indicates a complete failure of channel estimation.

Using the above channel estimation error model, the impact of channel estimation error on the ZF receiver for SM has been investigated in [17]. Using the perturbation theory, when the Frobenius norm $\|\hat{\mathbf{H}}^+ - \mathbf{H}^+\|_F$ is small, the perturbation of the channel matrix \mathbf{H} is approximated as a noise term $\tilde{\mathbf{n}}$ to the unperturbed system, i.e.

$$\widetilde{\mathbf{n}} = \mathbf{n} - \frac{\varepsilon \sqrt{E_s}}{\sqrt{N_t (1 - \varepsilon^{2})}} \widetilde{\mathbf{H}} \mathbf{s}.$$
(3.19)

The covariance of the approximated noise term can be derived as

$$\hat{\mathbf{R}} = \sigma^2 \mathbf{I}_{N_r} + \frac{\varepsilon^2 E_s}{N_t (1 - \varepsilon^2)} \mathbf{E}(\widetilde{\mathbf{H}} \widetilde{\mathbf{H}}^H), \qquad (3.20)$$

Where $E(\cdot)$ calculates the expected value of a random variable.

Similar to the SM scheme, using the perturbation theory and the effective channel transfer function \mathbf{H}_{eff} , the covariance of the approximated noise term can be derived as

$$\hat{\mathbf{R}}_{Ala} = \sigma^2 \mathbf{I}_4 + \frac{\varepsilon^2 E_s}{2(1 - \varepsilon^2)} \mathbf{E}(\tilde{\mathbf{H}}_{eff} \tilde{\mathbf{H}}_{eff}^{H}), \qquad (3.21)$$

where $\mathbf{\hat{H}}_{eff}$ has similar structure as \mathbf{H}_{eff} .

For hybrid transmission scheme with ZF receiver, using the perturbation theory and effective channel matrix \mathbf{H}_{hyb} , the covariance of the approximated noise term can be written as

$$\hat{\mathbf{R}}_{hyb} = \sigma^2 \mathbf{I}_{2N_r} + \frac{\varepsilon^2 E_s}{4(1 - \varepsilon^2)} \mathbf{E}(\tilde{\mathbf{H}}_{hyb}^H \tilde{\mathbf{H}}_{hyb}), \qquad (3.22)$$

where $\widetilde{\mathbf{H}}_{hyb}$ has similar structure as \mathbf{H}_{hyb} .

3.4.2 Simulation Results

Using the above channel estimation error model, we simulate the Alamouti scheme, the SM scheme, and the hybrid transmission scheme for 2×2 and 4×4 systems and compare the performance with each other.

The WINNER C1 channel model [14] is used to simulate the metropolitan suburban LOS scenarios, where the base stations are assumed to be located above rooftop and the coverage is ubiquitous. 1000 channel realizations are generated in the simulations. The distance between the neighboring antenna elements is set as 10 wavelengths at the transmit side and 0.5 wavelength at the receive side. The central frequency is set as 5.25 GHz. No path-loss and shadowing effect has been considered. Note that in this type of scenario, the correlations of the channel coefficients are relatively high due to the existence of the LOS component.

We further assume that the channels are stationary for 8 successive symbol periods, therefore each channel realization has been used 8 times for transmission. We use 4-QAM modulation for the SM scheme and 16-QAM modulation for the Alamouti scheme and the hybrid transmission scheme in our simulations in order to keep the same data rate in both transmission schemes.

3.4.2.1 Impact on Alamouti Scheme and Spatial Multiplexing

Figure 3.1 shows the performance of both the Alamouti scheme and SM with channel estimation error ($\mathcal{E} = 0.1$) at the receive side. Similar to the results reported in [17], we observe that the error floor for the nonlinear V-BLAST receivers is lower than the error floor for the linear ZF and MMSE receivers. Furthermore, the error floor for the Alamouti scheme is much lower than the SM scheme when the channel estimation error is small.

3.4.2.2 Impact on Hybrid Transmission Scheme

Figure 3.2 shows the results of using the SM scheme and the hybrid scheme using VBLAST ZF receiver. Again, the error floor for the hybrid scheme is much lower than the SM scheme. This is due to the fact that in the hybrid scheme, the Alamouti scheme is combined with the SM scheme. Since the Alamouti scheme achieves lower error floor than the SM scheme when the channel estimation error is small, it is clear that the hybrid scheme also has lower error floor comparing to the pure SM scheme.

Moreover, it is clearly shown that the hybrid scheme with antenna shuffling performs better than the scheme without antenna shuffling. The hybrid scheme without antenna shuffling achieves the error floor at around BER 4×10^{-4} , while the error floor for the hybrid scheme with antenna shuffling has the error floor at BER 10^{-4} .

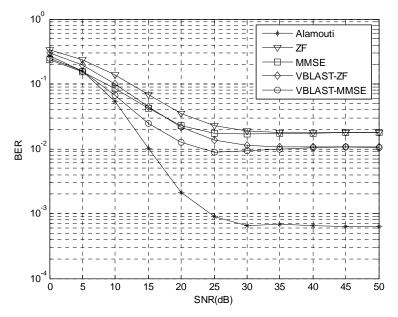


Fig. 3.1 Comparison of the Alamouti and SM schemes for 2×2 system with channel estimation errors $\varepsilon = 0.1$, WINNER C1 LOS scenario

3.5 Impact of Channel Quantization Errors at Transmitter

For dominant eigenmode transmission and the multi-user MIMO transmission, the CSI is required at the transmitter side. In FDD systems, this means a feedback link is necessary for the receiver to send the CSI information back to the transmitter. To do this, the CSI needs to be quantized at the receiver side. Therefore, the transmitter receives the imperfect CSI with quantization errors. In this section, we study the impact of quantization errors on these two transmission schemes. Note that another main source of CSI errors at the transmitter is the feedback delay for the CSI to be fed back to the transmitter. We will briefly discuss this at the end of this chapter.

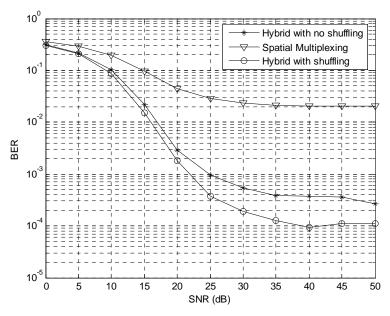


Fig. 3.2 Comparison of the hybrid transmission scheme and SM schemes for 4×4 system with $\varepsilon = 0.1$, WINNER C1 LOS scenario

3.5.1 Random Vector Quantization

In this subsection, we consider a simple and well-known random vector quantization (RVQ) scheme [12]. Note that the optimal quantizer for the MIMO system is not known in general, and the RVQ scheme provides a lower bound in terms of performance.

To quantize a vector \mathbf{h}_k , let us first obtain the direction of the vector as

$$\widetilde{\mathbf{h}}_{k} = \frac{\mathbf{h}_{k}}{\left\|\mathbf{h}_{k}\right\|},\tag{3.23}$$

where is the 2-norm.

We then quantize the direction of \mathbf{h}_k using a random codebook, where the 2^B quantization codewords are chosen from an isotropic distribution on the M-dimensional unit sphere independently, i.e.

$$C_{k} = \{ \mathbf{c}_{k,1}, \dots, \mathbf{c}_{k,N} \}$$
(3.24)

with $|C_k| = 2^B$.

By using the minimum chordal distance as the criterion, the quantized direction of channel \mathbf{h}_{ν} can be expressed as

$$\hat{\mathbf{h}}_{k} = \underset{\{\mathbf{C}_{k,j}\}_{j=1,\dots,N}}{\operatorname{arg\,max}} \left| \widetilde{\mathbf{h}}_{k} \mathbf{c}_{k,j} \right|.$$
(3.25)

3.5.2 Simulation Results

3.5.2.1 Impact on Dominant Eigenmode Transmission

We simulate the performance of the dominant eigenmode transmissions with channel quantization errors at the transmit side. 1000 channel realizations are generated for the WINNER C1 LOS channel model. 16-QAM modulation is used in the simulations. The right singular vector \mathbf{w} is quantized at the receive side using RVQ with 8 bits, and fed back to the transmitter. We further assume the receiver knows the CSI perfectly and a ZF receiver is deployed.

In Fig. 3.3, the results for the impact of channel quantization errors at the transmitter are plotted respectively for a 4×4 system. We observe that for the system with only the quantization errors at the transmit side, the BER performance degrades comparing to the case when perfect CSI at the transmitter. However, there has no error floor in the high SNR region. This is because the receiver knows the quantization errors at the transmit side, therefore a ZF receiver can be designed to avoid the error floor caused by the quantization error.

3.5.2.2 Impact on Multi-User Downlink Transmission Using ZF Precoding

In the following, we simulate one base station (equipped with 4 transmit antennas) communicates with 2 users each equipped with one receive antenna. We assume each user estimate the CSI perfectly and feedback the quantized CSI to the transmitter. The WINNER C1 LOS scenario is simulated with 10 wavelengths inter-element distance at the transmit side (base station) and 0.5 wavelength at the receive side (mobile users). We study both the system throughput and BER performance (assuming 4-QAM modulation).

Figures 3.4 and 3.5 show the BER performance and the system throughput using 8 bits RVQ respectively. The performance and throughput for the system with perfect CSI at the base station are also plotted as references. Unlike the results shown in the single user transmission case, the BER for the multi-user downlink transmission achieves an error floor

in the high SNR region. This is because the user only knows its own channel perfectly, but not the CSI for the other user. Therefore it is not possible to cancel the interference caused by the transmission to the other user using quantized CSI at the base station. Due to the same reason, the throughput (average capacity of two users) for the whole system achieves a floor at around 2.5 bit/s/Hz.

Figures 3.6 and 3.7 show the performance using different quantization bits (4/8/12/16/20 bits). It is clearly shown that as the quantization bits increases, both the BER performance and the system throughput improve.

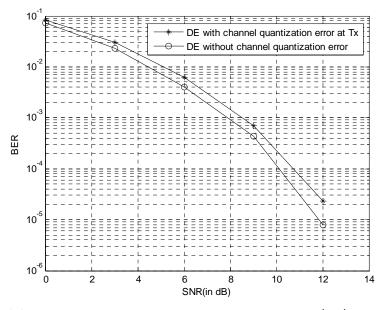


Fig. 3.3 Dominant eigenmode transmission schemes for 4×4 system with and without channel quantization errors at the transmit side, WINNER C1 LOS scenario, the number of RVQ bits is 8

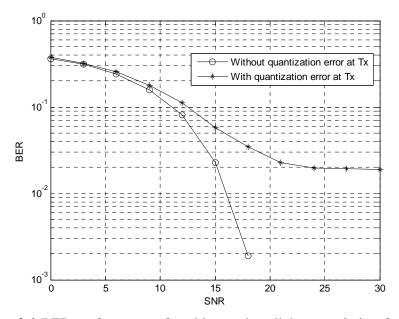


Fig. 3.4 BER performance of multi-user downlink transmission for two 4×1 channels, WINNER C1 LOS scenario, the number of RVQ bits is 8

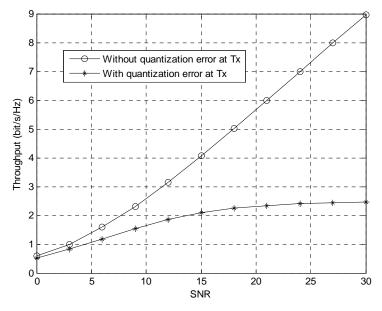


Fig. 3.5 Throughput of multi-user downlink transmission for two 4×1 channels, WINNER C1 LOS scenario, the number of RVQ bits is 8

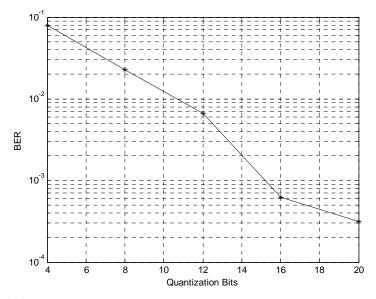


Fig. 3.6 BER performance of multi-user downlink transmission for two 4×1 channels with different RVQ bits, WINNER C1 LOS Scenario, SNR equals 21 dB

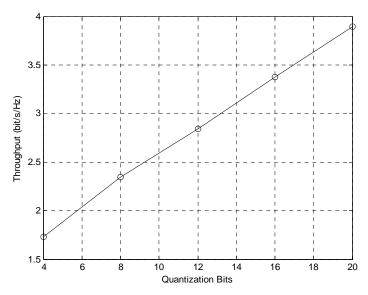


Fig. 3.7 Throughput of multi-user downlink transmission for two 4×1 channels with different RVQ bits, WINNER C1 LOS Scenario, SNR equals 21 dB

3.6 Concluding Remarks and Further Comments

In this chapter, we have studied the impact of channel estimation errors on various transmission schemes. A simple iid complex Gaussian error model has been used to model the channel estimation errors caused by the noise during the estimation process. Our simulations have shown that for small channel estimation errors, the Alamouti scheme is more robust than the SM scheme in the sense that its associated error floor of BER curve is lower than that of the SM scheme. In the case of 4 transmit antennas, it has been shown that the hybrid scheme outperforms the SM scheme. The performance of the hybrid scheme can be further improved by shuffling the antenna elements based on maximizing the effective channel capacity.

Using the random vector quantization method, we have also investigated the impact of channel quantization errors on the dominant eigenmode transmission and multi-user downlink transmission using ZF precoding. It has been shown that when the receiver knows the CSI perfectly, the quantization error only degrades the performance of dominant eigenmode transmission without an error floor. While for multi-user downlink transmission, both the BER performance and the throughput are limited at high SNR region due to interference among different users.

Another main source of channel estimation errors at the receiver is the interpolation errors, which exist even in the very high SNR region. For MIMO system, such errors can be modeled as complex Gaussian matrix with a scaled version of the MIMO channel covariance matrix [15]. Similar to the iid channel estimation error model studied in this chapter, the interpolation errors at the receiver lead to an error floor for spatio-temporal transmissions.

Besides the channel quantization errors at the transmitter, the feedback delay also generates errors at the transmit side. Again, assume no channel estimation errors at the receiver, in this case the receiver knows both the feedback CSI and the current CSI perfectly. For single link MIMO transmission schemes such as the dominant eigenmode transmission, the performance will be degraded but without error floor. For multi-user downlink MIMO transmission, there will be limitation on the performance due to the interference among different user. The performance of these schemes can be improved if an accurate channel prediction/tracking algorithm is deployed in the system [2].

References

- Alamouti SM (1998) A simple transmit diversity technique for wireless communications. IEEE Journal of Selected Areas in Communications. vol. 16, pp. 1451–1458.
- 2. Duel-Hallen A, Hu S, Hallen H (2000) Long range prediction of fading channels. IEEE Signal Processing Magazine, vol. 17, no. 3, pp. 62–65.
- Foschini GJ (1996) Layered space-time architecture for wireless communication in a fading environment when using multiple antennas. Bell Laboratories Technical Journal, vol. 1, no. 2, pp. 41–59.
- Foschini GJ, Gans MJ (1998) On limits of wireless communications in a fading environment when using multiple antennas. Wireless Personal Communications, vol. 6, pp. 311–335.
- Heath RW Jr., Paulraj A (2005) Switching between diversity and multiplexing in MIMO systems. IEEE Transactions on Communications, vol. 53, no. 6, pp. 962–968.
- 6. Paulraj A, Nabar R, Gore D (2003) Introduction to space-time wireless communications. Cambridge University Press, Cambridge.
- Peel CB, Hochwlad BM, Swindlehurst AL (2005) A vector-perturbation for near-capacity multiantenna multiuser communication – Part I: channel inversion and regularization. IEEE Transactions on Communications, vol. 53, no.1, pp.195–202.
- 8. Shim S, Kim K, Lee C (2005) An efficient antenna shuffling scheme for a DSTTD system. IEEE Communication Letters, vol. 9, no. 2, pp. 124–126.
- Tarokh V, Jafarkhani H, Calderbank AR (1999) Space-time block codes from orthogonal designs. IEEE Transactions on Information Theory, vol. 45, no. 5, pp. 1456–1467.
- Telatar IE (1999) Capacity of multi-antenna Gaussian channels. European Transactions on Telecommunications, vol. 10, no. 6, pp. 585–595. (Published in October, 1995 as a Technical Memorandum, Bell Labs)
- 11. Texas Instruments (2001) Improved Double-STTD schemes using asymmetric modulation and antenna shuffling. TSGR1#20(01)–0459.
- Trivellato M, Boccardi F, Tosato F (2007) User selection schemes for MIMO broadcast channels with limited feedback. Proceedings IEEE Vehicular Technology Conference, pp. 2089–2093.
- WINNER Deliverable D2.7 (2005) Assessment of advanced beamforming and MIMO technologies. IST-2003-507581, www.ist-winner.org.
- 14. WINNER Deliverable D5.4 (2005) Final report on link level and system level channel models. IST-2003-507581, www.ist-winner.org.
- 15. WINNER II Deliverable D2.3.3 (2007) Link level procedures for the WINNER system. IST-4-027756, www.ist-winner.org.
- Wolniansky PW, Foschini GJ, Golden GD, Valenzuela RA (1998) V-BLAST: an architecture for realizing very high data rates over the rich scattering wireless channel. Proceedings of URSI International Symposium on Signals, Systems and Electronics, pp. 295–300.

- 17. Zhang X, Ottersten B (2003) Performance analysis of V-BLAST structure with channel estimation error. Proceedings of IEEE Workshop on Signal Processing Advances in Wireless Communications, pp. 487–491.
- Zhuang X, Vook FW, Rouquette-Leveil S, Gosse K (2003) Transmit diversity and spatial multiplexing in four-transmit-antenna OFDM. Proceedings of IEEE International Conference on Communications, vol. 4, pp. 2316–2320.

4 On Multi-Cell Cooperative Signal Processing in Backhaul-Constrained Cellular Systems

Patrick Marsch and Gerhard Fettweis

4.1 Introduction

It is well known that inter-cell interference poses the main capacity limitation in today's and future cellular systems. To overcome this problem, multi-user joint detection or transmission across cell borders (often referred to as *virtual MIMO* or *network MIMO*) has been proposed by Weber et al. [1], offering the possibility of strongly improving network capacity and fairness. Optimistic capacity bounds for large clusters of cooperating cells have been determined for the uplink [2] and downlink [3], and corresponding detection and transmission schemes have been investigated in e.g. [4,5,6].

The main problem connected to such multi-cell signal processing is the vast amount of backhaul required for the exchange of information between cooperating base stations. This issue has e.g. been addressed by Sanderovich et al. [7] and Shamai et al. [8], where a circular Wyner model [9] is used to describe a cellular network and the required backhaul is minimized through signal compression. Our own research has started from a system perspective with the introduction of an optimization framework [10,11] that already achieves large portions of the possible capacity and fairness improvements under strong backhaul constraints by applying joint signal processing only to selected users, referred to as selective virtual MIMO. In the downlink, we have observed that there is an additional degree of freedom of either exchanging pre-processed and quantized signal values over the backhaul, or the uncoded, binary data of jointly preprocessed users, enabling a further reduction in the extent of backhaul required. In Marsch and Fettweis [12], we have extended our work and investigated methods of partitioning a cellular network into small subsystems, in which the stated optimization schemes and selective virtual MIMO can be applied locally,

requiring only local channel knowledge. In recent work [13,14], we have observed alternative schemes of information exchange between cooperating base stations, which however are beyond the scope of this chapter. We believe that multi-cell signal processing is in fact a realistic concept for next generation mobile communication systems, as attractive performance improvements can be obtained, while the cooperation between base stations – and the required availability of channel information at certain points in the network – can be limited to a feasible extent.

In this chapter, we summarize the key points of our previous work. We put a special emphasis on investigating the optimality of our proposed subsystem partitioning scheme by doing a general analysis of the potential extent of interference cancellation within subsystem partitioning concepts.

4.2 Notation

In general, if **X** is a matrix, then we refer to the *j*th *column vector* as **x**_{*j*}, and to the matrix elements as $x_{i,j}$, except for channel matrices **H**, where **h**_{*k*} refers to the row vector corresponding to user *k*. The operator • denotes element-wise multiplication and the operator \leq denotes element-wise inequality if applied to vectors or matrices. **A** < 0 states that matrix **A** is positive semidefinite, and perator Δ yields a square matrix with non-zero elements only on the diagonal, either extracted from a given square matrix or generated from a vector. The operator $\mathbf{Y} = \lfloor \mathbf{X} \rfloor$ yields $y_{i,j} = 1$ if $x_{i,j} > 0$, otherwise zero. The expressions $\mathbf{0}_{[i \times j]}$ and $\mathbf{1}_{[i \times j]}$ denote matrices with *i* rows and *j* columns, filled with zeros and ones, respectively. $\mathbf{I}_{[i]}$ denotes a size *i* identity matrix, operators (\cdot)^T and (\cdot)^H denotes expectation value. The operator $|\cdot|$ denotes vector norm, i.e. $|\mathbf{a}| = \mathbf{a}^{T}\mathbf{1}$, and eig(**B**, **C**) yields the unit-norm Eigenvector corresponding to the dominant generalized Eigenvalue of matrices **B** and **C**. The markers ^ and ~ are used to distinguish variables connected to the uplink or downlink, respectively, where necessary.

4.3 System Model

We consider a cellular mobile communications system, as depicted in Fig. 4.1. The system consists of M base stations with N_{bs} antennas each, and a total of K terminals, or *users*, with one receive or transmit antenna each. Each base station is connected to a certain cell area, and three base stations are always grouped together into one location, referred to as a site, where

the total number of sites is denoted by S. In each cell, R terminals are uniformly distributed over the cell area. We assume that a media access scheme is used that allows the transmission to take place over a large number of orthogonal and frequency-flat channels. This could e.g. be an OFDMA system where all involved devices are perfectly synchronized in time and frequency and the guard interval is designed sufficiently large to ensure that the system is free of inter-symbol interference (ISI). In such a system, we could define R sets of L sub-carriers each and assign each terminal to exactly one of these sets (also referred to as *resource blocks*), such that no two terminals in the same cell occupy the same resource block. Due to the resource orthogonality, we will experience no intra-cell interference, and the performance of the terminals will be limited solely due to inter-cell interference and thermal noise. Our setup assures that the system has an effective reuse factor of 1, i.e. the complete system bandwidth is reused in each cell.

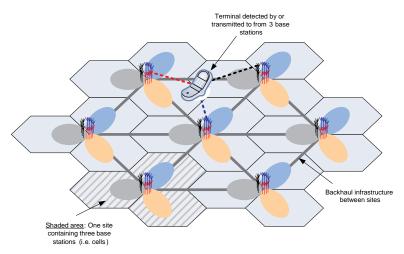


Fig. 4.1 The cellular system setup used throughout this chapter

The main extension of our observed system compared to conventional cellular systems is the fact that we consider additional links (possibly fibreoptic cables or directed microwave links) between sites to enable multicell cooperative signal processing. For example, multiple terminals can be served by multi-cell joint transmission or joint detection in order to combat inter-cell interference. We assume that the communication on these backhaul links is error-free but strongly limited in capacity, as a network operator will usually not be willing to invest a large amount of money into a high-bandwidth backhaul infrastructure. Due to these constraints, it appears

reasonable to only provide a subset of users with virtual MIMO schemes, one of the central aspects covered in our recent work.

In the downlink, we have two ways to perform a joint transmission from multiple base stations to multiple terminals, as we will discuss in detail later. First, we could consider that one of the base stations has the transmitter-side channel knowledge of the links from all involved base stations to all involved terminals, performs the required preprocessing, and then forwards preprocessed and quantized signals to the partnering base stations. In this case, the other base stations are basically degraded to *remote radio heads*, as they do not need to perform any additional baseband signal processing for the involved terminals. Alternatively, all base stations involved in a joint transmission could have the same transmitter-side channel information and perform beamforming independently. In this case, the data exchanged between base stations is uncoded binary user data. We will refer to these options as backhaul scenarios A and B.

For simplicity, we assume here that base station cooperation in the uplink is based solely on the exchange of received and quantized signals, and only observe linear precoding and postcoding schemes, though in our current work we are also investigating non-linear schemes where base stations may also exchange decoded [13] or partially decoded terminal messages for distributed interference cancellation, possibly in connection with multilevel coding [14]. We generally observe only the backhaul required for the exchange of data, and not the backhaul needed for the exchange of channel knowledge or other administrative overhead.

4.3.1 Optimization Framework

In Marsch and Fettweis [10,11,12], we have proposed an optimization framework to abstract the degrees of freedom and parameters involved in the optimization of a multi-cell cooperation enabled cellular system. In general, we observe this topic from an information theory point of view, and do not consider performance degradations in a practical system, e.g. due to imperfect synchronization of concurrently transmitting entities, the usage of practical modulation and coding schemes etc. We assume perfect channel state information at the base stations, which for the downlink in a practical system would have to be obtained through channel reciprocity and/or feedback from the terminals.

One of the degrees of freedom in such a system is the decision which terminals to assign to which of the R resource blocks, captured by

$$G \in \{0,1\}^{[K \times K]}, \ e.g. \ G = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$
(4.1)

where each column \mathbf{g}_k indicates which other terminals are occupying the same resources as terminal k and thus experiencing mutual interference. In this example, the first two and last two terminals are grouped onto the same resources, respectively. Obviously, grouping has to follow the law of transcivity and reflexivity, hence $\mathbf{G}^T = \mathbf{G}$ and $\forall i,j : \mathbf{g}_i^T \mathbf{g}_j \in \{0, \mathbf{g}_i^T \mathbf{g}_i\}$ has to hold. The next parameter specifies which antennas are actively employed for the communication with a user k. We refer to this as the *virtual MIMO configuration* matrix, as it implicitly determines which terminals are served by multi-cell signal processing. We define the matrix

$$\mathbf{V} \in \mathbf{N}_{0}^{+[N_{BS} \times K]}, \ e.g. \ \mathbf{V} = \begin{bmatrix} 16 & 16 & 0 & 0 \\ 0 & 0 & 8 & 8 \\ 10 & 10 & 16 & 16 \\ 0 & 0 & 8 & 8 \end{bmatrix},$$
(4.2)

where $N_{BS} = MN_{bs}$ and each element $v_{a,k} > 0$ determines that base station antenna *a* is actively involved in the communication with terminal *k*. In the uplink, and in a downlink based on backhaul scenario A, where received or preprocessed signals are relayed between base stations, the actual non-zero values in **V** state the number of quantization bits employed. Clearly, for two terminals *i* and *j* involved in a common virtual MIMO operation, it appears reasonable that the set of involved antennas and the quantization resolution should be chosen equally for both users, fulfilling $\forall i,j: \mathbf{g}_i^T \mathbf{g}_j > 0$ $\rightarrow \mathbf{v}_i^T \mathbf{v}_j \in \{0, \mathbf{v}_i^T \mathbf{v}_i\}$. In a downlink with scenario B, no quantization is necessary, and **V** can thus be constrained to $\mathbf{V} \in \{0, 1\}^{[N_{BS} \times K]}$.

The achievable rates of the terminals and the exact backhaul usage also depend on the choice of where the centralized signal processing is performed for each virtual MIMO operation. We capture this in parameter

$$\mathbf{s} \in \left\{1, \cdots, K\right\}^{[K \times 1]},\tag{4.3}$$

where each s_k states the index of the site doing the pre- or postprocessing for terminal k. A final parameter is the power allocation

$$\hat{\mathbf{p}}^{[l]}, \breve{\mathbf{p}}^{[l]} \in \mathsf{R}_{0}^{+[K \times 1]}$$
(4.4)

stating the transmit power assigned to each terminal on sub-carrier $1 \le l \le L$ in uplink and downlink, respectively. We can now state the transmission on each sub-carrier index *l* as

$$\hat{\mathbf{y}}^{[l]} = \left(\left(\mathbf{W}_{(\mathbf{G},\mathbf{V},\hat{\mathbf{p}}^{[l]})}^{[l]H} \mathbf{H}_{(\mathbf{G})}^{[l]H} \right) \bullet \mathbf{G} \right) \sqrt{\hat{\mathbf{P}}^{[l]}} \hat{\mathbf{x}}^{[l]} + \mathbf{W}_{(\mathbf{G},\mathbf{V},\hat{\mathbf{p}}^{[l]})}^{[l]H} \hat{\mathbf{n}}^{[l]}$$
(4.5)

$$\mathbf{\breve{y}}^{[l]} = \left(\left(\mathbf{H}_{(\mathbf{G})}^{[l]} \mathbf{W}_{(\mathbf{G}, \mathbf{V}, \mathbf{\breve{p}}^{[l]})}^{[l]} \right) \bullet \mathbf{G} \right) \sqrt{\mathbf{\breve{P}}^{[l]}} \mathbf{\breve{x}}^{[l]} + \mathbf{\breve{n}}^{[l]}$$
(4.6)

for uplink and downlink, respectively, where $\mathbf{y}^{[l]} \in \mathbf{C}^{[K \times 1]}$ are the received and postprocessed signals connected to the K terminals,

$$\mathbf{H}_{(\mathbf{G})}^{[l]} \in \mathbf{C}^{[K \times N_{BS}]} = \begin{bmatrix} \mathbf{h}_{1}^{[l]T} & \mathbf{h}_{2}^{[l]T} & \cdots & \mathbf{h}_{K}^{[l]T} \end{bmatrix}^{T}$$
(4.7)

denotes the channel between the base station antennas and terminals as a function of the chosen user grouping G,

$$\mathbf{W}_{(\mathbf{G},\mathbf{V},\mathbf{p}^{[l]})}^{[l]} \in \mathbf{C}^{[N_{BS} \times K]} = \begin{bmatrix} \mathbf{w}_{1}^{[l]} & \mathbf{w}_{2}^{[l]} & \cdots & \mathbf{w}_{K}^{[l]} \end{bmatrix}$$
(4.8)

are linear detection matrices (uplink) or precoding matrices (downlink), referred to as *beamforming matrices*, $\mathbf{P}^{[l]} = \Delta(\mathbf{p}^{[l]})$ are diagonal matrices with the uplink or downlink transmit power assigned to each terminal on subcarrier *l*, and $\hat{\mathbf{n}}^{[l]} \in \mathbf{C}^{[N_{BS} \times l]}$, $\mathbf{\bar{n}}^{[l]} \in \mathbf{C}^{[K \times l]}$ are instantaneous noise samples at the receiver side. For the downlink, we assume without loss of generality that $E\{\mathbf{n}^{[l]}\mathbf{n}^{[l]H}\}=\sigma^2\mathbf{I}$. In the uplink, arbitrary, yet uncorrelated noise with covariance matrix $\boldsymbol{\Phi}$ is assumed. $\mathbf{x}^{[l]} \in \mathbf{C}^{[K \times l]}$ contains the transmit symbols connected to the *K* terminals in uplink and downlink, in both cases fulfilling $E\{\mathbf{x}^{[l]}\mathbf{x}^{[l]H}\}=\mathbf{I}$. As shown in (4.5) and (4.6), the choice of beamforming matrices $\mathbf{W}^{[l]}$ depends on the user grouping **G**, virtual MIMO configuration **V** and power allocation $\mathbf{p}^{[l]}$, as we will discuss in more detail in Sect. 4.4.3.

We can now derive the achievable rates, or *capacities*, $\mathbf{c} \in \mathsf{R}_0^{+[K \times l]}$ of the users in bit/s/Hz/cell, as stated in (4.11) and (4.12), where the expectation values are calculated over the sub-carriers of each user, but the index *l* itself is omitted for notational brevity. The auxiliary variable

$$\mathbf{U} \in \left\{0, 1\right\}^{[K \times K]} = \left\lfloor \mathbf{G} \bullet \left(\mathbf{V}^T \mathbf{V}\right) \right\rfloor$$
(4.9)

states which users are commonly served by joint signal processing, and ξ_k is the relative quantization noise power (w.r.t. average receive or transmit power per antenna and subcarrier, in uplink or downlink), given by

$$\xi_{k} \in \mathsf{R}^{+[N_{BS} \times 1]} = \begin{bmatrix} \frac{1}{2^{\nu_{1,k}-2}} & \frac{1}{2^{\nu_{2,k}-2}} & \cdots & \frac{1}{2^{\nu_{NBS,k}-2}} \end{bmatrix}^{T}$$
(4.10)

$$\hat{c}_{k} = \underbrace{E}_{1 \leq l \leq L} \left\{ \log_{2} \left(1 + \underbrace{\frac{\mathbf{p}_{k} \mathbf{w}_{k}}{\mathbf{p}_{k}^{H} \mathbf{H}^{H} \Delta \left([\mathbf{G} - \mathbf{I}]_{k} \hat{\mathbf{p}}_{k}^{T} \right) \mathbf{H} \mathbf{w}_{k}}_{\text{Inter-cell interference from users on same resource block}} + \underbrace{\mathbf{w}_{k}^{H} \left(\Delta (\mathbf{H}^{H} \Delta (\hat{\mathbf{p}} \mathbf{g}_{k}^{T}) \mathbf{H}) \Delta (\xi_{k}) + \Phi \right) \mathbf{w}_{k}}_{\text{Quantization and thermal noise}} \right) \right\}$$

$$(4.11)$$

$$\tilde{c}_{k} = \mathop{E}_{1 \leq l \leq L} \left\{ \log_{2} \left(1 + \underbrace{\frac{\mathbf{\tilde{p}}_{k} \cdot |\mathbf{h}_{k} \mathbf{w}_{k}|^{2}}{\mathbf{h}_{k} \mathbf{W} \Delta \left([\mathbf{G} - \mathbf{I}]_{k} \mathbf{\tilde{p}}^{T} \right) \mathbf{W}^{ll} \mathbf{h}_{k}^{ll}}_{\text{Inter-cell interference from users on same resource block}} + \underbrace{\frac{\mathbf{\tilde{p}}^{T} \mathbf{u}_{k}}{\mathbf{I}^{T} [\mathbf{v}_{k}]} \mathbf{h}_{k} \Delta \left(\xi_{k} [\mathbf{v}_{k}^{T}] \right) \mathbf{h}_{k}^{ll}}_{\mathbf{Quantization noise (only backhaul scenario A)}} \right) \right\}$$

$$(4.12)$$

We here assume a simple separate quantization of the I- and Q- components at each base station antenna, knowing that vector quantization, possibly over all involved antennas and complete coding blocks [7] would further decrease quantization noise. The deviation between (4.11) and (4.12), and the capacity expressions in our original work [10,11] is due to the fact that we now model the interference caused by other, non-cooperative transmissions on the same resource block exactly, i.e. depending on the instantaneous choice of corresponding beamforming vectors \mathbf{w}_k . In our previous work, however, we modeled this type of interference statistically. When performing system level simulations over a large number of channel realizations, the result is asymptotically the same, though the statistical approach is of course computationally less complex. We can now also calculate the extent of backhaul for a certain choice of parameters. We assume that an infinite backhaul is available between base stations located at the same site, and are thus only interested in the backhaul required on the long distance links between different sites. We denote this quantity as a matrix $\mathbf{B} \in \mathsf{R}_{0}^{+[S \times S]}$, where each element $b_{i,j}$ states the backhaul required from site *i* to site *j* in bit/s. In the uplink, this can be derived as

$$\mathbf{B}^{[UL]} = \rho \sum_{k=1}^{K} \frac{\begin{bmatrix} \mathbf{0}_{[S \times s_k - 1]} & \mathbf{M}_{S} \mathbf{v}_{k} & \mathbf{0}_{[S \times S - s_k]} \end{bmatrix}}{\mathbf{u}_{k}^{T} \mathbf{1}}, \qquad (4.13)$$

where ρ is the effective per-user bandwidth, i.e. the number of quantized symbols per user, antenna and second. $\mathbf{M}_{S} \in \{0,1\}^{[S \times N_{gS}]}$ maps transmit antennas to sites. For downlink scenario A, where signals are precoded at a central site, quantized and relayed to cooperating sites, we can derive

$$\mathbf{B}^{[DL,A]} = \rho \sum_{k=1}^{K} \frac{\begin{bmatrix} \mathbf{0}_{[S \times s_k - 1]} & \mathbf{M}_{S} \mathbf{v}_{k} & \mathbf{0}_{[S \times S - s_k]} \end{bmatrix}^{T}}{\mathbf{u}_{k}^{T} \mathbf{1}}$$
(4.14)

For scenario B, where the involved sites exchange uncoded, binary terminal data and perform precoding redundantly, the required backhaul is

$$\mathbf{B}^{[DL,B]} = \rho \sum_{k=1}^{K} c_{k} \begin{bmatrix} \mathbf{0}_{[S \times s_{k}-1]} & \lfloor \mathbf{M}_{S} \mathbf{v}_{k} \rfloor & \mathbf{0}_{[S \times S-s_{k}]} \end{bmatrix}^{T}$$
(4.15)

For a fair comparison of backhaul scenarios A and B in practical cellular systems, however, it has to be noted that (4.15) actually expresses an upper bound on the required backhaul, whereas (4.14) yields an exact quantity. This is due to the fact that in scenario B, the backhaul traffic depends on the net throughput of the users whose binary data is relayed (which in a practical system will of course be significantly less that the information theoretical limits observed here), while the backhaul in scenario A is constant and independent of achieved throughput.

4.3.2 Optimization Problem

A network operator might now face an optimization problem such as

$$\left[\mathbf{G}^{*}, \mathbf{V}^{*}, \mathbf{s}^{*}, \mathbf{p}^{*}\right] = \underset{\mathbf{G}, \mathbf{V}, \mathbf{s}, \mathbf{p}}{\operatorname{arg\,max}} \left[Z\left(\mathbf{G}, \mathbf{V}, \mathbf{s}, \mathbf{p}\right) \right]_{\mathbf{D}, \mathbf{p}_{\max}}$$
(4.16)

under the constraints that the total capacity $\mathbf{D} \in \mathbf{N}_0^{+[S \times S]}$ of the given backhaul infrastructure may not be exceeded, i.e.

$$\mathbf{B}(\mathbf{V},\mathbf{s}) \le \mathbf{D} \,. \tag{4.17}$$

ī.

In the uplink, it is most reasonable to assume that the transmit power is constrained *per terminal*, hence

$$\sum_{l=1}^{L} \hat{\mathbf{p}}^{[l]} \le \hat{\mathbf{p}}_{\max} \in \mathsf{R}_{0}^{+[K \times 1]}, \qquad (4.18)$$

whereas in the downlink, a *per-base-station* power constraint over all transmissions to all terminals on all sub-carriers must be fulfilled, i.e.

$$\mathbf{M}_{\mathrm{B}} \sum_{l=1}^{L} \Delta \left(\mathbf{W}^{[l]} \mathbf{\breve{P}}^{[l]} \mathbf{W}^{[l]H} \right) \cdot \mathbf{1} \le \mathbf{\breve{p}}_{\mathrm{max}} \in \mathsf{R}_{0}^{+[M \times 1]}, \qquad (4.19)$$

where $\mathbf{M}_{B} \in \{0,1\}^{[M \times N_{BS}]}$ maps transmit antennas to base stations. Obviously, the choice of parameter **s** is also constrained through the backhaul infrastructure [10,11]. The function $Z(\cdot)$ can be any arbitrary function that takes the achievable rates and yields an overall performance metric, for example focusing on the sum rate of all users or the minimum common rate, where $Z(\cdot)$ would be defined as

$$Z(\mathbf{c}) = |\mathbf{c}| \quad or \quad Z(\mathbf{c}) = \min_{k=1\cdots K} \quad c_k, \qquad (4.20)$$

respectively. In our work, we define $Z(\cdot)$ such that it maximizes the average capacity of the 5% of weakest users.

4.4 Optimization Approach

Obviously, the optimization problem from the last section is high dimensional with a large solution space, prohibiting the usage of e.g. an exhaustive search. Furthermore, the power allocation is the only continuous parameter where convex optimization techniques can be applied. In Marsch and Fettweis [10,11], we have proposed a serialized and thus nonoptimal, but strongly simplified approach to the problem:

- 1. Use a heuristic, possibly terminal-location dependent grouping of terminals onto resource blocks (parameter **G**), as discussed in Sect. 4.4.1.
- 2. Find an optimal virtual MIMO configuration and selection of master sites (\mathbf{V}, \mathbf{s}) within the backhaul constraint, as addressed in Sect. 4.4.2.
- 3. Determine suitable beamforming matrices W, as shown in Sect. 4.4.3.
- 4. Optionally improve fairness through an optimization of power allocation (parameter **p**), as discussed in Sect. 4.4.4.

We will see later that it is not feasible to completely separate these optimization steps, as for example the selection of terminals for joint signal processing in step 2 requires an estimate of the performance under a certain precoding and power allocation strategy. The stated optimization steps and aspects discussed in the next sections are illustrated in Fig. 4.4.

4.4.1 User Grouping

As stated before, the grouping of users onto resources can be used to shape inter-cell interference such that selective virtual MIMO schemes are most effective. In Marsch and Fettweis [11], we have e.g. proposed to rank the users in each cell according to their *isolation*, i.e. a value close to one for cell-center users, and lower for cell-edge users, defined as

$$\eta_{k} = \sum_{\varphi \in \Psi_{k}} E\left\{\left|\mathbf{h}_{k,\varphi}\right|^{2}\right\} / \sum_{1 \le \varphi \le N_{BS}} E\left\{\left|\mathbf{h}_{k,\varphi}\right|^{2}\right\},\tag{4.21}$$

where Ψ_k contains the antenna indices of the default base station of user k. We then group the users according to their isolation, i.e. such that users with a similar isolation share the same resources. Then, users at cell edges experiencing strong mutual interference can be efficiently served with joint detection or joint transmission, while isolated users can be excluded from the further optimization process, as their achievable rates will already be fairly high without multi-cell signal processing.

In Marsch and Fettweis [12], we proposed a heuristic user grouping scheme that partitions a cellular system into subsystems that can be optimized independently, requiring only a minimal extent of information to be exchanged between subsystems. This also offers the benefit that transmitter-sided channel knowledge is only required within these subsystems. In general, we observed that if a cellular system is partitioned into subsystems, multi-cell signal processing has to be constrained to within these subsystems, unless

- an exchange of channel information between subsystems is enabled, and a mechanism exists for multiple subsystems to agree on joint signal processing across subsystems
- subsystems are defined such that they overlap, which again requires a mechanism for subsystems to agree on how to handle users in common

Both options would require a large extent of additional backhaul for negotiations between subsystems, and thus be contraproductive towards the actual aim of doing subsystem partitioning. Instead, we should define the heuristic user grouping (and inherent subsystem partitioning) such that the possible extent of interference cancellation within these subsystems already provides sufficient performance improvements. A useful input towards the design of subsystems is Fig. 4.2, which shows the main two interfering cells for any location in the central cell 1, based on the hexagonal cell setup and pathloss model in Marsch et al. [2]. As the interference pattern is the same in all cells, we can see that by grouping all users in e.g. cells 1, 2 and 3 close to the point where the three cells meet into one subsystem, we can assure that for these users the two strongest interferers can be cancelled – if they are assigned to the same resource block and enough backhaul is available.

We can similarly proceed for all users in cell those 1. except for which the main interference comes from cells 3 and 6. Here, we have an *asymmetrical* interference situation. i.e. we will not be able to find a user within cell 6 whose strongest interference comes from cells 1 and 3. To guarantee the cancellation of the two strong-

in cell 1 are very close to their own base station, and will thus usually have an acceptable SINR without any interference cancellation at all. According to these observations, we finally proposed scheme in а Marsch and Fettweis [12] that divides each cell spatially and resourcewise into 5 roughly equally-sized blocks, and groups users onto resources according to their

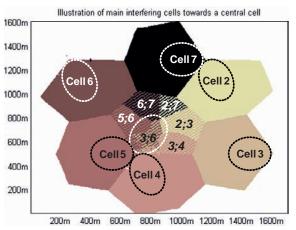


Fig. 4.2 Origin of main interference in the downlink

est interferers for all users involved, we would here have to establish a subsystem spanning at least 5 cells. However, exactly the mentioned users

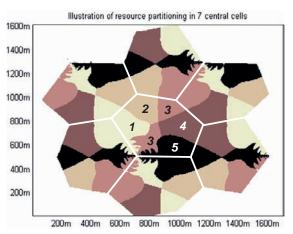


Fig. 4.3 Proposed resource block assignment

location, as illustrated in Fig. 4.3. Each of these regions spanning multiple cells can then be operated as an independent subsystem, where a selection of users for virtual MIMO can be performed locally. The global subsystem partitioning scheme can of course be combined with a ranking of users according to their cell isolation as stated at the beginning of this section, such

that each subsystem can assign users experiencing strong interference onto the same resources and invest a limited backhaul most effectively into these.

4.4.2 Virtual MIMO Configuration

As stated before, a smart selection of terminals for multi-cell signal processing, i.e. a choice of parameters (V, s) is an essential part of the optimization approach. In Marsch and Fettweis [10,11], we proposed an algorithm that initially links each terminal only to its default base station. Then, the achievable rates of the users are observed according to (4.11) or (4.12), and the algorithm determines additional base station antennas that could be used to improve the performance of the weakest users, hence by introducing additional extents of joint transmission or joint detection. From multiple options, the best one is chosen through a comparison of the ratio of metric improvement over additionally required backhaul, which can be seen as a gradient based approach. The algorithm continues iteratively until either the system metric $Z(\cdot)$ cannot be improved any more, or until no more backhaul capacity is available. The algorithm hence invests the available backhaul into the weakest users - according to our metric from Sect. 4.3.2 - and chooses the virtual MIMO operations that are most efficient in terms of metric improvement over required backhaul. The complexity of the algorithm depends strongly on how the terminal rates can be estimated for a given set of parameters, which we will discuss later. It cannot be guaranteed that the algorithm finds the global optimum, but we will see that the results are already very promising.

4.4.3 Beamforming Matrices

The calculation of receiver-side detection or transmitter-side precoding matrices, both referred to as *beamforming* matrices, has been widely studied in the context of the *multiple access channel* and *broadcast channel*, for which the achievable rate regions have been observed in e.g. [15,16,17,18]. Obviously, the calculation should be performed for each sub-carrier separately, and all equations in this section thus refer to the transmission on one sub-carrier per user, where the index l is omitted for brevity. Different from most publications in this field, our system model inherits the additional aspect that the precoding or postcoding vector connected to each user is constrained through the choice of the virtual MIMO

configuration matrix, i.e. it may only have a non-zero element for base station antennas that are actively involved in the communication with the user

$$\mathbf{W} = \mathbf{W} \bullet [\mathbf{V}]. \tag{4.22}$$

In the uplink, it is known that for a given power allocation, optimal beamforming vectors can be obtained for each user as the unit-norm Eigenvectors corresponding to the dominant Eigenvalues of the following generalized Eigenvalue decomposition [4]

$$\mathbf{w}_{k} = \left\lfloor \mathbf{v}_{k} \right\rfloor \bullet \operatorname{eig}\left(\tilde{\mathbf{h}}_{k}\tilde{\mathbf{h}}_{k}^{H}, \sum_{j=1,u_{j,k}=1}^{K} \hat{p}_{j}\tilde{\mathbf{h}}_{j}\tilde{\mathbf{h}}_{j}^{H} + \boldsymbol{\Phi}\right),$$
(4.23)

where $\mathbf{h}_k = \lfloor \mathbf{v}_k \rfloor^T \cdot \mathbf{h}_k$. This corresponds to the usage of a standard MMSE receive filter for each user, constrained according to (4.22). In the downlink, it is more difficult to determine the optimal beamforming vectors directly, as the performance of the terminals is not only cross-coupled through their assigned transmit power, but also through the beamforming

vectors themselves. In fact, it is known that the downlink problem typically is nonconvex. However, it has been shown in e.g. [4,17,19] that an optimal solution for such a downlink problem can be obtained by solving a dual uplink problem with more amenable mathematical properties. Precisely, it has been that proven the achievable rate region

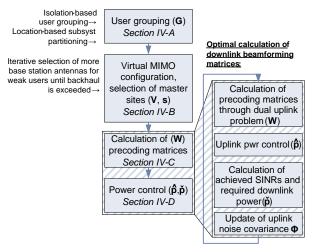


Fig. 4.4 Steps of the proposed optimization approach

of a downlink beamforming problem under a sum power constraint over the transmit antennas is equivalent to that of a corresponding uplink problem with the same sum power constraint and noise covariance $\mathbf{\Phi} = \sigma^2 \mathbf{I}$ [15,17,19]. Furthermore, any terminal rates achievable in the dual uplink must also be achievable in the downlink with the same matrix \mathbf{W} , but a different power allocation $\mathbf{\tilde{p}} \neq \mathbf{\hat{p}}$ [19].

In our case, we have to consider *per-base-station* power constraints, as defined in (4.19), for which the stated uplink-downlink duality does not

hold. The authors in Yu and Lan [20], however, have extended duality observations to downlink beamforming under per-antenna-(group) power constraints, and have shown that the achievable rate region under such constraints corresponds to that of a dual uplink problem with the same sum power and a different extent of noise at each receive antenna. They propose an iterative algorithm that finds the best downlink precoding matrix and power allocation, so that the gap between actually used power per base station and power limit is maximized, while target SINRs for the users are fulfilled, usually referred to as the solution of an SINR constrained optimization problem (SCO). In Marsch and Fettweis [21], we have modified the work from Yu and Lan [20] such that power-constrained optimization problems (PCO) can be solved directly, hence the best common multiple of target SINRs under fixed per-base-station power constraints are found. Adjusted to our system model based on selective virtual MIMO, we have proposed the following iterative algorithm in Marsch and Fettweis [22]. It is initialized with an uplink noise covariance matrix $\Phi = \sigma^2 \mathbf{I}$ and equal downlink power for all users and iterates through the steps (Fig. 4.4):

- 1. Calculate the optimal precoding matrix **W** for a given uplink noise covariance matrix $\mathbf{\Phi}$ and uplink power according to (4.23).
- 2. Perform power control in the uplink under the constraint $|\hat{\mathbf{p}}| = |\mathbf{\breve{p}}_{max}|$. This will be briefly addressed in Sect. 4.4.4.
- 3. Calculate the SINRs γ_k achieved in the uplink for the new power allocation, and knowing that under duality the same set of SINRs must also be achievable in the original downlink problem, calculate the downlink power allocation through [4]

$$\vec{\mathbf{p}} = (\mathbf{I} - \mathbf{T} \mathbf{\Psi})^{-1} \sigma^{2} \mathbf{T} \mathbf{1},$$
where $\mathbf{T} = \Delta \left(\frac{\lambda_{1}}{|\mathbf{h}_{1} \mathbf{w}_{1}|^{2}} \frac{\lambda_{2}}{|\mathbf{h}_{2} \mathbf{w}_{2}|^{2}} \cdots \frac{\lambda_{K}}{|\mathbf{h}_{K} \mathbf{w}_{K}|^{2}} \right)$
and $\mathbf{\Psi}_{jk} = \begin{cases} \left|\mathbf{h}_{j} \mathbf{w}_{k}\right|^{2} & k \neq j \land g_{j,k} = 1\\ 0 & \text{otherwise.} \end{cases}$

$$(4.24)$$

4. The uplink noise covariance Φ must have equal entries corresponding to any downlink transmit antennas connected to the same base station, i.e. $\tilde{\Phi} = \Lambda(\Phi, w, \Phi, \Phi, w, \Phi, w, \Phi, w, \Phi, w, \Phi, w)$ and $\tilde{\Phi}$ can be

 $\tilde{\Phi} = \Delta (\Phi_{11} \cdots \Phi_{11}, \Phi_{22} \cdots \Phi_{22}, \cdots, \Phi_{MM} \cdots \Phi_{MM})$, and $\tilde{\Phi}$ can be updated through a sub-gradient technique, i.e. [20]

$$\tilde{\boldsymbol{\Phi}} \to P\left\{\tilde{\boldsymbol{\Phi}} + t \cdot \Delta \left(\mathbf{M}_{\mathrm{B}} \Delta \left(\mathbf{W} \tilde{\mathbf{P}} \mathbf{W}^{H}\right) \mathbf{1}\right)\right\},\tag{4.25}$$

where $P\{\cdot\}$ is an Euclidian projection onto the nearest point fulfilling $tr\{\tilde{\Phi}\breve{P}_{max}\} \leq |\breve{P}_{max}|\sigma^2$ and $\tilde{\Phi} < 0$, and *t* is an appropriate step size.

Find more details in Yu and Lan [20] and Marsch and Fettweis [21,22]. Clearly, an optimal calculation of downlink beamforming vectors under per-base-station power constraints is computationally complex. Thus, it seams reasonable to use a strongly simplified, non-optimal calculation of beamforming vectors as a basis for the decisions in the optimization of the virtual MIMO configuration (see Sect. 4.4.2), and then perform an improved calculation for the actual transmission. A strongly simplified scheme was suggested in Marsch and Fettweis [10], where an equal transmit power per user and sub-carrier was assumed, and a transmit Wiener filter [23] was modified to our system model, i.e.

$$\mathbf{w}_{k} = \left\lfloor \mathbf{v}_{k} \right\rfloor \bullet \left(\sum_{j=1,u_{j,k}=1}^{K} \tilde{\mathbf{h}}_{j}^{H} \tilde{\mathbf{h}}_{j} + \frac{\mathbf{u}_{k}^{T} \mathbf{1} \sigma^{2}}{\mathbf{u}_{k}^{T} \tilde{\mathbf{p}}} \right)^{-1} \tilde{\mathbf{h}}_{k}^{H}, \qquad (4.26)$$

after which a normalization of the beamforming matrix was performed to fulfill a per-base-station power constraint *for each sub-carrier*, i.e.

$$\forall k : \mathbf{M}_{\mathrm{B}} \Delta \left(\sum_{j=1, g_{j,k}=1}^{K} \breve{p}_{k} \mathbf{w}_{k} \mathbf{w}_{k}^{H} \right) \mathbf{1} = \frac{\mathbf{u}_{k}^{T} \breve{\mathbf{p}}}{\mathbf{u}_{k}^{T} \mathbf{1}} \cdot \mathbf{1}_{[M \times 1]}.$$
(4.27)

Obviously, this distorts the original beamforming vectors obtained through (4.26), such that the mutual interference between users may increase. A trade-off between optimality and complexity can be achieved if the beamforming vectors are calculated as in (4.26), and normalized *over all users and sub-carriers*. In Sect. 4.6, we will compare the performance of the different non-optimal or optimal beamforming approaches.

4.4.4 Power Allocation

Power allocation can be used to reshape the distribution of terminal rates in the system according to the optimization metric $Z(\cdot)$. For achieving perfect fairness [4,24], have observed schemes for SINR balancing between users for uplink or downlink and under a sum power constraint. If perfect fairness is to be obtained in a downlink under *per-base-station* power constraints, it is possible to combine the concepts of Boche and Schubert with the duality observations of Yu and Lan [20] from the last section. Then, SINR balancing is performed in a dual uplink problem [21], the result is transformed into the downlink through (4.25), and the initially unknown uplink noise covariance updated through (4.26). Also, it is feasible to optimize the mean square error in both uplink and downlink, as investigated in Jorswieck and Boche [25] and Shi and Schubert [26]. An adaptation of these schemes to our scenario and simulation has been done in Marsch and Fettweis [22].

4.5 A Closer View on Subsystem Partitioning

The subsystem partitioning scheme from Sect. 4.4.2 has shown to enable a large extent of interference cancellation, while limiting the number of base stations and terminals involved into a virtual MIMO operation to 3. This also appears realistic from a practical point of view, as involving more devices into joint signal processing also increases the required effort for synchronization and channel estimation. In this section, however, we want to investigate theoretical gains from partitioning schemes that create clusters spanning 4 or more cells, thus allowing the cancellation of a larger number of interferers. To do this, we analyze the *average* SINR a terminal can experience in the *downlink* as a function of a choice of base station antennas \mathbf{v}_k actively transmitting to a terminal k, under the assumption that

- The frequency reuse factor is 1, hence all resources in all cells are used
- The interference from terminals involved in the same joint transmission operation is perfectly suppressed
- The array gain a terminal experiences always corresponds to N_{bs} , i.e. the total number of base station antennas involved in transmission divided by the number of terminals jointly served
- Each joint transmission operation uses a unique scrambling sequence, so that the interference a terminal sees from non-involved base stations is averaged over all terminal transmit powers in the respective cells

The mean SINR can then be stated as (again omitting index l for brevity):

$$\operatorname{SINR}_{k} = \frac{\breve{p}_{k} E\left\{ tr\left\{\mathbf{h}_{k}^{H}\mathbf{h}_{k}\Delta\left(\left\lfloor\mathbf{v}_{k}\right\rfloor\right)\right\}\right\}}{\frac{\breve{p}^{T}\mathbf{1}}{KN_{bs}} E\left\{ tr\left\{\mathbf{h}_{k}^{H}\mathbf{h}_{k}\Delta\left(\mathbf{1}_{\left[N_{BS}\times\mathbf{I}\right]}-\left(\left\lfloor\mathbf{v}_{k}\right\rfloor\right)\right)\right\}\right\}+\sigma^{2}}.$$
(4.28)

From (4.28), we can now derive the set of base station antennas that should be actively involved in the transmission to each terminal, such that a target

SINR λ is reached. while a *minimum* number of antennas is involved in the transmission (hence only the strongest links are exploited). In Fig. 4.5, we have plotted the number of additional base stations that have to be involved in а transmission to а terminal (hence equivalent to the number of interferers that have to be

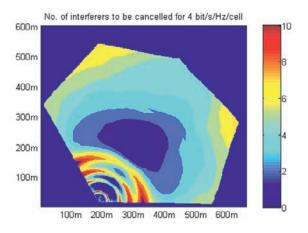
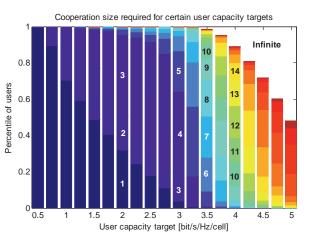


Fig. 4.5 Interference cancellation required for 4 bit/s/Hz

cancelled) in order to achieve a target capacity of 4 bit/s/Hz, as a function of possible terminal locations. Here, a large number of channel realizations according to the Okumura Hata pathloss model from Marsch et al. [2] have been observed, assuming an equal power is assigned to each terminal. But

apart from the set of base stations from which a certain terminal has to be supplied with transmit signals in order to achieve the SINR target, we have the problem that groups of users across different cells have to be found that can be commonly served with joint transmission such that all

achieve



h that *all* **Fig. 4.6** Required cooperation sizes for rate targets the SINR

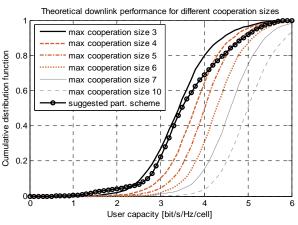
target. As already observed in Sect. 4.4.1, this can be a major issue for users with an *asymmetric* interference situation, where terminals are interfered by signals from a neighboring base station, but where there is only

marginal benefit for terminals in the corresponding cell to be involved in a common joint transmission. The result is that some users will have to be served by joint transmissions involving much more base stations than they would actually require to achieve their own SINR target, such that other terminals meet the SINR target. In Fig. 4.6, we can see the percentages of users being served by different numbers of base stations, such that a common spectral efficiency target is achieved *for all users*. Here, a brute force search has been applied over all reasonable groupings of users such that the number of base stations involved in the largest joint transmission operation, which we refer to as the maximum *cooperation size*, is minimized. We can see that for a common spectral efficiency target of 4 bit/s/Hz, around 12% of users would require an infinitely large cooperation size such that all users fulfill the spectral efficiency target.

If we now aim at finding the best possible grouping of users, while strictly limiting the maximum *cooperation size*, we can see the benefit of allowing more than 3 base stations to cooperate in joint transmission, as illustrated in Fig. 4.7. Obviously, there is a non-negligible advantage of larger cooperation sizes, until saturation can be observed for a cooperation size of beyond 10.

The results in Fig. 4.7 are theoretical, however, in the way that they are based on any arbitrary cooperation of base stations for different sets of terminals. Our aim of performing subsystem partitioning, however, is to divide the cellular network into clearly defined areas according to which terminals are assigned to subsystems, and according to which also the al-

cooperation lowed groups between of stations base is clearly defined. We have performed simulations with many Monte Carlo drops of users and for different allowed cluster sizes. in order to see whether the optimally chosen sets of cooperating base stations strongly correare locations. Unfortu-



lated to certain user Fig. 4.7 Gains of different cooperation sizes

nately, this was only the case for a cooperation size of 3, where the simulation results have suggested almost the same terminal location based subsystem assignment as proposed in Sect. 4.4.1. As shown in Fig. 4.7, our proposed partitioning scheme also shows a similar performance as a scheme based on an ideal grouping of users constrained to a cooperation size of 3. The only difference is that for a fixed subsystem partitioning, some terminals perform significantly worse (as they see major interference from outside the subsystem, which cannot be cancelled), while the majority of users performs slightly better (as they are not assigned to joint transmissions that are only beneficial to terminals in other subsystems). For any larger cooperation size, such a heuristic assignment is not possible, and base stations would have to agree on which sets of terminals to jointly serve by which base station antennas, leading to a large extent of overhead for negotiations on the backhaul.

The analysis in this section can be performed equivalently for the uplink, if *signal-to-leakage ratios* (SLR) are observed. Then, involving more base station antennas in the detection of a certain terminal can be interpreted as actively exploiting the signal of the terminal at these antennas, rather than letting the signals have a negative impact on the detection of another terminal. Basically, the same observations can be made as in the downlink, and this has lead us to the conclusion that in both cases, no heuristic terminal-location based subsystem partitioning scheme can be found for cooperation sizes larger than 3, such that the cancellation of the same number of major interferers can be guaranteed for a large majority of terminals. Obviously, the partitioning scheme from Sect. 4.4.1 could be chosen, with an additional potential cooperation between subsystems using the same resources, but requiring an additional negotiation overhead between subsystems while providing marginal gain.

Obviously, the simulation results in this section would be different if power control was considered, or if the signal propagation between cells would be influenced through e.g. a change in the carrier frequency or a stronger antenna downtilt etc. However, the main conclusions would remain the same. In fact, simulation results have shown that our proposed subsystem partitioning scheme performs even better if the antenna downtilt is increased. In this case, the benefit of doing joint detection or joint transmission within subsystems is still large, while the interference between subsystems using the same resources is further reduced.

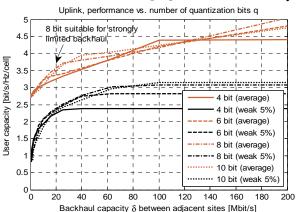
4.6 Simulation Results

We observe a cluster of 7 sites with 3 base stations each, surrounded by another tier of interfering sites. We assume $N_{bs}=2$ antennas per base station, and that each site within the cluster is connected to its partners via

bidirectional, errorless links of a common capacity δ , so that the backhaul infrastructure **D** is given as in Marsch and Fettweis [10]. We consider a fully

loaded 5MHz OFDMA system where 50 users per cell occupy L=6maximally spaced subcarriers each, leading to a per-user effective bandwidth $\rho=84$ kHz, and assume that the coherence bandwidth is small enough that each user's subcarriers are uncorrelated. We observe the ergodic behavior of

Rayleigh fading chan-



ergodic behavior of Fig. 4.8 Suitable number of quantization bits

nels with the pathloss model in Marsch et al. [2]. In general, we assume a standard receive signal based power control in the uplink, and equal power per user in the downlink, and only the rates of terminals in the central site are plotted. For uplink and downlink scenario A, we constrain the virtual MIMO configuration V to $v_{ok} \in \{0, q, 16\}$, i.e. the same number of quanti-

zation bits q is used throughout the system, and basically noiseless 16-bit quantization is used for backhaul links between base stations located at the same site.

In Marsch and Fettweis [10,11], we observed the performance of a centralized optimization approach, assuming that global channel knowledge is available in

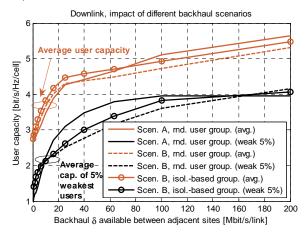


Fig. 4.9 Impact of different backhaul scenarios in DL

one central point of the network. From Fig. 4.8, we can e.g. see that q=8 quantization bits appear suitable under a strongly constrained backhaul,

which is also the case for the downlink [10]. Figure 4.9 compares backhaul scenarios A and B in the downlink, and also shows the benefit of using the

isolation-based user grouping from Sect. 4.4.1. Scenario B vields a superior average and 5th percentile performance for $\delta \leq 10$ Mbit/s/link. This is because the backhaul needed for the distribution of uncoded data among initially sites is lower, but increases quadratically in the number of sites involved. while for

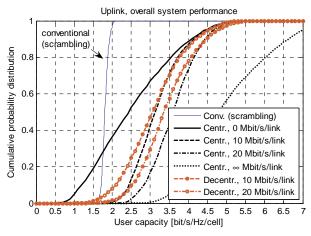


Fig. 4.10 Centralized / decentralized performance (UL)

scenario A, it increases linearly in the number of transmit antennas involved. The slight inferiority of scenario A for $\delta \leq 160$ Mbit/s/link is due to additional quantization noise. As stated before, it has to be kept in mind that the required backhaul is fixed for scenario A, whereas it scales down

according to the actual user traffic for scenario B. If joint transmission is limited to smaller numbers of involved entities – as in the case of the proposed subsystems – scenario B in general appears better.

Figures 4.10 and 4.11 from Marsch and Fettweis [12] show results where the optimization al-

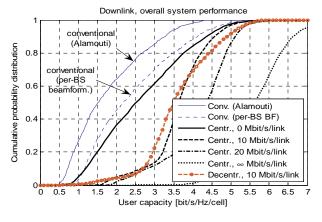


Fig. 4.11 Centralized / decentralized performance (DL)

gorithm from Sect. 4.4 is applied to small subsystems obtained through the partitioning scheme from Sect. 4.4.1, for uplink and downlink, respectively. The algorithm is started in 22 such subsystems around the central site, where each subsystem contains exactly 30 users equally taken from

three involved cells. We here use isolation-based user grouping from Sect. 4.4.1 and backhaul scenario B in the downlink. Obviously, for larger extents of available backhaul, the centralized optimization approach can achieve a much better performance, as large virtual MIMO operations can be performed involving many base stations. In the decentralized approach, however, there is no benefit of offering more than about 20 Mbit/s or 10 Mbit/s capacity per backhaul link, in uplink and downlink, respectively, as the number of users jointly served with virtual MIMO is limited to 3. We can see that in these low backhaul regimes, the decentralized schemes perform almost as well as the centralized approaches, again emphasizing that the proposed subsystem partitioning assures that for most terminals the two major interferers can be cancelled.

Figure 4.12 finally shows the cumulative probability distribution of user rates connected to different ways of computing precoding matrices W, as discussed in Sect. 4.4.3, here for a decentralized optimization approach isolation-based and grouping. user The previous simulation results for the downlink were all based on the precoding simple scheme from (4.26) and

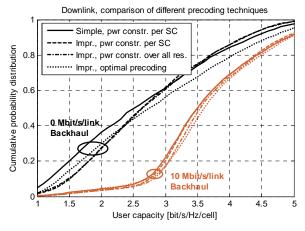


Fig. 4.12 Comparison of different precoding schemes

(4.27), where a transmit Wiener filter is used for terminals selected for joint transmission, normalized to a per-base-station power constraint for *each sub-carrier*, without explicit power control (see solid lines). The dashed lines show the benefit of calculating the beamforming matrices such that also other terminals in the same subsystem and group that are not involved in joint transmission with terminal k are taken into consideration, which can significantly improve the performance for the zero-backhaul case (where joint transmission is limited to adjacent base stations at the same site), at the price of a larger extent of channel knowledge needed. The dash-dotted lines show another slight performance improvement if the previous scheme is modified to assure the per-base-station power constraints by normalizing **W** over *all users and sub-carriers*, instead of normalization *per sub-carrier*. The last tuple of curves (dotted lines) shows

the performance if an exact calculation of precoders is done according to the duality scheme from Yu and Lan [20], as discussed in Sect. 4.4.3, yielding only a slight improvement in median capacity. The impact of different power allocation schemes in the downlink is discussed in Marsch and Fettweis [22].

In general, it appears that the benefit of using an optimal calculation of downlink precoding vectors is quite marginal, compared to a significant computational complexity connected to it. Thus, it seems reasonable to use a simple Wiener-filter based precoding scheme to estimate performance during the optimization of the virtual MIMO configuration, and to perform an optimal calculation of precoding vectors (at most) afterwards.

4.7 Conclusions

In this chapter, we have discussed a framework for optimizing the uplink or downlink of cellular systems, where multi-cell joint detection or joint transmission is possible, but constrained by a limited backhaul infrastructure between sites. Based on previous work, we have analyzed an approach that serializes the multi-dimensional optimization problem and yields good performance results at reasonable computational complexity. We have shown that the grouping of users onto resources can be used both to make selective virtual MIMO for subsets of users more effective, and also to partition the cellular network into individual subsystems that can be optimized in a decentralized way, requiring only local channel knowledge. A discussion on the potential benefit of having other subsystem partitioning schemes has shown that larger subsystems would definitely yield a non-negligible improvement, but it seems impossible to find a simple, user location dependent partitioning scheme that could guarantee the cancellation of a similar extent of interference for a large majority of users, as it is the case for our proposed scheme with cooperation size 3. Finally we have observed possible performance improvements in the downlink through an optimal calculation of precoding vectors. Results have shown that such schemes, despite a large computational complexity, only yield marginal gains, as the per-base-station power constraints seem to have a minor impact, especially when applied jointly to many resource blocks and sub-carriers.

In general, we are sure that selective virtual MIMO has the potential to play an important role in next generation mobile communications systems. A lot of research, however, remains in the actual implementation of joint signal processing techniques, for example w.r.t. time and frequency synchronization of communicating entities, and channel feedback for downlink transmission. As stated, our ongoing research is on alternative schemes of exchanging decoded information between base stations in the uplink [13], or partially decoded data if the terminals employ multi-level coding [14].

References

- Weber T, Maniatis I, Sklavos A, Liu Y (2002), Joint transmission and detection integrated network (JOINT), a generic proposal for beyond 3G systems. In: Proceedings of the 9th International Conference on Telecommunication (ICT'02), vol. 3, pp. 479–483
- 2. Marsch P, Khattak S, Fettweis G (2006), A framework for determining realistic capacity bounds for distributed antenna systems. In: Proceedings of the IEEE Information Theory Workshop (ITW'06)
- 3. Karakayali M, Foschini G, Valenzuela R, Yates R (2006), On the maximum common rate achievable in a coordinated network. In: Proceedings of IEEE International Conference on Communications (ICC'06)
- 4. Boche H, Schubert M (CISS2002), Solution of the SINR downlink beamforming problem. In: Proc. of the Conf. on Information Sciences and Systems (CISS)
- Boehnke R, Kuehn V, Kammeyer K (2005), Fast sum rate maximization for the downlink of MIMO-OFDM systems. In: Proceedings of the Canadian Workshop on Information Theory (CWIT'05)
- 6. Khattak S, Rave W, Fettweis G (2006), Multiuser turbo detection in a distributed antenna system. In: Proceedings of the 15th IST Wireless and Mobile Communications summit (IST'06)
- Sanderovich A, Somekh O, Shamai S (2007), Uplink macro diversity with limited backhaul capacity. In: Proceedings of the IEEE International Symposium on Information Theory (ISIT'07)
- Shamai S, Somekh O, Simeone O, Sanderovich A, Zaidel B, Poor V (2007), Cooperative multi-cell networks: Impact of limited-capacity backhaul and inter-users links. In: Proceedings of the Joint Workshop Coding and Communication (JWCC'07)
- 9. Wyner A (1994), Shannon theoretic approach to a Gaussian cellular multiple access channel. In: IEEE Transaction on Information Theory, vol. 40, no. 6, pp. 1713-1727
- 10. Marsch P, Fettweis G (EW2007), A framework for optimizing the downlink of distributed antenna systems under a constrained backhaul. In: Proceedings of the 13th European Wireless Conference (EW'07)
- 11. Marsch P, Fettweis G (ICC2007), A framework for optimizing the uplink of distributed antenna systems under a constrained backhaul. In: Proceedings of the International Conference on Communications (ICC'07)

- 12. Marsch P, Fettweis G (IST2007), A decentralized optimization approach to backhaul-constrained distributed antenna systems. In: Proceedings of the 16th IST Mobile and Wireless Communications Summit (IST'07)
- 13. Marsch P, Fettweis G (WCNC2008), On the rate region of a multi-cell MAC under backhaul and latency constraints. In: Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'08)
- 14. Marsch P, Fettweis G (ITW2008), On backhaul-constrained multi-cell cooperative detection based on multi-level coding, submitted for publication
- 15. Caire G, Shamai S (2003), On the achievable throughput of a multiantenna gaussian broadcast channel. In: IEEE Transaction on Information Theory, vol. 49, no. 7, pp. 802–811
- Vishwanath S, Jindal N, Goldsmith A (2003), Duality, achievable rates and sum-rate capacity of gaussian MIMO broadcast channels. In: IEEE Transactions on Information Theory, vol. 49, pp. 2658–2668
- Goldsmith A, Jafar S, Jindal N, Vishwanath S (2003), Capacity limits of MIMO channels. In: IEEE Journal on Selection Areas in Communication, vol. 21, no. 5, pp. 684–702
- Jindal N, Rhee W, Vishwanath S, Jafar S, Goldsmith A (2005), Sum power iterative waterfilling for multiple-antenna gaussian broadcast channels. In: IEEE Transactions on Information Theory, vol. 51, no. 4, pp. 1570–1580
- 19. Boche H, Schubert M (VTC2002), A general duality theory for uplink and downlink beamforming. In: Proceedings of the IEEE Vehicular Technology Conference (VTC) Fall
- 20. Yu W, Lan T (2007), Transmitter optimization for the multi-antenna downlink with per-antenna power constraints. In: IEEE Transactions on Signal Processing, vol. 55, pp. 2646–2660
- 21. Marsch P, Fettweis G (ITG2008), A direct solution for multi-user beamforming under per-antenna power constraints. In: Proceedings of the 7th International ITG Conference on Source and Channel Coding
- 22. Marsch P, Fettweis G (AT2008), On multi-cell cooperative transmission in backhaul-constrained cellular systems, submitted for publication
- 23. Joham M (2004), Optimization of linear and nonlinear transmit signal processing. Ph.D.-Thesis, Technische Universität München
- Schubert M, Boche H (2004), Solution of the multiuser downlink beamforming problem with individual SINR constraints. In: IEEE Transactions on Vehicular Technology, vol. 53, no. 1, pp. 18–28
- 25. Jorswieck E, Boche H (2003), Transmission strategies for the MIMO MAC with MMSE receiver: Average MSE optimization and achievable individual mse region. In: IEEE Transaction on Signal Processing, vol. 51, no. 11, pp. 2872–2881
- 26. Shi S, Schubert M (2005), MMSE transmit optimization for multiuser multiantenna systems, In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)
- 27. Yang W, Xu G (1998), Optimal downlink power assignment for smart antenna systems. In: Proceedings of the IEEE International Conference on Accoustics, Speech, and Signal Processing (ICASSP'98)

Part II Access

5 One-Shot Multi-Bid Auction and Pricing in Dynamic Spectrum Allocation Networks

László Kovács and Attila Vidács

5.1 Introduction

Radio spectrum is a valuable resource but usable frequencies are scarce. Spectrum should be distributed to the uses and users that generate the greatest value but existing management systems fail to do this. When designing usage rights, the focus tends to be on avoiding interference between users and uses and not so much on maximizing the economic benefits derived from the spectrum. As a result, valuable spectrum is left unused at any given time. This is the motivation for a more spectrum efficient technique, called Dynamic Spectrum Allocation (DSA), where the assigned spectrum blocks may vary in time and space [1,2,3,5,6,7]. Allocating the spectrum dynamically initiates a market-based spectrum allocation method, as well.

5.1.1 Trading and Liberalization

Creating a "market for frequencies" is a possible solution to distribute spectrum efficiently. There are two distinct policies that could be introduced separately or in combination: trading-the transfer of spectrum usage rights between parties in a secondary market, and liberalization-the relaxation of restrictions on services and technologies associated with spectrum usage rights. While trading alone allows the market to determine who has access to spectrum, liberalization lets the users decide how spectrum is used. Without liberalization, secondary market activity will be limited to transfers of existing usage rights. Without trading, liberalization will only enable existing users to switch services and technologies; alternative users would not be able to access spectrum. Introduction of both trading and liberalization could lead to more efficient use of spectrum [8]. Note, however, that trading and/or liberalization is not necessarily appropriate in all cases. For the foreseeable future, spectrum trading will co-exist alongside other spectrum management approaches.

5.1.2 Rights and Obligations

Spectrum usage rights are associated with a set of rights (which enable the user to use spectrum in certain ways) and obligations (which specify conditions that users must fulfill in order to maintain their rights). Usage rights can be defined in relation to four basic parameters [8]: geographical area (e.g., a country, a region or a defined area around a base station), duration and time of access (e.g., unlimited or defined length, access to spectrum throughout the entire day, or at a specific time of day only), spectrum block (i.e., the frequency range to which access is granted), and protection from interference (i.e., the right to receive signals without harmful interference from other spectrum users and the obligation not to cause such interference).

Holders of spectrum usage rights should be free to supply any type of electronic communication service (service neutrality) while ensuring that interference is appropriately dealt with, and to use any technology, abiding by common conditions (technology neutrality) [8].

5.1.3 Interference

Spectrum users in the liberalized environment will need to be confident that neighboring users will meet their obligations, especially in regard to interference, and that their own rights will be upheld. This requirement puts the additional burden on regulators and spectrum users of coordinating new and less predictable interference relationships. As a result, more flexible approaches to interference management would be necessary. Where possible, new, technology-neutral parameters for interference management will be required for adjacent frequencies and geographical areas.

The rest of the chapter is organized as follows. Section 5.2 overviews the related solutions for market driven spectrum distribution proposed so far. Section 5.3 defines a spatio-temporal DSA framework that forms the basis of our proposed allocation and pricing scheme given in Sect. 5.4. Sect. 5.5 gives an illustrative example on the achievable gains using our proposed solution. Finally, Sect. 5.6 concludes the paper.

5.2 Related Works

The authors of [9] introduced a DSA scheme in which a spectrum manager periodically auctions short-term spectrum licenses. The spectrum manager sells spectrum at a unit price, which may be plausible in certain scenarios, e.g., when there is a large number of spectrum buyers and none has enough power to influence the market clearing price. To overcome the situation when these assumptions are violated, in [10] the spectrum manager uses second price (or Vickrey) auctions instead. While earlier works only concentrated on CDMA providers, in [11] the situation is further extended with the presence of a DVB-T network provider. The focus was on the simplest non-trivial model capturing the following issue: a "two island" geography, in which each CDMA network has one cell per island but a single DVB-T cell covers both adjacent cells. This implies inter-cell interference issues and inter-related auctions, where license to use a spectrum band over one island has no value to the DVB-T operator unless it comes with a license to use the same band over the adjacent island.

Somewhat similar questions arise in wired communication networking environment as well. Lazar and Semret introduce the Progressive Second Price (PSP) Mechanism, an iterative auction scheme that allocates bandwidth on a single communication link among users [12] (see also the extensions in [13,14]). The allocations and prices to pay are computed based on the bids submitted by all the players. Users can modify their bids by knowing the bids submitted by the others, until equilibrium is reached. The main drawback of this scheme is that the convergence phase can be quite long. The mechanism was modified by Delenda, who proposed in [15] a one-shot scheme: players are asked to submit their demand functions, and the auctioneer directly computes the allocations and prices to pay without any convergence phase. In [16] they suggested an intermediate mechanism, which is still one-shot, but which does not suppose any knowledge about the demand functions. They considered quasi-linear utility functions, just like in [17], but here they allowed players to submit several bids like in [18], and used an allocation and pricing scheme that is close to the one described in [15]. Unlike in the PSP mechanism, they did not suppose that players know the bids submitted by the others before bidding.

Our solution differs from the previous DSA-pricing proposals, since those solutions used "islands geography" and have not dealt with interference whereas our model takes the interference between regions into consideration. Well elaborated proposals can be found for bandwidthsharing in wired networks. However, these could not be used to give the optimal allocation in our case due to the specialties of dynamic spectrum allocation, but the idea of determining the cost of the allocation is based on the exclusion-compensation principle, that lies behind all second-price mechanisms.

5.3 Spatio-Temporal DSA Model

In a previous paper [19] we defined a model for Dynamic Spectrum Allocation that handles interference issues in a flexible way. A brief summary of the model is presented here as the base for our pricing mechanism proposed in Sect. 5.4.

We consider regions within which we assume that the spatial distribution of the spectrum demand is homogeneous, only temporal changes are allowed. (For example, assume that the spectrum demand in the business quarter of a city, in the suburban region, or on a highway changes with time only.) Assume that the spectrum block to be distributed among all service providers, also called as Coordinated Access Band (CAB) [6], is the frequency range (\tilde{s}, \hat{s}) . The whole area is divided into K nonoverlapping regions (R_k) . Within the given region, M network service providers (NSPs) compete for the spectrum. The spectrum block allocated to the *m*th NSP within the *k* th region at time *t* is:

$$S_{m,k}(t) = (\check{s}_{m,k}(t), \hat{s}_{m,k}(t)).$$
(5.1)

The notations emphasize that the spectrum allocation is highly dynamic, each provider can be given different spectrum blocks at different regions and different time instants. (To ease the notations, the dependence on time *t* is not written explicitly in the followings.) Furthermore, let $|S_{m,k}|$ denote the size of the allocated spectrum block, i.e., $|S_{m,k}| = \hat{s}_{m,k} - \check{s}_{m,k}$.

5.3.1 Interference and Spectrum Efficiency

In our model interference is taken into account as a source for spectrum utilization degradation. "Noisy" spectrum cannot be fully utilized. First of all, spectrum utilization is decreased if the same frequency is used by different NSPs in nearby regions. The level of interference depends on the geographic location and size of the regions, as well as on the radio access technique used, the transmission power, and the positions and types of radio transmitters. This level of interference can be expressed by the *geographic coupling* parameter ε . Let $0 \le \varepsilon_{l,k}^{(m)} \le 1$ denote the "noise level" caused by provider *m* operating in region R_k that can be "heard" within region R_l . It is zero if there is no overhearing at all, and the value of one would mean that the radio transmission is heard undamped. The smaller the geometrical coupling the better from the interferences point of view.

From the NSPs point of view, the level of interference is the measure of how much their radio technology is affected by competing technologies. The level of disturbance (or jamming) between different NSP radio technologies is captured by the *radio technology coupling* parameter η . Let $0 \le \eta_{m,n} \le 1$ denote the coupling between the radio technologies used by the *m*th and *n*th NSPs. Looking at the two extremes, if the two NSPs have the same spectrum slice within the same region and $\eta_{m,n}$ is zero, NSP_n does not affect NSP_m at all, while when $\eta_{m,n}$ equals one means that the spectrum is ruined for NSP_m.

The cumulative effect of the geographic and radio technology couplings on NSP_m operating in region R_k from NSP_n in region R_l having the same spectrum is simply the product of the two factors, namely, $\varepsilon_{l,k}^{(n)} \cdot \eta_{m,n}$.

Having the appropriate model parameters to capture interference, let $\xi(S_{m,k})$ denote the *efficiency* of spectrum block $S_{m,k}$ that can be calculated as

$$\xi(S_{m,k}) = \frac{1}{|S_{m,k}|} \int_{S_{m,k}} \xi_{m,k}(\lambda) \, d\lambda, \qquad (5.2)$$

where $\xi_{m,k}(\lambda)$ is the efficiency of frequency λ from NSP_ms point of view in region R_k , that is

$$\boldsymbol{\xi}_{m,k}(\boldsymbol{\lambda}) = \prod_{i=1}^{M} \prod_{j=1}^{K} \left(1 - \boldsymbol{\varepsilon}_{j,k}^{(i)} \cdot \boldsymbol{\eta}_{m,i} \cdot \boldsymbol{I}_{\{\boldsymbol{\lambda} \in \boldsymbol{S}_{i,j}\}} \right).$$
(5.3)

Here $I_{\{\lambda \in S_{i,j}\}}$ indicates whether frequency λ is allocated to NSP_i in region R_j or not. The efficiency is one if no interference occurs and less than one if there is interference with neighboring regions.

5.3.2 Feasible Allocation

In our DSA scenario providers can have different capacity demands in different regions. Assume that the *m*th provider in the *k*th region has the capacity request $c_{m,k}$. We call an allocation $\mathbf{S} = (\mathbf{S}_1, ..., \mathbf{S}_M)$ with $\mathbf{S}_m = (S_{m,1}, ..., S_{m,K})$ feasible, if the spectrum blocks $\{S_{m,k}\}$ used by the NSPs satisfy the following conditions:

$$|S_{m,k}| \ge c_{m,k}, \quad \forall m,k, \tag{5.4}$$

$$\xi(S_{m,k}) \ge \beta_m, \quad \forall m, k, \tag{5.5}$$

$$\min_{\lambda \in S_{m,k}} \xi_{m,k}(\lambda) \ge \alpha_m, \quad \forall m, k.$$
(5.6)

In words, (5.4) makes sure that each allocated spectrum block is big enough to satisfy the NSPs capacity request. However, the real question is how the service of the NSP degrades if the allocated spectrum block for the service is "noisy", i.e., its efficiency is less than one. Thas is why conditions (5.5) and (5.6) assure that the spectrum "quality" is good enough for the given service. It can happen that robust techniques with error-prone encoding, or wideband solutions are more tolerant to noisy spectrum than others. In our model, parameters α and β try to answer this question. These two parameters can be seen as tolerance levels, i.e., to what extent the interference is tolerated by the provided service. We assume that from the operators point of view—when the spectrum efficiency is above the tolerable limit, the service can be provided adequately. Parameter β_m gives the minimum spectrum quality that must be met on the average (see (5.5)), while parameter α_m prescribes the minimal efficiency that must be available at all frequencies in the allocated block (see (5.6)).

Physically, parameter ε represents the attenuation of the signal between the regions. This can be calculated based on the signal strength, the distance between the regions and some kind of propagation model. Parameter η expresses co-existence and synchronization capability of the different radio technologies. The value of η can be estimated or determined via detailed simulations. Parameters α and β are thresholds for the signal to interference ratio.

5.3.3 Interference Tolerance

Knowing the real tolerance levels of an NSP would be a great help in the DSA framework. However, from the operators point of view, it would be much easier to say that interference is not welcome at all, "clear" spectrum block is needed to provide the service ensuring maximal user satisfaction. If all providers were intolerant to interference, then strictly disjoint spectrum blocks were needed even in neighboring (or coupled) regions everywhere. This would greatly reduce the effectiveness of DSA, and the result would look like nearly the same as the rigid spectrum allocation methodology used today. This would cause the spectrum to be more scarce, and thus *more expensive* at the end. In a DSA scenario where tolerance is much rewarded by increasing spectral efficiency, certain mechanisms need to be implemented to make it desirable for the NSPs to use all available techniques to tolerate co-existence as much as possible. A proper *pricing scheme* that charges providers who do not tolerate others and interfere to larger extent than necessary with other regions would be of great importance.

5.4 Pricing Scheme in the Proposed DSA Model

Our goal here is to propose a one-shot multi-bid pricing scheme that takes the special properties of the dynamic spectrum sharing into account. Although the progressive second price auction mechanism gives a suitable allocation for infinitely divisible resources, in case of dynamic spectrum allocation the size of the distributable spectrum cannot be explicitly determined, due to interference. It may happen that inside one region there will be carriers that cannot be distributed because of the interference arising from the neighboring regions, while other carriers can be allocated to more than one provider that do not disturb each other. This was the reason to suggest a proper pricing mechanism that satisfies the special needs of dynamic spectrum allocation and that charges providers that do not tolerate others and cause large interference to other regions.

5.4.1 Inputs

According to the proposed allocation scheme the spectrum is redistributed at given time intervals. Before the beginning of each time-period the providers send their multi-bids to a centralized spectrum broker entity (SB). This entity calculates the optimal feasible allocation—based on the DSA model—that maximizes social welfare, and also calculates the costs for the providers.

Let $I = \{1, ..., i, ..., I\}$ denote the set of *players*. Since the demands in different regions can be different, we handle the NSPs in each region separately, i.e., $I = M \cdot K$.

Player i submits a set of $N^{(i)}$ two-dimensional bids

$$B_i = \{b_{i,1}, \dots, b_{i,N^{(i)}}\},\tag{5.7}$$

where

$$b_{i,n} = (q_{i,n}, p_i(q_{i,n})), \quad n = 1, ..., N^{(i)},$$
(5.8)

where q represents the quantity of the demanded resource and p(q) is the offered price for the desired quantity.

The auctioneer (SB) collects all multi-bids to form the *multi-bid profile* as input.

$$B = (B_1, \dots, B_I).$$
(5.9)

5.4.2 Allocation and Pricing Rules

The multi-bid profile is used to compute the allocation a_i and its price c_i for each player $i \in I$, using the *allocation rule* A and related *pricing* scheme C.

An allocation rule A gives back the *allocation vector* as

$$A(B) == (a_1, ..., a_I), \tag{5.10}$$

where

$$a_i \in \{0, b_{i,1}, \dots, b_{i,N^{(i)}}\}, \quad i = 1, \dots, I,$$
(5.11)

i.e., the size of the allocated spectrum block for player i is chosen from the player's bid profile, or it is zero if none of its requests can be satisfied.

The pricing scheme C for a given allocation is formalized as

$$C(A(B)) = C(\mathbf{a}) = (c_1, ..., c_I), \quad i = 1, ..., I,$$
 (5.12)

where $c_i \leq p_i(a_i)$ is the price that player *i* must pay after given a_i amount of spectrum to use. The price imposed cannot be higher than the maximum offered price.

We define our allocation rule A for a given bid profile B as

$$A(B) = \arg \max_{\mathbf{a} \in Q^{f}} \sum_{i=1}^{I} p_{i}(a_{i}),$$
 (5.13)

where Q^f is the set of feasible allocations, i.e., for all $\mathbf{a} \in Q^f$ there exists a *spectrum allocation* $S(\mathbf{a}) = \{S_{1,1}, \dots, S_{M,K}\}$ where $|S_{m,k}| = a_{(m-1)K+k}$, which satisfies the feasibility conditions (5.5) and (5.6). In words, (5.13) finds the allocation vector that would maximize the total income *if* the players had to pay the maximum that they were willing to pay for the acquired spectrum amount. However, we propose to use a different pricing scheme that, instead of maximizing the total income, maximizes the allocation's *efficiency*.

The efficiency $\Theta(\mathbf{a})$ of an allocation is defined as

$$\Theta(\mathbf{a}) = \sum_{i=1}^{I} \Theta_i(a_i), \qquad (5.14)$$

where $\theta_i(a_i)$ is the *valuation* of player *i*. We call an allocation $\tilde{\mathbf{a}}$ optimal if it is efficient and feasible, that is

$$\tilde{\mathbf{a}} = (\tilde{a}_1, \dots, \tilde{a}_I) = \arg \max_{\mathbf{a} \in Q^f} \Theta(\mathbf{a}).$$
(5.15)

Comparing (5.13) and (5.15), our allocation rule would result in an optimal allocation only if the players bid $p_i(q) = \theta_i(q)$ as the price they offer for the requested quantity. This can be achieved with a proper pricing scheme that eventually leads the players to bid "honestly", i.e., tell the auctioneer how valuable the spectrum is for them. Using "second-price" (or Vickery) pricing scheme can be a solution.

The intuition behind is an exclusion-compensation principle which lies behind all second-price mechanisms: player i pays so as to cover the "social opportunity cost", that is to say the loss of utility he imposes on all other players by his presence. The price that player i has to pay can be calculated based on the declared willingness to pay (bids) of all players who are excluded by i 's presence. It can be shown that, by using this scheme, telling the truth (i.e., setting the bid price equal to the valuation) is a dominant strategy [12]. Thus, from now on we assume that all players bid $(q_i, \theta_i(q_i))$ pairs in their bid profiles. This results in that the allocation is economically efficient in that it maximizes total user valuation (social welfare).

The multi-bid profile obtained by deleting the bid of player i, also called as the *opponents' profile*, is defined as

$$B^{(-i)} = (B_1, \dots, B_{i-1}, 0, B_{i+1}, \dots, B_I).$$
(5.16)

Similar to (5.15), we can calculate the optimal feasible allocation for the $B^{(-i)}$ profile as well:

$$\tilde{\mathbf{a}}^{(-i)} = \left(\tilde{a}_1^{(-i)}, \dots, \tilde{a}_I^{(-i)}\right) = \arg\max_{\mathbf{a}^{(-i)} \in \mathcal{Q}^f} \Theta\left(\mathbf{a}^{(-i)}\right), \tag{5.17}$$

where

$$\mathbf{a}^{(-i)} = A\left(B^{(-i)}\right). \tag{5.18}$$

As explained above, in a second-price auction player i is charged a total price of

$$c_i(A(B)) = \Theta_i(\tilde{a}_i) - \left[\Theta(\tilde{\mathbf{a}}) - \Theta\left(\tilde{\mathbf{a}}^{(-i)}\right)\right] =$$
(5.19)

$$=\sum_{\substack{j=1\\j\neq i}}^{I} \left[\Theta_{j}\left(\tilde{a}_{j}^{(-i)}\right) - \Theta_{j}\left(\tilde{a}_{j}\right)\right].$$
(5.20)

In words, player *i* must pay exactly the amount that those players would pay who were excluded by player *i*'s presence. This price can be calculated by obtaining the allocations and prices with (i.e., using *B*) and without (i.e., using $B^{(-i)}$) player *i*'s bid, and taking their difference at the end.

5.5 Example

Consider a simple scenario with two regions as shown on Fig. 5.1. Two NSPs operate in both regions within the CAB. In region A there is also a DVB-T provider that covers region B, too. Furthermore, in region B a UWB provider is also present.

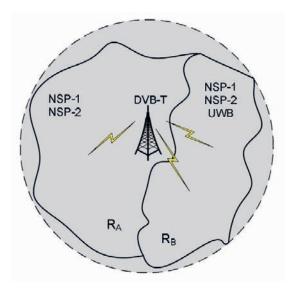


Fig. 5.1 Example Scenario

The DSA model is characterized by matrices η , ε^1 , ε^2 , ε^3 , and ε^4 (see Table 5.1 and 5.2).

Table 5.1 Radio technology coupling parameters η	1 _{m.n}
---	-------------------------

η	NSP-1	NSP-2	DVB-T	UWB
NSP-1	0	0.5	0.6	0.05
NSP-2	0.4	0	0.7	0.1
DVB-T	0.5	0.6	0	0.05
UWB	0.05	0.05	0.05	0

In words, matrix η shows the radio technology coupling parameters between the NSPs. Recall, that $\eta_{m,n}$ denotes the coupling between the radio technologies used by the *m*th and *n*th NSPs. The smaller this value, the better it is from the interferences point of view. By looking at the elements of matrix η in Table 5.1 we can see that the radio technology used by the UWB provider does not affect, and is not affected significantly by other providers ($\eta \ll 1$). **Table 5.2** Geographic coupling parameters $\mathcal{E}_{AB}^{(m)} = \mathcal{E}_{BA}^{(m)}$

$$\frac{\text{NSP-1 NSP-2 DVB-T UWB}}{\varepsilon_{AB}^{(.)} 0.2 0.3 1 0}$$

Table 5.2 shows the geographic coupling parameters between the regions. Recall, that $\mathcal{E}_{l,k}^{(m)}$ denotes the "noise level" caused by provider *m* operating in region R_k that can be "heard" within region R_l . By looking at the elements of matrix \mathcal{E}^4 we can see a strong coupling ($\mathcal{E} \approx 1$) between the regions ensuring the DVB-T provider to cover both regions. It has also a low tolerance level (see Table 5.3 for the interference tolerance parameters; $\alpha \approx 1$ and $\beta \approx 1$).

Table 5.3 Average (β_m) and maximum (α_m) interference tolerance parameters

	NSP-1	NSP-2	DVB-T	UWB
α	0.5	0.7	0.95	0.5
β	0.9	0.95	0.975	0.8

Table 5.4 Multibids

$B_1 : l$	$NSP-1(R_A)$	$B_4: NSP-1 (R_B)$
q	0 25 30	q 0 20 30
Θ	0 25 30 0 75 85	$\begin{array}{c cccc} q & 0 & 20 & 30 \\ \hline \Theta & 0 & 40 & 50 \end{array}$
	$NSP - 2 (R_A)$	$B_5: NSP - 2 \ (R_B)$
q	0 10 25	q 0 15 20
Θ	0 10 25 0 30 45	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
	$DVB - T(R_A)$	B_6 : UWB (R_B)
\overline{q}	0 5 10	$q \mid 0 \mid 20$
Θ	0 5 10 0 25 35	$\Theta \mid 0 \mid 20$

Table 5.4 lists the multi-bids of the providers

The optimal allocation and the total costs of the providers was calculated based on the allocation and pricing rules described in Sect. 5.4. A simulation tool was developed using Matlab to calculate the optimal values. The feasibility check of the solution is based on a Simulated Annealing process. Table 5.5 shows the optimal feasible allocation and the costs of the providers.

		R_A			R_{B}		
		NSP-1	NSP-2	DVB - T	NSP-1	NSP-2	UWB
	q	25	10	5	20	15	20
	с	15	10	20	10	10	0
С	/q	0.6	1	4	0.5	0.6	0

Table 5.5 Optimal allocation and the costs of the providers

The third row of the table shows the average cost of one spectrum unit for each provider.

The DVB-T provider pays far the most from among the providers, since it covers both regions and its level of tolerance is low.

Similarly, we can see that the cost assigned to the UWB provider in Table 5.5 is zero, since it does not interfere with the other providers. The zero cost can mean for example, that this provider does not have to pay any additional cost over the base spectrum unit price.

We also note, that the geographical coupling of NSP-1 is less than that of NSP-2, meaning that it causes less interference than the other, henceforth the unit price is also smaller than for NSP-2.

5.6 Conclusion

A market-driven Dynamic Spectrum Allocation (DSA) framework is a promising new approach to increase the usage efficiency of radio spectrum. In this chapter we proposed an auction and pricing method that can be used to distribute spectrum usage rights among competing radio network service providers (NSPs). The outlined solution is a one-shot multibid auction method, based on our spatio-temporal DSA framework [19] that takes into account interference issues in a flexible way. The suggested pricing mechanism satisfies the special needs of DSA, and charges providers that do not tolerate others, and cause heavy interference to other regions. The intuition behind this rule is that a particular user pays so as to cover the loss of utility he imposes on others by his presence.

References

- 1. "IST-DRiVE project," http://www.ist-drive.org.
- P. Leaves, S. Ghaheri-Niri, R. Tafazolli, L. Christodoulides, T. Sammut, W. Stahl, and J. Huschke, "Dynamic spectrum allocation in a multi-radio environment: Concept and algorithm," in *Proc., IEE Second International Conference on 3G Mobile Communication Technologies*, London, United Kingdom, pp. 53–57, 26–28 March 2001.
- P. Leaves, J. Huschke, and R. Tafazolli, "A summary of dynamic spectrum allocation results from DRiVE," in *Proc., IST Mobile and Wireless Telecommunications Summit*, Thessaloniki, Greece, 16–19 June 2002, pp. 245–250.
- 4. "IST OverDRiVE project," http://www.ist-overdrive.org.
- M. Buddhikot, P. Kolodzy, S. Miller, K. Ryan, and J. Evans, "DIMSUMnet: New directions in wireless networking using coordinated dynamic spectrum access," in *Position Paper in IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (IEEE WoWMoM 2005)*, Taromina/Giardini Naxos, Italy, June 2005.
- M. Buddhikot and K. Ryan, "Spectrum management in coordinated dynamic spectrum access based cellular networks," in *Proc., First IEEE International Symposium on New Directions in Dynamic Spectrum Access Networks*, Baltimore, MD, 8–11 November 2005.
- L. Kovács, J. Tapolcai, and A. Vidács, "Spatio-temporal dynamic spectrum allocation with interference handling," in *IEEE International Conference on Communications (ICC 2007)*, Glasgow, Scotland, UK, 24–28 June 2007.
- "Study on conditions and options in introducing secondary trading of radio spectrum in the european community," Study for the European Commission by consultants Analysys, DotEcon and Hogan and Hartson LLP, May 2004, summary of the report.
- 9. V. Rodriguez, K. Moessner, and R. Tafazolli, "Market-driven dynamic spectrum allocation: Optimal end-user pricing and admission control for CDMA," in *Proc.*, 14th European Information Society Technologies (IST) Mobile and Wireless Communications Summit, Dresden, Germany, June 2005.
- V. Rodriguez, K. Moessner, and R. Tafazolli, "Auction driven dynamic spectrum allocation: Optimal bidding, pricing and service priorities for multi-rate, multi-class CDMA," in *Proc., IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Berlin, Germany, September 2005.
- V. Rodriguez, K. Moessner, and R. Tafazolli, "Auction driven dynamic spectrum allocation over space and time: DVB-T and multi-rate, multi-class CDMA over a two-island geography," in *Proc., 15th European Information Society Technologies (IST) Mobile and Wireless Communications Summit,* Míconos, Greece, 4–8 June 2006.

- A. Lazar and N. Semret, "Design and analysis of the progressive second price auction for network bandwidth sharing," *Telecommunication Systems* - Special issue on Network Economics, 1999.
- 13. B. Tuffin, "Revisited progressive second price auction for charging telecommunication networks," *Telecommunication Systems*, vol. 20, no. 3, pp. 255–263, 2002.
- 14. P. Maillé and B. Tuffin, "A progressive second price mechanism with a sanction bid from excluded players," in *Proc., 3rd International Workshop on Internet Charging and QoS Technologies, ICQT 2003*, Munich, September 2003.
- 15. A. Delenda, "Mécanismes d'encheres pour le partage de resources télécom, France Telecom, Tech. Rep. 7831," 2002.
- 16. P. Maille and B. Tuffin, "Multi-bid auctions for bandwidth allocation in communication networks," in *Proc., The 23rd Conference of the IEEE Communications Society (INFOCOM 2004)*, Hong Kong, 7–11 March 2004.
- 17. N. Semret, "Market mechanisms for network resource sharing, Ph.D. dissertation, Columbia University," 1999.
- X. Wang and H. Schulzrinne, "Auction or tâtonnement finding congestion prices for adaptive applications," in *Proc.*, 10th IEEE International Conference on Network Protocols, ICNP 2002, Paris, France, November 12–15 2002.
- 19. L. Kovács and A. Vidács, "Interference-tolerant spatio-temporal dynamic spectrum allocation," in *Second IEEE International Symposium on New Directions in Dynamic Spectrum Access Networks*, Dublin, Ireland, April 2007.

6 Resource Allocation Strategies for SDMA/OFDMA Systems

Tarcisio F. Maciel and Anja Klein

6.1 Introduction

Space Division Multiple Access (SDMA) and Orthogonal Frequency Division Multiple Access (OFDMA) are promising technologies for the provision of flexible high-rate services in future mobile radio networks [1,2]. However, due to the large number of degrees of freedom in space, frequency, and time, SDMA/OFDMA systems need a sophisticated Resource Allocation (RA) strategy to efficiently assign resources to the users.

If Channel State Information (CSI) about users' channels is available at the Base Station (BS), either through direct estimation or feedback, SDMA can be used to serve a group of users simultaneously on the same radio resource. In the following, such a group is termed an SDMA group. The efficiency of the RA strategy can be influenced, first, by selecting which users to place in the same SDMA group, and second, by deciding whether and when to grant a resource to a group.

In general, performing optimum RA in space, frequency, and time is a very complex problem. For such optimum RA, the tasks of building SDMA groups and assigning resources to groups cannot be decoupled. Thus, it requires comparing the efficiency of assigning every possible SDMA group on every possible resource. This problem can be recognized as Non-Polynomial time Hard (NP-H) and has exponentially increasing complexity [3,4,5,6]. To simplify this problem, it is proposed here to suboptimally divide it into two tasks:

Task 1: Arranging users in SDMA groups. This is done by an SDMA algorithm that places spatially compatible users in the same SDMA group. Spatial compatibility, i.e., how efficiently users can be separated in space, is measured by a grouping metric computed using the available CSI.

Task 2: Assigning resources to users or SDMA groups. This is accomplished by a joint frequency/time assignment algorithm.

Finding the SDMA group that optimizes an adopted grouping metric on a given resource is known to be an NP-H problem [3,4] while joint frequency/time assignment of users or SDMA groups to resources can also become NP-H depending on the problem constraints [7]. Therefore, tasks 1 and 2 also ask for suboptimal solutions.

In order to solve the RA problem in space, frequency, and time, two suboptimal RA strategies are proposed in this chapter:

- The Space-Frequency/Time Resource Allocation (S-FT RA) strategy: as its name suggests, this strategy first handles the spatial component of the RA problem. It solves task 1 using a new low-complexity SDMA algorithm named Greedy Regularized Correlation-Based Algorithm (GRCBA) and builds with it a set of candidate SDMA groups. Then, it solves task 2 by formulating the frequency/time assignment of the candidate groups to the radio resources as a standard assignment problem [8], which is solved using Munkres' algorithm [9].
- The Frequency/Time-Space Resource Allocation (FT-S RA) strategy: as its name also suggests, this strategy first handles the frequency/time component of the RA problem. It solves task 2 by assigning an initial user to each radio resource. This task is also formulated as a standard assignment problem and solved using Munkres' algorithm. Then, starting with the initially assigned users, task 1 is solved by building SDMA groups on each resource using the Successive Projections Algorithm (SPA) [10,11].

Considerable work on SDMA grouping has been done to suboptimally solve task 1. In [12], greedy algorithms are proposed to iteratively build SDMA groups based on the Signal-to-Interference plus Noise Ratio (SINR) of the users. In [10,11,13], greedy algorithms are proposed, which base on null-space projections of the channels of the users being added to the SDMA group. In [14], several SDMA groups are built and compared afterwards in terms, e.g., of the group capacity, in order to determine the best group.

The SDMA algorithms in [3,10,11,13,14], among others, achieve high system capacity, but their grouping metrics depend on relatively complex operations, such as linear precoding or null-space projections, which increase the complexity of the algorithms when looking for the best SDMA group. In [4,5,6,15], SDMA algorithms whose grouping metrics are based on the spatial correlation are proposed and they are shown to be very efficient and to have low complexity.

The proposed GRCBA considered in the S-FT RA strategy to solve task 1 is a greedy variant of Regularized Correlation-Based Algorithm (RCBA) proposed by the authors in [6]. Its grouping metric is based on the correlation and gains of users' spatial channels. It has almost the same performance as RCBA, but lower complexity. The SPA from [10,11] considered in the FT-S RA strategy is also more complex than the proposed GRCBA.

Frequency/time assignment algorithms have been often studied in the context of resource assignment, bit-and-power-loading, cf. e.g. [16,17,18]. Such algorithms usually aim at efficiently allocating resources to SDMA groups while adjusting the number of bits and the power allocated to each user on the radio resources. Indeed, they are very efficient and enhance system throughput, but are also considerably complex.

For solving task 2, a new suboptimal frequency/time resource assignment algorithm is proposed here, which aims at maximizing the total weighted capacity of the system. Therefore, it takes into account both capacity and Quality of Service (QoS) aspects. It is formulated as a standard assignment problem and, for the S-FT RA strategy, it assigns groups to resources, while for the FT-S RA strategy, it assigns users to resources. However, its mathematical structure is the same in both cases.

In order to simplify the RA strategies, power is equally divided among subcarriers [4] and is allocated to users in the SDMA groups according to linear transmit Zero-Forcing (ZF) precoding [1,19], which at the same time increases throughput fairness in the system. The proposed RA strategies aim at providing a good trade-off between performance, complexity, and fairness.

The remainder of this chapter is organized as follows. Section 6.2 describes the adopted system model. Section 6.3 introduces the S-FT RA strategy. Section 6.4 describes the FT-S RA strategy. In Sect. 6.5 the performance and complexity of the proposed RA strategies is analyzed. Finally, Sect. 6.6 draws some conclusions.

6.2 System Model

This section describes the system model considered in this chapter. The downlink of a single BS is considered in the modeling. Interference from other BSs is assumed as Gaussian and is incorporated directly as part of the additive white Gaussian noise in the system. The BS has an n_T -element Antenna Array (AA) and there are *K* single-antenna users associated with the BS, thus totalizing $n_R = K$ receive antennas.

The BS uses OFDMA and its bandwidth is divided into *S* orthogonal subcarriers whose channel responses are assumed to be flat. A block of *Q* adjacent subcarriers, also called cluster or chunk [20], is defined as radio resource. A total number of $N = \lfloor S/Q \rfloor$ resources is assumed, where $\lfloor \cdot \rfloor$ is the nearest integer lower than or equal to the argument. *Q* is chosen such that channel responses do not vary much within a chunk, thus reducing the required signaling/estimation effort without much degradation of the system performance. Considering chunks of *Q* adjacent subcarriers reduces the number of resources on which SDMA groups have to be built, thus simplifying task 1, as well as it reduces the dimension of the frequency/time assignment problem and, therefore, also simplifies task 2.

On a subcarrier *s*, each link between the BS and a user *k* has an associated $1 \times n_T$ complex vector channel \mathbf{h}_{ks} . Denoting transposition by $(\cdot)^T$, the $K \times n_T$ channel matrix \mathbf{H}_s on subcarrier *s* of all users can be written by stacking the channel vectors \mathbf{h}_{ks} as

$$\mathbf{H}_{s} = \begin{bmatrix} \mathbf{h}_{1s}^{T} & \mathbf{h}_{2s}^{T} & \cdots & \mathbf{h}_{Ks}^{T} \end{bmatrix}^{T}.$$
(6.1)

Each resource is represented by its middle subcarrier, whose channel response is assumed to be perfectly known at the BS. The channel response \mathbf{H}_n of all users on resource *n* is written as in (6.1) using the referred middle subcarrier.

On resource *n*, building an SDMA group \underline{G} corresponds to adequately selecting a total of $n_G \leq n_T$ vector channels \mathbf{h}_{jn} , j = 1, ..., K, of \mathbf{H}_n , i.e., to optimally select n_G out of the *K* rows of \mathbf{H}_n according to the adopted grouping metric and problem constraints. n_G is the cardinality of \underline{G} , i.e., n_G is the number of users in the SDMA group \underline{G} .

Whenever <u>G</u> is scheduled for transmission on resource *n*, all the subcarriers of this resource are allocated to all users of the group. On each subcarrier *s* of the resource, the BS transmits data symbols d_{gs} , $g = 1, ..., n_G$, to all users in the group. The data symbols d_{gs} are assumed to be uncorrelated with average power σ_d^2 and are arranged in the $n_G \times 1$ input data vector \mathbf{d}_s , which is precoded using the $n_T \times n_G$ modulation matrix \mathbf{M}_n , transmitted through the $n_G \times n_T$ SDMA group channel \mathbf{G}_s , and distorted by noise, which is represented by the $n_G \times 1$ vector \mathbf{z}_s . \mathbf{z}_s is considered to be spatially white with average power σ_z^2 . The received signal is demodulated using the $n_G \times n_G$ demodulation matrix \mathbf{D}_n producing at the receivers the output data vector

$$\hat{\mathbf{d}}_{s} = \mathbf{D}_{n} \big(\mathbf{G}_{s} \mathbf{M}_{n} \mathbf{d}_{s} + \mathbf{z}_{s} \big). \tag{6.2}$$

Both \mathbf{M}_n and \mathbf{D}_n in (6.2) depend on the adopted precoding technique and are applied to all the subcarriers of resource *n*. Since the demodulation process is distributed among the users, \mathbf{D}_n is diagonal and decouples signals received by different users.

Let $(\cdot)^{H}$ denote conjugate transposition. Let \mathbf{m}_{jn} denote the *j*-th column of the modulation matrix \mathbf{M}_{n} on resource *n*. Let \mathbf{g}_{jn} denote the *j*-th row of the SDMA group channel matrix \mathbf{G}_{n} corresponding to the middle subcarrier of resource *n*. Assuming Gaussian signaling and perfect matched filtering at the receivers, the group capacity $C(\underline{G})$ of the SDMA group \underline{G}_{n} on resource *n* is estimated as

$$C(\underline{G}_n) = Q \sum_{i=1}^{n_G} \log_2 \left(1 + \frac{\mathbf{m}_{in}^H \mathbf{g}_{in}^H \mathbf{g}_{in} \mathbf{m}_{in}}{\sigma_z^2 + \sum_{j=1, j \neq i}^{n_G} \mathbf{m}_{jn}^H \mathbf{g}_{in}^H \mathbf{g}_{in} \mathbf{m}_{jn}} \right).$$
(6.3)

The spatial correlation between two vector channels \mathbf{h}_i and \mathbf{h}_j is measured by the maximum normalized scalar product [3,4,5,6]. Let $|\cdot|$ denote the absolute value of a complex scalar, $||\cdot||_2$ denote the 2-norm of a vector, and diag $\{\cdot\}$ denote a diagonal matrix whose diagonal elements are given in the vector argument. Using (6.1), a non-negative real $K \times K$ matrix \mathbf{R}_n containing the spatial correlation of every pair of channels \mathbf{h}_{in} and \mathbf{h}_{jn} from \mathbf{H}_n can be written as

$$\mathbf{R}_{n} = \left| \operatorname{diag} \{ \mathbf{a}_{n} \}^{1/2} \mathbf{H}_{n} \mathbf{H}_{n}^{H} \operatorname{diag} \{ \mathbf{a}_{n} \}^{1/2} \right|, \text{ with}$$
$$\mathbf{a}_{n} = \left[\left\| \mathbf{h}_{1n} \right\|_{2}^{-2} \quad \left\| \mathbf{h}_{2n} \right\|_{2}^{-2} \quad \cdots \quad \left\| \mathbf{h}_{Kn} \right\|_{2}^{-2} \right]^{T}, \tag{6.4}$$

where $|\cdot|$ is applied to \mathbf{R}_n element-wise. \mathbf{R}_n is used as input for the SDMA algorithms in the next sections.

6.3 Space-Frequency/Time Resource Allocation

This section introduces the S-FT RA strategy proposed in this chapter. Section 6.3.1 describes the proposed SDMA algorithm. Section 6.3.2 introduces the frequency/time assignment algorithm. Section 6.3.3 describes how the algorithms in Section 6.3.1 and Section 6.3.2 integrate into the S-FT RA strategy.

6.3.1 SDMA Algorithm: Greedy Regularized Correlation-Based Algorithm

In this section, the proposed GRCBA is introduced. Its grouping metric is based on the spatial correlation and gains of the spatial channels of the users. Consequently, it does not depend on the precoding matrices and has low complexity.

If spatial channels of the users in the same SDMA group are close to orthogonal, spectral efficiency gains are obtained. However, if they are spatially correlated, SDMA can even lead to spectral efficiency losses. Considering linear ZF precoding [19] and a fixed group size n_G , building an SDMA group *m* on resource *n*, indicated by <u>*G*</u>_{*mn*}, whose n_G channels are as uncorrelated as possible, is a good candidate solution for the SDMA grouping problem.

Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix, $[\cdot]_i$ denote the *i*-th element of a vector, and $[\cdot]_{ij}$ denote the element at the *i*-th row and *j*-th column of a matrix. In the GRCBA, shown in Table 6.1, an initial user vector channel \mathbf{h}_{cn} , indexed by *c*, is selected and a one-user SDMA group $\underline{G}_{mn} = \{c\}$ is built. Then, the spatial channel most compatible with respect to all the channels already admitted to \underline{G}_{mn} is added to the group. This procedure is repeated until the desired group size n_G is reached. $0 \le \alpha \le 1$ is a parameter controlling the preference for highly uncorrelated users or users with high channel gain when admitting a new user to the group \underline{G}_{mn} .

Table 6.1 Greedy regularized correlation-based algorithm

1. Set the SDMA group
$$\underline{G}_{mn} = \{c\}$$
.
2. For $g = 1$ to $n_G - 1$
Set $\underline{G}_{mn} = \underline{G}_{mn} \bigcup \arg\min_{c'} \left\{ \frac{1 - \alpha}{\|\mathbf{R}_n\|_F} \sum_{j} [\mathbf{R}_n]_{jc'} + \frac{\alpha}{\|\mathbf{a}_n\|_2} [\mathbf{a}_n]_{c'} \right\},$
with $c' \in \{1, 2, ..., K\} \setminus \underline{G}_{mn}$, and $j \in \underline{G}_{mn}$.

For a user, the lower his channel attenuation and his spatial correlation with the users in \underline{G}_{mn} are, the higher the chance of being admitted to \underline{G}_{mn} in the step 2 of the GRCBA in Table 6.1. GRCBA is a greedy variant of RCBA introduced in [6]. GRCBA and RCBA have almost the same grouping metric. However, GRCBA builds the SDMA group based on a simple algorithm while RCBA solves the optimization problem

$$\mathbf{x}_{n}^{opt} = \underset{\mathbf{x}_{n}}{\operatorname{arg\,min}} \left\{ \frac{1-\alpha}{\|\mathbf{R}_{n}\|_{F}} \mathbf{x}_{n}^{T} \mathbf{R}_{n} \mathbf{x}_{n} + \frac{\alpha}{\|\mathbf{a}_{n}\|_{2}} \mathbf{a}_{n}^{T} \mathbf{x}_{n} \right\}, \text{ subject to :}$$

$$\|\mathbf{x}_{n}\|_{1} = G, \qquad (6.5)$$

$$[\mathbf{x}_{n}]_{c} = 1, c \in \{1, 2, \dots, K\},$$

$$[\mathbf{x}_{n}]_{j} \in [0, 1], j = 1, 2, \dots, K,$$

where $\|\cdot\|_1$ is the 1-norm of a vector. The second constraint allows forcing a given user, indicated by *c*, to be in the SDMA group. The problem in (6.5) is the convex relaxation of the associated integer optimization problem and can be solved with non-exponential complexity using convex optimization methods [6,21]. While the RCBA builds a whole SDMA group at once, the GRCBA sequentially extends an initial group by admitting one new user per iteration. Anyway, the RCBA might require a considerable number of iterations of a convex optimization procedure and the GRCBA is proposed here to further simplify the SDMA grouping problem.

6.3.2 Frequency/Time Assignment Algorithm: Group-to-Resource Assignment

In this section, the proposed frequency/time assignment algorithm is described. Its objective is finding a scheduling of groups to resources that maximizes a weighted sum of revenues.

Consider a total number *M* of SDMA groups and a total number *N* of resources, with SDMA groups built using either RCBA or GRCBA. For each group *m*, three variables are defined:

- A non-negative group priority *w_{mn}*.
- A non-negative allocation revenue p_{mn} obtained by allocating the group *m* to resource *n*
- An assignment variable u_{mn} in $\{0, 1\}$ indicating whether group *m* is allocated on resource *n*.

The total weighted revenue of the system can be written as $\sum_{n=1}^{M} \sum_{j=1}^{N} w_{n} n_{j} u_{j}$

to
$$\sum_{m=1}^{\infty} \sum_{m=1}^{\infty} w_{mn} p_{mn} u_{mn}$$
.

By arranging w_{mn} , p_{mn} , and u_{mn} into an $M \times N$ weighted revenue matrix **W** and an $M \times N$ assignment matrix **U** as follows

$$\mathbf{W} = [\mathbf{W}]_{mn} = w_{mn}p_{mn}, \text{ and } \mathbf{U} = [\mathbf{U}]_{mn} = u_{mn}, \tag{6.6}$$

denoting by • the Hadamard product, and denoting by $\mathbf{1}_L$ an $L \times 1$ vector of ones, the problem of maximizing the total weighted revenue of the system can be written as

$$\mathbf{U}^{opt} = \arg\max_{\mathbf{U}} \left\{ \mathbf{1}_{M}^{T} (\mathbf{W} \bullet \mathbf{U}) \mathbf{1}_{N} \right\}, \text{ subject to :}$$

$$\mathbf{1}_{M}^{T} \mathbf{U} = \mathbf{1}_{N}^{T},$$
(6.7)

where the considered constraint ensures assigning only one group to each resource. The problem in (6.7) is a standard assignment problem [8] and is efficiently solved herein by applying Munkres' algorithm [9].

The allocation objective proposed here is to maximize the weighted capacity of the system. This allocation objective is termed further the Maximum Weighted Capacity (MWC) objective.

For the MWC objective, the group priority w_{mn} is defined as sum of the non-negative priorities v_{kn} of the users within the group, i.e.,

$$w_{mn} = \sum v_{kn}$$
, with $k \in \underline{G}_{mn}$, (6.8)

so that the higher the users' priorities are, the higher the priority of the SDMA group becomes. Moreover, user's priority management is kept reasonably decoupled from the SDMA grouping algorithm. Denoting by R_k and \overline{R}_k the contracted and measured average throughputs of user k, respectively, v_{kn} is defined herein as

$$V_{kn} = \frac{R_k}{\overline{R}_k} \,. \tag{6.9}$$

This makes group priorities frequency-independent and privileges groups containing users whose QoS levels are below the contracted ones.

In order to exploit multi-user diversity gains, the allocation revenue p_{mn} of the SDMA group \underline{G}_{mn} should be high if its current achievable rate on this resource is high. Therefore, p_{mn} is defined herein as

$$p_{mn} = C(\underline{G}_{mn}), \tag{6.10}$$

with $C(\underline{G}_{nn})$ given by (6.3). Therefore, groups whose users are in good channel conditions have increased chances of being allocated. As a result, the formulation in (6.7) incorporates a trade-off between the current achievable rates of the users in the SDMA groups and their QoS levels, thus providing some degree of weighted throughput fairness.

However, by suitably defining w_{mn} and p_{mn} , different allocation objectives can be pursued. In order to show this, the following allocation objectives are also considered:

- Maximum Capacity (MC): this objective aims at maximizing only the total capacity of the system and disregards QoS aspects. In this case, all the users' priorities v_k are made constant and equal, so that w_{mn} has no influence on the frequency/time assignment anymore, while the definition for the allocation revenues p_{mn} is kept. Considering the MC objective, the SDMA groups with highest capacities are always allocated to the resources.
- Round Robin (RR): this objective is a resource-fair one. In this case, all the users' priorities v_k are updated according to an RR policy, while the allocation revenues p_{mn} are defined to be constant and equal for all the SDMA groups. Thus, SDMA groups are allocated to the resources in an RR-like way.

The above objectives allow getting some useful insight on how capacity and fairness are handled by the proposed RA strategy.

6.3.3 Resource Allocation Strategy

This section describes how GRCBA and frequency/time assignment algorithm are combined in the new S-FT RA strategy proposed here.

Spatial compatibility is frequency-dependent and the best SDMA group on resource *n* might not be optimum for resource $n' \neq n$. Therefore, in the S-FT RA strategy first a candidate SDMA group is created for every user *k* on each resource *n*, thus totalizing M = KN candidate groups. Forcing user *k* to be in a group is done by setting c = k in (6.5) for the RCBA or in step 1 of Table 6.1 for the GRCBA. After that, considering the adopted allocation objective described in the previous section, the product $w_{mn}p_{mn}$ is computed for each group *m* on the resources on which it appeared. Otherwise it is set to 0 to avoid assigning a group to an unsuitable resource. Then, the frequency/time assignment algorithm solves (6.7) with Munkres' algorithm.

The S-FT RA strategy is applied for each allocation period, which corresponds in this chapter to one Time-Slot (TS). One TS contains multiple OFDMA symbols. Priorities w_{mn} are updated on a TS basis. Allocation revenues p_{mn} can be updated at a lower rate, since channel responses only change considerably after a few TSs. This is a reasonable assumption for short frame lengths, low to moderate users' speeds, and channel estimation at a low rate. Moreover, modifying (6.7) in the proposed S-FT RA strategy in order to apply it on a frame basis is straightforward.

In order to reduce the total number of candidate SDMA groups M = KNand consequently reduce computational costs, one can impose that a given SDMA group is not allocated on more than one resource during one allocation period. This is done by assuming that each of the *M* groups appears only once, i.e., that each SDMA group is unique, thus resulting in a number $M' \leq KN$ of candidate groups. Note that in the next allocation period, i.e., in the next TS, a given candidate SDMA group can appear again and be allocated to new resources. In Sect. 6.5 it will be shown that considering *M'* unique groups results in negligible performance losses compared to the case with *M* non-unique groups.

Assuming that Float Point Multiplications (FPMs) demand much more processing cycles than additions and logical operations, the complexity of the S-FT RA strategy can be estimated by its required number of FPMs.

The number $O_{\text{SF-T}}$ of FPMs required by the S-FT RA strategy is approximately of

$$O_{\rm S-FT} = N \left[\frac{K^2 - K}{2n_T} (n_T + 2) + Kn_T \right] + N K \left(\frac{2n_T^3 + n_T^2 + 3n_T}{2} \right), \qquad (6.11)$$

where the first term refers to the computation of the fraction of \mathbf{R}_n , given by (6.4) and which is required by the GRCBA, and the second term refers to the capacity calculations required for the *M* SDMA groups.

Since Munkres' algorithm only involves additions, subtractions, and logical operations, these are neglected in (6.12). However, it must be mentioned that S-FT RA involves a considerable number of such operations. Determining the exact number of required operations for the S-FT RA strategy is not within the scope of this chapter.

6.4 Frequency/Time-Space Resource Allocation

This section introduces the FT-S RA strategy, which is compared with the S-FT RA strategy described in the previous section. In Sect. 6.4.1 the SPA is reviewed. Section 6.4.2 introduces the frequency/time assignment algorithm of users to resources. Section 6.4.3 describes how the algorithms in Sect. 6.4.1 and Sect. 6.4.2 integrate into the FT-S RA strategy.

6.4.1 SDMA Algorithm: Successive Projections Algorithm

In this section, the SPA is shortly reviewed. In the SPA, an initial user vector channel \mathbf{h}_{cn} , indexed by *c*, is assigned on resource *n* and builds a one-user SDMA group $\underline{G}_n = \{c\}$. This initial assignment is performed by the frequency/time assignment algorithm, which is discussed in the next section.

Once the initial users are assigned, all the remaining users on a resource n have their channels projected on the null-space of \mathbf{h}_{cn} . After that, the user with the highest channel gain is added to \underline{G}_n . This procedure is repeated until the group size n_G is reached.

Let \mathbf{I}_L denote an $L \times L$ identity matrix. Then, the null-space projection matrix \mathbf{T}_{kn} of the *k*-th user admitted to \underline{G}_n is given by

$$\mathbf{T}_{kn} = \begin{cases} \mathbf{I}_{n_T}, & \text{for } k = 1 \\ \mathbf{T}_{(k-1)n} - \frac{\mathbf{T}_{(k-1)n}^H \mathbf{h}_{(k-1)n}^H \mathbf{h}_{(k-1)n} \mathbf{T}_{(k-1)n}}{\left\| \mathbf{h}_{(k-1)n} \mathbf{T}_{(k-1)n} \right\|_2^2}, & \text{for } 2 \le k \le n_G \end{cases}$$
(6.12)

Using (6.11), the SPA can be formulated as shown in Table 6.2.

 Table 6.2 Successive Projections Algorithm [10,11]

1. Set the SDMA group $\underline{G}_n = \{c\}$. 2. For g = 2 to n_G Set $\underline{G}_n = \underline{G}_n \bigcup_{c'} \arg \max_{c'} \left\{ \left\| \mathbf{h}_{c'n} \mathbf{T}_{(g-1)n} \right\| \right\}$, with $c' \in \{1, 2, ..., K\} \setminus \underline{G}_n$, and $j \in \underline{G}_n$.

Thus, the SPA algorithm follows the same principle as GRCBA in Sect. 6.3.1, but has a different grouping metric.

6.4.2 Frequency/Time Assignment Algorithm: User-to-Resource Assignment

This section describes the frequency/time assignment algorithm, which assigns to each resource an initial user so that the total weighted capacity is maximized considering this single user allocation. Consider a total number M = K of one-user SDMA groups $\underline{G}_{mn} = \underline{G}_{kn}$, where \underline{G}_{kn} contains only the user k. Considering the definitions in (6.9), (6.10), and (6.6), the initial users to be assigned to the N resources can be determined by solving a standard assignment problem of the same form of (6.7), where the matrices **W** and **U** are adequately defined considering the M = K one-user SDMA groups. Analogously to the S-FT RA strategy, different trade-offs between capacity and QoS can be pursued here by adequately changing the definitions of w_{mn} and p_{mn} . Herein, the same allocation objectives described in Sect. 6.3.2 are considered, i.e., the MC, MWC, and RR objectives.

6.4.3 Resource Allocation Strategy

This section describes how SPA and the frequency/time assignment algorithm of users to resources are combined in the FT-S RA strategy.

First, the frequency/time assignment algorithm is employed to assign one initial user to each resource. After that, the initial user on every resource is known and an SDMA group can be built using SPA. Whenever the MWC objective is considered, the FT-S RA strategy also aims at maximizing the weighted capacity of the system like the S-FT RA strategy. But differently from the S-FT RA strategy, it solves first task 2 and then task 1. Therefore, only the assignment of the initial user involves the MWC objective, which is not considered afterwards by the SPA algorithm in Table 6.2. Changing the grouping metric of SPA in Table 6.2 to include priorities could lead to groups composed of spatially incompatible users, and is therefore avoided here.

The number $O_{\rm FT-S}$ of FPMs required by the FT-S RA strategy is approximately of

$$O_{\rm FT-S} = N \left[\sum_{n=2}^{n_T} (K - n + 1) \left(\frac{5n_T^2 + 5n_T}{2} \right) + Kn_T \right] + N \left(\frac{2n_T^3 + n_T^2 + 3n_T}{2} \right), (6.13)$$

where the first term refers to the successive null-space projections and the second term refers to the capacity calculations.

6.5 Analysis and Simulation Results

The performance of the S-FT RA and FT-S RA strategies are studied through simulations in this section. A BS with a Uniform Linear Array (ULA) of $n_T = 4$ elements separated by half wavelength is assumed. The BS serves K = 16 single-antenna users. There are S = 96 subcarriers grouped in blocks of Q = 12 adjacent subcarriers, thus resulting in N = 8 resources. Equal transmit power and the same average noise power are assumed for all subcarriers. A fixed SDMA group size $n_G = 4$ is assumed. All users have the same rate requirement and it is assumed that the BS always has data to transmit to the users. For both S-FT RA and FT-S RA strategies, precoding matrices are determined applying linear transmit ZF [3,19]. Channel matrices are obtained using the WINNER Phase I Channel Model (WIM) [22]. Large scale fading is assumed to be ideally compensated by power control and only the fast fading is considered. Besides the performance curves achieved by the proposed RA strategies with the MC, MWC,

and RR objectives, the maximum achievable capacity, i.e., the Sato bound [23], is also provided. The most relevant parameters adopted in the simulations are given in Table 6.3.

Parameter	Value		
System bandwidth	937.5 kHz		
Center frequency	5 GHz		
Number of subcarriers	96		
Number of subcarriers per resource	12		
TS and frame durations	0.25 ms and 1 ms		
Update rate of w_{mn} and p_{mn}	1 TS and 1 frame (4 TSs)		
WIM scenario	C2		
Number of single-antenna users	16		
Users' speed	10 km/h		
Transmit ULA	4 omni elements separated by half		
	wavelength		
RA Strategy	S-FT RA	FT-S RA	
SDMA grouping algorithm	RCBA, GRCBA	SPA	
α parameter (see Table 6.1 and (6.5))	0.5	-	
SDMA group size	4		
Frequency/time assignment objectives	MC, MWC, RR		

Table 6.3 Simulation parameters

First, the average system capacity achieved by the S-FT RA strategy and the impact of considering unique groups instead of non-unique groups are investigated.

Figure 6.1 shows the achieved average system capacity in bps/Hz as a function of the average Signal-to-Noise Ratio (SNR) considering the different joint frequency/time allocation objectives, and either RCBA or GRCBA as the SDMA algorithm. In Fig. 6.1, it can be seen for both RCBA and GRCBA that the MC objective provides the highest capacity figures, as expected. It can also be seen that the MWC objective achieves almost the same average capacity achieved with the MC one. Indeed, the proposed MWC objective allows achieving over 90% of the capacity achieved when considering the MC objective with a better degree of fairness, as it will be shown later in this section. Because the RR objective is not capacity-oriented but resource-fair, it achieves the lowest capacity figures.

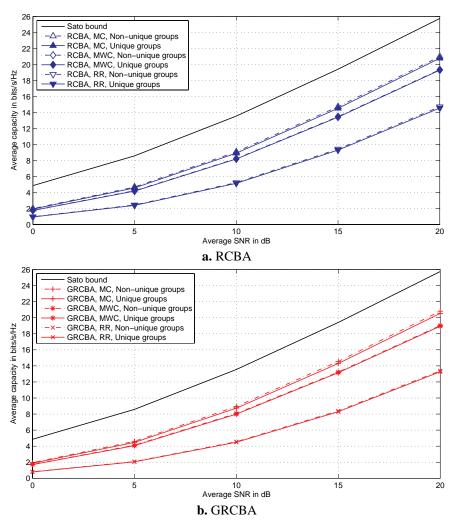


Fig. 6.1 Average capacity in bits/s/Hz of the S-FT RA strategy considering nonunique and unique SDMA groups

In spite of applying equal power allocation across frequencies and linear ZF precoding, which are not optimal for capacity maximization, it can be seen in Fig. 6.1 that the proposed S-FT RA strategy obtains at high SNR about 65% of the maximum achievable capacity drawn by the Sato bound. This performance could still be improved by using adaptive power allocation, e.g., water filling. However, this should not affect the relative performance among the different RA strategies and SDMA algorithms.

In Fig. 6.1, it can also be observed that similar results are achieved considering either non-unique or unique SDMA groups for the RCBA and GRCBA. Therefore, considering only unique groups and avoiding the same group to be assigned to multiple resources does not degrade capacity significantly. On the other hand, this significantly reduces the number of groups considered, thus reducing the complexity of the S-FT RA strategy.

Because the GRCBA is proposed as a less complex SDMA algorithm compared to RCBA, it is interesting to compare their performances in terms of average system capacity.

In Fig. 6.2, it is show for the case in which unique SDMA groups are considered that RCBA and GRCBA achieve similar results for all the considered allocation objectives. The results shown in Fig. 6.2 for RCBA and GRCBA are taken from Fig. 6.1a, b, respectively.

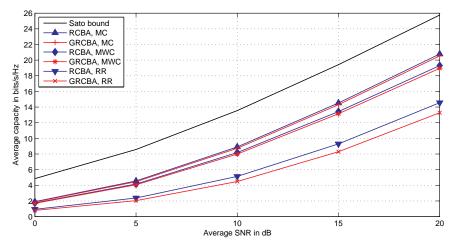


Fig. 6.2 Performance comparison of between RCBA and GRCBA in the S-FT RA strategy considering unique SDMA groups

As it can be observed in Fig. 6.2, the proposed GRCBA effectively approximates the performance of the RCBA, while being less complex than RCBA, since GRCBA is based on the simple greedy algorithm described in Sect. 3.2 and not on an iterative convex optimization procedure.

Based on the results in Fig. 6.1 and Fig. 6.2, it can be seen that the S-FT RA strategy considering GRCBA as SDMA algorithm, unique SDMA groups, and the MWC objective offers a good performance-complexity trade-off compared with the other configurations shown in the same figure.

Now, it is interesting to compare the performance of the S-FT RA strategy considering GRCBA and unique SDMA groups with the performance of the FT-S RA strategy, which uses the SPA as SDMA algorithm. In Fig. 6.3, the average capacity in bits/s/Hz achieved by the S-FT RA and FT-S RA considering the MC, MWC, and RR objectives is shown as function of the average SNR in dB.

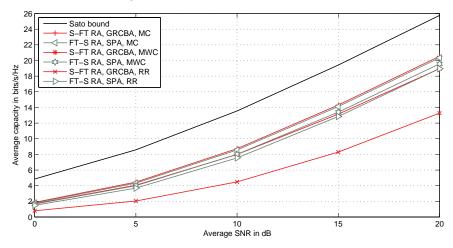


Fig. 6.3 Performance comparison between the S-FT RA and FT-S RA strategies for different allocation objectives

In Fig. 6.3, it can be seen that the S-FT RA and FT-S RA strategies attain almost the same average capacity when the MC and MWC objectives are considered. However, the S-FT RA strategy is considerably less complex than the FT-S RA strategy, as it can be derived from 6.11 and 6.13. With the values in Table 6.3, the FT-S RA strategy requires $O_{\rm FT-S} \approx 1.8 \times 10^4$ FPMs, while the S-FT RA strategy requires only $O_{\rm S-FT} \approx 1.2 \times 10^4$ FPMs.

Considering the RR objective in Fig. 6.3, it can be seen that the FT-S RA strategy has a much larger capacity than the S-FT RA strategy. This is due to the fact that, in the FT-S RA strategy, only the initial user assignment follows an RR objective, while the SPA is greedy and highly capacity-oriented. Thus, the FT-S RA strategy has not the same flexibility as the S-FT RA strategy to control the trade-off between capacity and QoS in the RA. These high capacity values achieved by the FT-S RA strategy come at the expenses of reduced throughput fairness, as shown in the sequel.

Because the proposed RA strategies should provide a good trade-off between performance, complexity and fairness, the Jain's Index of Fairness (JIF) [4]

$$\text{JIF} = \frac{\left(\sum_{k=1}^{K} \overline{R}_{k}\right)^{2}}{K \sum_{k=1}^{K} \overline{R}_{k}^{2}},$$
(6.14)

is computed with the average throughput of the users calculated for a varying number of frames in order to provide a fairness measure for the two RA strategies with the MC, MWC, and RR objectives. Figure 6.4 shows the average JIF for an increasing number of frames and an SNR of 10 dB.

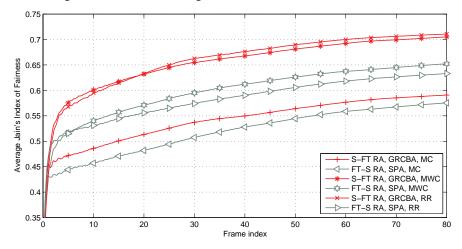


Fig. 6.4 Average jains' index of fairness for the S-FT RA and FT-S RA strategies with different allocation objectives

In Fig. 6.4, it can be seen that the S-FT RA strategy with the MWC allocation objective is much more fair than the MC and as fair as the RR one. This is due to the fact that MWC pursues weighted throughput fairness. It can also be noted that S-FT RA strategy with MWC obtains a JIF value above 0.7 just after a few frames.

From Fig. 6.3 and Fig. 6.4, and in summary, it can be seen that the proposed S-FT RA strategy is relatively efficient, providing considerably high capacity with high degree of throughput fairness among users. It is much more fair than the FT-S RA strategy and attains almost the same capacity than it. Most of the complexity of the proposed S-FT RA strategy resides on the group capacity estimation required to build **W**, which requires the computation of precoding matrices. Further simplifications of the studied S-FT RA strategy and the application of adaptive power allocation techniques are left for future investigation.

6.6 Conclusions

In this chapter, an S-FT RA strategy is proposed, which divides the RA problem into an SDMA grouping task and a joint frequency/time resource assignment task. These are solved by the proposed GRCBA and the frequency/time assignment of groups to resources, respectively. The proposed strategy has been shown to provide high average system capacity, reaching over 65% of the maximum achievable capacity at high SNRs. It considers an MWC allocation objective and is as fair as RR resource allocation. Compared to the case in which the MC allocation objective is considered, the proposed strategy with the MWC objective obtains almost the same capacity (over 90%) while providing a considerably higher degree of throughput fairness among users. Compared to the FT-S RA strategy, which employs the SPA for SDMA grouping, almost the same capacity figures are obtained, while the S-FT RA strategy is less complex than FT-S RA strategy. Thus, the proposed S-FT RA strategy is very flexible and offers a good trade-off between performance, complexity and fairness.

References

- 1. Paulraj A, Nabar R, Gore D (2003) Introduction to space-time wireless communications. 1st edn. Cambridge Univ. Press
- 2. Liu H, Li G (2005) OFDM-based broadband wireless networks: design and optimization. 1st edn. John Wiley & Sons
- Shad F, Todd TD, Kezys V, Litva J (2001) Dynamic Slot Allocation (DSA) in indoor SDMA/TDMA using a smart antenna basestation, IEEE/ACM Trans. Networking 1:69–81, vol 9
- 4. Calvo DB (2004) Fairness analysis of wireless beamforming schedulers. Ph.D. thesis, Technical University of Catalonia, Spain
- 5. Maciel TF, Klein A (2006) A low-complexity SDMA grouping strategy for the downlink of Multi-User MIMO systems. In: Proc. of the IEEE Personal, Indoor and Mob. Radio Commun. (PIMRC)
- 6. Maciel TF, Klein A (2007) A convex quadratic SDMA grouping algorithm based on spatial correlation. In: Proc. of the IEEE Internat. Conf. on Commun. (ICC)
- 7. Leung JYT (ed.) (2004) Handbook of scheduling: algorithms, models and performance analysis. 1st edn. Chapman & Hall/CRC
- Nemhauser G, Wosley L (1999) Integer and combinatorial optimization. Wiley & Sons
- Bourgeois F, Lassalle JC (1971) An extension of the Munkres algorithm for the assignment problem to rectangular matrices, Communications of the ACM 12: 802–804, vol 14

- Yoo T, Goldsmith A (2005) Optimality of zero-forcing beamforming with multiuser diversity. In: Proc. of the IEEE Internat. Conf. on Commun. (ICC), vol 1, pp 542–546
- Tejera P, Utschick W, Bauch G, Nossek JA (2006) Subchannel allocation in multiuser Multiple-Input-Multiple-Output systems. IEEE Trans. Inform. Theory 10:4721–4733, vol 52
- Shad F, Todd TD, Kezys V, Litva J (2001) Dynamic Slot Allocation (DSA) in indoor SDMA/TDMA using a smart antenna basestation, IEEE/ACM Trans. Networking 1:69–81, vol 9
- 13. Toelli A, Juntti M (2005), Scheduling for multiuser MIMO downlink with linear processing. In: Proc. of the IEEE Personal, Indoor and Mob. Radio Commun. (PIMRC)
- Fuchs M, Del Galdo G, Haardt M (2005) A novel tree-based scheduling algorithm for the downlink of multi-user MIMO systems with ZF beamforming. In: Proc. of the IEEE Internat. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP), vol 3, pp 1121–1124
- Spencer Q, Swindlehurst AL (2004) Channel allocation in multi-user MIMO wireless communications systems. In: Proc. of the IEEE Internat. Conf. on Commun. (ICC), vol 5, pp 3035–3039
- Koutsopoulos I, Tassiulas L (2002) Adaptive resource allocation in SDMAbased wireless broadband networks with OFDM signaling. In: Proc. of the IEEE Conf. on Computer Commun. (INFOCOM), vol 3, pp 1376–1385
- Zhang YJ, Letaief KB (2005) An efficient resource-allocation scheme for spatial multiuser access in MIMO/OFDM systems. IEEE Trans. Commun. 1: 107–116, vol 53
- Zhang Y, Totaro A, Costa E (2006) Adaptive resource allocation in chunkbased OFDMA/SDMA systems. In: Proceedings of the Internat. OFDM Workshop
- Joham M, Utschick W, Nossek JA (2005) Linear transmit processing in MIMO communications systems. IEEE Trans. on Signal Processing 8:2700-2712, vol 53
- Doettling M, Sternad M, Klang G, Von Hafen J, Olsson M (2006) Integration of spatial processing in the WINNER B3G air interface design. In: Proc. of the IEEE Vehic. Tech. Conf. (VTC), vol 1, pp 246–250
- Boyd S, Vandenberghe L (2004) Convex optimization, 1st edn. Cambridge Univ. Press 2004
- 22. Salo J et al. (2005) Matlab implementation of the WINNER phase I channel model ver1.5. IST-WINNER Tech. Rep. [Online]. Available: https://www.ist-winner.org/ChannelModelImpl.htm
- Jindal N, Rhee W, Vishwanath S, Jafar SA, Goldsmith A (2005) Sum power iterative water-filling for multi-antenna Gaussian broadcast channels. IEEE Trans. Inform. Theory 4: 1570–1580, vol 51

Part III Techniques and Technologies

7 Moment-Based Estimation of the Signal-to-Noise Ratio for Oversampled Narrowband Signals

W. Gappmair, M. Flohberger, and O. Koudelka

7.1 Introduction

Parameter estimation and synchronization in digital receivers has already reached a mature state [1,2]. But compared to symbol timing or carrier frequency/phase as parameters, not very much information is available about the estimation of the signal-to-noise ratio (SNR), although a lot of modern communication systems require the latter to be known for proper operation; power control for adaptive modulation [3] or iterative soft-decoding procedures [4] are most prominent examples in this respect.

Many SNR estimators for the additive white Gaussian noise (AWGN) channel, excellently reviewed and discussed by Pauluzzi and Beaulieu in [5], need the symbol timing to be recovered in advance. In addition, a lot of algorithms, in particular the data-aided and decision-directed solutions, need the carrier phase to be established as well.

However, to the best of the authors' knowledge nothing is reported in the open literature if both symbol timing and carrier phase are not available or reliable enough, as it might be the case in practice, e. g., in the low SNR regime of an adaptive coding and modulation concept. Motivated by this background, an appropriate algorithm will be derived in the sequel, for reasons of computational simplicity based on the evaluation of even-order moments as it has been successfully applied to algorithms using baud-rate sampling [6,7,8].

7.2 Equivalent Baseband Model

Let the zero-mean independent and identically distributed *M*-ary symbols $c_i = a_i + jb_i$ be normalized to unit variance so that $E[|c_i|^2] = 1$. Furthermore,

let the unit-energy baseband pulse h(t) be a root-raised cosine with symbol period *T* and roll-off factor α , where $0 \le \alpha \le 1$. Therefore, if affected by time shift τ , $|\tau| \le T/2$, and carrier phase offset $\theta \in [-\pi, \pi)$, the received signal can be written as

$$r(t) = \sqrt{S}e^{j\theta} \sum_{i} c_{i}h(t - iT - \tau) + \sqrt{N}w(t).$$
(7.1)

Real and imaginary parts of the complex zero-mean AWGN noise w(t) are assumed to be independent, each with the same variance of 1/2; signal and noise power are expressed by *S* and *N* such that the true SNR is simply defined as $\rho := S/N$.

As basically sketched in Fig. 7.1, r(t) passes the matched receiver filter $h^*(-t)$ providing an output signal

$$x(t) = r(t) \otimes h^{*}(-t) = \sqrt{S} e^{j\theta} s(t) + \sqrt{N} n(t) , \qquad (7.2)$$

where \otimes denotes convolution and n(t) stands for the (non-white) Gaussian noise at the matched filter output. With

$$g(t) = h(t) \otimes h^{*}(-t) = \frac{\sin(\pi t/T)}{\pi t/T} \frac{\cos(\pi \alpha t/T)}{1 - (2\alpha t/T)^{2}}$$
(7.3)

as the raised cosine function [9] used throughout this chapter, the signal part is immediately established by

$$s(t) = \sum_{i} c_i g(t - iT - \tau) .$$
(7.4)

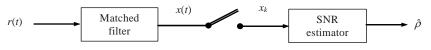


Fig. 7.1 Baseband model for SNR estimation of oversampled narrowband signals

7.3 Moment-Based SNR Estimation

As soon as the symbol timing has been recovered by appropriate means [1], second- and fourth-order moments are frequently applied for SNR estimation, not at least because the related estimator is fairly simple. Rederived in [6] and thoroughly discussed in [5], the principle was already introduced in 1967 by Benedict and Soong [7]. In the sequel, the approach will be extended to narrowband signals x(t) as they are available at the

output of the matched receiver filter, i.e., moment-based estimation of the SNR without knowledge about carrier frequency/phase and timing.

Note that $E[|n(t)|^2] = 1$ due to the unit-energy property of h(t). Therefore, with s(t) and n(t) assumed as independent zero-mean processes, the second-order moment can be directly computed as

$$M_{2} = E[|x(t)|^{2}]$$

= $S E[|s(t)|^{2}] + 2\sqrt{SN} \operatorname{Re}\{e^{j\theta}E[s(t)n^{*}(t)]\} + NE[|n(t)|^{2}]$ (7.5)
= $S E[|s(t)|^{2}] + N.$

Moreover, if the signal s(t) is adopted as cyclostationary, the expected operation in (7.5) appears after some algebra as

$$\Lambda_2 := E[|s(t)|^2] = \frac{1}{T} \int_{-T/2}^{T/2} \sum_i g^2(t - iT - \tau) dt = \frac{1}{T} \int_{-\infty}^{\infty} g^2(t) dt.$$
(7.6)

With Parseval's theorem [10], the relationship is solved as $\Lambda_2 = 1 - \alpha/4$ such that (7.5) converges to

$$M_2 = \Lambda_2 S + N . \tag{7.7}$$

Applying the properties of s(t) and n(t) mentioned before, the fourthorder moment develops after some tedious but straightforward manipulations as¹

$$M_{4} = E[|x(t)|^{4}]$$

$$= S^{2}E[|s(t)|^{4}] + 4\sqrt{S^{3}N} \operatorname{Re}\{e^{j\theta}E[|s(t)|^{2} s(t)n^{*}(t)]\}$$

$$+ 2SN \operatorname{Re}\{E[(e^{j\theta}s(t)n^{*}(t))^{2}]\} + 4SN E[|s(t)|^{2}|n(t)|^{2}]$$

$$+ 4\sqrt{SN^{3}} \operatorname{Re}\{e^{j\theta}E[s(t)|n(t)|^{2} n^{*}(t)]\} + N^{2}E[|n(t)|^{4}]$$

$$= S^{2}E[|s(t)|^{4}] + 4SN E[|s(t)|^{2}|n(t)|^{2}] + N^{2}E[|n(t)|^{4}].$$
(7.8)

Using the definition $\Lambda_4 := E[|s(t)|^4]$, henceforth denoted as correlation index, and $E[|n(t)|^4] = 2$, together with $E[|s(t)|^2] = \Lambda_2$ and $E[|n(t)|^2] = 1$, the fourth-order moment simplifies to

$$M_4 = \Lambda_4 S^2 + 4\Lambda_2 SN + 2N^2.$$
 (7.9)

On the other hand, for Λ_4 we have

¹ A typo has been detected in [11], which is now corrected in (7.8).

$$\Lambda_4 = \frac{1}{T} \int_{-T/2}^{T/2} \sum_{i,k,l,m} E[c_i c_k^* c_l c_m^*] g(t - iT) g(t - kT) g(t - lT) g(t - mT) dt . \quad (7.10)$$

Taking into account that the symbols c_i are independent and identically distributed, the expected operation yields

$$\Lambda_4 = \frac{K_4}{T} \int_{-\infty}^{\infty} g^4(t) dt + \lambda_0 \sum_{i \neq 0} \frac{1}{T} \int_{-\infty}^{\infty} g^2(t) g^2(t - iT) dt, \qquad (7.11)$$

where $K_4 := E[|c_i|^4]$ defines the symbol kurtosis and $\lambda_0 = 2$ for quadraturesymmetric modulation schemes like *M*-QAM or *M*-PSK, $M \ge 4$, whereas $\lambda_0 = 3$ for 2-PSK.

Solving Equations (7.7) and (7.9) with respect to S and N, the true SNR is finally estimated as

$$\hat{\rho} = \frac{S}{N} = \frac{S}{M_2 - \Lambda_2 S} = \frac{\sqrt{2M_2^2 - M_4}}{\xi M_2 - \Lambda_2 \sqrt{2M_2^2 - M_4}}$$
(7.12)

with

$$\xi = \sqrt{2\Lambda_2^2 - \Lambda_4} \ . \tag{7.13}$$

Apart from the immaterial constants Λ_2 and Λ_4 depending only on modulation scheme and pulse shape, equation (7.12) is formally not much different from the case with perfect symbol timing recovery [5,12], where $N = M_2 - S$ and $S = [(2M_2^2 - M_4)/(2 - K_4)]^{1/2}$. Note also that $K_4 = 1$ for *M*-ary PSK and 1 + 2/5 (M - 4)/(M - 1) for square QAM schemes.

7.4 Computation of the Correlation Index

Equations (7.12) and (7.13) require the evaluation of Λ_4 as formulated in (7.11). The problems related to a numerical solution can be circumvented if the relationship is shifted to the frequency domain. To this end, the auto-correlation of squared raised cosines is introduced as

$$R_1(m) := \frac{1}{T} \int_{-\infty}^{\infty} g^2(t) g^2(t - mT) dt .$$
(7.14)

Therefore, with $R_1(m) = R_1(-m)$, the correlation index is equivalently expressed by

$$\Lambda_4 = K_4 R_1(0) + 2\lambda_0 \sum_{m \ge 1} R_1(m) .$$
(7.15)

Applying to (7.14) the properties of the Fourier transform [10], $R_1(m)$ can be calculated most elegantly, because with

$$G(v) = \begin{cases} 1, |v| \le \frac{1}{2}(1-\alpha) \\ \cos^{2}\left[\frac{\pi}{2\alpha}(|v| - \frac{1-\alpha}{2})\right], \frac{1}{2}(1-\alpha) < |v| \le \frac{1}{2}(1+\alpha) \\ 0, |v| > \frac{1}{2}(1+\alpha) \end{cases}$$
(7.16)

as the bandlimited spectrum of raised cosines [9], we note that $G(v) \otimes G(v)$ is the Fourier transform of $g^2(t)$. Using this intermediate result substituted into (7.14), the auto-correlation develops as

$$R_{1}(m) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} [G(v) \otimes G(v)] e^{j2\pi v t} dv \right) g^{2}(t - mT) dt$$
$$= \int_{-\infty}^{\infty} G(v) \otimes G(v) \left(\int_{-\infty}^{\infty} g^{2}(t - mT) e^{j2\pi v t} dt \right) dv$$
(7.17)
$$= \int_{-\infty}^{\infty} G(v) \otimes G(v) \left(\int_{-\infty}^{\infty} g^{2}(t) e^{j2\pi v t} dt \right) e^{j2\pi n v} dv.$$

By taking into account the even symmetry of g(t), we finally arrive at

$$R_1(m) = \int_{-1-\alpha}^{1+\alpha} [G(v) \otimes G(v)]^2 \cos(2\pi mv) dv .$$
 (7.18)

Now, in contrast to (7.11), only the finite integral in (7.18) has to be evaluated. Table 7.1 lists the correlation index Λ_4 for different modulation schemes and roll-offs.

Table 7.1 Correlation index Λ_4 for different modulation schemes and roll-offs

	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1.0$
2-PSK	1.331	1.066	0.859	0.699
M -PSK, $M \ge 4$	1.106	0.920	0.768	0.646
16-QAM	1.316	1.120	0.956	0.819

7.5 Higher-Order Statistics

In the previous sections, an SNR estimator based on second- and fourthorder moments has been derived. However, a higher-order approach [12] might be used for this purpose as well.

With the results developed in [13], we can easily show that even-order moments are computed as

$$M_{2n} = E[|x(t)|^{2n}] = \sum_{m=0}^{n} \frac{(n!)^2}{(n-m)! (m!)^2} \Lambda_{2m} S^m N^{n-m}, \qquad (7.19)$$

where $\Lambda_{2m} := E[|s(t)|^{2m}]$. Therefore, with n = 3, the term M_6 is obtained accordingly. Together with M_2^3 and M_2M_4 , the corresponding sixth-order statistics is after some algebra established by

$$M_{2}^{3} = \Lambda_{2}^{3}S^{3} + 3\Lambda_{2}^{2}S^{2}N + 3\Lambda_{2}SN^{2} + N^{3},$$

$$M_{2}M_{4} = \Lambda_{2}\Lambda_{4}S^{3} + (4\Lambda_{2}^{2} + \Lambda_{4})S^{2}N + 6\Lambda_{2}SN^{2} + 2N^{3},$$

$$M_{6} = \Lambda_{6}S^{3} + 9\Lambda_{4}S^{2}N + 18\Lambda_{2}SN^{2} + 6N^{3}.$$

(7.20)

Normally, two out of the three equations would suffice to solve (7.20) with respect to *S* and *N*. Nevertheless, by introduction of an additional design parameter μ , which might be used in the sequel for optimization purposes, the third equation can be included as well according to

$$T = M_6 - (6 + 2\mu)M_2^3 + \mu M_2 M_4 \Big|_{N = M_2 - \Lambda_2 S} = d_3 S^3 + d_2 S^2 M_2, \qquad (7.21)$$

where

$$d_{2} = (\mu + 9)(\Lambda_{4} - 2\Lambda_{2}^{2}),$$

$$d_{3} = 12\Lambda_{2}^{3} - 9\Lambda_{2}\Lambda_{4} + \Lambda_{6}.$$
(7.22)

Finally, with $z := S/M_2$, equation (7.21) simplifies to

$$d_3 z^3 + d_2 z^2 - \frac{T}{M_2^3} = 0. ag{7.23}$$

Since $z = \rho/(\Lambda_2 \rho + 1)$ and $0 < \rho < \infty$, it is obvious that $z \in (0, 1/\Lambda_2)$ such that (7.23) can be solved straightforwardly by simple numerical techniques like secant method or bisectioning [14]. The related SNR estimate is then immediately calculated by $z/(1 - \Lambda_2 z)$.

What remains is the evaluation of $\Lambda_6 = E[|s(t)|^6]$. Recalling that the symbols c_i are independent and identically distributed, it can be shown after some tedious but straightforward manipulations that

$$\Lambda_{6} = \frac{K_{6}}{T} \int_{-\infty}^{\infty} g^{6}(t) dt + \lambda_{1} K_{4} \sum_{k \neq 0} \frac{1}{T} \int_{-\infty}^{\infty} g^{2}(t) g^{4}(t - kT) dt + \lambda_{2} \sum_{k \neq 0} \sum_{m \neq 0,k} \frac{1}{T} \int_{-\infty}^{\infty} g^{2}(t) g^{2}(t - kT) g^{2}(t - mT) dt$$
(7.24)

with $\lambda_1 = 9$, $\lambda_2 = 6$ for quadrature-symmetric modulation schemes and $\lambda_1 = \lambda_2 = 15$ for 2-PSK. On the other hand, $K_6 := E[|c_i|^6] = 1$ for *M*-ary PSK and $1 + 2/35 (M - 4)(23M - 53)/(M - 1)^2$ for square QAM schemes.

The evaluation of the integrals with infinite limits in (7.24) may be circumvented again by shifting the problem to the frequency domain and solving it by taking into account the properties of the Fourier transform, as already done with (7.11). To this end, we define

$$R_{2}(k,m) \coloneqq \int_{-1-\alpha}^{1+\alpha} [G(v) \otimes G(v)] \operatorname{Re}\{[(G(v) \otimes G(v))e^{j2\pi kv}] \\ \otimes [(G(v) \otimes G(v))e^{j2\pi mv}]\} dv$$
(7.25)

so that (7.24) is finally given by

$$\Lambda_6 = K_6 R_2(0,0) + 2\lambda_1 K_4 \sum_{k \ge 1} R_2(k,k) + 2\lambda_2 \sum_{k \ge 1} \sum_{m \ne 0,k} R_2(k,m) .$$
(7.26)

Table 7.2 lists the correlation index Λ_6 for different modulation schemes and roll-offs.

Table 7.2 Correlation index Λ_6 for different modulation schemes and roll-offs

	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1.0$
2-PSK	2.336	1.475	0.975	0.682
M -PSK, $M \ge 4$	1.512	1.064	0.775	0.587
16-QAM	2.274	1.7180	1.330	1.059

7.6 Simulation Results

In digital receivers, the matched filter output (see Fig. 7.1) has to be sampled appropriately for further processing. This means that $x_k := x(t = kT_s)$ will be used in order to assess the second- and the fourth-order moments in (7.12). Hence, the latter are suitably approximated by

$$M_2 \approx \frac{1}{K_s L} \sum_{k=1}^{K_s L} |x_k|^2, \ M_4 \approx \frac{1}{K_s L} \sum_{k=1}^{K_s L} |x_k|^4,$$
 (7.27)

where *L* denotes the estimator length in symbols and $K_s = T/T_s$ stands for the oversampling factor, which is assumed to be selected such that no alias effects occur.

Figure 7.2 visualizes the evolution of the mean estimator output $E[\hat{\rho}]$ using 4-PSK and 16-QAM schemes ($L = 1000, K_s = 4, \alpha = 0.25$). For comparison purposes, the output of the ideal estimator with $E[\hat{\rho}] = \rho$ is shown as well (dashed line). A bias, as it is typical for SNR estimators in general [5,15,16], occurs as soon as SNR < -1 dB. Note also that the mean values start to diverge from the ideal case if SNR > 15 dB (4-PSK) and SNR > 10 dB (16-QAM), respectively. This is mainly due to the fact that second- and fourth-order moments in (7.12) are only approximated by (7.27), introducing a bias for larger SNRs as well. Interestingly, even if the symbol timing is perfectly recovered, this phenomenon is observed with moment-based SNR estimators applied to modulation schemes with non-constant envelope [17]. In either case, however, this drawback can be mitigated by smoothing the jitter of (7.12) with larger values of L. Via simulation results, it could be verified that different values of K_s and α have solely a negligible influence on the evolution of $E[\hat{\rho}]$.

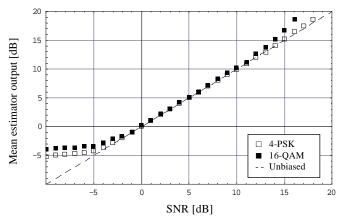


Fig. 7.2 Evolution of $E[\hat{\rho}]$: $L = 1000, K_s = 4, \alpha = 0.25$

Figure 7.3 illustrates the normalized mean square error (MSE) of the M_2M_4 estimate, i.e., $E[(\hat{\rho} - \rho)^2]/\rho^2$, evaluated for 4-PSK and 16-QAM ($L = 1000, K_s = 4, \alpha \in \{0.25, 0.5\}$). Also shown is the Cramer-Rao lower bound (CRLB) as the theoretical limit for unbiased estimation [5,18], assuming both data and carrier phase known to the receiver. Normalized to ρ^2 , it is given by

$$NCRLB = \frac{CRLB}{\rho^2} = \frac{1}{L} \left(\frac{1}{K_s} + \frac{2}{\rho} \right).$$
(7.28)

Due to bias effects, significant deviations from the NCRLB are observed for lower as well as higher SNRs.

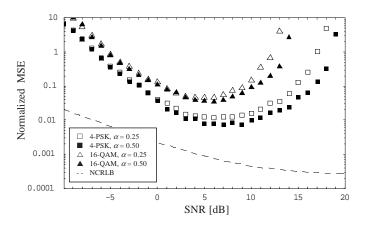


Fig. 7.3 Normalized mean square error: L = 1000, $K_s = 4$

Only in the medium range, the developed estimator provides useful results. The performance is a bit better if the roll-off becomes larger, whereas no improvement is identified when more samples per symbol are employed.

By detailed inspection of (7.12), it is clear that the estimation of the SNR fails when $\hat{\rho}$ is not a positive-real number, which might happen if $2M_2^2 - M_4 < 0$ such that the related root provides an imaginary value. In order to assess the discussed algorithm from this point of view, the success rate R_s , i.e., the quotient of successful and total SNR estimates, is introduced as a measure of efficiency. For 4-PSK and 16-QAM visualized in Fig. 7.4, it is seen immediately that the evolution of R_s is closely related to the results shown in Fig. 7.2, with a drastic loss of efficiency observed for those SNR estimates suffering from a significant bias effect.

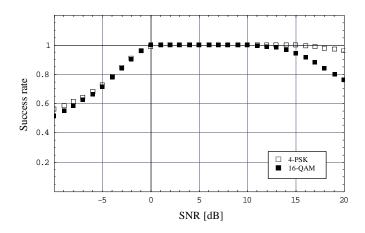


Fig. 7.4 Evolution of the success rate: L = 1000, $K_s = 4$, $\alpha = 0.25$

Finally, Fig. 7.5 illustrates the evolution of the mean square error of the SNR estimator based on the sixth-order statistics developed in the previous section (L = 1000, $K_s = 4$, $\alpha = 0.25$), in the legend indicated by M₂M₄M₆. In this context, the sixth-order moment M_6 is approximated in the same way as done with M_2 and M_4 in (7.27).

Even with μ optimized in order to achieve a minimum of the jitter performance, as it is shown in Fig. 7.5, no improvement over the M₂M₄ estimate is detected. More or less the same result has been obtained using an eighth-order model, apart from the fact that the evaluation of Λ_8 becomes fairly complicated in this case.

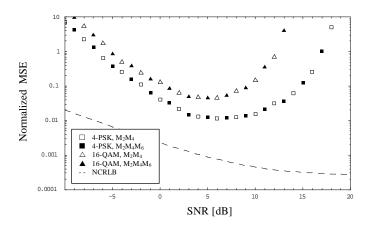


Fig. 7.5 Normalized mean square error: L = 1000, $K_s = 4$, $\alpha = 0.25$

7.7 Conclusions

Based on second- and fourth-order moments, a blind SNR estimator has been developed which does not require any knowledge about carrier and timing. Verified by simulation results, it is shown that the algorithm can be useful in the medium SNR range, whereas for small as well as large SNRs the mean estimator output exhibits a non-negligible bias effect. The impact of the latter can be mitigated by increasing the estimator length such that the SNR range, where the application makes sense, is extended.

The algorithm has been successfully tested in a real-time system [19]. No degradation was observed although the data symbols are no longer independent due to the implemented error correction scheme. Note again that the estimator is developed for narrowband signals such that front-end filter effects like group delay or linear distortions do not affect the performance.

Nevertheless, the comparison with the Cramer-Rao lower bound shows that more powerful solutions are possible. Unfortunately, statistics including moment orders higher than four turned out to be not that successful as initially expected. However, first results using a maximum-likelihood (ML) approach have been quite promising from the performance point of view; unfortunately, the required computational load is tremendous such that the algorithm is probably less useful in practice. Therefore, future research will focus on an appropriately modified ML solutions with reduced complexity.

References

- 1. Mengali U, D'Andrea AN (1997) Synchronization techniques for digital receivers. Plenum Press, New York
- 2. Meyr H, Moeneclaey M, Fechtel SA (1998) Digital communication receivers: synchronization, channel estimation, and signal processing. Wiley, New York
- 3. Chung TS, Goldsmith AJ (2001) Degrees of freedom in adaptive modulation: a unified view. IEEE Trans. Commun. 49: 1561–1571
- Summers TA, Wilson SG (1998) SNR mismatch and online estimation in Turbo decoding. IEEE Trans. Commun. 46: 421–423
- 5. Pauluzzi DR, Beaulieu NC (2000) A comparison of SNR estimation techniques for the AWGN channel. IEEE Trans. Commun. 48: 1681–1691
- Matzner R, Englberger F (1994) An SNR estimation algorithm using fourthorder moments. In: Proc. IEEE Int. Symp. Inform. Theory. Trondheim, Norway, p. 119
- 7. Benedict TR, Soong TT (1967) The joint estimation of signal and noise from the sum envelope. IEEE Trans. Inform. Theory 13: 447–454

- Wiesel A, Goldberg J, Messer H (2002) Non-data-aided signal-to-noise ratio estimation. In: Proc. IEEE Int. Commun. Conf. New York, pp. 197–201
- 9. Proakis JG (1989) Digital communications. McGraw-Hill, New York
- 10. Proakis JG, Manolakis DG (1996) Digital signal processing: principles, algorithms, and applications. Prentice Hall, Upper Saddle River, NJ
- Gappmair W, Flohberger M, Koudelka O (2007) Moment-based estimation of the signal-to-noise ratio for oversampled narrowband signals. In: Proc. 16th IST Mobile and Wireless Commun. Summit. Budapest, Hungary, Paper No. C4.1
- Gao P, Tepedelenlioglu C (2005) SNR estimation for nonconstant modulus constellations. IEEE Trans. Signal Processing 53: 865–870
- Lopez-Valcarce R, Mosquera C (2007) Sixth-order statistics-based non-data-aided SNR estimation. IEEE Commun. Lett. 11: 351–353
- 14. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1994) Numerical recipes in C: the art of scientific computing. Cambridge Univ. Press, New York
- 15. Beaulieu NC, Toms AS, Pauluzzi DR (2000) Comparison of four SNR estimators for QPSK modulations. IEEE Commun. Lett. 4: 43–45
- 16. Li B, Di Fazio R, Zeira A (2002) A low bias algorithm to estimate negative SNRs in an AWGN Channel. IEEE Commun. Lett. 6: 469–471
- Gappmair W, Koudelka O (2006) Moment-based SNR estimation of signals with non-constant envelope. In: Proc. 3rd Advanced Satellite Mobile Syst. Conf. Herrsching, Germany, pp. 301–304
- Alagha NS (2001) Cramer-Rao bounds of SNR estimates for BPSK and QPSK modulated signals. IEEE Commun. Lett. 5: 10–12
- Schmidt M et al. (2005) The software radio based L*IP satellite gateway. In: Proc. 11th Ka and Broadband Commun. Conf. Rome, Italy, pp. 105–112

8 Estimation of Rain Attenuation Distribution on Terrestrial Microwave Links with General N-State Markov Model

Balázs Héder and János Bitó

8.1 Introduction

Broadband radio communication systems operate at high carrier frequencies, i.e. microwave domain. In this frequency range wave propagation is highly influenced by precipitation, especially rain. For accurate planning of high frequency microwave links (e.g. feeder network of mobile operators) the reliable estimation of rain attenuation distribution on the designated links is essential.

In our previous works an N-state Markov Chain model was used to generate rain attenuation time series [5,6,8]. The model is applicable to estimating the first and second order statistics of rain attenuation. The model parameters were derived from fade slope statistics of attenuation measured on a given terrestrial microwave link only.

Rain attenuation or rain rate modeling with time series generation is of great interest in propagation models. Partitioned Fritchman model [4] is applied for modeling the duration of rain fade of terrestrial [2] and land mobile satellite links [3].

Present contribution demonstrates the novel application of our proposed general N-state Markov model to predict rain attenuation statistics of a planned microwave link. To obtain the proper parameters of our model, several measured yearly attenuation time series were considered simultaneously. The measurements were performed on different microwave links with different length, polarization and frequency.

It will be shown that stochastic rain attenuation process can be modeled by first order Markov chains. Thorough description of the model is given and our method for rain attenuation estimation is introduced. The model parameters were derived from high frequency measurements, where precipitation is the dominant attenuation factor. Hence, the presented model and the corresponding time series are only applicable to high frequency links. The parameters of the N-state Markov model can be adapted to other frequency bands and channel models as well, e.g. multi-path and shadowing fading [5].

The paper is organized as follows. In the first section the stationary behavior of rain attenuation process is discussed, then our proposed N-state Markov chain model is presented. Next sections describe the model parameterization, and show how to apply the Model for estimating first and second order statistics of rain attenuation on an arbitrary link. Simulation results and conclusions are given in the final two sections.

8.2 Stationary Examination of Rain Attenuation Process

In order to model rain fading process, information on stationary behavior is very important. In our measurements attenuation is only caused by precipitation, therefore path attenuation is not considered. Two types of precipitation can be distinguished: rain and sleet. In this work sleet events are not examined separately, i.e. rain attenuation events contain sleet events as well. Therefore, if the measured rain attenuation value is around zero there is no rain (precipitation) event, only scintillation and noise are present. Positive values mean valid precipitation attenuation.

The available measured rain (precipitation) attenuation time series are one year long realizations of a stochastic rain attenuation process, which contains several rain fading events. Due to the considerably long measurement period the measured realizations have the same properties as the real stochastic process.

To prove that the rain fading process is not stationary (even in wide sense) amplitude probability density functions of the measured realization in different time instants can be examined. The density functions are calculated from a 4 h period of our available measured attenuation time series.

The 4 h part can be defined in different positions in the whole one-year long data. In other words, density functions are calculated from measured data values located inside a 4-hour-long moving window at a given window position.

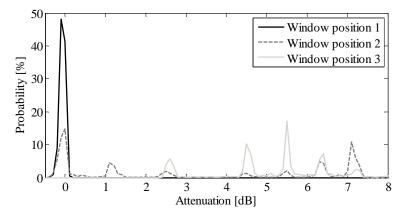


Fig. 8.1 Probability density of measured data in different positions of moving window

Figure 8.1 depicts density functions calculated in different positions of the moving window. The moving window in position 1 does not contain any precipitation event; therefore the mean of the probability density is zero. The moving window in position 2 and 3 contains valid rain (precipitation) attenuation data (the mean of the density function will be higher than zero) and the corresponding two density functions are highly different. These findings can be concluded as: not only the measured attenuation time series (which contains rain and non-rain events), but the rain fading process itself appears to be non-stationary as well, because the probability density function of the process varies in time.

Of course, the probability properties of a heavy rain fading, of a drizzle or of the scintillation during non-fading events are highly different.

The modeling of this non-stationary process with first order Markov model is only possible if information on the event type (non-fading, heavy rain, drizzle etc.) is considered. In our case an N-state Markov model is applied, where each state represents an attenuation level. Therefore different event types are separated on the basis of the degree of attenuation.

8.3 The N-State Markov Model

In the considered time discrete irreducible N-state Markov chain model there are numerous states according to the rain attenuation levels. The model includes discrete states; each state represents an attenuation level with 0.05 dB resolution (ΔA). The resolution can be chosen finer or coarser

as well; the optimal selection needs further investigation. According to our investigations 0.05 dB is appropriate for our goals.

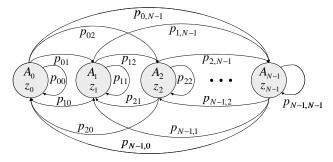


Fig. 8.2 Schematic representation of the N-state Markov Chain model

The schematic representation of the model is depicted in Fig. 8.2, where the number of states is N, the minimum and maximum attenuation levels are A_0 and A_{N-1} respectively. The values of the basic parameters are listed in Table 8.1.

Table 8.1 Basic parameters of the Markov model

A_0	-1.6 dB
$\Delta A = A_i - A_{i-1}$	0.05 dB
A_{N-I}	13.25 dB
N	298

The state probabilities z_i give the probabilities of attenuation levels A_i and can be arranged into the state probability vector \overline{z} , whereas the state transition probabilities p_{ij} can be arranged into the transition probability matrix $\overline{\overline{P}}$ (8.1) and (8.2).

$$z = [z_0, z_1, \dots, z_{N-1}]$$
(8.1)

$$\overline{\overline{P}} = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0,N-1} \\ p_{10} & p_{11} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ p_{N-1,0} & \cdots & p_{N-1,N-2} & p_{N-1,N-1} \end{pmatrix}$$
(8.2)

Using the transition probability matrix of the general N-state Markov chain model, the complement cumulative distribution function (CCDF) of

the generated rain attenuation time series as the steady state probability distribution of the Markov chain can be calculated according to (8.3) [1].

$$P(A \ge A_i) = \sum_{j=i}^{N-1} z_j, \ \bar{z} = \bar{P}^T \cdot \bar{z}$$
(8.3)

To calculate the fade and inter-fade duration statistics of the generated time series, the states of the N-state Markov model must be sorted into two classes: fading and inter-fading states. The derived two-state Markov chain model is depicted in Fig. 8.3, if fade and inter-fade duration are considered at a certain attenuation level A_i .

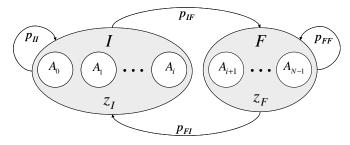


Fig. 8.3 State separation of the model to fading and inter-fading states

The z_F fading state probability, the z_I inter-fading state probability and the p_{IF} , p_{FI} , p_{II} and p_{FF} probabilities are determined from the model parameters using (8.4), (8.5) and (8.6). [7].

$$z_I = \sum_{k=0}^{i} z_k , \ z_F = \sum_{k=i+1}^{N-1} z_k$$
(8.4)

$$p_{IF} = \sum_{k=0}^{i} \sum_{m=i+1}^{N-1} \frac{z_k}{z_I} \cdot p_{km} , \ p_{FI} = \sum_{k=i+1}^{N-1} \sum_{m=0}^{i} \frac{z_k}{z_F} \cdot p_{km}$$
(8.5)

$$P_{II} = 1 - p_{IF}, \ P_{FF} = 1 - p_{FI} \tag{8.6}$$

Fade and inter-fade duration probabilities for different attenuation levels and duration intervals can be calculated by (8.7), where $p_{fd}(A_{ib}t)$ and $p_{id}(A_{ib}t)$ give the probability of fading and inter-fading, respectively, at an attenuation level of A_i with duration of exactly *t* seconds.

$$p_{fd}(A_i,t) = p_{FF}^t \cdot p_{FI}, \ p_{id}(A_i,t) = p_{II}^t \cdot p_{IF}$$
 (8.7)

To determine $p_{fd}(A_{i,t})$ or $p_{id}(A_{i,t})$ for different *t* durations at the investigated A_i attenuation level, the CCDF of fade or inter-fade duration can be calculated. Unfortunately p_{II} in (8.6) is very close to one, while p_{IF} is very close to zero irrespectively of the attenuation level. As a result the interfade duration statistics of the generated time series are almost identical at every attenuation level, therefore the inter-fade duration estimation with this method is not usable.

8.4 Model Parameterization

Our goal is to get a general model which can be applied during microwave link planning independently of the path length, polarization and operational frequency. Therefore by model parameterization several attenuation data series, which were measured on different microwave links, were considered. But the same rain rate causes different rain attenuation on microwave links which have different parameters. To eliminate this effect of variegation all of the measured data series had to be transformed into a hypothetical link operating at 23 GHz carrier frequency with vertical polarization and 1 km length. The transformation was performed using (8.8) and (8.9) based on the recommendation ITU-R P.530 [10]

$$A_{h}(t_{n}) = \frac{k_{h} \cdot L_{h}}{1 + L_{h} / d_{0}} \cdot \left(\frac{A_{m}(t_{n}) \cdot (1 + L_{m} / d_{0})}{k_{m} \cdot L_{m}}\right)^{\frac{\alpha_{h}}{\alpha_{m}}}$$
(8.8)
$$d_{0} = 35 \cdot e^{\left(-0.015 \cdot R_{0.01}\right)}.$$
(8.9)

In (8.8) and (8.9) $A_h(t_n)$ and $A_m(t_n)$ are time discrete attenuations on the hypothetical and measurement links in the nth time instant, k_h , α_h , k_m and α_m are polarization and operating frequency dependent variables described in [12], for the hypothetical and the measurement links, respectively. The lengths of the measurement links and the hypothetical link are L_m and L_h , respectively, while d_0 is the path reduction factor. The geographical location dependent *R* rain intensity is higher than or equal to $R_{0.01}$ in 0.01 percent of the year. In order to transmit data, $R_{0.01}$ must be known from ITU recommended $R_{0.01}$ values were utilized in the different geographical locations where the considered microwave links are set up. The $R_{0.01}$ related to the hypothetical link is determined considering one of the locations of our measuring nodes.

The transformation (8.8) and (8.9) is apparently inevitable; however it must be mentioned that after the transformation some important information about the dynamics of the fading (e.g. fade slope) are unfortunately lost. Nevertheless, sufficient information remains to enable proper determination of the Markov model parameters from the hypothetical link data.

Because of the parameters of the hypothetical link, the model can be applied from 15 GHz up to 38 GHz microwave links with horizontal or vertical linear polarization. The N-state Markov model has a very large transition probability matrix, therefore only the calculation method is presented instead of the exact elements. The probability parameters of the N-state Markov model can be calculated on the basis of fade slope statistics. Fade slope (ς) is a relevant second order statistical parameter for planning purposes e.g. for appropriate fade mitigation techniques, showing the gradient (in dB/s) of the fading at a given A_i attenuation level. The simulation time unit (STU) gives the time interval in seconds between two measured rain attenuation values. Considering the time discrete measured attenuation data and STU, the fade slope can be calculated with (8.10) and (8.11). The unit of fade slope is dB/STU, t_n is the nth time instant. In our case STU equals to 1 s because of the 1 Hz sampling frequency

$$\varsigma^{[dB/STU]} = \frac{A(t_{n+1}) - A(t_{n-1})}{2} |_{A(t_n) = A_i}$$
(8.10)

$$t_n = n \cdot STU \ , \ n \in N \ . \tag{8.11}$$

In our model the conditional probability density function (CPDF) of fade slope at each different attenuation level $(P(\varsigma|A_i))$ is estimated with simple Gaussian distribution functions as in (8.12) [8]. The estimation with Gaussian function is a heuristic idea. It is not stated that fade slope has Gaussian distribution, however this estimation provides satisfactory enough results

$$P(\varsigma|A_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{\varsigma}(A_i)} \cdot e^{-\frac{1}{2} \left(\frac{\varsigma}{\sigma_{\varsigma}(A_i)}\right)^2}$$
(8.12)

$$\sigma_{\zeta}(A_i) = \begin{cases} a \cdot e^{b \cdot A_i} & , A_i < 0 dB \\ c \cdot e^{d \cdot A_i} & , A_i \ge 0 dB \end{cases}$$
(8.13)

In (8.12) and (8.13) A_i is ith attenuation level in dB corresponding to the *ith* state, ς is the fade slope in dB/s. Because of the characteristics of the

fade slope, the expected value of the normal distribution is zero, whereas the attenuation dependent standard deviation parameter $\sigma_s(A_i)$ is approximated with exponential functions (8.13). The values of *a*, *b*, *c* and *d* experimental parameters are listed in Table 8.2.

Table 8.2 The experimental parameters of the Gaussian fade slope model

a	b	с	d
$8.914 \cdot 10^{-3}$	-1.018	$4.983 \cdot 10^{-3}$	0.2874

Determining the Conditional Probability Density Function (CPDF) of fade slope $P(\varsigma|A_i)$ using the Gaussian fade slope model for every A_i attenuation level as condition, the p_{ij} transition probability (from state A_i to state A_j) corresponds to the $P(\varsigma = (A_j - A_i)/2|A_i)$ value. In Fig. 8.4 a typical CPDF of fade slope at a certain A_i attenuation level $(P(\varsigma|A = A_i))$ is presented on the right side. Two states of the N-state Markov model with transition probabilities according to the CPDF of fade slope are also depicted in Fig. 8.4 on the left, where $\varsigma_j = (A_{i+j} - A_i)/2$. As it is also presented in the figure, the $p_{i,i+j}$ probability corresponds to the $P(\varsigma_j = (A_{i+j} - A_i)/2|A_i)$ value. If the CPDF of fade slope is a continuous function, we get the exact value of the transition probability with an integral around the proper fade slope value. In our discrete case a sum is used instead of the integral.

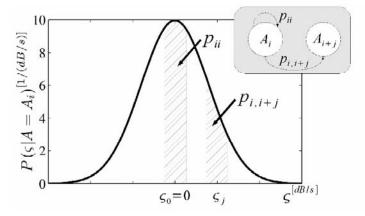


Fig. 8.4 Determination of state transition probabilities from the CPDF of fade slope (on the left) with two signed states of the Markov model

8.5 Applying the Proposed Model for a Designated Link

If the parameters of the planned link differ from those of our hypothetical link, the model parameters must be transformed in order to get time series for the appropriate link. The applied method is the same as given by (8.8). The suitable transformation $T\{.\}$ uses (8.14).

$$T\{A_h(t_n)\} = A_p(t_n) = \frac{k_p \cdot L_p}{1 + L_p / d_0} \cdot \left(\frac{A_h(t_n) \cdot (1 + L_h / d_0)}{k_h \cdot L_h}\right)^{\frac{\alpha_p}{\alpha_h}} \quad (8.14)$$

Similarly to (8.8), $A_p(t_n)$ in (8.14) and $A_h(t_n)$ are the time discrete rain attenuation values on the planned and hypothetical links, L_p and L_h are the path lengths, k_p , α_p , k_h and α_h are polarization and operating frequency dependent variables for the planned and the hypothetical links, respectively. In Fig. 8.5 the principle of parameter transformation is depicted.

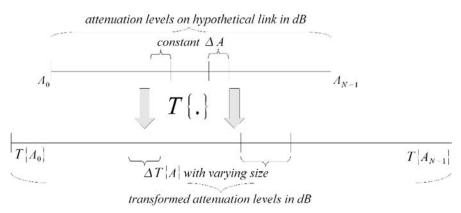


Fig. 8.5 Principle of parameter transformation

The attenuation level interval from the minimum (A_0) up to the maximum $(A_{N.1})$ attenuation level must be transformed to the planned link with (8.14). The transformed attenuation levels are signed by $T\{A_i\}$.

The attenuation quantization step (ΔA) is a constant 0.05 dB in case of the hypothetical link, but after the non-linear transformation it will be varying in case of the planned link. The varying quantization step values are important input parameters for the time series generator. They indicate which attenuation levels belong to the individual Markov model states.

Considering the transformed attenuation levels, the rain attenuation statistics of the proposed link can be determined with expressions (8.3), (8.4), (8.5), (8.6) and (8.7). Assigning the transformed attenuation levels to the Markov model states time series realizations can be generated as well.

If we can estimate the maximal and minimal rain attenuation on a proposed link, a better prediction can be achieved. Considering these extreme attenuation values, only the appropriate part of the transition matrix must be considered, which can be calculated using (8.15) and (8.16).

$$\overline{\overline{P}}' = \left\{ P_{ij}; m < i, j < M \right\}$$
(8.15)

$$m = \left\lceil \frac{A_m - T\{A_0\}}{T\{\Delta A\}} \right\rceil, \ M = \left\lceil \frac{A_M - T\{A_0\}}{T\{\Delta A\}} \right\rceil$$
(8.16)

In the expressions *m* is the minimal and *M* is the maximal state number corresponding to the smallest and highest estimated rain attenuation level occurring on the proposed link. If the maximal and minimal values of the prospective rain attenuation cannot be estimated, $T\{A_0\}$ and $T\{A_{N-1}\}$ values must be used as A_m and A_M respectively. Now, from $\overline{\overline{P}}$ the rain attenuation

and the fade duration CCDF of the planned link can be calculated with (8.3), (8.4), (8.5), (8.6) and (8.7).

8.6 Results

In this section an example is presented which explains how the above described model and method can be used to estimate first and second order statistics of a microwave link in the planning phase. The $R_{0.01}$ value is necessary to know related to the location where the new link will be deployed. In this example the parameters of the proposed link are listed in Table 8.3.

Table 8.3 Parameters of the proposed link

Location	<i>R</i> _{0.01} [mm/h]	Frequency [GHz]	Polarization	Length [km]
Szeged	35.97	15	V	15.17

It must be mentioned, that this is a realistic link that belongs to our countrywide measurement system, therefore the measurement can be used for checking the results. First of all, the model parameters must be transformed using (8.14) according to the link parameters. The maximum rain attenuation was estimated as 31 dB, whereas the transformed A_0 was considered as the minimal attenuation.

In Fig. 8.6 the resulted expected attenuation CCDF with the consideration of the maximal expected attenuation value and the CCDF of the measured one-year-long rain attenuation time series realization are depicted. Please observe, that the CCDF prediction accuracy is very good and we only used the given link parameters for the transformation of the Markov model parameters.

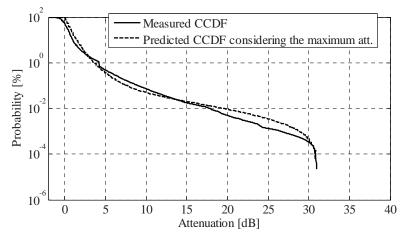


Fig. 8.6 Rain attenuation CCDF prediction with the N-state Markov model on the proposed link in Szeged and the CCDF of the attenuation on the implemented link in a one year period

As a consequence, if the maximum expected attenuation value is considered, the CCDF prediction works very well, otherwise, if the maximal attenuation level cannot be estimated, the CCDF prediction still remains excellent for attenuation values lower than approximately 15 dB.

The fade duration CCDF of the Markov model can be determined from the transition matrix of our Markov model for a given attenuation level as it is described in Sect. 8.2. Afterwards the expected fade duration CCDF related to the proposed link can be determined with transforming the attenuation threshold using (8.14) with the consideration of the link parameters. In Fig. 8.7 the resulting expected fade duration CCDFs and the fade duration CCDFs of the one-year-long measured attenuation values are depicted for different attenuation levels.

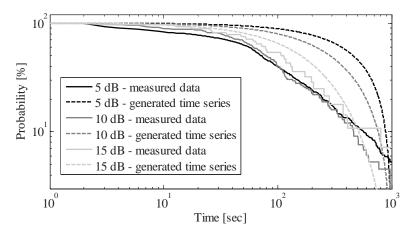


Fig. 8.7 Fade duration statistics prediction with the N-state Markov model on the proposed link in Szeged and the second order statistics of data measured on the implemented link in a one-year period

Please observe that the fade duration CCDF prediction is considerably good at 15 dB attenuation level, especially if we consider that only the proposed link parameters and our general N-state Markov model were used to predict the fade duration CCDF. For lower attenuation levels the prediction of the fade slope CCDF is not so accurate.

8.7 Conclusion

In this paper we focused on presenting how to apply our general N-state Markov Chain model to estimate rain attenuation statistics of a microwave link in the planning phase. The knowledge about probable rain attenuation is highly important when planning microwave connections (e.g. the feeder network of a cellular mobile communication system), because – among other features – it can provide the setting of fading margin correctly. Very accurate approximation of the physical fading process can be achieved with his model.

An N-state Markov model is proposed with detailed description to predict the first and second order statistics of a proposed microwave link in the early planning phase. A comparison was given to demonstrate the accuracy of our generated time series and measured data: both the rain attenuation and fade duration CCDF values were compared. First and the second order statistics of the generated time series were determined accurately from the Markov model parameters. We found, that the attenuation CCDF estimation is excellent and the fade duration CCDF estimation is fairly good at medium attenuation levels. The inter-fade duration CCDF estimation is unfortunately not applicable because of the characteristics of the Markov model.

The proposed model is planned to contribute to the relevant ITU-R recommendation as a potential candidate for attenuation time series generator.

Future work may include developing further methods to optimize and fit the parameters of Markov model, by minimizing the r.m.s error between the measured and the estimated CCDFs [9], which would lead to another parameter set. The selection of optimal Markov model resolution is highly recommended to get simpler model and to mitigate the amount of modeled noise.

Acknowledgement

This work was carried out in the framework of IST FP6 SatNex NoE project and was supported by the Mobile Innovation Center Hungary.

References

- Bitó J (1996) Digitale Mobilfunk-Kanalmodelle unter besonderer Berücksichtigung von adaptiven digitalen modellen (in German). Ph.D. thesis, TU Berlin, Berlin, Germany
- Csurgai-Horváth L, Bitó J (2006) Fade Duration Modeling of Terrestrial Links Applying Markov Chain. In: Loughborough Antennas and Propagation Conference – LAPC, Loughborough, UK, pp. 457–460
- Csurgai-Horváth L, Bitó J (2006) Land Mobile Propagation Fade Duration Modeling by Markov Chains. In: International Workshop on Satellite and Space Communications (IWSSC), Madrid, Spain, pp. 242–246
- 4. Fritchman BD (1967) A binary channel characterization using partitioned Markov chains. IEEE Trans. Inform. Theory 13:221–227
- Héder B, Bitó J (2006) Markov Chain Modelling of Attenuation Time Series of Land Mobile Satellite Channel. In: ASMS 2006 Conference, Herrsching am Ammersee, Germany, CD Proceeding
- Héder B, Bitó J (2005) Rain Attenuation Time Series Generation Applying N-State Markov Model Parameterised from Hungarian Measurement. In: ESTEC 2005 Conference, Noordwijk, The Netherlands, CD Proceeding
- Héder B, Bitó J (2006) Second Order Statistics of Rain Attenuation Time Series Generated With N-State Markov Chain Model. In: EuCAP 2006 Conference, Nice, France, CD Proceeding

- Héder B, Horváth P, Bitó J (2006) Attenuation Time Series Generation at 38 GHz with Time and State Discrete Markov Model. In: IST Mobile Summit 2006, Myconos, Greece, CD Proceeding
- 9. ITU-R P.311-11 Recommendation (2003) Acquisition, presentation and analysis of data in studies of tropospheric propagation. ITU, Geneva, Switzerland
- 10. ITU-R P.530-11 Recommendation (2005) Propagation data and prediction methods required for the design of terrestrial line-of-sight systems. ITU, Geneva, Switzerland
- 11. ITU-R P.837-4 Recommendation (2003) Characteristics of precipitation for propagation modeling. ITU, Geneva, Switzerland
- 12. ITU-R P.838-2 Recommendation (2003) Specific attenuation model for rain for use in prediction methods. ITU, Geneva, Switzerland

9 An Investigation of the Applicability of Fade Duration Markov Model in Attenuation Time Series Synthesis for Multipath Fading Channel

László Csurgai-Horváth and János Bitó

9.1 Introduction

The high frequency millimeter band radio links in satellite or terrestrial communication are particularly affected by the attenuation on the link path. Different type of precipitation, multipath propagation, interference or scintillation may cause short and long term attenuation which should be taken into account during the design of this kind of radio connections [1]. Attenuation or received power time series needs to be measured for long periods – several months or years – to be able to calculate relevant first or second order statistics to figure out the system outage time, unavailability periods, etc. Synthetic attenuation time series are widely used to substitute the long term measurements and generate data to calculate the above mentioned link parameters and statistics [2].

In this contribution a novel method of attenuation time series synthesis for land mobile satellite (LMS) radio channels will be introduced for those cases in which the fading process cannot be considered as stationary.

The proposed time series synthesizer is based on stochastic methods applying Markov chains parameterized from a measurement of a LMS channel [3,4]. The examination of the measured data in point of stationarity [5] shows that the attenuation process cannot be modeled with a single first order, homogenous Markov chain. Therefore the process of the time series synthesis will be distributed in two parts: a two-state Markov chain will be applied to model the fade/non-fade events [6], afterwards the fine structure of fades will be generated with a Markov model of fade duration [7]. The fade duration model is based on a partitioned Fritchman's Markov chain. From this model the required parameters for the two-state fade/non-fade model can be derived. Finally, the fast fluctuation of the attenuation [8] will be modeled with a hidden Markov model (HMM) giving to the time series synthesizer the capability of realistic channel propagation modeling [9].

9.2 Description of the Measured Data

The parameters of the time series synthesizer are based on the measurement data of a land mobile satellite channel received on board of a moving vehicle with a speed of 60 km/h on highway. To investigate the stationarity of the LMS fading process a second set of measurement has been also processed. This is the same LMS link measured in city environment with the speed of 10 km/h. The measurements are performed by DLR during 1984-87 [10].

The measured analog time function of the received power has been sampled with 300.5 Hz frequency and digitalized by an A/D converter. A normalization of the data was carried out to set the second moment of the fading amplitude process (power) equal to 1 (0 dB). The parameters of the measurement are detailed in Table 9.1. The fading events which can be observed in the measured time series are mostly caused by the multipath propagation due to buildings, vegetation and other shadowing obstacles.

Satellite	MARECS (d=39150 km, geostationary)		
Elevation	24°		
Frequency	1.54 GHz		
Sampling rate	300.5 Hz		
Vehicle speed	60 km/h (highway)	10 km/h (city)	
Measurement duration	81.2 min	27.8 min	

Table 9.1 Description of the measured satellite link

9.3 Stationarity Investigations of the Attenuation Process

The object of this contribution is to develop a stochastic method to synthesize attenuation time series by applying homogeneous, first order Markov models. This kind of Markov chains are widely used tools to model stationary stochastic processes [11]. We will show that the non-stationarity of the fading process necessitates an application of combined Markov chains to generate the time series. The attenuation process with fading on a LMS radio link is a stochastic process; an investigation of its stationarity is needed before selecting the right Markov modeling method for time series generation.

In [5] a structure function has been introduced to classify a fading process in the aspect of stationarity. The structure function of a stochastic process r(t) is defined by Clarke with (9.1), as the function of t and the $\tau \log$, where $E\{\}$ means the expected value:

$$D_{r}(\tau) = E\{[r(t+\tau) - r(t)]^{2}\}.$$
(9.1)

Taking into account that the building of the expected value is a linear operation, the square of (9.1) can be calculated as in the following equation:

$$D_{r}(\tau) = E\{[r(t)]^{2}\} + E\{[r(t+\tau)]^{2}\} - 2E\{r(t+\tau) \cdot r(t)\}.$$
 (9.2)

If r(t) process is a real-valued stationary process, the autocovariance function can be written with (9.3) as the following:

$$R_r(\tau) = E\{r(t+\tau) \cdot r(t)\}.$$
(9.3)

If r(t) is a stationary stochastic process, than the function R_r depends only on the lag τ . Therefore by applying (9.3) and due to the stationarity $E\{[r(t)]^2\} = E\{[r(t+\tau)]^2\}$, the structure function can be expressed for stationary processes as it follows,

$$D_r(\tau) = 2 \cdot [R_r(0) - R_r(\tau)]\}, \qquad (9.4)$$

where $R_r(0)$ is the autocovariance function if $\tau = 0$.

This result means that if a process is stationary and $\tau \rightarrow inf$, the structure function is converging to the constant value of $2R_r(0)$. Therefore the structure function can be applied to test the stationarity of a stochastic process. In this case the normalized structure function $D_r(\tau)/R_r(0)$ approaches the horizontal asymptote at 2, and it shows a significant deflection from this constant value when the process is not stationary.

To prove the applicability of the structure function in the stationarity analysis of time series, we calculated the normalized structure function for the two different measurement environments (city and highway).

In Fig. 9.1. the normalized structure function of the fading amplitude of the two LMS links is depicted.

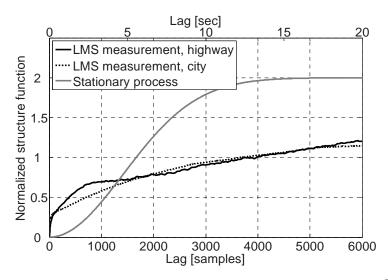


Fig. 9.1 Structure function of the LMS fading amplitude normalized with $R_r(0)$ (highway and city environment), and the structure function $D_g(\tau)$ of a stationary process

The maximal number of lag was during the calculation 6000 samples, which means approximately 20 s taking into account the 300.5 Hz sampling rate of the signal. The figure also depicts the structure function for a stationary random variable with Gaussian autocovariance function, $\exp(-\tau^2/\tau_0^2)$ where τ_0 is a constant [5]. This function is given with (9.5):

$$D_g(\tau) = 2 \cdot [1 - \exp(-\frac{\tau^2}{\tau_0^2})].$$
(9.5)

The $D_g(\tau)$ structure function was calculated with the $\tau_0 = 2000$ constant. The function approaches the horizontal asymptote for large τ due to the stationarity. Contrarily, the structure function both of the two different LMS fading amplitude time series are showing a rising tendency even at higher values of τ .

In the case of the measurement performed on the board of a moving vehicle with the speed of 60 km/h in highway environment, the typical obstacles (bridges, buildings, isolated trees) may cause fade events in the duration of 1–5 s. Therefore the fading process on the LMS channel cannot be considered as a stationary process, because the structure function shows a rising tendency even at $\tau = 20$ s.

In the case of the city measurement the nature of the structure function is similar to the case of highway. It can be observed that the rising speed of the structure function is slower in the city measurement. This can be construed with the higher number of shadowing obstacles in the city which may cause fade events more frequently.

These results necessitate the application of combined Markov modeling methods instead of a single Markov chain to overcome this problem and allow modeling the short as well as the long-term fading process with appropriate precision.

9.4 Event Modeling with Two-State Markov Chain

To model the long-term behavior of the channel a time and state discrete first order Markov chain can be applied [6]. The two states of this Markov chain are representing the fading and non-fading events (Fig. 9.2.). The duration of the fade events and the inter-fade periods is determined by the q_{ij} transition probabilities of the Markov chain:

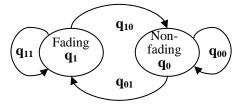


Fig. 9.2 Two-state Markov chain to model fade/non-fade events

The state transition matrix \overline{S} of the Markov chain is:

$$\stackrel{=}{S} = \begin{pmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{pmatrix} = \begin{pmatrix} 1 - q_{01} & q_{01} \\ q_{10} & 1 - q_{10} \end{pmatrix}.$$
(9.6)

Consequently, due to the stochastic feature of the transition matrix two independent parameters i.e. the q_{01} probability of non-fading \rightarrow fading transition and the q_{10} probability of fading \rightarrow non-fading transition determine the process.

The state probability vector of the model at time *n* consists of the fading probability of $q_1(n)$ and the non-fading probability of $q_0(n)$:

$$s(n) = [q_0(n) \quad q_1(n)] = [1 - q_1(n) \quad q_1(n)].$$
(9.7)

As the state of a first order Markov process at time n+1 depends only on the state at n we can write:

$$s(n+1) = s(n) \cdot \overline{\overline{S}} . \tag{9.8}$$

At steady state i.e. s(n)=s(n+i) for $\forall i$ (9.8) becomes:

$$(1-q_1,q_1) = (1-q_1,q_1) \cdot \begin{pmatrix} 1-q_{01} & q_{01} \\ q_{10} & 1-q_{10} \end{pmatrix}.$$
(9.9)

By solving (9.9) q_{01} can be expressed:

$$q_{01} = \frac{q_1}{1 - q_1} \cdot q_{10} \,. \tag{9.10}$$

The consequence is that according to (9.6) and (9.10) the two-state Markov chain is determined by to parameters, i.e. q_1 and q_{10} .

The q_1 probability of fading and the q_{10} transition probability from fading state to non-fading state will be obtained from the fade duration model as described in the next section.

9.5 The Attenuation Threshold Dependent Fade Duration Model

According to the definition, fade duration is the time interval over which the received signal level remains below a certain attenuation threshold, taking into account that level crossings may occur both upward and downward directions. [1] In this section we introduce a Markov model of fade duration with attenuation threshold dependent model parameters. From this model the complementary cumulative distribution function (CCDF) of fade duration can be calculated for any desired threshold. Therefore it allows generating a synthetic event duration process at each threshold by Monte Carlo simulation according to the given CCDF.

The fade duration process will be modeled with an N-state partitioned Fritchman's Markov chain according to [3]. The original model has been developed to characterize the error-free run and the error-cluster distribution of a binary channel. The model is also applicable in the modeling of the fading statistics of the mobile radio channels [12]. In this contribution we apply this kind of Markov chain to model the fade duration distribution of the LMS channel. The state diagram of the applied 5 state Markov chain is depicted in Fig. 9.3. The number of states in this model is determined by the measured fade duration distribution as it will be explained later.

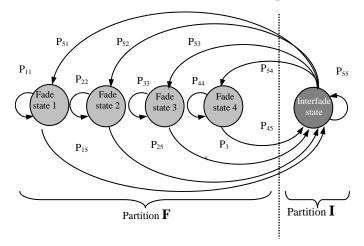


Fig. 9.3 5-state Markov-chain applied for fade duration modeling

The partition F contains those states which are representing different lengths of fades. The interfade events are represented in partition I with only one state. Because the CCDF of the partitions can be analytically expressed with the transition probabilities, this Markov chain is applicable to model the fade duration [13].

The transition matrix of this Fritchman's Markov chain is a 5*5 stochastic matrix. In the Fritchman's model there are no transitions between the states in the same partition thus the matrix contains non-zero elements only in the diagonal, in row 5 and due to the stochastic feature of the matrix in column 5, because the sum of the elements in a row is always 1:

$$= \begin{pmatrix} p_{11} & 0 & 0 & 0 & 1 - p_{11} \\ 0 & p_{22} & 0 & 0 & 1 - p_{22} \\ 0 & 0 & p_{33} & 0 & 1 - p_{33} \\ 0 & 0 & 0 & p_{44} & 1 - p_{44} \\ p_{51} & p_{52} & p_{53} & p_{54} & 1 - \sum_{i=1}^{4} p_{5i} \end{pmatrix}.$$
 (9.11)

As the CCDF of the partition F can be expressed for any discrete number of n with (9.12) [3], from this equation the fade duration CCDF can be calculated:

$$F_F^C(n) = \sum_{i=1}^{N-1} \frac{p_{Ni}}{p_{ii}} p_{ii}^n .$$
(9.12)

where N=5 is the number of states, p_{ij} is the transition probabilities. The physical meaning of *n* in this equation is the duration. A single transition of the Markov chain represents the time between two samples of the fading amplitude process.

The state transition probabilities can be determined with the gradient method [14] applied on the fade duration statistics (CCDF) of the original measurement data. The number of states in the Fritchman's model is determined by the nature of the fading process to be modeled. In the gradient method we approximate the logarithmic CCDF of the fade duration with lines. The number of lines which is required for the correct approximation is determining the state number in the Markov chain, the gradient and *y* axis cross points of the lines are giving the transition probabilities.

The steady state probabilities can be expressed with (9.13):

$$Z_{N} = \frac{1}{1 + \sum_{i=1}^{N-1} \frac{p_{Ni}}{p_{iN}}}, \text{ and } Z_{i} = \frac{p_{Ni}}{p_{iN}} Z_{N} \qquad 1 \le i < N.$$
(9.13)

The fading and inter-fading partition probabilities are:

$$Z_F = \sum_{i=1}^{N-1} Z_i$$
, and $Z_I = 1 - Z_F$. (9.14)

As an example, the parameterization of the fade duration Markov model has been performed at 2 dB threshold. The measured and modeled CCDF curves are graphed in Fig. 9.4. The value of the root mean square error (RMSE) is equal to 0.0042, which means a good approximation.

Our proposed attenuation time series synthesizer requires the CCDFs of the fade duration for attenuation threshold levels between the minimal and maximal fade depth. Therefore we have to determine the fade duration CCDF with appropriate fine resolution of the levels. In [13] is presented that there is a relationship between the transition matrix elements of the fade duration Markov models at different attenuation thresholds.

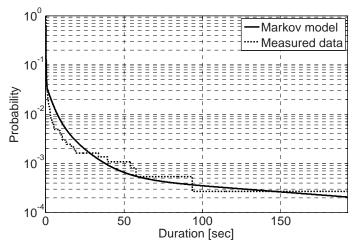


Fig. 9.4 Markov modeled and measured CCDF of the fade duration at 2 dB. RMSE = 0.0042

The expressions in (9.15) and (9.16)

$$p_{ii}(A) = a_{p_{ii}}A^3 + b_{p_{ii}}$$
(9.15)

$$p_{5i}(A) = a_{p_{5i}}A^3 + b_{p_{5i}}$$
(9.16)

can be used to approximate the attenuation threshold dependence of p_{ii} and p_{5i} where A denotes the attenuation threshold, a_{pij} and b_{pij} are constant parameters.

This method allows the calculation of the fade duration CCDF for any desired threshold. It gives us not only the opportunity to determine the two-state fade/non-fade Markov chain parameters, but allows to simulate the fine structure of a fading event as we will discuss it in the next sections.

9.6 From the Fade Duration Model to the Two-State Fade/Non-fade Model

The CCDF of the fade duration at a certain threshold gives the probability that the duration of fades is longer than the specified value. In the following we define a fade event if the attenuation is higher than 2 dB, because this threshold is above the scintillation level. The minimal duration of the fade event has been chosen as 1 s, taking into account the average and minimal fade duration calculations on the investigated link. However, these values can be slightly modified to achieve better modeling results. An investigation of the threshold dependence can be found in the literature [6].

The probability that the fading depth is higher than a given threshold is identical with the CCDF of the fade duration at the same threshold. According to this fact the q_1 parameter (fading probability) of the two-state fade/non-fade model (See Fig. 9.2.) can be set up from the fade duration CCDF.

Figure 9.5. shows the Fritchman modeled fade duration CCDF curves between 2–10 dB thresholds for the highway environment.

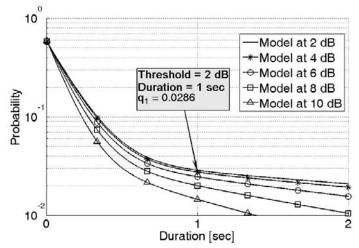


Fig. 9.5 Fade duration CCDF at 2–10 dB thresholds, calculated with Markov model

If we apply the above definition for a fade event, in the two-state fade/non-fade model the value of $q_1 = 0.0286$ as explained in Fig. 9.5.

The q_{10} parameter of the two-state fade/non-fade model (probability of fade \rightarrow non-fade transition) is given by (9.17) and it can be calculated from the transition matrix of the Fritchman's model at the selected 2 dB threshold:

$$q_{10} = \sum_{i=1}^{N-1} p_{iN} \cdot \frac{Z_i}{Z_F}.$$
(9.17)

as the sum of transition probabilities weighted with the Z_i steady state probabilities and normalized with Z_F fading partition steady state probability (see (9.13) and (9.14)).

From the N=5 state fade duration model at 2 dB, according to (9.17) the value of $q_{10} = 0.0018$. Together with q_1 the parameters of the two-state fade/non-fade Markov-chain are fully computable applying (9.10). The state sequence of the two-state Markov chain models the fade/non-fade event series with one second resolution for any desired duration.

A typical simulation example for 80 min duration is shown in Fig. 9.6:

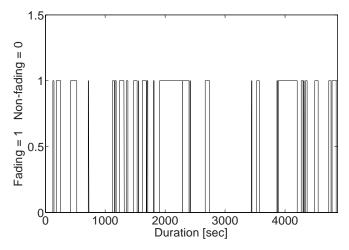


Fig. 9.6 Simulated fading/non-fading event series

In this section the two-state fade/non-fade Markov model and its parameterization process from the fade duration model has been presented. This method allows the modeling of the long term behavior of the propagation channel.

9.7 Simulate a Single Fading Event

After the generation of long term fade/non-fade event series the next step is to achieve a synthetic attenuation time series with fine (one sec) resolution by filling each single fading state with independently simulated attenuation time functions.

The attenuation threshold dependent Fritchman's fade duration model is applicable to generate a single fading event with the required duration and attenuation limit. From the model the CCDFs of fade duration at any fade level can be determined with Eq. (9.12).

Before starting with the synthesis of the single fading events a database of the CCDFs should be set up with appropriate fine attenuation threshold resolution. The usefulness of the partitioned Fritchman's Markov model is that it gives analytical expressions to calculate the fade duration CCDF's, which simplifies and speeds up the time series synthesis compared with the method if we build them directly from the measured data.

The two state fade/non-fade Markov model determines the D_t duration of the individual fade events. During a real fade event the attenuation increases from the fading-free level up to the maximal attenuation level and then it decreases until the fading-free level is reached. This is not a monotonic process, during the fade event there are local changes in the direction of the attenuation level change.

Let us define the elementary fading event as a microscopic part of the fading process, where the attenuation is monotonically increasing and then decreasing in the range of a randomly determined A_{max} fade depth. The attenuation time function generation process synthesizes elementary fading events which will be linked together until the desired D_t is achieved.

The flowchart of the elementary fade event generation process is depicted in Fig. 9.7. The process is a series of inverse transformations applied on the fade duration CCDF at each A_i threshold. The method results a set of fade durations between A_{min} and A_{max} with the appropriate probability distribution.

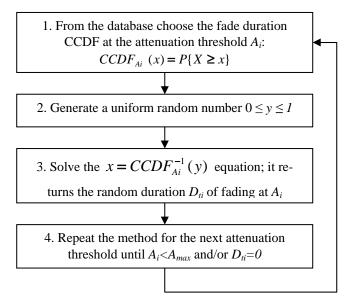


Fig. 9.7 Elementary fading event generation process

In this manner random duration of fading can be generated at the attenuation levels between $A_{min} = 2$ dB to A_{max} . The duration of the fading is decreasing while the attenuation level is increasing according to the

physical phenomena of the fading process. Therefore, by sub-ordering and decentralizing the randomly distributed fade duration length at each level a single elementary fade event can be generated as it is depicted in Fig. 9.8.

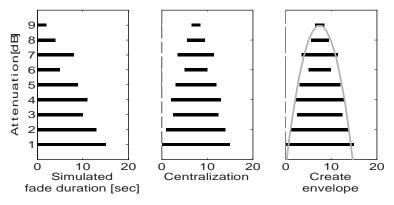


Fig. 9.8 Generating a single elementary fading event

A typical attenuation time function for the duration of 60 s and $A_{max} = 30$ *dB* is depicted in Fig. 9.9. It was constructed from elementary fade events according to the method described above. Please notice that the generated maximum of attenuation can be less then A_{max} because the probability of fade duration longer than a time quantization can be zero at higher attenuation levels.

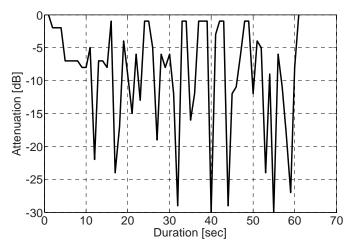


Fig. 9.9 Synthesized fading event for 60 s

9.8 Modeling the Scintillation

Scintillation is the rapid fluctuation of the different received signal characteristics. To improve our time series synthesizer the scintillation of the received signal amplitude will be modeled with a hidden Markov model (HMM).

Scintillation is a stationary process on duration of several minutes [8]; therefore this process with a first order homogenous Markov chain can be modeled. We apply a HMM where each state represents a different level of attenuation with given transition probabilities to the other states. The parameterization of the HMM can be performed with the Baum-Welch algorithm [4] which requires an appropriate training data series. This training data series can be generated by filtering a Gaussian white noise [9] with the parameters given in Table 9.2:

Table 9.2 Filter parameters for the training data generation

Initial data	Gaussian white noise
Filter type	4th order Butterworth
Cutoff frequency	0.1 Hz
Stopband slope	-24 dB/decade

The HMM is applicable to generate scintillation time series for any desired duration, giving the fine structure of the attenuation process and it can be added to the large scale synthetic time series of attenuation generated with the combination of the two-state and the fade duration models.

9.9 Evaluation of the Synthesized Time Series

A typical synthesized attenuation time series for the 80 min. duration can be seen on the Fig. 9.10. The time series generation has been performed with 0.1 dB attenuation resolution and a 2–30 dB attenuation range. The scintillation was generated with the HMM method and added to the previously generated time series.

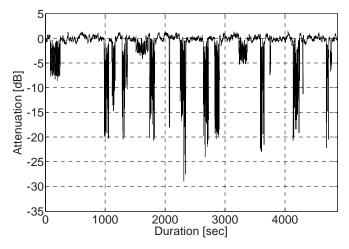


Fig. 9.10 A realization of the synthesized attenuation time series

To qualify the attenuation time series first and second order statistics are calculated. The CCDF of the measured and synthesized time series is depicted in Fig. 9.11.

The agreement of the two curves is quite good; however each realization of the generated time series may give a different result.

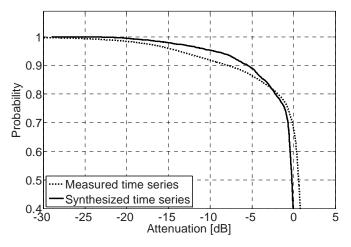


Fig. 9.11 CCDF of the measured and generated time series. RMSE = 0.0211

As the CCDF is invariant to the time series permutation, the fade duration statistics has also been calculated to compare the measured and synthesized time series. In Fig. 9.12. the number of fadings are depicted as the function of duration for 5 and 10 dB thresholds. At longer durations a deviance of the synthesized fade duration time series statistics can be observed from the measured time series statistics. As the fade length is determined by the two-state Markov model, further investigations are needed to refine the parameters of this model and achieve better statistical agreement at the long durations.

Nevertheless, Fig. 9.12. has been calculated from one realization of the synthetic time series. By generating multiple synthetic time series and averaging them a better result may be achieved.

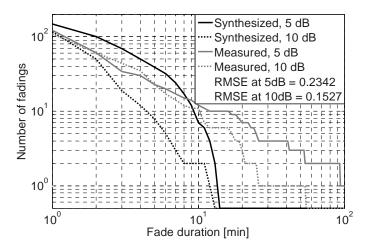


Fig. 9.12 Fade duration of the measured and generated time series

9.10 Summary

In this contribution a method to generate attenuation time series for land mobile satellite channel with multipath fading has been presented. After the investigation of the stationarity of the measured LMS attenuation time series, the need of combined Markov modeling method has been assessed. The synthesize process has two main steps: at first with a properly parameterized two-state fade/non-fade Markov chain the long term behavior of the channel will be simulated by fade/non-fade events. It simulates the time between two consecutive fading and the duration of the single fade events. The next step is to simulate the fine structure of a fading event with the desired duration and maximal attenuation. The combination of the two steps generates a long term attenuation time series. In the parameterization of the two-state fade/non-fade model and fade event generator an attenuation threshold dependent Fritchman's Markov model has been used. Finally, the scintillation has been modeled with a HMM, resulting a complete representation of the original behavior of the LMS attenuation process. The proposed time series generator has been evaluated with comparing the first and second order statistics (CCDF and fade duration) of the synthesized and real measurement time series. A realization of the attenuation process generated with our stochastic method shows sufficient agreement with the measurement, however further improvement is needed to achieve better matching of the second order statistics. This can be done by parameter sensitivity analysis, particularly in the case of the two-state Markov model parameters.

Future extension of the time series synthesizer could be the implementation of the seasonal variability of the two-state model with taking into account the local climatic variability in the model parameters.

Acknowledgment

This work was carried out in the framework of IST FP6 SatNex NoE project and supported by the Mobile Innovation Center, Hungary (Mobil Innovációs Központ).

References

- 1. G.E. Corazza (Ed.) (2007) "Digital Satellite Communications", Springer
- L. Castanet (Ed.) (2007) E-book: "Influence of the Variability of the Propagation Channel on Mobile, Fixed Multimedia and Optical Satellite Communications", www.satnex.org
- 3. B. D. Fritchman (1967) "A binary channel characterization using partitioned Markov chains," IEEE Trans. Inform. Theory, vol. 13, pp. 221–227
- 4. Welch L. R (2003) "Hidden Markov Models and the Baum-Welch Algorithm", IEEE Information Theory Society Newsletter, pp. 1–13
- R.H. Clarke (1968) "A statistical theory of mobile radio reception," Bell Syst. Tech. J., vol. 47, pp. 957–1000
- 6. L. Castanet, T. Deloues, J. Lemorton (2003) "Methodology to simulate long-term propagation time series from the identification of attenuation periods filled with synthesized events", COST 272-280 International Workshop on Satellite Communications from Fade Mitigation to Service Provision, Noord-wijk, The Netherlands
- L. Csurgai-Horváth, J. Bitó (2007) "Attenuation Time Series Synthesis for Land Mobile Satellite Links", 16th IST Mobile and Wireless Communications Summit, Budapest, Hungary, CD proceeding #292

- 8. V. I. Tatarskii (1961) "Wave Propagation in a Turbulent Medium", translated from Russian by R. A. Silverman (McGraw-Hill, New York)
- 9. E. Matricciani, M. Mauri, C. Riva, (1996) "Relationship between scintillation and rain attenuation at 19.77 GHz", Radio Science, vol.31, no2, pp. 273–279
- E. Lutz, D. Cygan, M. Dippold, F. Dolainsky, W. Papke (1991) "The Land Mobile Satellite Communication Channel-Recording, Statistics, and Channel Model", IEEE VT-40, pp. 375–386
- 11. Gardiner C.W. (2003) "Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences", Third Edition, Springer
- J. Bitó (1996) "Adaptive digitale Kanalmodelle für Mobilfunkkanäle" (Adaptive digital channel models for mobile radio channels), Frequenz (Germany), Vol. 50, No. 11–12, pp. 261–267
- L. Csurgai-Horváth, J. Bitó (2006) "Fade Duration Modeling of Satellite Links Applying Markov Chain", 3rd Advanced Satellite Mobile Systems Conference, Herrsching am Ammersee, Munich, Germany, pp. 76–83
- J.-Y. Chouinard, M. LeCours, G.Y. Delisle (1988) "Estimation of Gilbert's and Fritchman's models parameters using the gradient method for digital mobile radio channels," IEEE Trans. Veh. Technol., vol. 37, pp. 158–166
- 15. E. Lutz (1996) "A Markov model for correlated land mobile satellite channels", Int. Journal of Satellite Communications vol. 14, pp. 333–339

10 Cost-Optimised Active Receive Array Antenna for Mobile Satellite Terminals

Oliver Lücke, Alberto Pellon, Pau Closas, and Juan A. Fernandez-Rubio

10.1 Introduction

In the following we present results from research performed in the ESA funded project "Cost-Optimised High Performance Active Receive Phase Array Antenna For Mobile Terminals (CORPA)" (ESTEC/Contract No. 19814/06/NL/JD).¹

The project objectives are the design and prototyping of a complete DBF platform providing multimedia reception in the S-band (frequencies around 2.6 GHz) within the Satellite Digital Multimedia Broadcasting (S-DMB) system.

The complexity of the design of such antennas requires an extensive theoretical analysis of feasibility w.r.t. cost optimisation and functionality.

The design of the antenna is performed in four steps: (i) analysis of the environmental conditions, (ii) array design (shape/geometry and radiating elements), (iii) DBF techniques and algorithms design (iv) DBF platform selection. In the following we will provide a detailed discussion of these design steps.

10.2 System Scenario

The system scenario to be considered is defined by the Digital System E of the S-DMB standard [1]. The service link is in S-band, specifically, a 25 MHz bandwidth and a centre frequency of 2642.5 MHz have been proposed.

¹ Project partners: Space Services CE Lda. (Portugal), Instituto de Telecomunicacaos (Portugal), TriaGnoSys GmbH (Germany), Universitat Politècnica de Catalunya (Spain), Satellite Services BV (Netherlands).

Digital System E deploys Code Division Multiplex (CDM) based on QPSK or BPSK modulation with concatenated Reed–Solomon (RS) and convolutional error correcting coding. The chip rate is 16.384 Mcps and the processing gain amounts to 64 for the data channels (QPSK) and 2048 for pilot symbols transmitted on the pilot channel (BPSK).

Following the Digital System E specifications, initial system requirements according to Table 10.1 were established, where the antenna agility requirements (field of view, maximum angular velocity and acceleration) assume a land mobile scenario with the antenna being mounted on, e.g., a passenger car.

Parameter	Value
Frequency	2630–2655MHz
Polarisation	RHCP or LHCP
Field of view	Azimuth: 0°–360°
	Elevation: 8°–72°
G/T	$-8 dB/K^{\circ}$
Interference Level	C/I > 15 dB
Gain	> 15.1dBi
Noise figure	0.8dB
Polarisation isolation	< -15dB to < -20 dB
Sidelobes	< 10 dB
Maximum Angular velocity	45°/s
Maximum angular acceleration	$45^{\circ}/\mathrm{s}^2$

Table 10.1 Antenna specifications

RHCP Right-hand Circular Polarised, LHCP Left-hand Circular Polarised

10.3 Simulation Assessments

10.3.1 Environmental Conditions

The antenna design has to be driven by the analysis of the environmental conditions. Each possible area of application for the DBF antenna, namely aeronautical, maritime, and land mobile, poses particular requirements on the antenna system.

The land mobile scenario is considered the most demanding scenario due to severe multipath propagation, frequent line-of-sight shadowing/blocking, and dynamics of vehicle movements. The focus of the antenna design will thus be on the land mobile scenario. In the land mobile case it needs to be addressed that the signal environment may be highly time variant, where the definition of the multi-path propagation channel characteristics is of particular importance.

10.3.2 Directional Land Mobile Satellite Channel Model

The land mobile satellite (LMS) channel is characterised by multi-path propagation and frequent blocking/shadowing of the line-of-sight (LOS) path typically due to vegetation, buildings, hills, etc. To assess the performance and algorithm requirements of an DBF antenna array by means of software simulations, realistic modelling of the LMS is needed. This requires in particular a channel model that captures the angles of arrival (AoA) of the multipath signal components.

In the frame of the project described here, an easy to implement and low complexity geometry-based stochastical model (GSCM) of the LMS channel was realised using MATLAB for assessing DBF performance (cf. [2] which deals only with the terrestrial mobile channel; similar channel modelling techniques for the LMS channel are proposed in, e.g., [3,4]).

The GSCM is based on a statistical distribution of scatterers, which act as a diffuse reflectors of the satellite signal. The angularly resolved impulse response for a given geometric distribution of scatterers is computed with a simple ray tracing algorithm assuming only single scattering.

The distribution of the scatterers has to be defined such that the resulting power delay profiles (PDP) and the angular power spectrum (APS) agree reasonably well with the observations obtained by measurements.

In the standard implementation, near scatterers are positioned around the mobile terminal. Here it is assumed that the scatterer are uniformly distributed around the mobile terminal up to a certain visibility radius, cf., e.g., [2,5]. As the mobile terminal moves, new scatterers appear and disappear at the edge of the visibility radius.

Further, the non-uniform scatterer cross section (NSCS) method is applied, where the scatterers are uniformly distributed but the scatterer cross section (i.e. the reflectivity factor) being a function of the radial distance to the terminal [2,6]. Here, weighting of the scatterers' cross section according to a negative exponential function is applied which directly results in the desired negative exponential PDP. Due to the geometry of satellite, terminal and scatterer plane the iso-delay curves are ellipses and accordingly equal scatterer cross section is applied to all scatterers on the same iso-delay ellipse.

Further, to better agree with measurement data, far scatterers (representing high-rise buildings, mountains) are introduced in the GSCM [2]. Far scatterers are concentrated in clusters being sufficiently far located from the mobile terminal to reproduce the large excess delays observed in measurements.

Finally, dynamics of interruption and regain of the LOS is independent of the scatterers (in contrast to [3]) and described by a 3-state Markov model (e.g. [7]). Switching between the states is "hard", i.e. effects such as knife edge diffraction that would lead to oscillations of signal power and a smoother transition region between states are currently not implemented but could be included. The channel simulator tool implemented for this project allows also to include land mobile vehicle mobility and platform movements e.g. being obtained from measurements. In the following the channel characteristics obtained from a test scenario will be analysed and verified against expected results as far as they are available.

Figures. 10.1 and 10.2 shows the time series of the elevation and azimuth angles, respectively, of the LOS and the multi-paths. Variation of the LOS elevation is mainly due to vehicle movements (roll and pitch). Very low elevation angles, even negative, occur due to the far scatterers or vehicle movements.

Due to the lack of applicable measurement data for the spatial characteristics of the LMS channel, verification of the channel simulator has to focus on the case that an omni-directional antenna is used, i.e. when spatial information is not relevant.

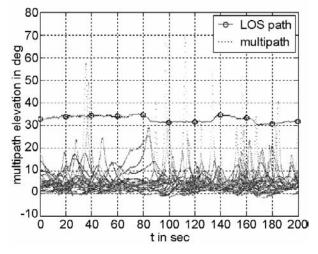


Fig. 10.1 Time series of elevation angles of LOS and multipath signals

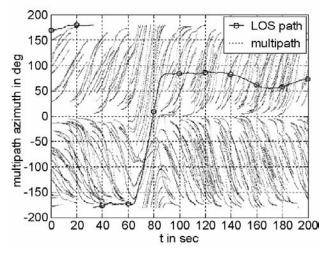


Fig. 10.2 Time series azimuth angles of LOS and multipath signals

Figure 10.3 shows the normalised squared absolute value of the timevariant channel impulse response $h(t, \tau)$ obtained with the spatial channel simulator assuming an omni-directional antenna (no far scatterers).

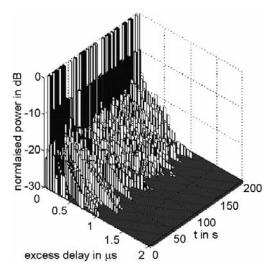


Fig. 10.3 Normalised (LOS path power is 0 dB) squared absolute value of the time-variant channel impulse response $h(t, \tau)$

Assuming a symbol duration being much larger than the delay spread for the further analysis (frequency flat or multiplicative fading), the multi-path components for a given t are summed over all excess delays

 τ resulting in the channel impulse response h(t). The statistics of h(t) as obtained from the simulation can now be compared with the known expected statistics; the PDF of the total receive signal power agrees very well with theory (cf. Fig. 10.4).

The PDF of excess delay is not negative exponential due to the uniform distribution of near scatterers but an accurate negative exponential PDP is obtained due to the NSCS model (cf. Fig. 10.5).

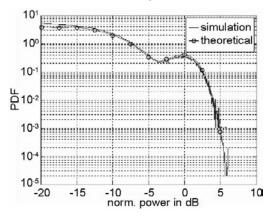


Fig. 10.4 PDF of the power of the sum receive signal (LOS path plus multipath) assuming an omni-antenna and flat fading

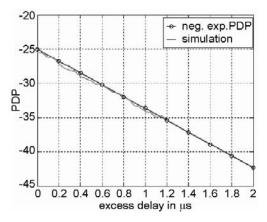


Fig. 10.5 PDP for near scatterers with NSCS

Assessment of the DBF algorithms performance in the LMS channel can now be performed based on the model described above. It remains to verify the directional LMS channel simulation model with measured data as soon as they are available.

10.3.3 Conformal Array Simulations

DBF algorithms suitable for the respective mobile scenario and for a conformal array shall be investigated within this study. To evaluate realistically the performance of adaptive beamforming algorithms, an accurate conformal array model should be deployed taking into account not only the geometry of the array but also the array elements' radiation characteristics.

Numerical methods such as the method of moments are usually deployed for accurate analysis of the electromagnetic properties of antennas (scattering matrix, far-field radiation pattern, etc.). The advantage of this approach is the high accuracy and the ability to capture all details of an antenna design, e.g., mutual coupling between adjacent elements of an array, radiating element geometry, dielectric substrate, antenna feeds, etc.

The 3D electromagnetic solver WIPL-D [8], which uses the method of moments, is used in this project to investigate various conformal array geometries and radiating element designs. This software is suitable for this kind of structures due to its ability to deal with 3D blocking systems. However, additional software must be developed to compute phase compensations during the simulation. Relative phase differences between the elements must be compensated coming from

- relative propagation delay differences due to different array element phase centre locations,
- element amplitude and phase pattern (including polarisation) differences due to different arrival angles for different elements.

The high level of accuracy and detail comes at the cost of increased simulation times, depending on the complexity of the investigated antenna structure. The computation times typically are around a few minutes for determination of the radiation pattern for antenna set-ups comprising a few patch antennas, which seems acceptable if the radiation pattern needs to be computed only once. However, considering investigation of adaptive beamforming algorithms in particular in a time dependent signal environment it is usually required to compute many times the radiation pattern to simulate convergence of adaptive beamforming weight computation to a static or time-variant signal environment.

Therefore, an alternative approach based on the isolated element pattern (e.g. obtained from WIPL-D simulation) and taking into account the exact array geometry and element orientations is proposed. This approach allows implementation of a simulation code for fast computation of the radiation pattern, providing also a software interface for an adaptive beamforming algorithm code. Instead of using the pattern of the isolated element, neglecting mutual coupling, the embedded element patterns could be computed using WIPL-D to include also the effect of coupling.

A MATLAB simulation tool was implemented that uses the element patterns pre-computed in WIPL-D (or any other similar electromagnetic (EM) solver) and a given array geometry to compute the individual array element responses (in amplitude and phase) to a plane wave of given polarisation impinging on the conformal array.

As a simple example Fig. 10.7 shows the comparison of RHCP and LHCP (right/left hand circular polarised) radiation pattern for 6 elements arranged as shown in Fig. 10.6; the main lobe is pointed towards $\vartheta = 75^{\circ}, \varphi = 0^{\circ}.^{2}$

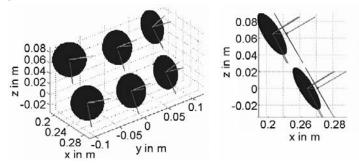


Fig. 10.6 6 array elements arranged on a cone. Right: side view

As expected, the CA simulation tool using the isolated element patterns, neglecting mutual coupling, cannot reproduce exactly the pattern computed with an EM-solver.

However, the accuracy is adequate to assess in the required detail the performance of adaptive beamforming algorithms because the co-polar (here: LHCP) main lobe is sufficiently accurate reproduced in terms of peak value and pointing direction, and beamwidth (cf. theta- and phipattern-cuts shown in Fig. 10.8).

The obvious difficulties in accurately reproducing sidelobes and the cross-polarised pattern are not considered critical; any (adaptive) beamforming algorithm must perform mostly independent of the particular shape of sidelobes, location of pattern nulls, and the cross-polarised pattern because in an implementation of the antenna these depend also on factors such as manufacturing inaccuracies.

² Note for the WIPL-D simulation that the main lobe is not exactly pointing towards $\varphi = 0^{\circ}$ due to mutual coupling; thus the stronger asymmetry of the LHCP pattern as shown in Fig. 10.6; however, we can neglect for the rough comparison considered here.

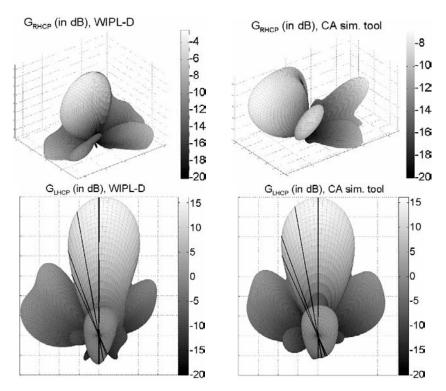


Fig. 10.7 Comparison of gain patterns (3D patterns are in log-scale, values below –20dB are not shown). *Top:* RHCP. *Bottom:* LHCP (top view). *Left:* WIPL-D results. *Right:* CA simulation tool using isolated element patterns. (also indicated are cuts for $\varphi = 0^{\circ}, 10^{\circ}, 20^{\circ}, 24^{\circ}$)

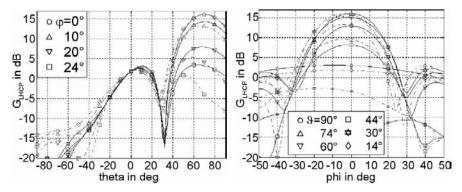


Fig. 10.8 Comparison of LHCP gain patterns as obtained with WIPL-D (*dot* and *dash line*) and the CA simulation tool (*solid line*) implemented for CORPA. *Left:* theta-cuts for $\varphi=0^\circ$, 10° , 20° , 24° . *Right:* phi-cuts for $\vartheta=90^\circ$, 74° , 44° , 20° , 10°

10.4 Antenna Design

Although a planar antenna sub-system seems to be most attractive solution for reaching a low profile terminal, its efficient realisation especially while antenna points to lowest elevation angle would be faced with numerous problems. Regarding the scan range needed two conformal antenna configurations appeared as potential candidates, sphere or cone and have been traded-off (cf. Fig. 10.9).

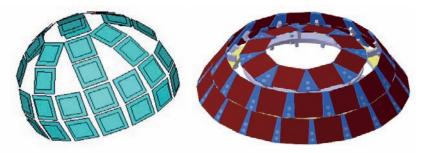


Fig. 10.9 Spherical section vs. double cone option

Both configurations were analysed comparing number of radiating elements, number of elements forming the main beam, overall dimensions and obviously compliance to technical requirements. In particular, a dual cone configuration has been selected thanks to its advantage in terms of number of radiating elements, overall dimensions (profile) and easier integration. The baseline solution is designed to provide coverage from 8° to 90° and its overall dimensions will be around 14 cm of height and 56 cm of diameter.

Both commercial and in house developed 3D software tools were used for the design of the array. Nevertheless, an extensive trade-off has been conducted between different kinds of radiating elements. Aperture coupled elements, coaxial-fed patches and planar inverted-F antenna (PIFA) elements have been simulated and breadboarded in order to make the proper selection. Following breadboarding and testing activities, a PIFA element on air (cf. Fig. 10.10) has been chosen in order to increase efficiency, and reduce the number of elements at array level, a key point to reduce processing power in digital domain and the number of LNBs (low noise block converters) needed.

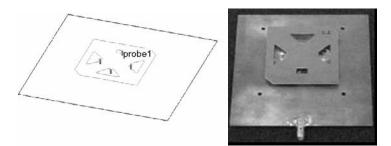


Fig. 10.10 Single radiating element on ground plane; *Left:* WIPL-D model; *Right:* real element

The measured pattern of a single PIFA element is depicted in Fig. 10.11, showing satisfying performance in terms of gain and cross-polarisation suppression.

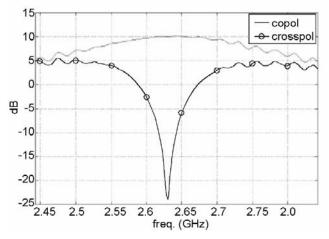


Fig. 10.11 Anechoic chamber measurement of a single PIFA element (*x*-axis: frequency in GHz; *y*-axis: co-/cross-polar gain in dB)

A complete front-end has been developed, including Low Noise Amplifier (LNA), filters, demodulator and down converters. The main drivers of the design were low noise figure (NF), suitable gain and proper interface to the digital board. In fact, an intermediate frequency (IF) signal will be delivered to analog-to-digital (A/D) converters for processing purposes. A customized LNA based on transistor has been developed to achieve both, low NF and low power consumption, key point in a mobile system. To avoid saturation of the I&Q demodulator a miniature printed band-pass filters have been designed, breadboarded and tested. As sketched in Fig. 10.12, the full radio frequency (RF) front-end could be integrated below the radiating element.

Requirements for NF (below 0.8 dB), phase noise, filtering rejection, ripple and group delay are a challenge due to the hard restrictions related to cost and power supply maximum allowed for a mobile system.

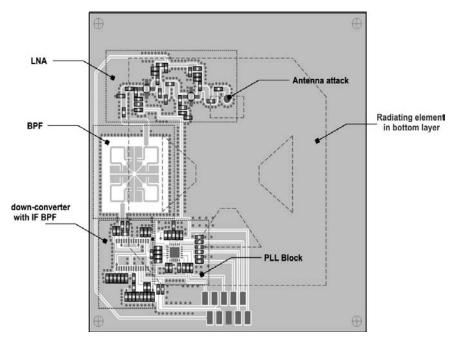


Fig. 10.12 Schematic view of RF front end of a single radiating element

In addition to this a mechanical subsystem was developed to hold the whole system, including the digital platform (cf. Fig. 10.13). This structure was developed in a plastic material to reduce cost and weight.

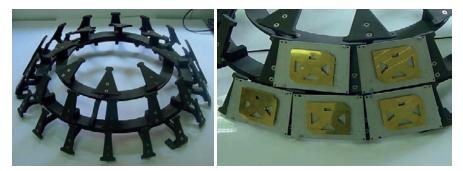


Fig. 10.13 Mechanical subsystem. *Right:* with 5 radiating elements including RF front-end

10.5 Digital Beamforming

A particular challenge in a mobile environment is the design of the adaptive DBF algorithm that has to cope with multipath fading effects, blocking and shadowing of the line-of-sight, and interference.

Several techniques can be considered depending on the side information available (namely temporal or spatial a priori knowledge) to compute the weights of the antenna array (w). Thus, the array output is computed as $y[n] = w^{H} \cdot x[n]$.

The use of spatial information yields to the Minimum Variance Beamforming (MVB), which minimizes the array output power constraining the pattern to point to the desired signal. This approach is robust against synchronization errors. However, the main drawback is that the steering vector of the desired signal must be perfectly known. Maladjustment of the steering vector can drastically degrade the performance of the antenna array, even nulling the desired signal. A possible source of error in that sense is a bad calibration process. Hence, calibration is a key aspect to be considered when using MVB.

In contrast, Temporal Reference Beamforming (TRB) avoids the need of a perfect calibration of the antenna array arms (although desirable), since the information of the steering vector is not considered. The side information consists in a reference signal highly correlated with the desired signal and uncorrelated with the interference signals, i.e. a known signal structure like a preamble or a pilot signal as the S-DMB system includes [1]. Its major constrain is the need of a perfect synchronization.

When perfect synchronization is delivered, the TRB technique exhibits the optimum signal-to-interference-plus-noise ratio (SINR) at the array output. This technique minimizes the mean square error between the array output and the reference [9,10] yielding the following weights expression: $\mathbf{w}_{\text{TRB}} = \mathbf{R}_{xx}^{-1} \cdot \mathbf{P}$. Where \mathbf{R}_{xx} stands for the autocorrelation matrix of the received signal at each element and vector \mathbf{P} is the correlation between the received signal and the reference, which provides an estimate of the steering vector of the reference signal.

Notice that the reference and the desired signal directions of arrival are the same since both are transmitted by the same satellite. For the case under study, we aim at mitigating interference without perfectly calibrating the antenna array. Thus, the technique considered hereafter is the TRB.

In addition, the spread spectrum nature of the received S-DMB signal [1] makes necessary to decide whether to despread the reference signal before DBF or after. Notice that data signals are not considered in the

computation of weights, and thus they are not despreaded. Indeed, they are delivered transparently to the communications receiver, and only the reference pilot channel is considered.

When performing Digital Beamforming (DBF) before despreading the pilot signal, the main advantage is that the digital processing demanded by the despreading process is only required at the array output, in contrast to despreading the signal before DBF where each arm of the antenna array must be processed. In contrast, the main drawback is that the weight vector update must operate at a higher rate and that the construction and inversion of the autocorrelation matrix becomes cost consuming.

When weight vector update is performed after despreading the pilot signal, the rate is reduced by a factor given by the processing gain of the CDM system, which alleviates this computation. The TRB technique can be implemented by means of an adaptive algorithm (e.g. LMS or RLS) or considering the Sample Matrix Inversion (SMI). The latter computes the inverse of \mathbf{R}_{xx} directly without an iterative algorithm and will be considered for implementation in the digital platform, thanks to the data rate reduction achieved after despreading the pilot signal.

The operation is as described in Fig. 10.14, the pilot channel is despreaded and TRB weights are computed at a low rate (bit rate). These weights are copied in the non-despreaded part of the array to electronically steer the pattern to improve the SINR at the array output. Hence, the output of the array is the spread spectrum S-DMB signal with an optimum SINR.

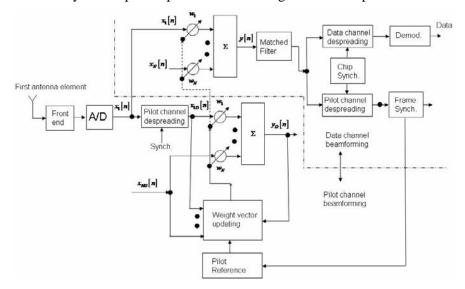


Fig. 10.14 Block diagram of DBF processing

10.5.1 Implementation Aspects

Regarding the DBF platform design, it is of interest to achieve a solution based on commercial available components, especially in the digital processing part (DSP/FPGA devices) of the terminal.

A mixed solution of state-of-the-art FPGA with embedded hard/soft DSP devices can cope with the digital requirements of the system and the designed DBF techniques. In particular, the system has been implemented using Xilinx FPGA technology XC5VLX220.

As argued before, the TRB technique relies on a good synchronisation to properly operate. In a cold start, this synchronism cannot be assumed known and a strategy should be considered to lock it. To this aim, the designed system is composed of 2 operation modes: Acquisition and Tracking.

On the one hand, Acquisition mode considers the acquisition of synchronism (cf. Fig. 10.15). Basically, within this mode, sectors of the conformal array are considered and the pattern is electronically steered to roughly point somewhere in the space. Notice that we allow maladjustments in the steering vectors, since we are not really interested in perfectly pointing to a given direction but to feed data from the actual sector to the receiver. Thus, calibration is still not considered. The array output feeds the Communications Receiver, which is in charge of acquiring synchronism, when the delivered signal achieves a certain signal-to-noise ratio threshold. Several sectors are scanned until synchronism is locked.

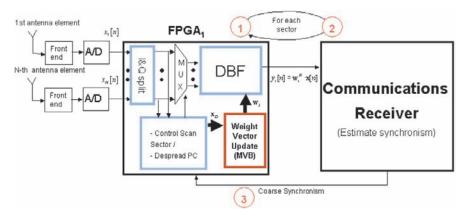


Fig. 10.15 Acquisition phase in the DBF processing

On the other hand, tracking mode starts when synchronism is locked and the TRB operates to feed the communications system with reliable data (cf. Fig. 10.16). With proper synchronism, pilot channel can be despreaded and weights can be computed. As long as synchronism is locked, the antenna array will be able to track the satellite with no need of knowing its steering vector.

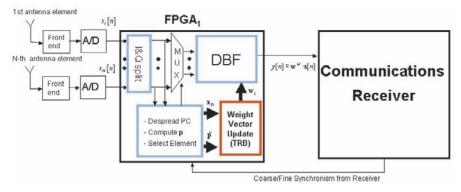


Fig. 10.16 Tracking phase in DBF processing

The implementation of the most interesting blocks in the Tracking mode is presented hereafter. Namely, the inphase and quadrature (I&Q) splitting block, the despreading and selection of illuminated elements and the computation of DBF weight vectors are studied.

10.5.1.1 A/D Conversion and IF-Sampling

The A/D conversion can be performed either in baseband or using IFsampling. The conventional approach consists in splitting the received signal and multiplying each arm by a local oscillator (LO) and a 90° shifted version. This procedure provides I&Q components after lowpass filtering and quantization. A number of errors may appear when considering the conventional approach, mainly caused because the two arms must be closely matched for correct demodulation, e.g. gain balance, quadraturephase balance or DC offsets.

In the proposed design, the IF-sampling philosophy is considered. In particular, the design considers: $f_s = 65.536$ MHz, $f_{IF} = 49.152$ MHz and $R_c = 16.384$ MHz, being the sampling, intermediate and chip frequencies respectively.

After IF-sampling, a digital frequency translation can be performed shifting the spectra to zero frequency. This is achieved in the digital domain multiplying the digitalized signal by $e^{in\pi/2}$. Note that the signal is cyclically multiplied by $\{1, j, -1, -j\}$ which consists in taking the even samples alternating the sign as the I component and the odd samples, also alternating the sign, as the Q component. The odd samples of the I component are null as well as the even samples of the Q components.

Figure 10.17 depicts how the digital I&Q demodulator has been implemented in the digital platform, it can be observed that the number of inputs to the FPGA has been reduced to a half w.r.t. a conventional approach since the I&Q stream splitting is performed in the digital platform.

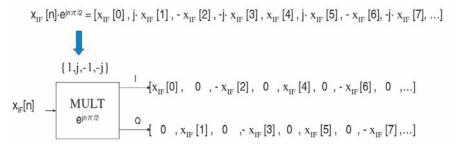


Fig. 10.17 Schematic diagram of the implementation of I&Q-sampling on the FPGA

10.5.5.2 Pilot Despreading and Selection of Illuminated Elements

The despreading of pilot signal and the computation of vector P are related processes. Figure 10.18 shows a schematic representation of this block for the *i*-th antenna element, being its input the I&Q data streams. Basically, a correlation is performed with the known spreading sequence of the Pilot Channel, but in a sequential manner. Each sample of the stream is multiplied by the corresponding spreading code chip (since synchronism is available) and an accumulator is used to add all samples of the stream. When $64 \cdot N_{sc}$ samples have been added, the block outputs a despreaded bit, and when $2048 \cdot N_{sc}$ samples have been added, the block outputs the *i*-th element of the **P** vector. In addition, when the **P** vector is computed, the N_{p} largest values correspond to the illuminated elements, which are then selected and used in the Weight Vector Update block.

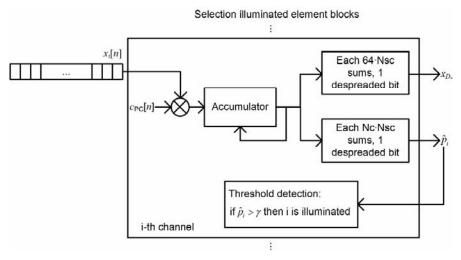


Fig. 10.18 Despreading of a single antenna element signal

10.5.5.3 DBF Weight Vector Update

The implementation of the Weight Vector Update is discussed. The inputs for this block are the complex despreaded pilot signals of each illuminated element and its corresponding complex P vector. Considering that in the S-DMB system the processing gain is 64, the despreaded signal rate is 256 KHz. Despreaded data has a low rate, which alleviates the processing power required to compute the correlation matrix and its inversion. However, the operations involved in the weight vector update are best suited to floating point arithmetic and to processor-like set of instructions. Thus, we consider the use of an embedded soft-processor in the FPGA device, whose main drawback w.r.t. hard-processors is that it consumes area in the FPGA. This 32-bit RISC processor is known in Xilinx devices as Micro-Blaze. Recalling from previous section that the weight equation is $\mathbf{w}_{\text{TRB}} = \mathbf{R}_{xx}^{-1} \cdot \mathbf{P}$, the most consuming operation to be performed is the computation and inversion of the autocorrelation matrix. In order to alleviate the computational cost of this operation, the QR decomposition has been considered.

Finally, in order to assess the correct operation of the digital platform, the system is fed with synthetic S-DMB signal generated using MATLAB software. In particular, a scenario where the line-of-sight-signal impinges the array at 0° and an interference at 9° is considered. Also, the FPGA communicates with a PC Workstation MATLAB shell via serial port, delivering computed weights. This allows us to plot the resulting radiation

pattern of the digital antenna array for any given scenario simulated. As shown in Fig. 10.19, the computed radiation pattern exhibits the expected behaviour: pointing to the LOS and nulling the interference.

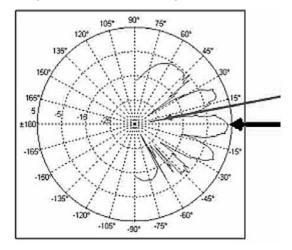


Fig. 10.19 Radiation pattern adapted to the desired signal at 0° (*bold arrow*) and an interfering signal at 9° (*narrow arrow*)

10.6 Conclusions

After finalisation of the simulative performance analysis of the antenna and selection of the DBF platform beginning of 2007, the project is now (October 2007) still in the implementation phase (antenna, DBF platform and software). Integration of all components is planned towards the end of this year, followed with measurements in an anechoic chamber. The CORPA project is expected to be finalised early 2008.

References

- "ITU-R BO.1130-4: Systems for digital satellite broadcasting to vehicular, portable and fixed receivers in the bands allocated to BSS (sound) in the frequency range 1 400–2700 MHz," Tech. Rep., International Telecommunications Union – Radio Communications (ITU-R), 2001.
- A.F. Molisch, A. Kuchar, J. Laurila, K. Hugl, R. Schmalenberger, "Geometrybased directional model for mobile radio channels – principles and implementation". In: European Trans. Telecomm. 14, pp. 351–359, 2003.

- P.R. King, B.G. Evans, S. Stavrou, "Physical-statistical model for the land mobile-satellite channel applied to satellite/HAP-MIMO". In: Proceedings 11th European Wireless Conference, vol. 1, April 2005, pp. 198–204.
- F. Amoroso, W.W. Jones, "Geometric model for DSPN satellite reception in the dense scatterer mobile environment." In: IEEE Trans. Commun., vol. 41, no. 3, pp. 450–453, Mar. 1993.
- R.B. Ertel, P. Cardieri, K.W. Sowerby, T.S. Rappaport, J.H. Reed, "Overview of spatial channel models for antenna array communication systems". In: IEEE Personal Communications, vol. 5, no. 1, pp. 10–22, Feb 1998.
- J. Laurila, A. F. Molisch, E. Bonek, "Influence of the Scatter Distribution on Power Delay Profiles and Azimuthal Power Spectra of Mobile Radio Channels," in Proc. ISSSTA'98, Sept. 1998, pp. 267–271.
- F.P. Fontan, M. Vazquez-Castro, C.E. Cabado, J.P. Garcia, E. Kubista: "Statistical Modelling of the LMS Channel". In: IEEE Transactions on Vehicular Technology, vol. 50, no. 6, pp. 1549–1567, Nov 2001.
- 8. WIPL-D, "Electromagnetic Modelling and Simulation", 2006. [Online]. Available: http://www.wipl-d.com/index.php. [Accessed: Jul. 28, 2006].
- 9. Harry L. Van Trees, Optimum Array Processing. Detection, Estimation and Modulation Theory, Part IV. Wiley Interscience, 2002.
- 10. Robert A. Monzingo and Thomas W. Miller, Introduction to Adaptive Arrays, John Wiley & Sons, 1980.

11 Scheduling Techniques for Mobile Broadcast and Multicast Services

Michael Knappmeyer, Robin Chiang, Nigel Baker, and Ralf Tönjes

11.1 Introduction

With the increasing capabilities of mobile terminals, the reception of multimedia content is becoming more and more attractive for customers. Improving high bandwidth provision enables 3G cellular systems to deliver such data to the end user. The 3G mobile experience is following a trend towards personalized services and applications which are used not only for voice communication but also for entertainment. Customers favour short video/audio clips rather than watching complete movies on their small screen. The deployment of new broadcasting networks like Digital Video Broadcasting – Handheld (DVB-H) results in large costs both in terms of capital as well as operating expenditure. Hence it is an enormous commercial risk for providers to install such systems.

Multimedia Broadcast Multicast Service (MBMS) is considered a promising technology to distribute multicast and broadcast services efficiently and at reasonable costs within existing 3GPP cellular networks. Enriched interactive multimedia services can be offered providing, for example, the latest news or publishing the current top ten music video clips. The trend towards user generated content distribution which is observed in the Internet will no doubt increase the number of available services and content.

Set against this background of predicted increase both in terms of number of services/content and users, the competition for limited resources will only intensify. MBMS is regarded as a bearer technology and not a complete service provisioning system therefore efficient scheduling for optimal utilization of radio and network resources is essential. Scheduling in MBMS comprises several levels of abstraction which require different strategies. This chapter presents several scheduling techniques and approaches for MBMS content distribution. It explains the concepts of (1) a *Dynamic MBMS Resource Scheduler* located in the Radio Access Network (RAN), (2) a *Carousel Service Scheduling* approach and (3) means for an optimal scheduling of *Multicast Streaming Services*. The latter two schedule services and content while the first deals with IP packets.

The remainder of this chapter is structured as follows. First, the problem is analyzed in-depth by introducing MBMS, explaining fundamentals of scheduling mechanisms with special regard to the context of MBMS. The three elaborated concepts are described afterwards. Finally, they are evaluated by showing individual simulation results before the conclusions are drawn.

11.2 Problem Analysis

The section details the problem definition and defines the need for scheduling of MBMS services.

11.2.1 Introduction of MBMS

In its 6th Release the 3rd Generation Partnership Project (3GPP) standardized MBMS to provide efficient broadcast and multicast service delivery in existing UMTS and General Packet Radio Service (GPRS) networks [1, 2, 3]. Limited to the downlink direction, broadcast/multicast IP packet delivery is introduced in the Core Network (CN) and point-to-multipoint (ptm) transmissions are specified for the Radio Access Network (RAN) in order to save radio resources. Without MBMS, the delivery of the same content to a group of users necessitates multiple subsequent transmissions. MBMS requires only one single transmission, serving all recipients simultaneously.

MBMS allows two modes of operation: the *broadcast mode* and the *multicast mode*. In broadcast mode, transmissions take place regardless of user presence in the defined target area, whereas in multicast mode only areas with subscribed users need to be served. MBMS multicast mode is associated with subscription and authorization prior to group joining. While in multicast mode users have to join a group before the service reception is possible, broadcast mode services are locally activated by the User Equipment (UE). Moreover, MBMS classifies three basic types of user services according to the method used to distribute their contents: (1) *streaming services* providing a stream of continuous media, e.g. video and audio, (2) *file download services* used to deliver binary files and (3) *carousel services* whose content (either download files or streaming media) is

periodically retransmitted. Subscription is always per service. Each service may contain several contents provided by a Content Provider (CP).

The MBMS reference architecture is shown in Fig 11.1. The Broadcast/Multicast Service Centre (BM-SC) is added as a completely new functional entity serving as central controlling unit. It is connected to the Gateway GPRS Support Node (GGSN) by the interfaces Gmb and Gi. The former provides access to the control plane functions, the latter to the bearer plane.

Inside the UMTS Terrestrial RAN (UTRAN), the MBMS data is delivered from the NodeB to the UEs using the logical channels (1) *MBMS Control Channel* (MCCH), (2) *MBMS Transport Channel* (MTCH) and (3) *MBMS Scheduling Channel* (MSCH). All three of them are mapped onto the Transport Channel Forward Access Channel (FACH) which utilizes the Secondary Common Control Physical Channel (S-CCPCH). It is important to note that in the current specification of 3GPP the power of the S-CCPCH is not controlled, i.e. the data is transmitted with full power regardless of whether recipients at the cell edge need to be supplied or not.

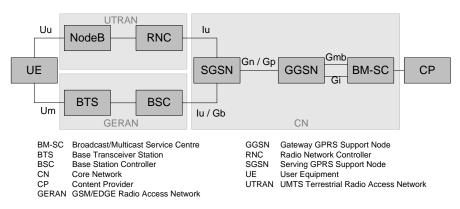


Fig. 11.1 MBMS reference architecture

11.2.2 Scheduling and Congestion Control in MBMS

11.2.2.1 Fundamentals

Generally speaking, scheduling is always necessary, if limited resources need to be shared, e.g. processing power in the CPU. The set of rules for the decisions are defined by the *scheduling discipline*. In packet switched communication networks, scheduling and queuing are important means for congestion control and for ensuring a defined Quality of Service (QoS)

level [4]. Congestion control algorithms can be subdivided into different categories [5]. *Open loop control* algorithms comprise *Source Control* and *Destination Control* algorithms. Their decisions do not depend on feedback information from the congested parts of the network. Hence the status of the network is not monitored. On the contrary, decisions of *Closed Loop Congestion Control* algorithms are based on such feedback information. The feedback can be either implicit (global), or explicit. Explicit feedback in turn can be further categorized into persistent and responsive in case of occurring congestion.

11.2.2.2 Scheduling in the Context of MBMS

When analyzing scheduling the context of MBMS, several categories of autonomous scheduling mechanisms can be identified. They differ with respect to (1) the entity which is performing the scheduling, (2) the objects which are arranged and (3) the discipline which applies. Scheduling is maintained on different abstraction levels and hence on different layers of the Open Systems Interconnection (OSI) reference model.

The first to mention is the scheduling of transport blocks and radio frames at the Data Link Layer. This scheduling is executed by the NodeB in collaboration with the Radio Network Controller (RNC) in order to utilize the available radio resources efficiently. It is part of the Radio Resource Management (RRM). As a basis for decisions the channel state information is used.

Since the user plane of the UMTS packet switched domain is based on the Internet Protocol (IP), the classical IP packet scheduling applies in the wired Core Network (CN) of the architecture. The network nodes can be considered as IP routers. With the introduction of the IP Multimedia Subsystem (IMS) for session control, combined approaches of Differentiated Services (DiffServ) and Integrated Services (IntServ) have been specified [6]. The IMS entities *Policy Decision Function* (PDF, Release 6) and *Policy Control and Charging Rules Function* (PCRF, Release 7) are responsible for resource reservation and QoS provision, respectively. The scheduling of IP packets operated by the NodeB might also include radio resource and channel feedback information in addition to QoS requirements. The usage of such information originating from other OSI layers results in cross-layered approaches. MBMS services are delivered unidirectional using the connection less User Datagram Protocol (UDP). Hence the data flow cannot be controlled at the Transport Layer.

For MBMS to be considered as a complete subscription based service and content provisioning system (not just bearer technology), services and content must be scheduled to optimize system resources and to maximize economic profit. This type of scheduling is maintained by the BM-SC as central entity. The RAN is always the bottleneck of the system. Compared to wired network paths, the radio links provide much less bandwidth. This chapter proposes two approaches which consider feeding RRM information back to the BM-SC. This way, the service/content scheduler is able to avoid congestion and to utilize the resources most efficiently.

MBMS is a dynamic system in terms of (1) user mobility, (2) varying cell capacity due to dynamic resource utilization, e.g. by classical voice telephony, (3) addition and removal of services and content, (4) dynamic service subscription and cancellation. Scheduling algorithms also need to cope with various QoS requirements. In the case of Streaming Services, a constant bandwidth needs to be provided and jitter should be as small as possible. Regarding Carousel Services, the average waiting time should be minimized whereas for Download Services the throughput has to be maximized.

This current work focuses on the scheduling of IP packets at the radio link and on central scheduling of content and services because these two categories are the most critical for the overall efficiency of MBMS. The scheduling of services and content for usage in subscription based systems like MBMS has not received much attention to date.

11.3 Concepts and Algorithms

In the following subsections, three complementing concepts for scheduling in MBMS are explained. The first approach aims at an efficient scheduling of IP packets in the RAN, whereas the second and third concepts deal with the task of optimizing the scheduling of content. The latter two distinguish between scheduling for Carousel Services and for Streaming Services.

11.3.1 Dynamic MBMS Resource Scheduler

The Forward Access Channel (FACH) was chosen to be the channel to transport MBMS traffic to its subscribers. However, the common channel FACH lacks power control. This is crucial in the delivery of broadcast and multicast services. We propose to optimize the common channel by adding a lightweight feedback mechanism in order to incorporate the power control function. The channel state feedback information from the receiving UEs is used to adjust the transmission level and hence to save unnecessary power resources. This reduces the interference and allows the provision of more services in an interference-limited system. There is currently no specific scheduling algorithm recommended by 3GPP to efficiently provide MBMS services to its subscribers.

Utilising queuing theory and Call Admission Control the scheduling decisions are improved in order to provide enhanced resource utilization and fairness to individual user groups. Moreover, the requirements of the three MBMS service types are considered.

11.3.1.1 Consideration of Channel State Information

The MBMS control function inside the RNC establishes an MBMS Radio Access Bearer (RAB) by sending service specific signaling messages to all the UEs in the cell listening to the MCCH. MBMS data is then transferred via the MTCH. The MSCH carries MBMS service transmission schedule information. It indicates when the specific MBMS service is expected to be transmitted. These three logical channels are then mapped onto the transport channel FACH which utilizes the S-CCPCH.

One way to incorporate a light weight feedback considering the channel state information of the associated UEs without overloading the uplink is to attach a dedicated channel to the multicast group. The power control mechanism of the dedicated channel can be used to control the inner loop power control of the common channel. Multicast users can hence roam independently from their initial open loop power control footprint without disrupting the service delivery.

The instantaneous SIR S_i of each user *i* can be written as shown in (11.1) where G_i is the spreading factor, h_i is the path loss, P_i is the total power, *W* is the total system bandwidth and η_0 is the background noise [8].

$$S_i = \frac{G_i h_i P_i}{\sum_{j \neq 1} h_j P_j + \eta_0 W}$$
(11.1)

11.3.1.2 Dynamic Resource Scheduling

Dynamic Resource Scheduling (DRS) is a centralized and adaptive framework providing resource scheduling by optimal power assignment and code hopping. The proposed Dynamic MBMS Resource Scheduler (DMRS) is illustrated in Fig. 11.2.

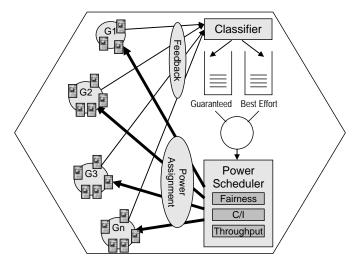


Fig. 11.2 Dynamic MBMS resource scheduler

The power control technique is used in order to keep the Signal Interference Ratios (SIRs) for all transmission sessions above a defined threshold at a certain time period. Equation 11.2 should hold to satisfy the required Minimum QoS γ_i for session *i* [8].

$$\sum_{i=1}^{N} g_i < 1; \quad g_i = \frac{\gamma_i}{\gamma_i + G_i}$$
(11.2)

$$P_{i} = \frac{\eta_{0} W g_{i}}{h_{i} \left(1 - \sum_{j=1}^{N} g_{j} \right)}$$
(11.3)

The Power Index g_i is based on the fixed spreading gain G_i of session *i*. Each multicast session is assigned a transmission power level P_i as shown in (11.3). Equation (11.2) analyzes the availability of the allocated FACH transmission power catered for MBMS traffic to admit an incoming session for the given interference conditions and required QoS. The group is only created if the inequality is satisfied. The spreading gain can be changed with regard to the information rate and the power allocated to the group to adjust to the required QoS level. This enhancement makes it possible to adapt to the fluctuation of the channel experienced by the multicast group. The goal is to find the power vector $\overline{P} = [P_1, P_2, ..., P_N]$ which minimizes the total power (11.4) such that the QoS requirements are satisfied (11.5) while satisfying all QoS requirements.

$$\sum_{i=1}^{N} P_i \tag{11.4}$$

$$S_i \ge \gamma_i; \quad i = 1, 2, ..., N$$
 (11.5)

The FACH transmission power is limited to \tilde{P}_i . When groups are created, their requests for subscription are classified into two groups as shown in Fig. 11.2 and buffered in the guaranteed queue or in the best effort queue according to the traffic characteristics of the requested service. The guaranteed queue stores requests for services that have to be delivered at a predefined rate. It includes real time and non-real time Variable Bit Rate (VBR), Constant Bit Rate (CBR) and Available Bit Rate (ABR) service data. The best effort queue is used for Unspecified Bit Rate (UBR) services. The scheduler serves requests from the guaranteed queue prior to those in the best effort queue. Both queues are internally processed in a First Come First Serve fashion.

11.3.1.3 Power Scheduling

The emphasis is placed on the power scheduler. It is located inside the NodeB and determines the spreading gain for each session request while checking the availability of Orthogonal Variable Spreading Factor (OVSF) codes. The power index is computed using (11.2). Next, the Node B employs the CAC test:

$$\sum_{j=1}^{N} g_j < 1 - \frac{\eta_0 W}{\min_i \left(\frac{\tilde{P}_i h_i}{g_i}\right)}$$
(11.6)

It is carried out at the head of the guaranteed queue. ABR, CBR and VBR services use minimum, average and maximum symbol rate respectively. The quantity of multicast sessions can be determined and the spreading code be assigned to each session. The DMRS scheme is performed every radio frame to serve each traffic class efficiently.

11.3.2 Scheduling of Carousel Services

11.3.2.1 Basic Principles

The content of a carousel service is defined to be repeated periodically for a guaranteed reception even in case of turned off or temporarily not supplied end devices. When the equipment is connected again, it automatically receives the content. The scheduling of carousel services and their contents reflects the problem of arranging an optimal broadcast circle. The number of instances per circle can be varied in order to achieve best performance. The most important efficiency metric for the algorithms is the average mean waiting time, i.e. the time between the user demand and the actual data reception.

The proposed concept is based on several deliberations: (1) The content is either broadcasted more often for services that more users are subscribed to (multicast mode) or the higher the public interest is (broadcast mode). (2) The smaller the content size is, the more often this content should be broadcasted. (3) Congestion should be avoided by not transmitting further on already jammed network paths and radio cells. Instead, the scheduler should wait until the resources become free again. (4) Content priority should be considered. (5) New contents should be added to the broadcast cycle automatically and immediately. The required interaction between the network entities is depicted in Fig. 11.3 below.

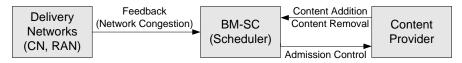


Fig. 11.3 Interaction for service/content scheduling

11.3.2.2 Basic Algorithm

According to previous work by Vaidya and Hameed [7] the following properties per content are primarily used for the scheduling decisions.

- *p_i*: demand probability
- l_i : required time duration (transmission length) for delivery
- R(i): point of time when the content was last recently broadcasted

Their Online algorithm is based on the square root rule given in (11.7). For every content *i* the value of the weighting function W(i) is calculated. The item with the highest resulting W(i) is chosen for being broadcasted next. *Q* denotes the current time. The difference Q-R(i) thus refers to the

spacing s_i between two consecutive instances of content *i*. *M* items are stored in the content database. As a result, the algorithm is able to construct an optimized broadcast cycle by adapting the frequency of an item. Vaidya's and Hameed's work can be further extended and improved to match the special needs of Carousel Service scheduling in MBMS as described below.

$$s_i \propto \sqrt{\frac{l_i}{p_i}} \Longrightarrow W(i) = s_i^2 \frac{p_i}{l_i} = \left(Q - R(i)\right)^2 \frac{p_i}{l_i}$$
(11.7)

11.3.2.3 Extension with Priority Scheme

The demand probability is associated with the recipients' demands whereas the priority x_i offers the CP the possibility of influencing the scheduling decisions. This so-called priority scheme can be used for the following three purposes.

(a) Newly added high priority contents can be broadcasted more immediately than low priority contents. This can be achieved by initializing the instant of time when the item was last recently broadcasted corresponding to the priority class. Table 11.1 shows an example of three different classes.

Table 11.1 Initialization of R(i)

Priority class	Initial value R(i)
1	R(i) = -1
2	$R(i) = Q - \Delta t; Q < \Delta t < (Q+1)$
3	R(i) = Q

(b) The frequency of a high priority content can be increased according to (11.8). This results in a smaller spacing and a shorter access time for high priority content. A minimum amount of fairness is guaranteed because low priority content is not excluded from being broadcasted. This approach leads to a worse performance regarding the overall mean access time of all content but it decreases the mean access time for high priority items at the cost of the access time of low priority items.

$$W(i) = \left(Q - R_i\right)^2 \frac{p_i}{l_i x_i} \quad \forall i, 1 \le i \le M$$
(11.8)

(c) A modification of the *Online algorithm* is based on a *Packet Fair Queuing* approach and on the calculation of the optimal spacing between two consecutive instances of a content. It can be further extended to guarantee a defined Mean Access Time t_i per priority as shown in (11.9). The carousel is filled with high priority items and only with those of lower priority if there is still enough space/time for keeping within the defined time boundary.

$$t_{i} = \frac{1}{2}s_{i} = \frac{1}{2} \left(\sum_{j=1}^{M} \sqrt{p_{j}l_{j}}\right) \sqrt{\frac{l_{i}}{p_{i}}}$$
(11.9)

11.3.2.4 Consideration of Congestion

The presented Online algorithm can be modified in order to react to congestion. The approach is based on the provision of persistent explicit feedback information. We assume that the carousel scheduler receives the level of congestion in the format of a single parameter per service and hence per content: $0 \le c_i \le 1$, $c_i = 1$ denoting maximum congestion, whereas $c_i = 0$ implies that congestion does not occur and the full "basic" bandwidth is available. A congestion of $c_i = 0.5$ implicates that only half of the bandwidth is available and that it takes twice as long to deliver the same content.

Hence, the transmission time l_i (measured in abstract time units [7]) can be substituted as shown in (11.10) in order to take congestion into account. Parameter a_i denotes the amount of data and *b* refers to the basic bandwidth, *z* is used as conversation coefficient. This assumption is feasible since the basic bandwidth is supposed to be constant. For MBMS carousel service delivery the UMTS QoS traffic class *background* is used.

$$l_{i}\left[\text{time units}\right] = \frac{a_{i}\left[\text{Bits}\right]}{b\left[\frac{\text{Bits}}{s}\right] \cdot \left(1 - c_{i}\left[\right]\right)} z\left[\frac{\text{time units}}{s}\right]$$
(11.10)

This equation can be used in two ways, namely the adaptation of the frequency and the restriction to high priority content in case of congestion.

(a) The dynamic adaptation of the frequency can be achieved by modifying the square root rule as given in (11.11). This results in a more frequent transmission of content which do not have to be sent on congested links. Since the basic bandwidth b and the conversion factor z remain constant independent from the content, they can be omitted.

$$f_i \propto \sqrt{\frac{p_i}{l_i}} = \sqrt{\frac{p_i b(1-c_i)}{a_i z}} \implies f_i \propto \sqrt{\frac{p_i (1-c_i)}{l_i}}$$
(11.11)

(b) Regarding the restriction to high priority content, the level of congestion can also be mapped to the priorities, i.e. the level of congestion directly determines which content is allowed to be broadcasted. This strategy results in non fair scheduling decisions. Table 11.2 presents a possible mapping of congestion and priorities.

 Table 11.2 Mapping of congestion and priority

Level of congestion	Allowed priority classes
$0 \le c \le 1/3$	1, 2, 3
1/3 < c < 2/3	1, 2
$2/3 \le c \le 1$	1

A second approach tries to restrict the broadcast cycle to a total maximum. The time needed for broadcasting the entire cycle is t_{cyc} , t_{cyc_max} denotes the defined limit. First, the algorithm takes only high priority (class 1) content into account and determines the resulting cycle time. If there is space left in the cycle, it adds the contents of the next lower priority and continues until either the maximum cycle time is exceeded or content of all specified priorities x=1. x_{max} are included into the set *S*. Only elements of set *S* are considered for being broadcasted.

$$S = \left\{ i \mid x_i \le x_{\max}, 1 \le i \le M \right\}$$
(11.12)

$$t_{cyc} = \sum_{i \in S} l_i f_i = \alpha \sum_{i \in S} \sqrt{p_i l_i} = \alpha \sum_{i \in S} p_i \frac{a_i}{b(1 - c_i)}$$
(11.13)

11.3.3 Scheduling of Streaming Services

The most important requirement for Streaming Services is the provision of a continuous data stream without major disruption. The aim of the proposed centralized Streaming Service Scheduler is to optimally utilize the available radio resources in order to maximize the efficiency, while avoiding congestion. It will maximize the number of supplied users and/or the amount of entirely delivered contents.

The interactivity depicted in Fig. 11.3 also applies to Streaming Services. The CP sends its request for publication of a new content for an already existing service to the BM-SC where this request is stored in a queue. The scheduler is designed to arrange an optimized order of content distribution. It considers the spatial distribution of the recipients amongst various areas of the cellular network and is able to take RRM feedback into account. Moreover, it tries to estimate the potential benefit of each content provision by weighting.

11.3.3.1 Concept Overview

The entire scheduling algorithm is based on a reservation of RAN resources per target area and hence covers two domains, namely the temporal domain (discretized into time slices) and the spatial domain (discretized into target areas). The algorithm comprises three basic phases performed at the beginning of each time slice.

- 1. Read the requests out of the queue and calculate the weighting factor for each request on the basis of the service and content properties
- 2. Sort the requests by their weighting factor
- 3. Process the requests in the list from beginning to the end; For each request:
 - Check availability of RAN resources (stored inside the *Resource Matrix*)
 - If available: Reserve RAN resources by reducing the capacity of the target areas for the whole transmission period. Write Content-ID into the *Delivery Matrix*.

In parallel, another thread reads the values from the Delivery Matrix and initializes the delivery of the content. The required duration for content delivery as well as its required bandwidth is assumed to be known.

11.3.3.2 Weighting

The requests in the queue are weighted according to these properties:

- *t_{arr,i}*: arrival time of request, per content
- *t*_{start,i}: earliest time for delivery, per content
- $t_{end,i}$: latest time for delivery, per content
- *l_i*: required time duration for delivery, per content
- b_i : bandwidth requirement, per content
- x_i : priority, assigned by CP, per content
- *S_j*: number of subscribers, per associated service
- C_j : number of target areas, per associated service

The overall weighting factor W(i) is then calculated as shown in (11.14) below. Q denotes the current time. Consequently, content is preferred if (1) it is highly prioritized by the CP, (2) has been in the queue for a long time, (3) there is not much time left until it has to be delivered, (4) its delivery does not take much time, (5) the bandwidth requirements are low, (6) the quantity of recipients is high and (7) if there is only a small amount of target cells, i.e. if the recipients are concentrated in an area of the topology.

$$W(i) = x_i \frac{Q - t_{arr,i}}{t_{end,i} - l_i - Q} \cdot \frac{1}{l_i b_i} \cdot \frac{S_j}{C_j}$$
(11.14)

11.3.3.3 Consideration of Spatial User Distribution and RAN Resources

A *Resource Matrix* is introduced for resource monitoring and reservation. It contains a row per target area. A target area is defined to comprise several cells in order to support macro diversity mechanisms. Each column of the matrix corresponds to a time slice. Every value of the matrix represents the available capacity inside a target area for the provision of MBMS streaming services, i.e. it represents the available bandwidth for MBMS service provision of the most utilized cell inside that area.

The matrix is periodically refreshed by the scheduler taking RRM feedback and previously reserved resources into account. Figure 11.4 depicts the Resource Matrix and the Delivery Matrix graphically. The capacity available for provision of MBMS resources changes dynamically (light grey) due to non MBMS-traffic. The other bars represent resource reservation for the delivery of content (one content per pattern).

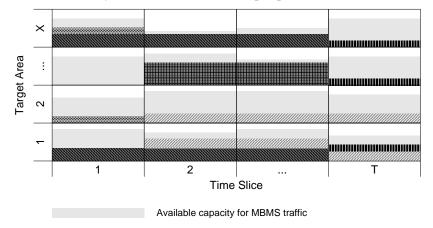


Fig. 11.4 Resource matrix

11.3.3.4 Splitting of Reception Groups

In order to achieve a higher RAN utilization the reception group can be split, i.e. the scheduler does not wait until resources in all target areas are available but supplies the available areas first and the remaining ones later. The number of splits can be adjusted by setting a *split threshold* value. A delivery session is split if the ratio of *not available* areas to *available areas* is below the defined limit.

Regarding the CN this splitting results in multicast instead of unicast transmissions. But it still remains ptm in the RAN. The advantage is the increase of supplied recipients. If simultaneous delivery into all areas is intended splitting can be switched off.

11.4 Performance Evaluation

The performance of the proposed scheduling techniques is evaluated by discrete event simulation. The results are presented in the following subsections.

11.4.1 Evaluation of Scheduling for Carousel Services

In order to analyze the performance of the carousel scheduling algorithm, an abstract simulation model was developed. The fact that it consists of three network nodes (Content Provider, BM-SC and Recipients) makes it reusable for other broadcast/multicast systems as well, but without losing accuracy. The demand probability per content was modelled according to the Zapf distribution with various access skew coefficients [7]. The delivery time and the delivery priority were chosen to be equally distributed. A minimum of 1000 simulation runs were carried out per data point. The obtained results prove the efficiency of RAN feedback mechanisms. The scheduler manages to maintain a defined Mean Access Time per priority class. High priority items (class 1) are preferred to low priority items (class 3), so their frequency is adapted at the expense of the remaining low level services. The temporal congestion variation results in the exemplary Mean Access Times illustrated in Fig. 11.5. The maximum per priority class is given in Table 11.3. Hence, the RAN feedback allows for an optimisation of the Mean Access Time. For further results, the interested reader may refer to [9].

Table 11.3 Defined thresholds

Priority class	Allowed priority classes	
1	$t_{\text{limit}}(1) = 700 \text{s}$	
2	$t_{\text{limit}}(2) = 900 \text{s}$	
3	$t_{\text{limit}}(3) = \text{undef.}$	

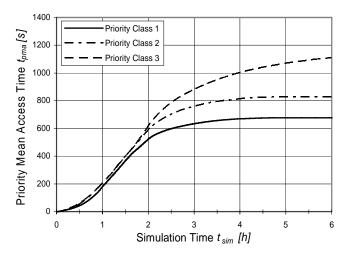


Fig. 11.5 Resulting mean access time per priority

11.4.2 Evaluation of Scheduling for Streaming Services

The *discard rate* and *number of supplied users* are chosen as efficiency metrics of the algorithm. The discard rate represents the ratio of completely distributed contents (in all target cell areas) to the total number of content distribution requests sent by the CPs. The number of supplied users is determined by adding up the number of users in a supplied target area. This distinction is important when the splitting of user groups is taken into account.

Figure 11.6 shows the dependence of discard rate with standard deviation. The deviation is used to assign the requests to a specific service and hence to a specific target area. Therefore, the lower the standard deviation chosen the more contents are delivered into the same area. Obviously, the algorithm version with both sorting and splitting performs best in case of a larger deviation. The number of supplied users is the chosen exemplary scenario is given in Fig. 11.7. The figure reveals that sorting the requests influences the performance most.

11.4.3 Evaluation of Dynamic MBMS Resource Scheduler

The MBMS Resource Scheduler is divided into two major work items. The first investigates how to incorporate the lightweight feedback mechanism. The second focuses on the dynamic power scheduler. The current status of simulation is restricted to analyzing the user feedback.

A multicast group of users (consisting of 15 UEs) and 15 Unicast UEs were placed in a cell moving randomly in a rectangular grid of $(300 \text{ m})^2$, $(200 \text{ m})^2$ and $(100 \text{ m})^2$ at 3 km/h. The difference in E_b/N_0 experienced by the UEs as compared to the randomly chosen UE is illustrated in Fig. 11.8. It can be seen that averaging the E_b/N_0 over 3 random UEs does not significantly improve the E_b/N_0 difference. The estimation of the E_b/N_0 is similar for all three analyzed mobility matrices. Hence, the feedback will be taken from one UE chosen at random. Future work will continue to investigate which UEs are most suitable for providing the feedback. The performance of the proposed DMRS concept will be modelled and simulated.

11.5 Conclusions

This chapter presented some approaches and scheduling techniques for mobile broadcast and multicast services. They were primarily designed for subscription based systems like the multicast mode of MBMS. Scheduling will be one of the key components of service provisioning systems as the quantity of multicast and broadcast content starts to increase. It enables cellular communication networks with means for efficient and affordable mass distribution of content.

The proposed algorithms were proven to control and reduce congestion in the Radio Access Network while ensuring a defined Quality of Service. The given simulation results imply the positive benefit of deployment within a real system.

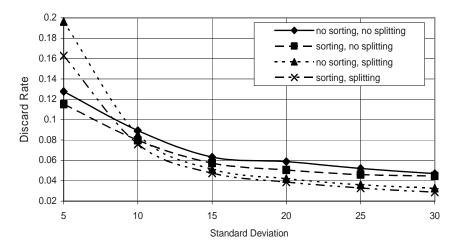


Fig. 11.6 Discard rate in dependence of standard deviation

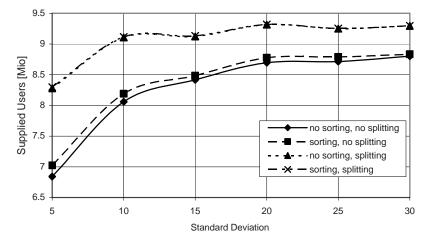


Fig. 11.7 Number of supplied users in dependence of standard deviation

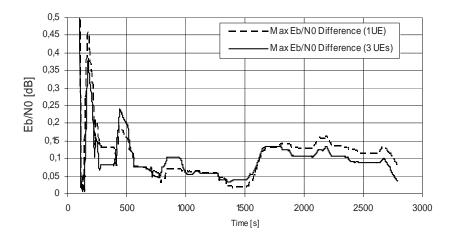


Fig. 11.8 Average Difference in Eb/N_0 in an area of 200 m² from 1 and 3 randomly chosen UEs

Acknowledgment

This work was supported by the European IST project C-MOBILE which is investigating MBMS enhancements for systems beyond 3G. The discrete event simulator *OPNET Modeler* was provided by OPNET Technologies, Inc.

References

- 1. 3GPP (2007) TS 23.246 Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description (Release 6). V6.12.0, June 2007
- Parkvall S, Englund E, Lundevall M, Torsner J (2006) Evolving 3G mobile systems: broadband and broadcast service in WCDMA. IEEE Communications Magazine, vol 44, Issue 2, pp. 30–36
- Rummler R, Aghyami AH, Yun Won Chung (2006) A multicast mechanism for UMTS. IEEE Wireless Communications, vol 13, Issue 4, pp. 94–103
- 4. Tanenbaum AS (2003) Computer networks. ed 4, Pearson Education International, Upper Saddle River, NJ
- Cui-Qing Yang, Reddy AVS (1995) A taxonomy for congestion control algorithms in packet switching networks. IEEE Network, vol 9, Issue 4, pp. 34–45
- 6. Camarillo G, Garcia-Martin MA (2006) The 3G IP multimedia subsystem (IMS): merging the Internet and the cellular worlds. ed 2, J. Wiley, Chichester

- Vaidya NH, Hameed S (1996) Data Broadcast Scheduling: On-line and Offline Algorithms. Technical Report 96-017, Texas A&M University, Texas, USA
- Sampath A, Kumar PS, Holtzman (1995) Power Control and Resource Management for a Multimedia CDMA Wireless System, IEEE Personal Indoor Mobile Radio Communication Conference, pp. 21–25
- Knappmeyer M, Tönjes R (2007) Adaptive Data Scheduling for Mobile Broadcast Carousel Services, 65th IEEE Vehicular Technology Conference, pp. 1011–1015

Part IV Networks

12 Body Area Network and Its Standardization at IEEE 802.15.BAN

Huan-Bang Li and Ryuji Kohno

12.1 Introduction

Wireless communication network technologies have fundamental importance in supporting modern ubiquitous networks to connect a person at anywhere, at anytime, and with anybody. Some well known wireless networks include intelligent mobile telecommunication (IMT) networks, wireless metropolitan area networks (WMAN), wireless local area networks (WLAN), wireless personal area networks (WPAN), etc. As wireless communication and networking approaches to personal areas, WPAN plays a more and more important role in order to meet the different technical requirements compared to other traditional wireless technologies. As a result, a number of new wireless technologies were developed for supporting WPAN. Some examples of these technologies are Bluetooth, ZigBee, and Ultra-wideband (UWB) [1,2,3,4]. WPAN is considered to be able to greatly increase the quality and efficiency in office and home automation, wireless audio/video and data streaming or exchanging, as well as in wireless sensor networks.

Various communication and network technologies have been used in supporting medical and healthcare services in past years. One example is the computer/network aided clinical record management system. This system provides a common platform among diagnosing, nursing, and dosing. Thus, it can greatly improve the total management efficiency in a hospital. As the aging population increases, setup of effective healthcare system as well as solution for shortage of working power become important issues in more and more countries. Communication and network technologies are expected to provide more active roles in supporting medical and healthcare services so as to mitigate aging population related problems. Recently, there are strong demands on introducing body area network (BAN) technology from various parties such as medical and healthcare societies as well as information and communications technology (ICT) industries. Actually, cooperation among medical and healthcare societies and ICT industries is of central importance in promoting medical ICT (MICT). Currently, there are intensive activities world wide on applying ICT in a more active and direct way to support medical and healthcare services. Continua Health Alliance was formed as a global tycoon in June 2006 [5] to accelerate the R&D on MICT. Similar regional efforts can be seen from the MAGNET Beyond project [6] of European Union and from the MICT Consortium of Japan [7].

IEEE 802 is an international standardization committee that has been working on various wireless technology standards. Many famous wireless technology standards, such as 802.11 WLAN, 802.15.1 Bluetooth, etc., were developed from this committee. As partly to response to the strong interest on BAN, a study group referred to as IEEE 802.15.BAN (SG-BAN) was set up within IEEE 802.15 [8]. The latter is the working group 15 (WG15) for WPAN in IEEE 802. The main mission of SG-BAN is to conduct preliminary studies towards establishing an IEEE standard of BAN for supporting medical and healthcare services, though some nonmedical applications can also be benefitted or supported by BAN. This chapter summarizes the activities of this SG-BAN so as to give a general guidance of BAN. Main issues and current discussion are overviewed. Moreover, a prototype BAN developed is illustrated.

12.2 SG-BAN and BAN Definition

Figure 12.1 shows the status of IEEE 802.15 by September 2007. A number of WPAN standards including 802.15.1 and 802.15.4, which defines the physical layer (PHY) for Bluetooth and ZigBee respectively, had been developed from this working group. The most recent standard developed from WG15 is 802.15.4a, which was approved in March 2007 and mainly uses UWB technology to provide an alternative PHY for 802.15.4 for low data rate applications.

BAN was selected from a list of candidate technologies considered for standardization [9]. The candidate technologies includes

- Extreme low power radios and Energy efficient data transfers
- Use of MIMO for WPAN
- Software defined radios
- Multi-band chirp spread spectrum

- Near field communication (NFC) and RFID
- Multi-packet / Contention-Free MAC / PHY

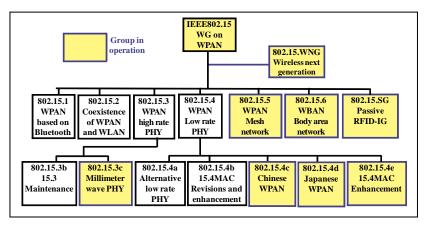


Fig. 12.1 Current status of IEEE 802.15 working group (by September 2007)

- Methods to improve wireless reliability for PAN networks
- Body area network (BAN)
- New wireless network architectures
- Mobile, Nomadic Mesh-networks

BAN was first started as an interest group (IG-BAN) at Jacksonville meeting, FL, USA, in May 2006. It was formally approved as a study group by the IEEE 802 executive committee (EC) in November 2006, because of the increasing interests of the group members.

What is BAN? What is the difference between BAN and other existed networks? To answer these questions, the definition and scope of BAN is summarized as follows [10]. BAN will provide short range, low power and highly reliable wireless communication for use in close proximity to or inside body. Data rates, typically up to 10 Mbps, will be offered. The typical communication range of BAN is around 2 m, which can be optionally extended to 5 m. Because that BAN is operated with close connection to human body, it is considered as a technology of high potentiality to provide enhanced high quality and convenience for people in a more direct manner than other networks. BAN can be further divided into two categories depending on their operating environments. One is so-called wearable BAN, which is mainly operated on the surface or in the vicinity of body. Another is so-called implant BAN, which is operated inside body. Wearable computers can be regarded as a kind of wearable BAN, which has been proposed and studied for many years [11,12]. In comparison, implant devices

with communication ability have not been studied very well although there are some researches tackle this subject [13].

	Other 802 standards	BAN
Configuration	15.3, 15.4 MAC	Single scalable MAC with reliable delivery
Power consumption	Low power consumption	Extremely low power while communicating to protect human tissue
Power source	Conventional power source	Compatible with body energy scavenge operation
Requirements (QoS)	Low latency	Guaranteed and reliable response to external stimuli
Frequency band	ISM	Regulatory and/or medical authorities approved bands for in and around human body
Channel	Air	Air, vicinity of human body, inside human body
Safety for human body	None	Required (e.g. SAR)

Table 12.1 Comparison between BAN and other IEEE 802.15 standards

It is obvious from the above description that the main focus of BAN is laid on medical and healthcare related applications. However, some members are looking at non-medical applications with BAN such as consumer electronics (CE) including entertainments. On one hand, there is a possibility that both medical and non-medical applications are able to be supported by a same PHY. On the other hand, requirements on MAC, security, and QoS for medical and non-medical applications are different for these two types of applications. Therefore, attention must be paid carefully to deal with these two types of applications. After intensive discussion on the project authorization requirement (PAR) and five criteria (5C), which are the two required documents to define the group for forming a task group, SG-BAN summarized its uniqueness as shown in Table 12.1 [14]. As can be predicted, because that BAN operates in vicinity, on, or inside human body, it will have much different channel models compared to other IEEE standards. The body will show strong effect on channel models for both wearable and implant BAN. Generally, channel model for wearable BAN should be basically of multipath characteristics, while channel models for implant BAN will show heavy attenuation characteristics. Due to the operating environment of BAN, safety to human body has a higher priority than the other wireless systems. Parameters like specific absorption rate (SAR) need to be taken into consideration to protect human tissues. Because of the short communication range and requirement of protecting body tissues, extremely low power radios are potential PHY candidates for BAN. Moreover, when operated inside human body, scavenge power supply is desired to support long term operation. There are also high QoS and low latency requirements for vital data related transmissions. These issues should be simultaneously considered for PHY and MAC design.

Finally, as the authors prepare the final version of this manuscript, the PAR and 5C of BAN-SG has passed a three-level approval: by EC, by New Standards Committee (NesCom), and by Standard Board, by the end of 2007. SG-BAN becomes a new task group referred to as 802.15.6. With this approval, standardization process of BAN was formally started.

12.3 BAN Applications and Usage Models

Like other wireless communication networks, the basic function of BAN is to provide effective and reliable networking for transmission of data, voice, or picture through wireless communication links. BAN applications can be classified with different categorizations. Here, we divide BAN applications into the following three categories [15]. (1) Medical and healthcare applications, (2) Applications for assisting persons with disabilities, and (3) Entertainment applications.

12.3.1 Medical and Healthcare Applications

BAN can be used to provide assistance to automatic medical treatment, automatic dosing, and vital signal monitoring. Figure 12.2 shows an intuitive view of automatic medical treatment/dosing process. The automatic medical treatment/dosing process. The automatic medical treatment/dosing can be described as a closed loop control process. There are three steps in the process. At the first step, various vital and healthcare data are collected using various medical sensors attached to a person. These collected data are forwarded to a command unit automatically. At the second step, the command unit decides the corresponding treatment method or correct dosing based on the received vital and healthcare data. Based on the decision it made, the command unit generates corresponding commands and sends these commands to the action unit. At the third step, the action unit conducts the treatment or dosing to objectives according to the commands it received. When the treatment or dosing is finished, sensors will collect updated vital and healthcare data on a preset time interval and the closed loop control enters a new circulation.

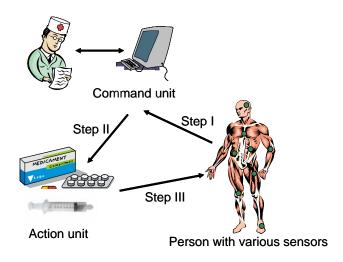


Fig. 12.2 Closed loop control for automatic medical treatment or automatic dosing

The concept of the closed loop control shown in Fig. 12.2 corresponds to many real applications in assisting medical treatment or dosing. For example, pacemaker is an electronic device that helps person's heart to beat regularly. To adjust the heart beat to a correct rhythm, signals that decide the correct heartbeat need to be extracted from a sympathetic nerve and other related nerves. Commands are generated according to the collected nerve signals and sent to a pacemaker controller. The pacemaker controller adjust heart beat according to the commands. The above process forms a closed loop control among detectors of nerve signals, a pacemaker controller, and a pacemaker. They play the roles of the sensor, the command unit, and the action unit, respectively. Another example is automatic insulin injection for diabetes patients. By detecting glucose in the blood, corresponding amount of insulin can be the insulin pump to conduct the injection. Therefore, a closed loop control is formed among glucose detector, insulin controller, and insulin pump, which respectively act as the sensor, the command unit, and the action unit.

The basic function of BAN is to collect and transmit data gathered by various sensors. In the above two examples, sensors, command unit, and action unit are laid on or put inside body. Generally, BAN can be used for transmission of vital and healthcare data from sensors to command unit as well as for forwarding treatment or dosing commands from command unit to action unit. Some typical vital and healthcare data and their required data rates are summarized in Table 12.2. It can be seen that except for

electro cardio gram (ECG) and electro encephalo graphy (EEG), the other data including heartbeat, blood pressure, body temperature, and glucose can be transmitted at a data rate smaller than 100 bps. For ECG and EEG,

Vital and healthcare data	Data rate (bps)
Heartbeat	< 100
Blood pressure (BP)	< 100
Electro cardio gram (ECG)	~ 2500
Electro encephalo graphy (EEG)	~ 540
Body temperature	< 100
Glucose	< 100

Table 12.2 Examples of vital and health data

the required data rate may be higher depending on the number of simultaneously operated sensors. There are different usage models for wearable BAN and implant BAN. A usage model for wearable BAN is vital and healthcare data collection and monitoring for inpatients in a hospital. BAN nodes, which are BAN radio transceivers with different built-in sensors, are distributed and attached to a group of inpatients. Built-in sensors of BAN nodes may be different depending on functions or purposes. BAN nods collect various vital and healthcare data of inpatients. Then, these data are transmitted to a nurse center in real time. This provides a possibility for nurses to get those simple or regular data automatically without going to patient and performing measurement one by one. It is obvious that by performing with BAN the working load of nurses can be reduced. Thus, operation efficiency of a hospital can be increased, which is helpful to reduce cost as an additional outcome. In addition, BAN can also be implemented with ID-TAG like function, which is desired for reducing the risk of unintentional wrong treatment or wrong medicine.

12.3.2 Applications for Assisting Persons with Disabilities

The second category of BAN applications is assisting persons with disabilities. There are different fields to adopt BAN. Figure 12.3 shows an example of using BAN to provide assistance to a person with visual disability. Small cameras or video cameras are attached to the sunglasses of the person. These cameras are used to detect objectives, stairs, vacant seat in a train, and so on. Radars are also attached to the stick worn by the person or on the body of the person. The radars provide ranging and position ability to locate an objective. Pictures captured by cameras and ranging/positioning data gathered by radars are sent through BAN to a portable signal processor carried by the person himself in real time. The signal processor simultaneously interprets those pictures and ranging/positioning

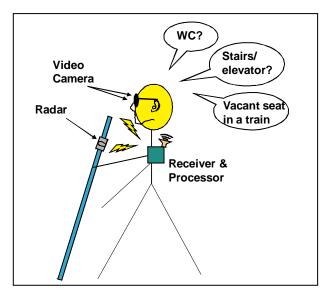


Fig. 12.3 Assistance to people with visual disability

data into voices. These information and data can be used by the person for objective finding, route guidance, and so on. Because pictures or videos are dealt, the required data rate may be up to several Mbps.

Other examples of this category of applications include muscle tension sensing and stimulation as well as fall down detection. For a person in rehabilitation with hurt on legs, his legs may not strong enough to support him. By attaching muscle tension sensors on the related position, data of muscle tension can be gathered. Based on the data, decision can be made if an artificial supporting implement should be triggered up during the rehabilitation process. Fall down detection is an important function in assisting a person with walking disability or an elder person. Combined with other vital or health sensors, necessary protection measures can be triggered up.

In a further deliberate design, BAN may also be expected to be used for assisting speech-impaired person. Using sensors attached to fingers and hands that can detect the movement of finger and hands as well as relative positions between fingers and hands, BAN can collect these data and send them to a signal processor in real time, where finger language is interpreted to vocal language.

12.3.3 Entertainment Applications

Because of the body centric characteristics, BAN is capable of providing support for entertainment and gaming applications. However, we'll skip this part as it is not within the scope of this chapter.

Before concluding this section of discussion on BAN applications, let's draw a conceptual block diagram of BAN in supporting medical and healthcare services in Fig. 12.4. A single BAN consists of a BAN server, a number of sensor nodes, and several relay nodes. BAN server act as the coordinator of BAN to perform channelization and other network control. BAN server aggregates data collected by all sensor nodes in the network. Then, it sends these data to a medical/healthcare server through existing wireless networks such as WPAN, WLAN, mobile phone network, etc.. Each BAN can use different infrastructure depending on its own possibility. Because of the sensitivity of the BAN data, high security measures are required. Only permitted medical experts or healthcare advisers can refer to the related data by accessing the medical/healthcare server. For sensitive vital data in some situations, BAN server is required to be able to generate warning signs for unusual vital data. In most cases, sensor nodes should communicate with BAN server directly. In case that direct link between sensor nodes and BAN server is not available, relay nodes are used to establish connection between BAN server and sensor nodes. It should be noted that relay function can be implemented within sensor nodes. That is, no independent relay nodes are needed although we illustrate relay nodes as independent ones for the purpose of convenience in Fig. 12.4.

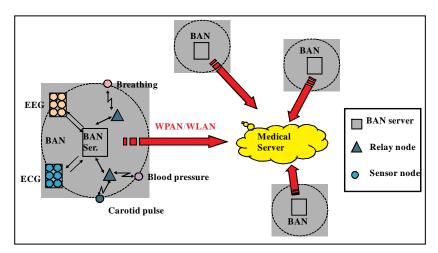


Fig. 12.4 Conceptual block diagram of BAN

12.4 Some Short-Range Technologies and a Prototype BAN

12.4.1 BAN Requirements and Some Short-Range Wireless Technologies

A principle requirement on BAN is low emission power due to the following two reasons. First, because BAN is mainly carried by persons, the emission power must be restricted to a level that it is not harmful to human tissue or organs. International commission on non-ionizing radiation protection (ICNIRP) is an organization that focuses on radiation protection safety levels [16]. Guidelines issued by ICNIRP, such as SAR, should be referred to at a maximum scope. A second reason for low emission power is because of the coexistence requirement between BAN and other wireless systems. Low emission level means that interference level of BAN will be in a reasonable low level so that it is tolerable by other wireless systems. This is also an important issue to be addressed by the upcoming standard.

Another important requirement on BAN is reliability. Because a large number of applications deal with medical and healthcare data, in some cases, errors in these data may cause serious results. Robust data transmission and secure networking need to be designed carefully, so that the probability of error occurrence is controlled at a minimum level. Other requirements on BAN include issues such as low cost, simplicity, high security, large scalability, low power consumption, small formation size, etc.

From the requirement of low emission power, it is obvious that possible candidates in implementing BAN should be short range wireless communication technologies. In the following, we look at some short range wireless technologies from a point of view on their suitability for BAN.

12.4.1.1 Specific Low Level Radio

Specific low level radio operates at 430 MHz frequency band and can be used without license. It is also favorable in the sense that it has low emission power as well as low power consumption. The drawback of this technology is that the available data rate is only 2400 bps, which can not meet the demand of some BAN applications.

12.4.1.2 Zigbee Using IEEE 802.15.4

IEEE 802.15.4 can operate at three different frequency bands of 2.4 GHz (worldwide), 800 MHz band (regional), and 900 MHz band (regional). When operating at 2.4 GHz band, it achieves a maximum data rate of 250 kbps. The drawback of IEEE 802.15.4 is also the limited data rate.

Moreover, interference may be an issue when operating at 2.4 GHz band, because several wireless systems exist at the 2.4 GHz frequency band.

12.4.1.3 Bluetooth Using IEEE 802.15.1

Bluetooth also operates at 2.4 GHz band and can provide data rates of more than 1 Mbps. One disadvantage of Bluetooth is that it has biggest emission power and power consumption among the technologies discussed here. Moreover, the number of simultaneously operated piconets (SOP) can be supported by Bluetooth is smaller or equal to seven. This will restrict some BAN applications.

12.4.1.4 Low Rate UWB of IEEE 802.15.4a

IEEE 802.15.4a is a new WPAN standard, which is approved by IEEE Standard Review Committee (RevCom) in March 2007. When UWB band (3.1–10.6 GHz) is used, IEEE 802.15.4a has the smallest emission power density and power consumption among the technologies discussed here. The nominal data rate of IEEE 802.15.4a is 850 kbps, which can be increased to as high as 26 Mbps in option. A drawback of UWB band is that the transmission loss is so significant by human body that it is difficult to be used for implantable BAN.

12.4.2 A Prototype BAN System

A prototype system of wearable BAN based on UWB is developed. UWB is used because of its extremely low power density. Figure 12.5 shows the transceiver structure of the prototype system. Each BAN transceiver operates at a center frequency of 4.1 GHz and the UWB signal occupies a bandwidth of 1.2 GHz. Average emission power density of UWB transmitters is -45 dBm/MHz, which is 3.7 dB lower than -41.3 dBm/MHz. The latter is the power density function (pdf) mask defined by FCC of USA and later adopted in most UWB regulations world wide. To simplify the

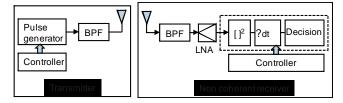


Fig. 12.5 A prototype UWB transceiver for wearable BAN

implementation, an on-off keying (OOK) modulation is adopted at the transmitter and an energy detector is adopted at the receiver. Energy detector doesn't need synchronization operation, thus it can reduce the device cost. Pulse shape generated at the transmitter is Gaussian pulse. The built-in sensors used with the UWB transceiver is three dimensional accelerate sensors. The date rate of the transceiver is 66 kbps which is high enough to transmit the three dimensional accelerate data. However, there is no difficulty to raise the data rate up to more than 850 kbps.

The overall structure of the prototype BAN system is shown in Fig. 12.6. There are three types of devices, BAN coordinator, relay nodes, and sensor nodes. BAN coordinator controls the network and decides channelization for all other nodes. Sensor nodes collect obtained data and send those data to BAN coordinator. Relay node are used to make connection between sensor nodes and the BAN coordinator to pass data from a sensor node to the BAN coordinator when there is no available direct link between sensor nodes and the BAN coordinator. Although these three types of devices are assigned with different roles, they are implemented with the same UWB transceiver structure as given in Fig. 12.5. Because that sensors used in the prototype BAN are acceleration counter, which can detect three-dimensional acceleration data. By combining with local magnetic field sensors, fall detection can be conducted. Various network structures with the prototype BAN system are under investigation.

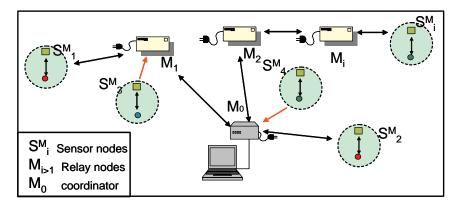


Fig. 12.6 A prototype BAN with UWB transceivers

12.5 Issues in Discussion and Future Work

12.5.1 Wearable BAN and Implant BAN

Depending on if it is operated outside or inside body, BAN can be divided into wearable BAN and implant BAN. While wearable BAN is considered for both medical and non-medical applications, implant BAN is mainly considered for medical and healthcare applications. There is an argument that if wearable BAN and implant BAN should be supported by a single PHY or they can be supported by two separated PHYs with proper intersection at a coordinator. Many researches investigate the propagation characteristics on wearable BAN and implant BAN [17, 18]. Because wearable BAN operates in vicinity or on human body while implant BAN operates inside body, wearable BAN suffers decaying during transmission. The following differences exist between wearable BAN and implant BAN.

- Different requirements on frequencies due to different operating environment and channels.
- Implant BAN is more power limited and it requires much smaller form factor compared to wearable BAN.
- Both wearable BAN and implant BAN need to consider tissue protection (e.g., SAR) while implant BAN may subject to much strong restriction.

From the above argument, it seems reasonable to reach a conclusion that two separate PHYs, that respectively focus on wearable BAN and implant BAN, should be designed. However, this should not exclude the possibility of a single PHY solution that is common for both wearable BAN and implant BAN. More investigation and study are needed before a concrete decision can be made.

It should be noted that wearable BAN and implant BAN should have some kinds of interoperability or inter networking function even they are implemented with different PHYs. This is to guarantee effective signal follow or data exchange within BAN. One way to do so is that at least wearable BAN and implant BAN should have a common BAN server. The common BAN server is capable of communicating with wearable BAN as well as implant BAN.

12.5.2 Frequency Regulations

Frequency regulations for medical applications were issued in many countries or regions. Figure 12.7 summarizes the available frequency bands for wireless medical telemetry system (WMTS) and for medical implant communication system (MICS) in EU, Japan, and USA. The definition of WMTS is as follows.

The measured physiological parameters and other patient related information are transmitted via RF communication between a patient-worn transmitter and a remote monitoring unit.

It is clear that WMTS is assigned with different frequency bands in Japan and USA. In comparison, frequency bands for MICS in most countries or regions are assigned with 402–405 MHz. There is also expectation to extend the MICS frequency band to 401–406 MHz in some countries and regions. A common problem for WMTS and MICS is that bandwidth of a single channel is usually narrow in current regulations. That limits high data rate applications.

Besides the above WMTS and MICS bands, there are other potential candidates to be considered as frequency bands for BAN applications. These candidates include industrial, scientific, medical (ISM) band and ultra-wideband (UWB) frequency band. ISM band at 2.4 GHz is available world wide. However, there are many wireless systems that already operate at ISM band such as WLAN of IEEE 802.11b, Bluetooth of 802.15.1, and Zigbee of IEEE 802.15.4. Coexistence among different wireless systems needs to be carefully considered. For UWB bands, an IEEE standard of 802.15.4a was formally approved in March 2007, which is designed for low data rate and high precision ranging applications. It also should be noted that UWB regulations are different among countries or regions. Selection of frequency band for BAN will affect the PHY design. Intensive investigation and study must be carried out before give a decision.

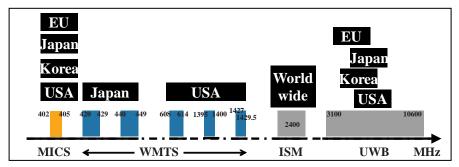


Fig. 12.7 Frequency regulation for WMTs and MICS

12.6 Conclusion

BAN is expected to be a much useful technology in supporting medical and healthcare services, disability assistance, as well as consumer electronics. In this chapter, some inspiring fields that BAN can play important roles are overviewed with emphasis being laid on medical and healthcare with specific usage models. Generally, BAN can be divided into wearable BAN and implant BAN. Due to different operating environments and characteristics, wearable BAN and implant BAN may require different technologies for implementation, while a single solution for both is desired. More studies and evaluations are underground before giving a concrete conclusion.

As a general review, several short range communication technologies are compared for their suitability in supporting BAN. A prototype BAN using UWB is developed and described. The main reason of using UWB is the low emission level and potential ability of high data rate. With this prototype system, the authors intend to give an intuitive concept of BAN structure. The standardization activities of IEEE 802.15.BAN group were introduced. Main progress made and main issues discussed in this group were reviewed.

References

- D. Rosener, "Introduction to bluetooth engineering," John Wiley & Sons Inc., 2007.
- F. Eady, "Hands-on zigbee: Implementing 802.15.4 with microcontrollers," Newnes, 2007.
- 3. K. Siwiak and D. McKeown, "Ultra-wideband radio technology, "John Wiley & Sons, Ltd., 2004.
- 4. M. Ghavami, L. Michael, and R. Kohno, "Ultra wideband signals & systems in communication engineering, "John Wiley & Sons, Ltd., 2004.
- 5. http://www.continuaalliance.org/home.
- 6. http://www.ist-magnet.org/.
- 7. http://www2.nict.go.jp/pub/whatsnew/press/h18/070305-1/070305-1.html (in Japanese).
- 8. http://grouper.ieee.org/groups/802/15/.
- 9. Erik Schylander, "15WNG Guidelines for New Work Items," IEEE 802.15-06-0002-01, Jan. 2006.
- W. Astrin, "BAN Project Authorization Request (PAR) draft," IEEE 802.15-07-0575-06-0BAN, July 2007.
- T. Martin, E. Jovanov, D. Raskovic, "Issues in wearable computing for medical monitoring applications: a case study of a wearable ECG monitoring device," Conference Records, The Fourth International Symposium on Wearable Computers, pp. 43-49, Atlanta USA, Oct. 2000.

- 12. G. Kortuem, Z. Segall, "Wearable communities: augmenting social networks with wearable computers," IEEE Pervasive, vol. 2, no. 1, pp. 71–78, Jan.–March 2003.
- F. Graichen, R. Arnold, A. Rohlmann, and G. Bergmann, "Implantable 9-channel telemetry system for in vivo load measurements with orthopedic implants," IEEE Trans. On Biomedical Engineering, vol. 54, no. 2, pp. 253–261, Feb. 2007.
- 14. W. Astrin, H.-B. Li, and R. Kohno, "SG BAN project draft 5C," IEEE 802 15-06-0488/r7, July, 2007.
- H.-B. Li, K. Takizawa, B. Zhen, K.Y. Yazdandoost, S. Hara, and R. Kohno, "Response to IG-BAN's Call For Applications," IEEE 802.15-06-0241-00-0ban, May 2006.
- 16. http://www.icnirp.de/
- Y. Hao, A. Alomainy, Y. Zhao, C. G. Parini, Y. Nechayev, P. Hall, and C. C. Constantinou, "Statistical and deterministic modellingof radio propagation channels in WBAN at 2.45 GHz," Proc. IEEE Antennas Propag. Soc. Int. Symp., pp. 2169–2172, 2006.
- J. A. Ruiz, J. Xu, and S. Shimamoto, "Propagation characteristics of intra-body communications for body area networks," Proc. CCNC, pp. 509–513, 2006.

13 Generic Abstraction of Access Performance and Resources for Multi-Radio Access Management

Joachim Sachs, Ramón Agüero, Khadija Daoud, Jens Gebert, Georgios Koudouridis, Francesco Meago, Mikael Prytz, Teemu Rinta-aho, and Haitao Tang

13.1 Introduction

The rapid evolution of wireless communications technologies increases the diversity of deployed radio access technologies (RATs). Forthcoming scenarios will be characterized by a great heterogeneity, not only from the presence of the subjacent technologies, but also considering the perspective of new business models and cooperation between networks. Taking these scenarios into consideration, it becomes a real challenge on how to manage different radio interfaces in a uniform way. Some network cooperation solutions are under standardisation: 3GPP is defining an architecture that integrates different access technologies in a common packet core network and allows inter-system handover [1]; similarly, IEEE 802.21 is currently defining a framework to support Media Independent Handover services [2]. However, these activities do not look in depth how radio resources are managed efficiently between the inter-worked systems.

This chapter focuses on multi-radio resource management aspects and presents an approach pursued within the Ambient Networks research project [3, 4, 5] that enables to deal with different radio accesses by the use of abstraction. This work extends earlier work in [6]. Using an innovative, distributed, and flexible multi-access architecture, it is demonstrated how different metrics about the subjacent resources can be generalized and illustrate furthermore how access selection algorithms may benefit from this generalization.

In this chapter Sect. 13.2 introduces a Multi-Radio Access architecture. Section 13.3 presents the service requirements that need to be fulfilled for access selection, followed by a description of a service specification in Sect. 13.4. Sect. 13.5 discusses how the Generic Link Layer provides the performance abstraction of RATs, while Sect. 13.6 maps the service requirements onto access resource abstractions. Sect. 13.7 describes how these abstractions can be used in access selection algorithm. Finally, Sect. 13.8 summarises and concludes this chapter.

13.2 Multi-Radio Access Architecture

The Multi-Radio Access (MRA) architecture [4, 5, 7] developed by the Ambient Networks project is based on two main functional entities; Multi-Radio Resource Management (MRRM) and Generic Link Layer (GLL). These components are part of the larger Ambient Networks architecture, in particular its so called Ambient Control Space [3]. Figure 13.1 shows an overview of the architecture in one ambient network including functional entities, interfaces, the user-plane data path (red lines marked as *U-Plane*) and inter-network control signalling (green lines marked as *ACS* and *ANI*).

MRRM is the key control entity in the multi-access system. It monitors the available accesses for a given user terminal (or user network) and allocates one or more of them to the on-going connections at a particular time. That is, MRRM performs access selection (which possibly includes execution of handovers between accesses). The decisions are based on different input parameters, e.g., the performance and quality of an access, the resource costs and current resource availability, the operator and user policies, the service requirements, and the terminal capability and behaviour. They can be triggered from many sources: session setup, modification, or release; detection of new access; change in performance or resource status of current access; change in preferences or policies; etc. MRRM also provides functionality for generic access advertisements and discovery, which makes the access scanning and detection processes more efficient in terms of energy and resource consumption. Note that for simplicity Fig. 13.1 shows a single MRRM entity, but in practice the functionality can be distributed among multiple MRRM entities that may take on different roles in their joint operation.

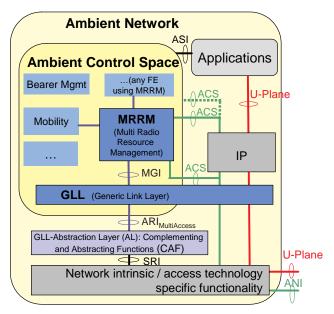


Fig. 13.1 Multi-Radio access functional architecture in the ambient networks architecture

GLL supports MRRM for the access selection and handover processes with:

- 1. A generic abstraction of the access link performance and resource status, shown in Fig. 13.1 as "GLL-Abstraction Layer (AL): Complementing and Abstracting Functions (CAF)", which interact with the network intrinsic and access technology specific functionalities;
- 2. Functionality to assist and "smoothen" inter-access handovers (for example, link layer context transfers between accesses), shown in Fig. 13.1 as "GLL" (primarily the part that stretches outside of the Ambient Control Space);
- 3. Control functionality for configuration (thresholds, triggers) and reporting of the abstracted link performance and resource status values, also part of "GLL" in Fig. 13.1 (mainly the part inside of the Ambient Control Space).

This chapter concerns the first component and its implications on the MRRM access selection algorithms. Note that the GLL functionality also can be distributed among multiple entities.

As mentioned above, MRRM monitors accesses and chooses one (or more) of them for the user terminal. But what is really monitored and what

is selected? Different accesses manage user data flows and monitor access link performance and resource status in different ways and in different nodes (e.g. access anchors, resource control servers, access points and base stations). An abstraction model of the connectivity for user data flows [5, 8] is therefore necessary, see Fig. 13.2. A full description of the connectivity abstraction model is outside the scope of this chapter (see [5, 8] for details). For the present purposes it is sufficient to know that *end-to-end flows* carry the application-level service performance requirements (of relevance for multi-access operation) and that the MRRM and GLL functionality operates on (and monitors) *access flows*. Hence, MRRM selects access flows to support end-to-end flows. To be able to compare the performance of an access flow with the service requirements the two elements should be based on the same model and descriptors. The next section will discuss the service requirements model in detail.

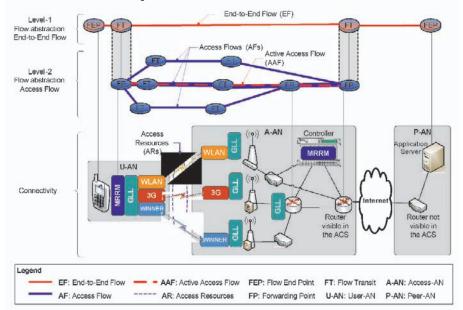


Fig. 13.2 Multi-Radio access connectivity abstraction model and architecture

13.3 Service Requirements

End-to-end flows carry application-level service requirements in the connectivity abstraction model developed in [5, 8]. The role of MRRM is then to provide a suitable access flow for an end-to-end flow, which matches the end-to-end flow requirements and thus the negotiated service needs. It is thus necessary to find a radio link performance abstraction by which the metrics can be matched into the end-to-end flow requirements; these are the three fundamental link quality related parameters: reliability, data rate, and transmission delay (latency).

13.3.1 Reliability Requirements of Applications

In general, a data stream can be corrupted when being transmitted due to transmission errors or during inter-system handovers. It is important to understand what requirements are put on reliability. Applications can be grouped into error-tolerant and error-sensitive applications.

(1) Error-tolerant Applications

Error-tolerant applications can handle a certain amount of corruption of the transmitted data. The errors can, for example, be residual bit errors within transmitted data packets or packet errors from lost data packets, normally introducing some noise. Typical examples of error-tolerant applications are speech, audio and video applications.

(2) Error-sensitive Applications

Error-sensitive applications cannot tolerate any transmission errors in the data. A single error event could cause a substantial degradation of the data quality already. For example, a large file transmitted may become invalid due to a single bit error.

It is important to note that communication protocols influence the reliability required from the transmission chain. A certain communication protocol can turn an error-tolerant application into an error-sensitive application and vice versa, e.g., layered encapsulation of data, as well as some security protocols (e.g. IPSec). As a result, error-sensitive applications constitute the most significant class of applications to be used in IP networks.

13.3.2 Rate Requirements of Applications

An application has a specific requirement on the data rate. All applications require at least a minimum rate, such that the transmission and/or session do not exceed a sensible limit.

(1) Discrete-rate Applications

These applications typically require data rates centred around one or more fixed discrete rates. Speech applications are the examples that have a more or less constant service rate; audio or video transmission can be constantrate or variable-rate encoded around different encoding rates. Higher link rates than the encoding rate are not exploited, while persistently too low link rates cannot support the service.

(2) Elastic Applications

Elastic applications have no upper rate requirement if a minimum data rate is provided. With increasing data rates the service performance improves. An example application is file transfer.

13.3.3 Delay Requirements of Applications

An application has a requirement on the tolerable transmission delay; it is always upper bound by a maximum acceptable delay, which can depend on user/application behaviour (like conversational interactions) or on system parameters (e.g. battery lifetime).

(1) Delay-sensitive Applications

The delay-sensitive applications require their datagrams being transmitted from transmitter to receiver within a certain delay bound. For example, they are interactive media sessions like (voice/video) telephony, multiplayer on-line games, and streaming applications, while the streaming application has a much more relaxed delay requirement due to a large playout-buffer at the receiver.

(2) Delay-insensitive Applications

Delay-insensitive applications (e.g. file transfer) have their delay bounds that are orders of magnitude larger than typical transmission delays. These bounds are given, for example, by higher-layer protocol timers (e.g. TCP).

13.4 Service Specification

The goal of access selection is to select the best suited access(es) for a particular service among the available access systems. It is necessary that the access selection function knows the service requirements, and that this service specification is sufficient for the access selection to judge the suitability of an access system. A service specification, as depicted in Fig. 13.3, contains the service requirements as discussed in Sect. 13.3 separately for downlink and uplink direction. In addition the service specification contains the service type and a service priority. The service type classifies the service according to its requirements in terms of reliability, delay and rate as described in Sect. 13.3. Alternative service types could be the UMTS QoS classes as defined in [9]. In some cases an access provider restricts the usage of access resources for certain service types. For example, the amount of resources that may be used for best-effort data services may be limited to 70% of the available resources while the remaining resources are exclusively reserved for conversational services like telephony. The service type is used in order to determine to what extend an access system is useable for the particular service. Moreover, a priority level describes the priority level of a user service to obtain access to the transmission resources. This priority level is typically determined in the service level agreement between the end user and the access provider. For example, some users may subscribe to a cheap best-effort telephony service with low priority, while other users subscribe to a more costly premium telephony service that guarantees preferred service. Service priority levels could, for example, be classified as "bronze", "silver" or "gold." The service priority is part of the service specification; it allows a resource management function to determine whether the requirements of a particular service can be met at a given traffic load and traffic mix.

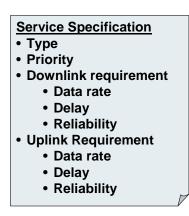


Fig. 13.3 Service specification

13.5 Generic Access Performance Abstraction

The GLL¹ derives performance metrics of an access system by abstracting access technology-specific link performance measures. The resulting metrics, when provided to MRRM, must allow comparison between different access systems considering, in addition, the characteristics and requirements of the end-to-end data flow. Consequently, they need to describe the reliability, rate and delay characteristics of the access system, resulting in the following metrics (see Fig. 13.4):

- Link rate,
- Delay,
- Residual Bit Error Rate (BER), and
- Residual Packet Error Rate (PER).

For each of these metrics different values can be derived for the expected average, the minimum and the maximum values, and, the expected variation. For link rate and delay, an instantaneous value can also be provided.

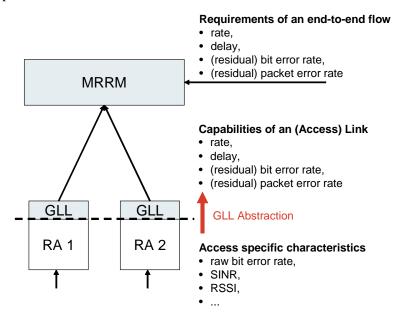


Fig. 13.4 Generic link performance abstraction

¹ Or more accurately the "GLL-Abstraction Layer (AL): Complementing and Abstracting Functions (CAF)", see Fig. 13.1.

The abstracted link quality values are derived from access specific measures, for example:

- Received Signal Strength Indicator (RSSI),
- Signal to Interference and Noise Ratio (SINR),
- Channel Quality Indicator (CQI), or
- Raw bit error rate (bit error rate before channel decoding).

An example for the approximation of the access link performance has been described in [10]: based on a modified Shannon capacity the performance of an access system is:

$$C_i = B_i \cdot \min\left\{\log_2\left(1 + \frac{SINR_i}{\Delta SINR_i}\right); \varepsilon_{\max,i}\right\}.$$

The maximum achievable data rate Ci for the communication channel of an access system i according to Shannon depends on the spectral channel bandwidth Bi and the signal-to-noise-and-interference ratio SINRi, while any realistic access system cannot achieve this theoretical limit. According to [10] the difference can be described by two access system specific parameters $\Delta SINR_i$ and $\varepsilon_{max,i}$. $\Delta SINR_i$ describes the loss and overhead of an access system compared to the Shannon capacity; $\varepsilon_{max,i}$ reflects the upper bound of the achievable data rate which is limited by the highest coding and modulation scheme that can be used.

In real systems, the link performance abstraction is limited by implementation characteristics and restrictions on which parameters are measured and at what precision. For example, a simple receiver and a complex receiver of the same access technology may be able to achieve different link performance in terms of data rate for the same received signal strength. Therefore, technology and implementation-specific methods are used to abstract the link performance. The abstraction function is performed by the GLL Abstraction Layer – Complementing and Abstraction Functions (CAF) (see Figs. 13.1 and 13.2). The mapping function needs an abstraction that meets the flow requirements. A suitable abstraction can be given at IP packet level. This means that the aforementioned capabilities of an access link will be expressed in terms of IP packet attributes including: throughput/goodput, packet delay, and packet loss.

13.6 Generic Access Resource Abstraction

13.6.1 Generic Access Resource Metrics

Different RATs have widely different mechanisms for using and sharing the available resources among users, for example, by division into time/frequency slots or codes, as well as, shared statistically using contention-based schemes. The notions of the amount of total resource occupied (the load level) and the amount of resource that a particular user session occupies are therefore quite different for different RATs. Different RATs may further be deployed with full or only spotty/hotspot coverage. For MRRM to exploit radio resource information from heterogeneous RATs in its operation (e.g., for load management), it is necessary to have a mechanism that can derive comparable measures also for the resource status.

Some notion of the scope of a particular resource is necessary in a resource status abstraction model. For wireless accesses the geographical coverage is of particular interest. Hence, the resources in the multi-access system are divided into Access Resource Areas (ARA), which can differ for different RATs. Typically an ARA corresponds to a cell area, but it could also be a smaller or larger unit. For one RAT in the multi-access system the resources in each cell area could be visible to MRRM, whereas for another RAT MRRM may only see the aggregated resources over multiple cell areas. In the latter case the aggregate of cell areas is seen as one ARA. It may also be that MRRM can see the resources in individual cell areas even if they are managed by RAT-intrinsic radio resource management functions.

MRRM is assumed to be aware of all ARAs for all RATs in the multiaccess system. MRRM needs to know which ARAs a user terminal (UT) is connected to. This is reported by the UT or GLL via abstracted link quality/link performance reports as described in Sect. 13.5.

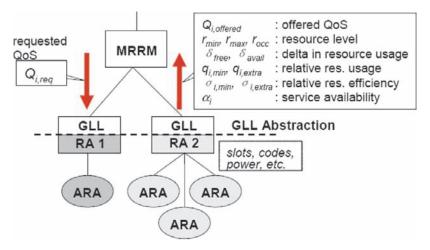


Fig. 13.5 Metrics for generic access resource abstraction

For every ARA, which provides sufficient link performance, the GLL abstracts and computes resource levels according to a relative resource level as shown in Fig. 13.6. Depending on what is the limiting resource(s) in the RAT, one can compute, for example, the relative number of (time) slots or codes, the relative amount of (downlink) power, the relative occupied bandwidth, the average collision ratio, or combinations thereof such as power and slots.

The computation is done by GLL-AL (Fig. 13.2). It can be reported to the GLL/MRRM periodically or when RAT-specific events occur; in order to reduce signalling load, the values are communicated only when certain thresholds have been reached.

The following resource levels are computed per ARA, as depicted in Fig. 13.5:

- *r_{min}* is the current minimum required amount of resources for all active users/sessions in the ARA.
- r_{occ} is the current occupied amount of resources for all active users/sessions in the ARA. This can be larger than r_{min} whenever extra "elastic" resources are provided for users/sessions that can benefit from it. It is assumed that all of the extra resources ($r_{occ}-r_{min}$) can be reclaimed (either instantaneously or after some delay) and assigned to other users.
- r_{max} is the current maximum amount that can be used in the ARA. The "headroom" $(1-r_{max})$ is the margin required to cope with changing resource usage of active users in the ARA due to, for example, user mobility. If the time to free up extra, "elastic" resources is very small or zero,

 r_{occ} resources could be allowed to grow beyond the r_{max} limit as these can be freed up to give room for increasing resource usage of active users.

The following resource levels are derived from the above:

- $\delta_{avail} = r_{max} r_{min}$ is the relative amount of currently available resources, including those that can be (instantaneously) freed up.
- $\delta_{free} = r_{max} r_{occ}$ is the relative amount of currently free resources.

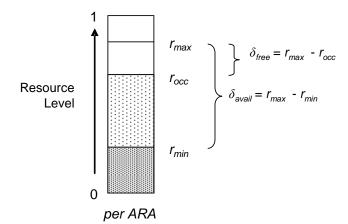


Fig. 13.6 Metrics for generic access resource abstraction

In addition to the general resource levels, further (abstracted) resource values are derived for a specific service $Q_{i,req}$, which is captured in a service specification as described in Sect. 13.4. Consider an access selection decision (admission control, load management) for a user/session *i* requesting quality $Q_{i,req}$ in an area covered by at least one ARA. $Q_{i,req}$ is defined by the service requirements on minimum bit rate and maximum delay and error tolerance, as well as, how much additional quality (typically extra bit rate) is beneficial for the service. For each ARA, MRRM transmits $Q_{i,req}$ to the associated GLL and requests a RAT-specific mapping of $Q_{i,req}$ to generic resource measures. The GLL retrieves resource information from the underlying RAT-specific entities, computes the resource levels r_{min} , r_{occ} , and r_{max} , and finally computes the following generic MRRM resource measures per user/session:

- $Q_{i,offered}$ is the offered quality, in terms of bit rate, maximum delay and additional quality (typically extra bit rate). This is typically only reported back to MRRM when $Q_{i,offered}$ differs from $Q_{i,req}$.
- $q_{i,min}$ is relative RAT-specific resource usage if user/session *i* was "served" in the ARA with minimum quality requirements. E.g., if there

are 10 slots in the ARA and 1 slot is required then $q_{i,min}=0,1$. Note that if $q_{i,min}>1$ then it is not possible to meet the minimum quality requirements in the ARA because the minimum amount of requested resources exceeds the available resources and the request is typically rejected.

- $q_{i,extra}$ is as above, but where user/session *i* is given some extra quality. So $q_{i,extra} > q_{i,min}$ as it contains additional spare resources.
- $\sigma_{i,min}=q_{i,min}/\delta_{avail}$ is relative resource efficiency/impact of serving user/session *i* in the ARA with minimum quality requirements. For example, if $\delta_{avail}=0.4$ and $q_{i,min}=0,1$ then $\sigma_{i,min}=0,25$, meaning that 25% of the remaining resources would be used by allocating the service data flow to this resource. Note that if $\sigma_{i,min}>1$ then the amount of resources requested exceeds the available resources and such a request is typically rejected.
- $\sigma_{i,extra} = q_{i,extra}/\delta_{avail}$ is as above, but where user/session *i* receives extra quality beyond $q_{i,min}$.
- α_i is the service availability which describes the number of services requests of type *i* that can still be supported by the access resource. Note that it is related to $1/\sigma_{i,min}$. However typically the amount of required resources depends non-linearly on the load level. The non-linear relationship depends on the access technology. In the determination of α_i also the resource usage policy of the access provider needs to be considered. For example, an access provider can limit the amount of resources that may be occupied by a certain service type. In this case the service availability does not depend on the overall remaining resources, but on the available budget of service specific resources.

Note that all measures apart from $Q_{i,req}$ and $Q_{i,offered}$ are given as relative values, as shown in Fig. 13.6. MRRM receives the (absolute) service request $Q_{i,req}$ from a higher level according to a service specification as described in Sect. 13.4, and forwards $Q_{i,req}$ to the GLLs, which convert it into access-specific relative resource requirements q_i . The other relative measures are calculated and reported back to MRRM.

13.6.2 Access Resource Structures and Combined Access Resource Metrics

In Sect. 13.6.1 it is presented how the status of a particular access resource can be described in abstract metrics, which enables to generically evaluate the suitability of an access resource to be used for a specific data service. In a realistic access system an access resource consists of several components. All of these need to be considered in the evaluation of an access system. Most notably an access resource has an uplink and a downlink component. A service flow can only be served if the resources in both uplink and downlink are sufficient for the corresponding service requirements. The uplink and downlink resource components can be independent as for access technologies using frequency division duplex. For time division duplex systems the partitioning of the access resource into uplink and downlink component can be dynamically adapted.

The main component of the access resource is the *physical resource*; in a wireless communication system it corresponds to the transmit power and the interference head room. An access resource is said to be suitable for a service only if sufficient transmit power is available at the given interference and noise level so that a link budget (i.e. SINR) is achievable that provides sufficient link performance. However, the usage of the physical access resource requires the availability of a further resource component, which is denoted as channelization resource. This channelization resource provides a data flow access to the physical resource; it can be considered as a multiplexing identifier within a limited name space. The channelization resource depends on the access technology. In circuit-switched time-division multiple access based access technologies (e.g. GSM/EDGE) the channelization resources corresponds to the time slots; in code division multiple access based access technologies (e.g. UMTS) the channelization resource is equivalent to the channelization codes; in access technologies based on dynamic multiplexing (e.g. EGPRS, LTE, WiMAX) the channelization resource corresponds to the logical connection identifiers that can be assigned.

Depending on the structure of an access network and the traffic load, the access resource can be either limited by the physical resource or the channelization resource. Let us consider a network with a large number of users with data intensive services. Users in the same cell or neighbouring cells cause a significant amount of interference and when exceeding a certain load level not enough transmit power remains available to achieve a sufficiently high link budget. This scenario is limited by the physical resource. In another example, a large number of users with low data rate services like voice-over-IP or chat are assumed. In this scenario the low traffic volume causes little interference and sufficient transmission power is available. However, when exceeding a certain number of active users the available channelization resources are exhausted and no new services can be admitted any more. In this case the system is limited by the channelization resource.

For multiple components of the access resource a combined resource abstraction has to be determined, as depicted in Fig. 13.7. The combined resource abstraction is determined either as the maximum (for r_{min} , r_{occ} , q_i ,

 σ_i) or as the minimum (α_i , δ_{avail} , δ_{free} , r_{max}) of the resource metrics of the different resource components.

For a service request $Q_{i,req}$ multiple suitable service realisation options can exist. For example, a service request for an elastic service type has a minimum data rate requirement but can benefit from obtaining a higher throughput. For such a service multiple realisation options of different data rate $\{Q_{i,l}, Q_{i,2}, ...\}$ can be provided by GLL with the corresponding resource abstractions, as shown in Fig. 13.7.

	Downlink	Uplink
Service Request	Rate/Delay/Reliability: Q _{i,req}	Rate/Delay/Reliability : Q _{i,req}
Resource component 1 (e.g. physical resource)		
Resource component 2 (e.g. channelization resource)		
Combined resource abstraction	$ \begin{aligned} & \& \text{Resource Status} \\ & \{\min(r_{max}^{(x)}), \max(r_{min}^{(x)}), \max(r_{occ}^{(x)}), \min(\delta_{free}^{(x)}), \min(\delta_{avail}^{(x)}) \} \\ & \& \text{Service cost options} \\ & & \geq \{Q_{i,l}, \max(q_{i,l}^{(x)}), \max(\sigma_l^{(x)}), \min(\alpha_l^{(x)}) \} \\ & & \geq \{Q_{i,2}, \max(q_{i,2}^{(x)}), \max(\sigma_2^{(x)}), \min(\alpha_2^{(x)}) \} \\ & & \searrow \dots \end{aligned} $	

Fig. 13.7 Combination of access resource metrics

13.7 Access Selection Process

Figure 13.8 depicts the access selection process, which is divided in two major parts: (i) policy-based access selection operates on a set of pre-defined parameters and validates the compatibility of access resources (AR) according to a set of policies and, (ii) dynamic access selection operates on a set of time-varying parameters of compatible and validated access resources.

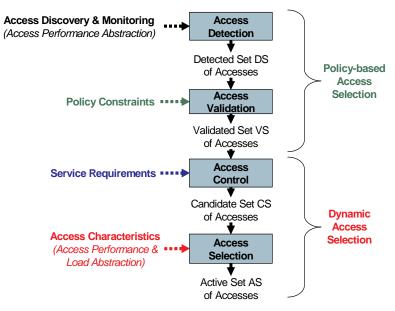


Fig. 13.8 Access Selection Process

13.7.1 Policy-based Access Selection

For a user terminal the Detected Set (DS) of ARs is determined from link monitoring and scanning of radio access networks via the available radio interfaces. The user device may either actively discover its neighbours or passively listen to advertisements. The Validated Set (VS) of ARs is determined based on policy rules; it only comprises ARs of acceptable networks and RATs. The criteria for determining the VS mainly consist of user and network preferences, the radio access configuration settings including terminal/access network capabilities, supported QoS classes, security support, type of mobility support including protocols and handover delay, context transfer support, supported radio interfaces and service price/resource costs.

13.7.2 Dynamic Access Selection

Dynamic access selection is applied for a given service flow request. It can operate in the range of minutes to seconds or even milliseconds. Therefore parameters derived from link performance and access resource abstractions are needed.

(1) Link-performance Based Access Selection

In the access selection process of Fig. 13.8, those elements of the VS which match the requirements of the flow are selected for the Candidate Set (CS) of ARs. Hence there is one CS per service. Within the CS the access resources are sorted according to their performance. In this step the abstracted averaging or expected performance metrics of data rate, transmission delay and reliability as described in Sect.13.5 are used. For link-performance based access selection, the best suited element from the CS is selected for data transmission, i.e. it is added into the Active Set (AS) of ARs. In special situations where a GLL entity controls two or more tightly integrated radio accesses, fast access selection can be performed [11]. In this case more instantaneous values of data rate, transmission delay and reliability metrics need to be considered.

(2) Resource Based Access Selection

Resource based access selection complements link-performance based access selection in order to achieve an optimised usage of the available radio resources and avoiding dropped sessions due to congestion. This is achieved by assigning a weight to the elements of the CS based on the resource situation and thus re-sorting the CS according to a combination of link performance and resource weight. In [12] a relative load measure (similar to r_{occ}) is defined. However, it is also noted in [12] that this measure is insufficient for load balancing due to varying cell capacities of different RATs. This problem is overcome by adding a relative resource impact (σ_i) in the abstraction, which allows comparing radio cells with different capacity.

Once computed, the MRRM resource measures and possibly the resource levels are signalled from GLL to MRRM, which then uses the values to determine access selection weights.

An example of MRRM decisions is the following: For admission control a first step is to check whether $\sigma_{i,min} \leq l$ for the ARA; if not then the user/session *i* cannot be admitted there.

Initial access selection or load management can be based on many different algorithms. Examples can be:

• Choosing the RAT/ARA with maximum amount of available resources δ_{avail} for a user session:

$$i = \arg \max_{j} \left\{ \delta_{avail} \left(ARA_{j} \right) \right\}$$

• Choosing the RAT/ARA with maximum amount of free resources δ_{free} :

$$i = \arg \max_{j} \left\{ \delta_{free} \left(\text{ARA}_{j} \right) \right\}$$

• Choosing the RAT/ARA with minimum amount of relative required resources *q_i*:

$$i = \arg\min_{i} \{q_{j}(ARA_{j})\}$$

• Choosing the RAT/ARA with the maximum service availability α_i :

$$i = \arg \max_{i} \{ \alpha_{j} (ARA_{j}) \}$$

• Choosing the RAT/ARA with minimum resource usage efficiency *σ_{i,min}*, that is:

$$i = \arg\min_{i} \left\{ \sigma_{j,\min} \left(\text{ARA}_{j} \right) \right\}$$

• Choosing the RAT/ARA with minimum resource usage efficiency $\sigma_{i,extra}$, that is:

$$i = \arg\min_{j} \left\{ \sigma_{j,extra} \left(\text{ARA}_{j} \right) \right\}$$

Note that if the current load is high in all ARAs, i.e. $\sigma_{i,extra}>1$ for all RATs, then alternatively the RAT / ARA could be selected which minimizes $\sigma_{i,min}$.

Many other MRRM decision algorithms can be considered based on the MRRM resource measures above. The key purpose of the measures is to provide sufficient, comparable radio information on current radio resource state and resource usage efficiency for various heterogeneous RATs for effective MRRM operation.

Note that in some cases additional interactive negotiation between GLL(s) and MRRM can be performed, when GLLs provide additional information about the best service performance that can be provided (without guarantees). For this a translation from relative measures back to an absolute measure $Q_{i,offered}$ is required, as exemplified below:

- MRRM gets a request for (absolute) resources $Q_{i,req}$. E.g. $Q_{i,req}$ is a request for 150 kb/s service.
- MRRM passes the request $Q_{i,req}$ on to GLLs, where it is translated to relative resources $q_{i,min}$.
- GLLs reply to MRRM relative load values for load balancing, to determine the relative resource costs *q*_{*i*,*min*}.

- If spare access resources are available, it may be be beneficial to know what absolute service level $Q_{i,offered}$ could be provided to the service by the access (i.e. what maximum Q can be provided by $q_{i,extra} > q_{i,min}$). This requires that GLL would not only make a translation $Q_{i,req} \leftrightarrow q_{i,min}$ but also $q_{i,extra} \leftrightarrow Q_{i,offered}$.
- Then MRRM would get the information from GLL: "Your request $Q_{i,req}$ can be handled, it costs the relative resources $q_{i,min}$. But the access could even support the service request at level $Q_{i,offered}$, which would cost the resources $q_{i,extra}$."

(3) Access Priorities

So far, only abstracted values of resource availability and resource costs have been considered. In a realistic scenario, these resource abstractions can be weighted in the MRRM access selection decision according to a priority of different RATs/ARAs. In this case the generic resource metrics need to be adapted by a priority weight factor.

The priorities can be set for several reasons:

- To reflect the operator or terminal priorities of RAT/ARA usage.
- Some RATs/ARAs may be provided by other cooperating operators. In this case additional roaming/cooperation charges may exist for the usage of those RATs/ARAs.
- Some RATs/ARAs may have less efficient operation, e.g. they require more signalling for handover or AAA signalling.
- Some RATs/ARAs may provide less security.

13.8 Conclusion

For access selection in multi-access networks a key problem is how to efficiently manage radio resources independently of the number of access systems and their nature. Access selection has to determine the best suited access for a given service. It has been shown how service requirements are appropriately described. In order to understand the suitability of an access for a service, abstract RAT-specific information on link performance and resource measurements are required to make different access systems comparable. Abstraction models have been developed, which derive generic, sufficiently detailed, and accurate RAT descriptions. MRRM can thus make access selection without knowledge of technology-specific characteristics. The abstraction models are derived for link performance in terms of rate, delay, and bit error rate. Further a resource abstraction model has been provided, where available access resources are divided into resource areas (for example, cells in mobile systems) and the resource state is described in terms of relative resource levels, the relative resource usage, and the relative usage efficiency for a given flow. Typically access resources are made up of several resource components, like radio resources and channelization resources for both uplink and downlink. It has been described how these different resource components are combined to a common resource description. A number of different access selection algorithms are given to demonstrate how these abstracted metrics can be used. By combining a number of generic link performance and access resource metrics, a wide variety of access selection algorithms can be realized.

Acknowledgments

This work has been partly funded by Ambient Networks, a research project supported by the European Commission under its Sixth Framework Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Ambient Networks project or the European Commission.

References

- 1. 3GPP TR 23.882 (2006) 3GPP system architecture evolution (SAE): Report on Technical Options and Conclusions. December
- 2. IEEE 802.21 (2005) Draft IEEE Standard for Local and Metropolitan Area Networks: Media Independent Handover Services. July
- Niebert N, Schieder A, Abramowicz H, Malmgren G, Sachs J, Horn U, Prehofer C, Karl H (2004) Ambient Networks – An Architecture for Communication Networks Beyond 3G. IEEE Wireless Communications 11: 14–22
- Lundsjö J, Agüero R, Alexandri E, Berggren F, Cedervall C, Dimou K, Gebert J, Jennen R, Jorguseski L, Karimi HR, Meago F, Tang H, Veronesi R (2005) A Multi-Radio Access Architecture for Ambient Networking. In: Proc. 14th IST Mobile & Wireless Communications Summit, Dresden, Germany, 19–23 June
- Sachs J, Prytz M, Gebert J (2007) Multi-Access Management in Heterogeneous Networks. Wireless Personal Communications, http://dx.doi.org/ 10.1007/s11277-007-9431-8

- Sachs J, Agüero R, Daoud K, Gebert J, Koudouridis GP, Meago F, Prytz M, Rinta-aho T, Tang H (2007) Generic Abstraction of Access Performance and Resources for Multi-Radio Access Management. In: Proc. 16th IST Mobile & Wireless Communications Summit, Budapest, Hungary, 1–5 July
- Ambient Networks (2005) Multi-Radio Access Architecture. Project deliverable D2-4, December. Available at http://www.ambient-networks.org/ (accessed October 2007)
- Ambient Networks (2006) Multi-Access & ARI design & initial specification. Project deliverable D2-C1, December. Available at http://www.ambientnetworks.org/ (accessed October 2007)
- 9. 3GPP TS 23.107 (2007) Quality of Service (QoS) concept and architecture (Release 7). September
- Mohr W (2003) Spectrum demand for systems beyond IMT-2000 based on data rate estimates. Wireless Communications and Mobile Computing 3: 817–835
- Dimou K, Agüero R, Bortnik M, Karimi R, Koudouridis GP, Kaminski S, Lederer H, Sachs J (2005) Generic Link Layer: A Solution For Multi-Radio Transmission Diversity in Communication Networks Beyond 3G. In: Proc. 62nd IEEE Semiannual Vehicular Technology Conference, Dallas, Texas, USA, September 25–28
- 12. 3GPP TR 25.881 (2001) Improvement of Radio Resource Management (RRM) across RNS and RNS/BSS. December

14 A Decentralized RAT Selection Algorithm Enabled by IEEE P1900.4

J. Pérez-Romero, O. Sallent, R. Agustí, J. Nasreddine, and M. Muck

14.1 Introduction

Heterogeneous radio access networks (RANs) concept, also known as Beyond 3G (B3G) systems, is intended to propose a flexible and open architecture for a large variety of wireless access technologies, applications and services with different Quality of Service (QoS) requirements, as well as different protocol stacks. RANs differ from each other by air interface technology, cell-size, coverage, services, price and ownership. The complementary characteristics offered by the different radio access technologies (RATs) make possible to exploit the trunking gain leading to a higher overall performance than the aggregated performances of the stand-alone networks. Clearly, this potential gain of B3G systems can only turn into reality by means of a proper management of the available radio resources. Joint Radio Resource Management (JRRM) refers to the set of functions that are devoted to ensure an efficient and coordinated use of the available radio resources in heterogeneous networks scenarios [1,2,3]. More specifically. JRRM strategies should ensure that operator's goals in coverage and QoS levels are met while providing as high as possible overall capacity (i.e. the sum of the capacities achieved in every single RAN of the operator) by using the available resources. Within JRRM, the RAT selection, responsible for the allocation of connections to specific RANs either at session initiation or during the session life-time (i.e. switching on-going connections from one RAT to another leading to inter-system or vertical handovers), is the key enabler to properly manage heterogeneous radio access network scenarios [3].

Different works have been published in the open literature dealing with RAT selection in heterogeneous networks. In [4] the benefits of JRRM in terms of inter-system handover and inter-system network controlled cell

reselection are analyzed in a heterogeneous UMTS/GSM network. In [5] the Analytic Hierarchy Process and Grey Relational Processes are jointly used as a tool to introduce priorities among user preferences, service applications and network conditions in heterogeneous networks involving UMTS and WLAN. Similarly, in e.g. [6] and references therein, the problem is discussed from a more general perspective, comparing several substitution policies and including the multi-mode terminal dimension with speech and data services.

However, all of these works address the problem from a centralized perspective. Indeed, Radio Resource Management (RRM) functions in a wireless cellular network are mainly centralized, i.e. the functions are implemented in a central network node such as RNC (Radio Network Controller) in UTRAN (UMTS Terrestrial Radio Access Network). This can be justified because a central network node may have a more complete picture of the radio access status than a particular node, so that RRM decisions can be made with more inputs. However, a centralized RRM implementation has some drawbacks in terms of increased signalling load or transfer delay of the RRM algorithm's inputs to the central node. This prevents an efficient implementation of short-term RRM functions such as packet scheduling and explains why wireless cellular technology evolution (e.g. High Speed Downlink Packet Access – HSDPA) exhibits the trend towards implementing RRM functions on the radio access network edge nodes (i.e. base stations).

Additionally, the terminal also keeps relevant information that could be of great interest for making smarter RRM/JRRM decisions. This is why some RRM functions, although typically implemented in the network side (either on central or edge nodes), are assisted by mobile terminal measurement reports. Handover algorithm is a clear example, since the knowledge of the propagation conditions from the terminal to the different surrounding cells is a key aspect for making the proper decision on what cell(s) the terminal should be connected to.

This chapter goes one step beyond in this trend towards distributed RRM/JRRM functions by proposing RAT selection strategies executed in mobile terminals. This approach has been found to be inefficient in the past because of the limited information available at the terminal side (e.g. the terminal does not know the cell load). Nevertheless, this can be overcome if the network is able to provide some information or guidelines to the terminal assisting its decisions. In this way, while a mobile-assisted centralized decision making process requires the inputs from many terminals to a single node, the network-assisted decentralized decision making process requires the input from a single node to terminals, which can be significantly more efficient from a signalling point of view. In this respect,

the on-going IEEE P1900.4 [4] standardization effort could provide the necessary support to this network-assisted mechanism.

The objective of IEEE P1900.4 is to define standardized protocols and corresponding reconfiguration management system architecture for the optimization of resource management, in order to provide improved capacity, efficiency and utility within a heterogeneous wireless network wherein devices support multiple air interfaces, with multi-homing and dynamic spectrum access capabilities in licensed and unlicensed bands [4]. IEEE P1900.4 provides the necessary management functions and standardized rules to allow these devices making decisions in a distributed fashion whilst providing operators with fair and effective exploitation of network resources thanks to an exhaustive set of rules to be followed by user equipments, thus enabling decentralized RAT selection mechanisms.

Under this framework, this chapter will support the proposed decentralized RRM/JRRM approach with an illustrative example focusing on RAT selection functionality. For that purpose, it is taken as a reference the RAT selection algorithm presented in [7] for heterogeneous CDMA/TDMA scenarios and its applicability in the framework of IEEE P1900.4 is proposed. The rest of this chapter is organized as follows. Section 14.2 discusses the RAT selection enablers defined in IEEE P1900.4. Section 14.3 presents a study case regarding how interference can be reduced through decentralized RAT selection. This strategy is evaluated with the simulation model described in Sect. 14.4 and results are presented in Sects. 14.5 and 14.6. Particularly, Sect. 14.5 considers a situation with a single type of traffic and Sect. 14. 6 extends the work to the multi-service case. Finally, conclusions are summarized in Sect. 14.7.

14.2 RAT Selection Enablers Defined by IEEE P1900.4

RAT selection strategies are devoted to decide the adequate RAT that a given user should be connected to in a heterogeneous network as illustrated in Fig. 14.1. This decision is taken at session initiation (i.e. initial RAT selection procedure) as well as during session lifetime, which can trigger a vertical handover procedure in case the current RAT must be changed. RAT selection strategies may respond to different principles, like e.g. service-based policies (i.e. allocating the RAT according to the service characteristics) or load balancing principles (i.e. try to keep similar load levels in the different RATs). When this decision should be taken in a decentralized way at the terminal, this requires enabling functionalities on the

network/terminal side and a corresponding transport channel, which is one of the aspects covered by IEEE P1900.4.

More specifically, on the network level, IEEE P1900.4 is proposing to introduce a "Network Reconfiguration Manager" (NRM) entity covering the following functionalities [4]:

- The "Information on Dynamic Spectrum Allocation" module is providing management protocols giving indications on Dynamic Spectrum Allocation rules from the network to the user device. The network (meta-) operator is communicating its spectrum assignment to the devices that will choose their resource selection strategies correspondingly.
- The "Radio Resource Selection Policies" module is deriving optimization constraints to be imposed onto user terminals. The goal is to constraint the resource selection optimization process in the terminals so that a global system objective is achieved (e.g. maximum system capacity utilization, etc.).
- The "Recovery of Context Information" module represents the interfacing of the NRM with the network equipment for recovery of operational information from the RATs as well as the optional link to user equipment providing feed-back on observed QoS, etc.

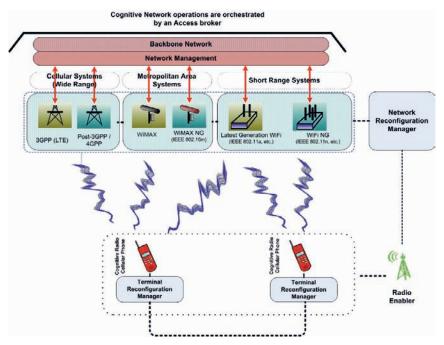


Fig. 14.1 IEEE P1900.4 heterogeneous system vision

- The "Representation Definition of Context and Policy Information" module is presenting the effort on context and policy related ontology and management protocol definitions.
- The "Control of Resource Selection Strategy Change in User Equipment" module is expected to trigger the resource selection strategies within the user devices such that the distributed optimization is performed in a controlled manner.
- The "Security Issues" module is expected to provide suitable security means in order to assure the ownership of the policies and context information provided by the network.

Further entities introduced by IEEE P1900.4 include:

- The "Terminal Reconfiguration Manager" which performs distributed decision making based resource selection, subject to network constraints (policies), and controls the information flow from the terminal to the network.
- A "Radio Enabler for Reconfiguration Management" which acts as signalling link between the network and the terminal and may be deployed as a dedicated physical or logical channel [4].

14.3 Case Study: Interference Reduction through Decentralized RAT selection

In general, wireless systems are interference-limited and, consequently, any engineering technique devoted to either reduce interference or to improve the robustness of the system to bear interference will readily increase network capacity and operator's revenue. In this context, the RAT selection can exploit the different sensitivities to interference that diverse RATs may exhibit so that a smart JRRM follows. In particular, in TDMAbased access systems (e.g. GSM/GPRS) there is no intra-cell interference. In turn, inter-cell interference is caused by a single user in every co-channel cell and therefore there is no inter-cell interference in neighbouring cells, as illustrated in Fig. 14.2, because they operate in different carriers. In contrast, in CDMA-based systems (e.g. UMTS) the intra-cell interference is caused by every single user transmitting in the cell. Furthermore, inter-cell interference is also originated by all simultaneous users in all neighbouring cells, since a complete frequency reuse is considered, as shown in Fig. 14.2. Consequently, CDMA systems are much more sensitive to multi-user interference than TDMA ones.

Taking this into account, the underlying idea of the proposed decentralized JRRM approach is to take advantage of the coverage overlap provided by the existence of several RANs using different access technologies in a certain service area in order to improve the overall interference pattern generated in the scenario for the CDMA-based systems and, consequently, to improve the capacity of the overall heterogeneous network. This can be achieved through appropriate RAT selection algorithms that avoid the connection of the more interfering users to CDMA. In that sense, notice that the users generating more interference in CDMA will be those located farther from their serving base station, because they will be transmitting a higher power level seen as interference by neighbouring base stations.

As illustrated in Fig. 14.2, the interference *I* measured by the neighbouring base station depends on the power transmitted by the terminal P_T , which in turn depends on the path loss L_p of this user to the serving base station due to power control. In TDMA the transmitted power also depends on the path loss but since the neighbouring cells operate with different frequencies no inter-cell interference is generated. Taking this into account, an interference reduction can be achieved by forcing some terminals with a high path loss in CDMA to be connected to the TDMA-based RAT while CDMA keeps only the terminals with low path loss.

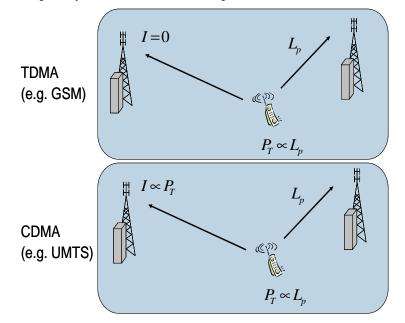


Fig. 14.2 Intercell interference between neighbouring cells in TDMA and CDMA systems (L_p denotes the path loss, *I* the inter-cell interference and P_T the transmitted power)

The resulting RAT selection strategy is illustrated in Fig. 14.3 and Fig. 14.4. The decisions are taken autonomously by the terminal from its path loss measurements $L_p(t)$ to the best CDMA cell. These measurements can be obtained from the downlink received power of a common control channel whose transmit power is known, e.g. the Common Pilot Channel in UMTS. Measurements are averaged over periods of *T* seconds. Then, at session initiation, in case that the resulting averaged path loss $L_p(t)$ in dB is above a given threshold PL_{th} , the selected RAT will be TDMA, while if the path loss is below PL_{th} the selected RAT will be CDMA, as shown in Fig. 14.4. The threshold PL_{th} is provided by IEEE P1900.4 radio enabler to execute the algorithm autonomously at the terminal. In case that there is no capacity available for the new session in the selected RAT (i.e. admission control is not passed), the other RAT will be selected instead. Finally, if no capacity is available in any of the two RATs, the session will be blocked.

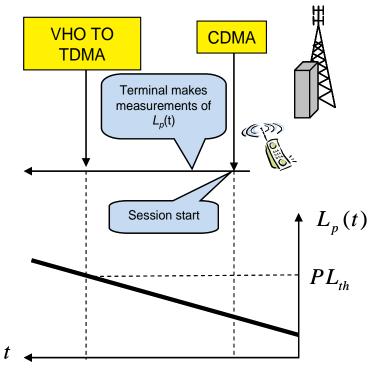


Fig. 14.3 Decentralized RAT selection strategy

In the example of Fig. 14.3, the user at session start selects the CDMAbased RAT (e.g. UMTS), then it continuously measures the path loss to the CDMA cell and when it is above the threshold PL_{th} , a vertical handover to the TDMA-RAT is triggered (assuming that there is TDMA coverage at that point). In order to avoid undesired ping-pong effects leading to continuous RAT changes for users with path loss close to the threshold PL_{th} , a hysteresis margin Δ (dB) is introduced, meaning that, when the mobile is connected to TDMA, a vertical handover to CDMA will be triggered if $L_p(t)$ is below PL_{th} - Δ during M_{down} consecutive samples. Similarly, when the mobile is connected to CDMA, a vertical handover to TDMA will be triggered if $L_p(t)$ is above PL_{th} + Δ during M_{up} consecutive samples.

From a practical point of view, the different parameters PL_{th} , Δ , M_{up} and M_{down} are fixed by the network based on the collected information by the specific RRM procedures of the different RATs, such as the statistical path loss distribution. The parameters are sent to the users via IEEE P1900.4 protocols, as illustrated in Fig. 14.4, and then the users can execute the algorithm periodically to take the corresponding decisions based on their own measurements.

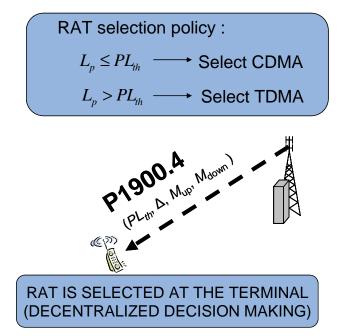


Fig. 14.4 Information to be transmitted through P1900.4

14.4 Simulation Model

The proposed scheme has been evaluated by means of system level simulations in a scenario with a CDMA-based and a TDMA-based RAT. Specifically, UTRAN and GERAN (GSM/EDGE Radio Access Network) are considered as examples of these two technologies. Seven omni-directional cells for GERAN and seven for UTRAN are considered. The cells of both RANs are collocated. The separation between base stations is 2 km. In GERAN, the seven cells represent a cluster so that all cells operate with different carrier frequencies. The parameters of the UE (User Equipment), UTRAN and GERAN cells are taken from [7]. It is assumed that all terminals have multi-mode capabilities, i.e. they can be connected either to UTRAN or to GERAN. Three carriers per cell in the 1800 MHz band are assumed in GERAN and a single carrier is considered in UTRAN. The urban macro-cell propagation model in [8] is considered for both systems, where the path loss is a function of the distance d between mobile and base station given by: $L_p(dB)=128.1+37.6\log [d(km)]+S(dB)$, where S(dB) corresponds to the log-normal shadowing with 10 dB standard deviation. The mobility model described in [9] is considered with constant mobile speed of 3 km/h and shadowing decorrelation distance of 20 m.

UTRAN RRM PARAMETERS	
UL admission threshold (η_{max})	1.0
DL admission threshold (P _{max})	42 dBm
Measurement time	1 s
Active Set size	1
Replacement hysteresis	3 dB
Time to trigger handover	0.64 s
Minimum Ec/Io	-16 dB
GERAN RRM PARAMETERS	
Measurement period	0.48 s
Minimum access power	-105 dBm
Minimum received power to trigger handover (UL or DL)	-100 dBm
Samples below minimum power to trigger handover	3
RAT SELECTION ALGORITHM PARAMETERS	
Measurement interval (T)	1 s
Hysteresis margin (Δ)	1 dB
$M_{\rm up}/M_{\rm down}$	3/3
PL_{th}	Variable

Table 14.1 RRM parameters

In UTRAN, an iterative perfect power control procedure is considered to simulate the inner loop power control aiming at achieving the target (E_b/N_0) that ensures the required Block Error Rate (BLER). With respect to GERAN, a slow power control is simulated in the uplink, so that the transmitted power is changed in steps of 2 dB every measurement period of 0.48 s in order to reach a specific sensitivity level. No power control is simulated in the downlink, and all the channels are transmitted with maximum power.

A summary of the main RRM parameters is given in Table 14.1, together with the parameters of the RAT selection algorithm. With respect to the admission control procedure in UTRAN, three conditions are checked [3], namely the uplink load factor should be below threshold η_{max} , the downlink transmitted power below P_{max} and there must be available OVSF (Orthogonal Variable Spreading Factor) codes in the base station. In GERAN, voice users are accepted if there are available time slots. With respect to the admission control for horizontal handovers, the availability of OVSF codes for UTRAN or time slots for GERAN is checked in the new cell. If admission is not passed, a vertical handover will be tried, and if it is not possible, then the call will be dropped. On the other hand, whenever a call is about to be dropped due to propagation/interference conditions in a given RAT, a vertical handover will also be tried.

For comparison purposes, the proposed strategy is compared to a classical centralized Load Balancing (LB) strategy, in which the network allocates the user to the RAT having the lowest load level. Load measurements are averaged within periods of 10 s to smooth load fluctuations and are obtained from the base stations having the lowest path loss among those of each RAT. Whenever a horizontal handover is required in the current RAT, the suitability of executing a vertical handover instead is evaluated, so that the mobile is again served by the lowest loaded RAT.

14.5 Results in a Single Service Scenario

This section analyzes the system performance obtained by means of the proposed strategy in a scenario where only real time (RT) users (i.e. voice) are considered. Calls are generated according to a Poisson process with call rate of 10 calls/h/user and exponentially distributed call duration with an average of 180 s. In UTRAN, the Radio Access Bearer (RAB) is the 12.2 kb/s speech bearer defined in [9], considering a dedicated channel (DCH) with spreading factor 64 in the uplink and 128 in the downlink. In

GERAN, each voice user is allocated to one TCH-FS (traffic channel fullrate speech), i.e. one time slot in each frame of a given frequency.

One of the critical parameters to be set is the path loss threshold, PL_{th} , which has high impact on the performance of the proposed algorithm, as detailed in the following. On the one hand, the value of PL_{th} affects the QoS levels of users in the RATs in the sense that low PL_{th} values will tend to reduce UTRAN interference thus improving the performance of users connected to this RAT. On the other hand, it also controls the traffic distribution between the considered RATs, in the sense that low PL_{th} values will tend to increase the number of users allocated to GERAN while high values will tend to reduce these number of users and to allocate more users in UTRAN. Consequently, the setting of the path loss threshold PL_{th} results from the trade-off between how much the UTRAN interference can be reduced while avoiding an excessive load unbalance. To illustrate these effects, three different representatives values of PLth have been selected, namely PL_{th} ={115 dB, 120 dB, 125 dB}, corresponding, approximately to the 40-th, 60-th and 80-th percentiles of the path loss distribution, respectively.

From the point of view of load distribution in the two RATs, Fig. 14.5 plots the average uplink load in UTRAN and GERAN with the different PL_{th} values and for the LB strategy (similar trends not shown here for the sake of brevity are also obtained for the downlink case). Notice that the case $PL_{th}=120$ dB achieves the better load balancing between both RATs, while for $PL_{th}=115$ dB there is a higher load in GERAN and for $PL_{th}=125$ dB the load is higher in UTRAN. Then, the proposed algorithm with $PL_{th}=120$ dB achieves a load distribution similar to the LB case, so that with this setting load balancing considerations are also included in the proposed algorithm.

From a performance point of view the total aggregated throughput (i.e. including UTRAN and GERAN) is depicted in Fig. 14.6 for the downlink (similar performance improvements are also observed for the uplink). The highest throughput is provided by PL_{th} =120 dB, revealing to be the most suitable solution from both QoS and load balancing points of view. Compared to a pure LB, the achieved gain can be up to about 24% for heavy load conditions. The origin of the gain comes from the fact that the decentralized RAT selection algorithm with PL_{th} =120 dB also achieves load balancing between RATs through a more intelligent and efficient user distribution, reducing the overall interference in the system. Compared to the other settings, i.e. PL_{th} =115 dB or 125 dB, the gain comes from the benefits of the better load balancing obtained with PL_{th} =120 dB.

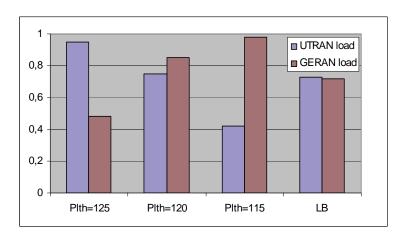


Fig. 14.5 Uplink load in UTRAN and GERAN for the different values of PL_{th}

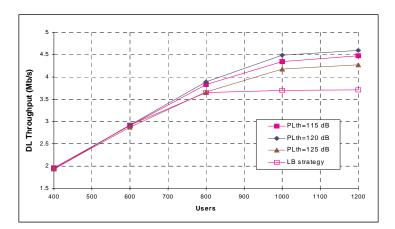


Fig. 14.6 Downlink throughput for different values of PL_{th}

14.6 Results in a Multi-service Scenario

This section analyzes the impact of having different types of traffic in the scenario. For that purpose, let assume that RT (i.e. voice) users coexist with non-real time (NRT) users (i.e. www navigation). In such a case, the proper RAT selection should account not only for the interference reduction and load balancing considerations, but also the way how the service is provided in each of the RATs. For example, in the case of UTRAN, NRT

traffic can be served by means of dedicated channels, while in the case of GERAN NRT traffic is served by means of shared channels, meaning that the available bit rate will be reduced when the number of users sharing the same time slot increases. Furthermore, it is also usual that in GERAN RT traffic has priority over NRT users, so that when increasing the number of RT users connected to this RAT, there will be less resources available for NRT traffic, which eventually will reduce the bit rate even more.

With the above considerations, and assuming a scenario with UTRAN and GERAN RATs, the proposed algorithm is only applied by RT users, assuming that the service they can receive is similar in the two RATs, while the NRT users will always select by default UTRAN, and they will only be served through GERAN in case there is not capacity available in the other RAT. In that respect, notice that the parameters transmitted through P1900.4 depicted in Fig. 14.4 will be the same as in the previous case. On the other hand, notice that a simpler approach could be to split traffic in accordance with the service (i.e. RT users in GERAN and NRT in UTRAN), however, this would achieve a poorer performance because no interference considerations would be made.

In the results presented here, voice users follow the same model that was explained in the previous section, while NRT users follow the www browsing model explained in [10], with 5 pages per session, an average reading time between pages of 30 s, an average of 25 packets per page, and interarrival packet time of 0.125 s for the uplink and 0.0228 s for the downlink. The average packet size is 366 bytes. A session rate of 24 sessions/h/user is assumed. WWW browsing service is provided in UTRAN by means of dedicated channels (DCH) making use of the transport channel type switching procedure. The considered RAB assumes a maximum bit rate of 64 kb/s in the uplink and 128 kb/s in the downlink [9]. In turn, in GERAN, the www service is provided through a PDCH (Packet Data Channel) with a round robin scheduling algorithm to allocate transmissions to users sharing the same time slot. On the other hand, a link adaptation mechanism operating in periods of 1s is used to select, for each user, the highest modulation and coding scheme (MCS) that ensures the specific sensitivity requirements. The highest modulation scheme considered here is MCS-7.

In this scenario, the control of the parameter PL_{th} will allow modifying the amount of RT traffic that is served by each technology, which will have an influence over the performance observed by both RT and NRT users. To illustrate this effect, Fig. 14.7 plots the percentage of total throughput improvement achieved with the proposed strategy with respect to a purely service-based selection in which the RT users were served through GERAN and the NRT users through UTRAN. The results are provided for three representative situations depending on the amount of load of each service, namely low RT load and high NRT load, high RT load and low NRT load, and high load of both RT and NRT. The case with low loads of both RT and NRT traffic is not included because it does not present significant differences with any of the considered settings. On the other hand, it is worth mentioning that the LB strategy that was considered for comparison in the previous section is not presented here because it does not take into account service aspects in the selection.

From the results presented in Fig. 14.7 the following observations can be made, which lead to defining the proper setting of the algorithm parameters to be transmitted through IEEE P1900.4 depending on the corresponding traffic mix:

- The situation with low RT load and high NRT load provides the smallest improvements, and in this case the best setting is a low value of PL_{th} like 110 dB, corresponding to approximately the 20-th percentile of the path loss distribution. Notice that, with this value, the algorithm will serve most of the voice users through GERAN and will keep in UTRAN only those voice users located close to the base station, thus leaving a large room in this technology for NRT users. Consequently this setting basically captures the service component, which is more relevant than having a balanced load in a scenario with a high number of NRT users.
- In the situation with high RT load and low NRT load the highest improvement is achieved with the setting PL_{th} =120 dB. Notice that in this case, the situation is very similar to the one analyzed in Sect. 14.5, where no NRT load was considered, and therefore the interference reduction together with load balancing considerations allow improving the overall performance.
- Finally, in the situation with a high load from both RT and NRT users, the highest improvement is observed with $PL_{th}=115$ dB. In such a situation high values of the threshold (e.g. 125 dB) lead to an excessive number of RT users in UTRAN, consuming part of the capacity that could be used by NRT traffic, thus degrading its performance. On the other hand, due to the high number of RT users, it is also beneficial to try to distribute them more or less equally among the two RATs, so that interference for users experiencing the worst propagation conditions can be reduced. As a result of that, setting PL_{th} at an intermediate value between those considered in the two other cases reveals to be the most adequate solution.

Figure 14.8 summarizes the above considerations by presenting the adequate setting of PL_{th} for the different service mixes together with the principles governing each setting.

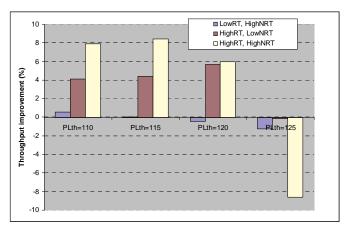
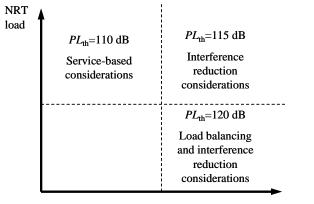


Fig. 14.7 Throughput improvement with respect to a purely service-based selection for different values of PL_{th} and different loads of RT and NRT traffic



RT load

Fig. 14.8 Appropriate setting of PL_{th} for different load conditions

14.7 Conclusions

This chapter has addressed the decentralized implementation of RAT selection strategies for heterogeneous wireless networks based on the functionalities enabled by IEEE P1900.4. A case study corresponding to a RAT selection that reduces the interference in CDMA through a smart allocation of users to RATs according to the measured path loss has been analyzed. In this case, only a minimum set of configuration parameters should be transmitted with the help of IEEE P1900.4 so that the RAT selection decision is taken autonomously by the mobile terminal. In this way, signalling can be reduced with respect to the centralized scheme. Results have shown that with this approach load balancing considerations can be retained while at the same time achieving a higher throughput than if a pure load balancing strategy was used. In turn, when considering the algorithm operating in a multi-service environment, where the abilities of each RAT to provide a given service may be different, it has been obtained that the setting of the path loss threshold used by the algorithm should be made depending on the existing traffic mix, so that service, load and interference reduction considerations can be balanced.

Acknowledgement

This work was performed in project E2R II which has received research funding from the Community's Sixth Framework program. This work reflects only the authors' views and the Community is not liable for any use that may be made of the information contained therein. The contributions of colleagues from E2R II consortium are hereby acknowledged.

References

- 1. 3GPP TR 25.881 v5.0.0 "Improvement of RRM across RNS and RNS/BSS".
- 3GPP TR 25.891 v0.3.0 "Improvement of RRM across RNS and RNS/BSS (Post Rel-5) (Release 6)".
- 3. J. Pérez-Romero, O. Sallent, R. Agustí, M. Díaz-Guerra, *Radio Resource Management Strategies in UMTS*, John Wiley & Sons, 2005.
- A. Tölli, P. Hakalin, H. Holma, "Performance Evaluation of Common Radio Resource Management (CRRM)", *Proceedings of the IEEE International Conference on Communications (ICC'02)*, Vol. 5, April, 2002, pp. 3429–3433.
- 5. Q. Song, A. Jamalipour, "Network Selection in an Integrated Wireless LAN and UMTS Environment Using Mathematical Modeling and Computing Techniques", *Wireless Communication and Mobile Computing*, June 2005.
- S. Lincke-Salecker, "The Benefits of Load Sharing when Dimensioning Networks", *Proceedings of the 37th Annual Simulation Symposium (ANSS'04)*, April, 2004.
- 7. J. Pérez-Romero, O.Sallent, R.Agustí, "Network Controlled Cell Breathing in Multi-service heterogeneous CDMA/TDMA scenarios", *Proceedings of the*

IEEE Vehicular Technology Conference Fall (VTC-Fall'06), Montreal, September, 2006.

- 8. 3GPP TR 25.942 "Radio Frequency (RF) system scenarios".
- 9. 3GPP TS 34.108 "Common Test Environments for User Equipment (UE); conformance testing".
- 10. UMTS 30.03 v3.2.0 "Selection procedures for the choice of radio transmission technologies of the UMTS", ETSI, April, 1998.
- 11. IEEE P1900 Home Page, http://www.ieeep1900.org

Part V Applications

15 Business Models for Local Mobile Services Enabled by Convergent Online Charging

Frank C. Bormann, Stephan Flake, and Jürgen Tacken

15.1 Introduction

This article is concerned with business models for a new kind of blended personalized and context-aware services. Currently most business models for context-aware services are network operator-centric. They provide the access network and charging infrastructure and create value by offering different mobile services based on their pricing strategies. Today only huge third party service and content providers have the opportunity to bring their services into the portfolio of a network operator.

In contrast, this article is about enabling business for small- and medium-sized enterprises (SMEs) via the access network of large operators. There are a lot of SMEs in different segments (e.g., health, tourism, publishing, maintenance), which can certainly benefit when they are enabled to offer their local mobile services via a mobile network infrastructure. They can offer typical Internet services (Instant Messaging, chat, video download, local search, local news, etc.) enriched by context information of the users (e.g., presence, location). Additionally, these services (including contents) can be personalized based on the user preferences.

The work presented here is part of the ITEA project "Local Mobile Services" (LOMS). The LOMS project investigates methods and tools for development, deployment and use of context-aware mobile services. The main aim is to allow for easy creation of smarter services by non-expert service providers. Within the project, case studies of currently deployed

local mobile services were investigated and market potentials in different segments have been analyzed [1].

In the following section, we first address the business models of the different players, which deliver mobile services. Note that the key factors in the business models are the charging and billing processes. The processes described in this article are based on the principle of context-aware convergent online charging [2]. As an example, a context-dependent payment flow is described in the third section, indicating the requirements for flexible, convergent online charging mechanisms. Afterwards, the necessary extensions for existing standard charging and billing mechanisms are described.

15.2 Business Models for Mobile Services

The following definition of the term "business model" is taken from R. Hawkins:

A business model is a description of the commercial relationship between a business enterprise and the products and/or services it provides on the market. More specifically, it is a way structuring various cost and revenue streams such that business becomes viable, usually in the sense of being able to sustain itself on the basis of the income it generates.

A taxonomy of different business models for mobile services has been presented in studies on the mobile multimedia [3] and 3G markets [4].

There are three basic business models according to the dominant players:

- Content provider (CP), which delivers digital content to the mobile user.
- Service provider (SP), which delivers mobile services to the mobile user.
- Network Operator (**NO**), which operates the access network infrastructure and delivers basic services to the user (e.g., Voice, SMS, Internet Access).

Currently some network operators already tend to take over the role of service providers or content providers to extend their business.

15.2.1 The LOMS Role Model

Within LOMS, two additional roles are considered to analyze business relations between the players in more detail (see Fig.15. 1). The Platform Operator (**PO**), which operates a Service Delivery Platform including a Charging and Billing System (CBS), and the Service Operator (**SO**), which provides a service creation tool. The PO provides enabling services like charging, billing, profile management of mobile users to the SO. On the other hand, he has contact to different NOs to reach a large base of mobile users.

The SO provides dedicated service templates for SPs operating in a certain segment (e.g., publishing, maintenance). Thus the SO is aggregating a lot of similar SMEs and supports them to easily create their services. The non-expert SP fills out predefined service templates and deploys them on the LOMS Platform. Both the templates and the deployed services can make use of external services that offer contents, context information, or other external services like news from a CP.

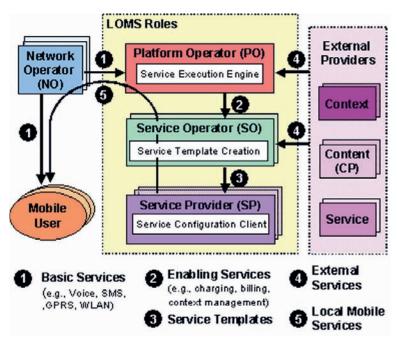


Fig. 15.1 LOMS role model

To analyze the business between the LOMS players, it is important to see which LOMS role is responsible to define the pricing of the service usage and the models for revenue sharing. Three types of charges can be differentiated:

- Traffic charges for accessing the network.
- Service charges for using services.
- Content charges for downloading content.

A key challenge within the LOMS platform is to provide an open interface to existing Charging and Billing systems (CBS) and enable the SPs and SOs to define their usage tariffs and pricing. This has to be in line with the different value elements each partner adds to the service [5].

15.2.2 Categories of Mobile Services and Different Charge Types

There are four main categories for mobile services today: Voice, Messaging, Internet and Content. We added a fifth category "Local Mobile Services" to consider context-sensitive services separately (see Table 15.1). The switching of voice calls is the classical mobile service offered by a Mobile Network Operator (NO). Usually this is offered together with some **basic messaging** services like SMS/MMS. For these services, the NO defines the traffic charges for resource usage of his access network.

Mobile Services	Charge	LOMS
Voice Calls	type	role
National, International	traffic	NO
Messaging		
SMS, MMS	traffic	NO
Mail, Instant messaging, Chat, Blog	service	PO/SO
PoC, Video conferencing	service	PO/SO
Internet Access		
GPRS, UMTS (Cellular)	traffic	NO
WLAN	traffic	NO / PO
Content centric service		
Operator portal (closed community)	service	PO/SO
Mobile portal (open community)	content	SP/ CP
Local Mobile Service (LOMS)		
Local search (e.g., Point of Interest,)	service	SP
Local news (e.g., weather, traffic jam)	service	SP
FindAFriend	service	SP
Blended service with context	service	SP

Table 15.1 Mobile services, charge types and LOMS roles

There are **additional messaging** services like E-Mail, Instant Messaging, Chat, or Blog, which are services already known from the Internet, e.g., via a fixed telephone line (DSL). To make this kind of services available for mobile users, two conditions have to be met:

- 1. The mobile user needs to have Internet Access.
- 2. The additional messaging service has to be accessible for the user.

Typically there are two options for the first condition (1):

- (a) The NO offers Internet Access based on GPRS or UMTS.
- (b) The NO or PO offers Internet Access via WLAN.

Internet access is a basic service for the SO to build enabling services for the SP. For Internet access, typically traffic charges apply per volume (e.g., 10 ct/KByte) or time (e.g., 10 ct/min), which are defined by the NO or PO offering the Internet access.

To meet the second condition (2), the additional messaging service needs to be offered by a player. This could be the NO or PO. With regard to the LOMS role model, this is preferably the SO, because the additional messaging services are regarded as enabling services. The SO or respectively the SP has the option to define dedicated service charges per usage or a monthly subscription fee.

The same applies for more advanced messaging services like Push to talk over Cellular (**PoC**) or mobile Video Conferencing. These services are typically implemented based on the IP Multimedia Subsystem (IMS) [6]. Typically, *IMS-services* allow direct IP connectivity between mobile terminals and application servers using the Session Initiation Protocol (SIP) over any packet-switched mobile network, supporting a multitude of different network access technologies. For LOMS, we assume that these IMS-services are offered either by the PO or SO and are blended with additional content and context-sensitive services by the SP.

Content-centric services are defined by the kind and source of digital content offered to the mobile user. Most mobile NOs have a portal (e.g., T-Zone, Vodafone Live!, i-mode) through which they provide access to digital content like ring tones, images, news, TV trailers, etc. Typically the content is only available for mobile users, which have a subscription to access the portal (**closed community**). Some content is for free, included in the subscription, and other content has to be paid by a dedicated service charge (e.g., $1.00 \notin$ download). The NO will ensure the revenue sharing with the content provider. In terms of the LOMS role model, the NO can also adopt the role of the PO to define the service charges and conditions for revenue sharing.

Another kind of content-centric services are mobile portals, which focus on an **open user community**. These portals are offered by a CP (or SP), which aggregates content for his user community. For a mobile user, there is no dedicated service charge to access the portal, but a content charge is applied per use directly by the CP or SP. This additional payment process is inconvenient, because the user needs an additional payment method (e.g., Credit card, PayPal, etc.) or has to rely on sending a Premium-SMS with fixed charges.

Generally, **local mobile services** can be blended services where a SP combines different services from the other previously mentioned mobile service categories to create a new service for the mobile user. In particular,

messaging services or content can be adapted to the user's contextual situation (location, presence) or his preferences.

Typical examples for local mobile services are a local search, which offers a list and more detailed information about Points of Interest (PoI) near by, or a local news service that offers information on local weather conditions or traffic situations. Another service "FindAFriend" can locate other mobile users registered to this service and provides a route to meet them.

The SP is in this case responsible to define the usage tariffs and pricing for the service usage. The challenge is to correlate the different charges and revenue sharing between different partners when blending the services. For a local mobile service, everything can additionally be dependent on the users' actual contexts.

15.3 A News Publishing Scenario

In this scenario, the SP is a local newspaper publisher, which enriches the delivery of its print media by providing latest news on a mobile portal dedicated to an event (i.e., a film festival). He can connect to a CP, which delivers multimedia news (audio, video) for the film festival.

The SO is a news agency operating a Service Creation Environment and keeps the connection to the PO offering enabling service. For simplicity, we assume that the PO has an arrangement with the main mobile NOs, so that many mobile users can access the mobile portal and benefit from the context-aware service.

One key element to stimulate usage of a "new" local mobile service is to control the pricing for the mobile user. In the ideal case, the SP can configure the pricing dependent on the user's contextual situation (e.g., the current location). For example, the SP may temporarily reduce all applicable charges (e.g., by means of a "Free Zone") so that the user can become acquainted with the offered services.

15.3.1 Charging of Service Usage and Revenue Sharing

Two different kinds of mobile services (i.e., WLAN Internet Access and local Media News Service) are offered in the scenario with different charge types:

- 1. The network traffic will be charged based on the consumed volume (KBytes). The **base rate is 1 ct** / **10 Kbyte**.
- 2. The Media News Service will be charged based on the consumed Service Units (SU). In the scenario, the **base rate is 1 ct / SU**.

As context parameters, we consider the current user **location**, which can be "outside" or "inside" the event area or "free zone". An additional context parameter is the user preference to watch an **advert** before getting the video news message, which can be "yes" or "no".

Examples for rules to grant a context-dependent discount per service are:

(1) WLAN Internet Access Service

{**IF** the user location is "inside" the event area, **THEN** apply a discount of 40% on the base rate}

(2) NEWS Service

{**IF** the user location is "inside" the event area **AND** the user preference is "yes" to watch an advert before the video news, **THEN** apply a discount of 30% on the base rate}

The SP needs to define a set of discount rules, which specifies the actual rates to be taken for the charging requests of each service. The discount for the WLAN Internet Access Service depends on the current user location. No discount is granted for users outside the film festival area. The discount for the Local News Service depends on the user location and his preference to watch adverts. Typically, discounts are higher inside the event area and when the advert option is set to "yes".

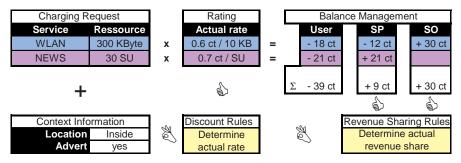


Fig. 15.2 Example for context-aware charging and revenue sharing

The handling of a context-dependent charging request is shown in Fig. 15.2. In the rating process, actual rates are determined by applying the discount rules. In the balance management, the rules for revenue sharing can depend on the discount rules:

{**IF** a discount is granted for the WLAN traffic charge of the user, **THEN** the remaining charge is debited on the SP balance}

Thus the SP can subsidize the user for using the WLAN, but has to compensate the SO.

In the following, we present in detail the **context-dependent payment flow** among all actors. For simplicity, we assume that in all cases 90 % of the traffic charges collected by the SO are forwarded to the PO.

15.3.2 Mobile User Outside the Event Area

The user is outside the event area where the film festival takes place and accesses the Internet via WLAN with his mobile device (Fig. 15.3). He will be fully charged for the network traffic (base rate 0.1 ct/KByte) when he browses the Internet. When accessing the Latest News Service of the SP via the portal, he will be charged on the basis of Service Units for the selected multimedia news (1 ct per Service Unit). For the used resources, we assume that the user has consumed 300 KByte of traffic for downloading a video at a price of 30 Service Units.

Under this condition, the user will be charged 30 ct for the traffic and 30 ct for the selected video news. The SP will earn the service charges and the SO the traffic charges.

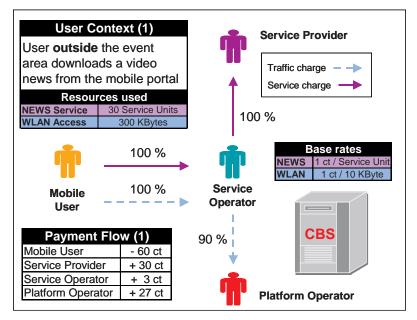


Fig. 15.3 User outside the film festival area fully charged

15.3.3 Mobile User Inside the Event Area

Inside the event area, the user again accesses the portal via WLAN to use the Media News Service of the Service Provider. The context information of the user in this case will lead to a partial charge of his account. The SP will take over 40% of the traffic charges to boost usage of mobile services of users in the event area. Additionally, the SP provides discounts to the service charges for latest news. For example, multimedia news are subsidized by 20% (Fig. 15.4).

The motivation for the SP is that he may have some indirect revenue from sponsoring or branding activities going along with the content. For the mobile user, a context-dependent discount can be a motivation to use the service.

The situation can be even more inviting for the user if he can change his preference to watch adverts before the video news message and get a further 10% discount on the service charge.

The SP can in this case try to get some direct revenue from the sponsor. It can be seen that the definition of pricing and revenue sharing, which is performed by the SP and the SO, is one key to stimulate the service usage and to create a sustainable business for both actors.

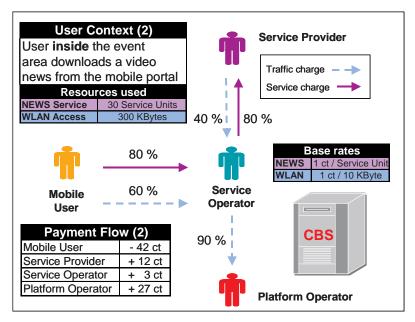


Fig. 15.4 User inside event area receiving a discount for traffic and service charges

15.3.4 Requirements for Convergent Online Charging

The scenario described above shows that a success factor for contextaware services is the ability to perform also the charging of these services depending on the context. But only if the actual charges are transparent to the user a successful business is supposable. This leads to the following requirements for **Convergent Online Charging**:

- Because the pricing of different services can depend on the context situation, the user needs to get an **Advice of Charge** (AoC) prior to service usage.
- Additionally, the user may want to set a monetary threshold value for a dedicated service session to get a notification when the money is spent (**Online Charging Indication**).
- **Credit Authorization** prior to service usage is required by the SP and the SO to ensure that the user is able to pay for the service to be used (fraud prevention and revenue assurance).

The online monitoring of service usage by each partner (SP, SO, PO) should be supported by a **real-time revenue sharing** mechanism.

Additional requirements lead to the term **Convergent** Online Charging or Convergent Online Mediation:

- The charging system should support both circuit-switched (IN) and packet-switched (IP) services in parallel. Independent charging of different bearers for the traffic should be enabled (e.g., GSM, GPRS, WLAN). But also bearer-independent charging of value added services on top of the bearer should be enabled.
- The charging system should support different payment methods and account types. For prepaid accounts, credit control is needed, but also for postpaid accounts a credit limit can be defined.
- The charging system should support session-based and event-based charging mechanisms.

15.4 Extending Charging and Billing Systems

The 3rd Generation Partnership Project (3GPP) specifies the general architecture and principles for offline and online charging of services in mobile networks [7]. There are several approaches for platforms supporting *offline charging* for future mobile services [8], but it must be noted that they don't support interaction during the service usage. A dedicated standard is

published for the application and interfaces of an Online Charging System (OCS) [9], which serves as the basis for our implementation.

A **Credit Control Server** (CC-server) receives charging requests from a Credit Control Client (CC-Client), which resides in the network. A charging request or, more generally, an input event is based on the following parameters:

- 1. **User:** Identifier for the user, usually his calling number (MSISDN) or e-mail address
- 2. **Service**: Identifier of the used service (e.g., Mobile originated Call, SMS, WLAN, NEWS)
- 3. **Resource**: A value for the resources, which should be reserved or have been used (e.g., duration in seconds, count of SMS, volume in KBytes, Service Units).

For such an input event, a **Rating Function** (RF) determines if a valid subscription (account) exists for the user and if the corresponding balance still has coverage for the requested resource usage. Usage Tariffs containing usage rules and cost functions are used for this.

The CC-Server sends a response to the CC-Client containing the granted or consumed resources for the requested service.

There are two different charging functions: Event-based and sessionbased charging. Session-based charging is normally used for bearer or traffic charging. Several intermediate requests are generated during a session for credit control. For credit control of applications, typically sessionbased charging and event-based charging with unit reservation (SCUR, ECUR) are used. Thus a reservation is always required before the service access is allowed and a charging request can report used resources.

The Account & Balance Management Function (ABMF) is used to maintain the actual balance on each account according to the charging requests processed by the Rating Function.

The **Charging Gateway Function** (CGF) generates Call Detail Records (CDRs) for post processing of the requests by a Billing System. Additionally, recharge operations for prepaid accounts will be handled via this gateway. The collection of payments for postpaid accounts is the core function of the **Billing System**. Bills will be generated containing the history of all detailed items for each chargeable event occurred in a certain period of time, typically one month.

To register a new user and sell subscriptions to different services with appropriate usage tariffs, a **customer care** function is needed which is provisioning the RF and ABMF accordingly.

15.4.1 Design of the Online Charging Interfaces

The proposed design and implementation of new interfaces for a convergent online charging system is shown in Fig. 15.5. At the CC-Server, we suggest two different interfaces for charging, i.e., the Parlay X API for Payment [10] and the Diameter standard for Credit Control Application (DCCA) [11] based on the Diameter base protocol [12].

- The bearer (WLAN traffic) is charged via a session-based charging function on the Diameter interface (DCCA).
- The value added service (Media NEWS) is charged via an event-based charging function on a Web Service Interface (Parlay X Payment). In the scenario, this is the Media News Service.

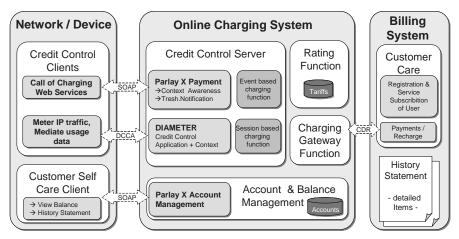


Fig. 15.5 Functional view of the interfaces for the online charging system

To enable credit control of a context-aware mobile service, an additional parameter is needed in the charging request between by the CC-Clients and CC-Server. The *contextInfo* is a list of attribute/value pairs whose format has to be previously agreed upon between the SO and the SP.

The context-aware charging Web Service (**CWS**) is built upon the volume-related payment operations proposed in the Parlay X Web Service Payment API and basically extends them by the additional parameters *serviceID* and *contextInfo*. The *serviceID* is used to identify different services created by the SP with the help of a Service Creation Environment provided by an SO (e.g., Media News Service).

The parameter *contextInfo* is used to evaluate which discount has to be applied according to the discount rules specified by the SP during service creation.

Two additional functions are offered by the CWS to fulfill the above requirements:

First, **Advice of Charge** (AoC) before the actual service usage is able to rate the service usage *prior* to really using the service (AoC-P, see [13]). The implemented AoC-P Web Service calculates service charges based on a provided volume expressed in, e.g., Service Units, while also taking additional provided contextual information into consideration.

Secondly, a function **Online Charging Indication** allows users to set a threshold for service usage. This threshold determines the maximum amount of money a mobile user is willing to pay for using a service. When service usage reaches the specified limit, the user will be notified (e.g., a beep or a direct SMS).

The use of Web Services loosely coupled via SOAP protocol follows the SOA paradigm and enables easy integration into other business processes. Within the LOMS project, the CWS is used along with a composition of other Web Services in BPEL. A detailed description of the underlying Web Service infrastructure is given in [2].

In general, the CWS could also be used for session-based charging of the bearer, but it is more appropriate to use the **Diameter Credit Control Application** [11] for the following reasons.

The IP traffic is metered and usage data mediated by an IP proxy, which normally already supports the Diameter protocol for credit control. To add context awareness, existing optional Diameter attribute/value pairs can be reused. For example, in one of our demo implementations, we take the WLAN Access Point IP address to derive the user location. Of course, some integration is needed to allow the application of discount rules in front of the Rating Function.

Additionally, there are performance limitations when using verbose SOAP messages for session-based charging with real-time credit control.

15.4.2 System Design and Benefits

Besides the design of new interfaces for the Online Charging System, the key challenge is to apply the discount rules, which have been defined by the SP during service creation.

We have chosen an approach on having a generic tariff configuration inside the Rating Function, which is independent of the context ontology and loosely coupled with the discount rules.

For each serviceID, one usage tariff with the applicable base rate is generated by the SO or PO. The serviceID and the base rate are the essential parameters inside the Service Template given to the SP for service creation. Additionally, a concept of **Charging Zones** is used to realize discounts as fractions of the base rate. In the current implementation, 10 different Charging Zones are defined to enable discounts in steps of 10% and have one "Free Zone". For each zone, a dedicated usage tariff is created. The mapping of context information to the Charging Zone is performed with an existing zone mapping mechanism in the CC-Server.

For the SP, this has the benefit that he can dynamically change the discount rules to stimulate the service usage without requiring changes to the usage tariffs, which otherwise could only be carried out by the SO or PO.

In the current implementation, Orga Systems' convergent Billing Systems OPSC GoldTM was used in combination with Orga Systems' Media Control Point MCPTM for the Diameter Credit Control Application (DCCA). But due to the context-agnostic tariff definition, any other already existing Real-time Billing System of a NO can be used to integrating the CC-Server at the front end and configuring the zone mapping.

15.5 Conclusion and Outlook

In future service-oriented business architectures, value is generated on all service levels, i.e., content, service and network. Not everything can be paid via the traffic charges collected by a network operator, as additional charges from different actors apply. The key is to keep a transparent pricing to the user and ensure viable business models for all involved players.

Small Service Providers can have a high impact by generating network traffic and service and content consumption in several user communities. Service Operators can work as aggregators in particular segments (e.g., tourism, health, maintenance, publishing) to increase the multiplying effect on actively consuming mobile users.

We see convergent online charging and billing as a promising solution for reliable third party payment operations able to satisfy all involved parties. The corresponding enabling services suggested here allow contextaware Advice of Charge and charging operations. Summarized, the features are:

- 1. Context-aware online charging of service usage allows the SPs, SOs and POs to define the pricing and revenue sharing.
- 2. Providing Advice of Charge *before* an actual service usage to mobile users will increase their confidence.
- 3. Setting a charging limit to allow a mobile user to define a maximum amount he is willing to pay for a dedicated service usage.

The prototype implementation shows the practicability of our interface specifications and the overall system design.

For the future, it is expected that mobile consumers will become mobile "prosumers" actively creating content or services. Thus everybody can become a SP or CP, which will further challenge the business modeling and charging mechanisms to be developed.

Another important topic is the integration of a settlement procedure between a PO and several NOs into the CBS. In [14] we describe extensions to the Parlay X standards to support authentication and authorization with different roles. This is the basis for the settlement on the Web-Service layer. The integration based on the DIAMETER protocols is still under investigation.

References

- LOMS Project Consortium. Deliverable D2.1 Collection, Analysis and Comparison of Local Mobile Services. January 2007. http://www.dloms.org/ deliverables/ LOMS_D2.1_v1.0.pdf.
- F. C. Bormann, S. Flake, J. Tacken. Convergent Online Charging for Contextaware Mobile Services. In: 2nd International IEEE Workshop on Service Oriented Architectures in Converging Networked Environments (SOCNE 2007), Niagara Falls, Canada, May 2007.
- 3. V. Poulbére. Impacts of New Business Models on the Mobile Multimedia Market, www.hhs.se/cic/roundtable2003.
- 4. S. Panagiotakis, M. Koutsopoulo, A. Alonistioti. Business Models and Revenue Streams in 3G Market. In: 14th IST Mobile and Wireless Communications Summit, Dresden, Germany, June 2005.
- 5. R. Kumar, I. Grgic Gjerde, T. Smura, H. Hämmäinen, T. Tesch. Business Models in Telecomunications, ECOSYS Project, October 2004.
- 6. Gonzalo Camarillo, Miguel A. Garcia-Martin. The 3G IP Multimedia Subsystem. Second Edition. John Wiley, New York, 2006.
- 7. 3GPP TS 32.240 Group Services and System Aspects; Telecommunication Management; Charging management; Charging Architecture and Principles.
- M. Koutsopoulou, A. Kaloxylos, A. Alonistioti, L. Merakos. A Platform for Charging, Billing & Accounting in Future Mobile Networks. In: 1st International Workshop on Next Generation Networking Middleware (NGNM04), Athens, Greece, May 2004.
- 9. 3GPP TS 32.296 Online Charging System (OCS): Application and Interfaces.
- 10. European Telecommunications Standards Institute. ETSI202 391-6 V1.1.1, Parlay X Web Services; Part 6: Payment, The Parlay Group, March 22, 2005.
- H. Hakala, L. Mattila, J-P. Koskinen, M. Stura, J. Loughney. Diameter Credit-Control Application, Internet Proposed Standard RFC 4006, August 2005, Chapter 5, Session based credit-Control.

- P. Calhoun, J. Loughney, E. Guttman, G. Zorn, J. Arkko. Diameter Base Protocol, Internet proposed standard RFC 3588, September 2003. Chapter 9, Accounting.
- F. C. Bormann, S. Flake, J. Tacken, C. Zoth. Towards Context-Aware Service Discovery: A Case Study for a New Advice of Charge Service. In: 14th IST Mobile and Wireless Communications Summit, Dresden, Germany, June 2005.
- 14. F. C. Bormann, S. Flake, J. Tacken, C. Zoth. Third-party-initiated Contextaware Real-time Charging and Billing in an Open SOA Platform. To be presented at: 3rd Int. IEEE Workshop on Service Oriented Architectures in Converging Networked Environments (SOCNE 2008), Okinawa, Japan, April 2008.
- 15. The Parlay Group, Web Service Working Group. Parlay Web Services Business Models, October 2002.

16 Rights Management for User Content

György Kálmán and Josef Noll

16.1 Introduction

With the spread of always-online Internet connections, digital user content and upcoming tools or services for content sharing induced a change in user behaviour. Traditional user and provider roles are not separated any more. The end user is creating his content and sharing it over the network. Rights management in the home area is composed from two, possibly disjunct problem areas. One is the management of content, which the user has been purchased from commercial sources, the other is the management of user-owned content.

User created content induces the need for a rights management solution, which keeps the user and his content in focus. Current solutions offer services only for industrial customers and are only dealing with the traditional consumer role of the user. The more active, content producing users need a tailor-made solution to handle their content. Such a solution will enable fine grained rights control over distributed material in an easy and secure way.

Social life over the internet is becoming more important, in an interconnected world, network presence – visibility and acknowledgement of other people – is a major driving force. A user may want to share pictures with his friends and family, with a society or just with one person. Currently, the user has a wide variety of possibilities to share content, but until now, no fine grained right management solution was designed with user needs in mind, and with support of home content.

16.2 Background

The bigger user base results in the average knowledge in IT technologies is sinking. This raises the need for easy-to use systems, since this is not longer a matter convenience, but rather a limiting factor for success. Wide variety of terminals and extending range of user devices are requiring content adaptation (nearly the same user experience on different terminals) and integrated security methods (hard to set up public-key authentication or to type in 15 character passwords on a media player). For some purposes, it would be beneficial to use seamless authentication, like it is implemented for Wireless Application Protocol (WAP) services of the operator and selected third parties. Because no user interaction is needed it is recommended to use it in personalisation and content adaptation services [1]. If security is not critical, methods similar to cookies could be used, where after one successful authentication, the system is keeping the user logged in for a certain period. Alternatively, Single Sign On (SSO) solutions can be used, where after the user authenticates himself towards the SSO service, further authentication requests will be handled by the system.

Content adaptation has a growing importance in pervasive computing, since terminals with very different capabilities are used to access the same information sources. Beside the technical problems associated with conversion, the commercial content protection solutions usually do not provide a method for content transformations.

To provide good user experience, these incompatibilities may be hidden with deploying a DRM broker into the home network, which can cooperate with user devices and is able to distribute licenses in a secure and easy way. A central device for controlling a home network was introduced in the IST ePerSpace [2] project, which provides service discovery and content adaptation services for compatible devices. However, the ePerSpace solution lacks support for content management.

Although there are solutions for key distribution and license management, most of them are not optimized for the special circumstances in a home environment. Either the service is not user friendly (e.g. key directories of PGP) or use a third party pay service (e.g. VeriSign) and in general, are not designed for the constraints of the user environment: mobility, battery use, computational power and trustworthiness. Mobility can be addressed with secured transport protocols to provide secure and easy access to home content from the internet side.

Entertainment devices usually have limited computing capability, thus they might be supported through a specific network device which is able to carry out complex cryptographic operations and exchange the generated information with other parties using a secure and easy method. A solution to computational problems and trusted devices could be to deploy smartcard based authentication in the home environment [3] (Fig. 16.1).

In this paper we show a solution which can bridge the gap between the DRM solution shown in Popescu et al. [4] and the smartcard based authentication architecture in Pujolle et al. [3] in order to enable cheap, easy and secure user authentication and personalization services [5]. As an extension of the original concept, a new user scenario is shown, where the user is sharing his own content to other users through his own home network and with the possibility of integrating external services (MySpace, Face-Book, etc.).

As a new functionality, the concept of an interconnecting service is shown, which is using a (preferably) trusted entity for rights object generation. The objective is to provide a flexible solution, which can interoperate between the different service providers and make it possible for the user to select users of the different services from one interface and share content for them. The system then distributes the appropriate rights objects using the third party services or other solutions.

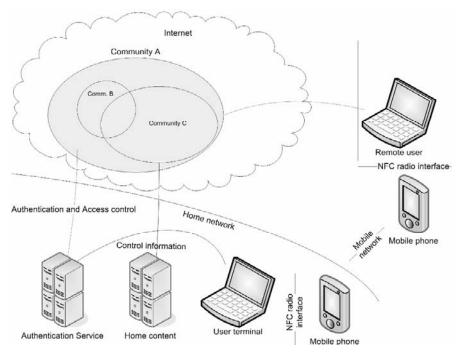


Fig. 16.1 Home network with access control and out-of-band key distribution

First, an overview is given about typical devices and their capabilities in a home network, then an overview about smartcards, their possible area of use and possible candidates for widespread use. A possible secure and widely available device, which can help to overcome the problematic deployment of a smartcard infrastructure, the mobile phone and its Subscriber Identity Module (SIM) capabilities are shown.

Two user scenarios are explained, where the first shows the typical case of commercial content distribution and the use of a DRM broker for a better user experience. The second shows the viewpoint of a content creator user, who wants to share his own content for his friends: photos from the last hiking trip for his friends, a video clip, where he is asking for advice from a workshop and sharing the unfiltered catalogue of his pictures within his family.

Possible solutions for these scenarios are shown and an evaluation of the proposed architecture is given and areas of further investigation described.

16.3 Devices in the Home Network

A home network can be composed of PCs, media players, mobile phones, storage units, STBs or other devices. Most of these devices are mobile and move between different networks. Wireless networks also made it easier to welcome guests on the home network.

The problems begin with securing access to a home network. Mostly there is no or just weak security applied on those, so they are wide open for malicious intruders. For example, lots of WLANs use no encryption at all or employ the compromised WEP standard. Networks with open access or even with WEP secured access open for malicious attacks on the user data from any place within the coverage of the wireless cell.

Setting up a secure network may be a hard task, since keys have to be transmitted and devices have to authenticate themselves. This may be done by using out of band key delivery methods (like using an USB stick or in an SMS via the mobile network), in case of automatic key delivery, only the communication is secured, but the client does not prove, that he is allowed to connect. Even if the user is able to do this process, convenience considerations might cause him to neglect security. Also, currently, the user may decide to grant access or not, but inside the network it is extremely rare to use some kind of additional access restriction. This means, that either no access is given or the guest can access practically all network resources. While keeping secure access, content adaptation is becoming more important. In order to ensure good representation of content, profile management methods, such as UAprof (WAP Forum, 2001) were introduced. This enables content creators to define content representation based on generic rules and the serving system can adapt these based on the transmitted terminal profiles.

Content stored in a home network may be also adapted to the different devices, to ensure good results (e.g. creating lower resolution video for a portable media player from a digital satellite stream). Content adaptation can be problematic, because current DRM solutions usually do not allow changes in the content. If a device could provide connection between the content providers rules and user needs, the adapted and legal content would be available on any user device. Such a device could act as an end entity for the content provider and hide the inner network of user devices.

16.4 Rights Management

Since users are starting to create and share content with others, the home infrastructure has to support some kind of rights management. This includes not only storage of acquired licenses from content provider companies, but taking care of own content. Home networks store a great deal of personal information which should be secured.

Based on various roles of a user in a certain context, a need to share with a specified group of users arises. This can be done by introducing community content access, based on group authentication. A design with the end user in the focus is needed to enable secure and easy sharing of content over the internet. This means that while preserving ease of use, the system has to use strong encryption, group authentication and efficient key management. Group authentication is essential to enable sharing between different user groups based on various properties, like friends, school classes or other interests.

The basic problem of home DRM is, that these systems usually rely on *compliant devices*. A device needs to meet certain requirements in order to get accepted by the system. Compliance raises a problem with the restricted and optimized nature of home devices. If individual authentication is used, public key operations need to be carried out, because mutual authentication is required between the DRM system and the terminal. This could be problematic for simple devices, like an MP3 player and resource consuming for a device like a PDA.

A DRM solution for home networks is proposed in Popescu et al. [4], where creating device domains in the home environment is shown. This paper points out, that problems associated with the mobile environment (battery powered consumer devices in particular), and the possibility of reducing the number of expensive calculations. So, the use of a designated cryptographic device would be beneficial.

Content adaptation has a growing importance in pervasive computing, since terminals with very different capabilities are used to access the same information sources. Beside the technical problems associated with conversion, the commercial content protection solutions usually does not provide a method to make transformations without quality loss.

The use of group authentication can help to overcome the problems associated with content adaptation and personal content sharing. This solution fits much better to the general use of home devices, because in this scenario, a device has only to prove, that it is part of a group, which can be done by simple hash calculations for example.

After authenticating the devices, also securing of the transmission environment is advised. This could be done by negotiating symmetric session keys or calculating hash values for example.

It cannot be assumed, that all devices have cryptographic hardware and tamper resistant hardware. This can be solved by adding a smartcard into the system.

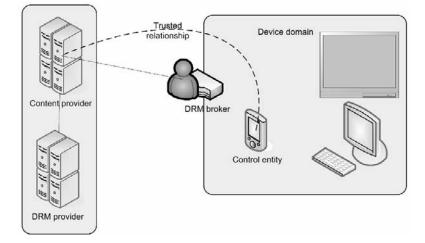
16.5 Usage Scenarios

In the introduction, two possible problem areas were shown.

- *Commercial Content:* The content is purchased from a commercial source, it is equipped with a DRM solution to enforce the owner's rules over the usage in the end user's system.
- *User Content:* The content is created by the user and shared over the internet. The local (home) infrastructure is playing a key role, no commercial DRM is available.

These problems can be framed into a user scenario, where possible problems and solutions can be shown. First, we will show a scenario for the commercial content. This one is more familiar and has more constraints.

16.5.1 Commercial Content



The main actors of this scenario are well-known (Fig. 16.2):

Fig. 16.2 User scenario with commercial content

- *Content provider:* The entity, which is making the content available for purchase for the end user. It represents the retailer. For easier presentation, the content creator/owner is represented by the retailer.
- *DRM provider:* A supporting entity, which delivers a software solution to ensure, that the content is used as it was defined by the Content provider.
- *End user:* The consumer, gets certain rights for the purchased content delegated by the Content provider and enforced by the DRM solution of the DRM provider

In this classical view, the user faces severe limitations. The lesser problems are for example, that in many countries, it is a customer right to create backups from content purchased (like backing up a CD) and the DRM solutions usually are not permitting to copy the content without quality loss.

The more important problem of compatibility arises, when the user is buying content from different providers and wants to use it on different devices. As the home networks are getting more complex, and digital media is being used in nearly all kinds of entertainment devices, the user faces problems with using legal content on various devices.

An example could be a song with Apple's FairPlay, where the song can be uploaded to an iPod, but cannot be converted to MP3 without quality loss. This means, that the user cannot play the song on a networked media player, a home theatre PC or a different mobile device. Also, it cannot be guaranteed, that the user will be able to play a piece of content, for example, 10 years after the purchase. Since he is not able to convert it without quality loss, a new standard can draw producer attention, and support may be stopped in years. The same problem arises, if the manufacturer of a specific device is shutting down and stopping support services. In a bad case, the user won't be able to enjoy the content after the (usually short) lifetime of the device.

Our concept in this scenario introduces new entities into the system, and is using the capabilities of the current home networks. In order to enable a more flexible content use, the following entities are introduced:

- *Device domain:* A group of devices formed based on their relationship to the user, for example devices of a Personal Area Network (PAN) or devices of a home network (media players, HTPCs, MP3 players, etc.).
- *Control entity:* A designated device, which controls a device domain, preferably on-line and equipped with easy-to-use user interface.
- *DRM broker:* A device, which plays the role of the end device for the Content provider and masks the internal devices while respecting the provider defined rules of content use.

The DRM broker can mask the internal network, so the user would be able to use the purchased content in any of his devices, irrespective of the provider of the content or the manufacturer of the device. This functionality can be integrated into different devices based on the provider's preference or other requirements. A natural solution could be to include this service into the home gateway of the user (e.g. ~modem, router). Depending on the implementation, this can be done with the current devices.

The easier problem is to solve the additional resource needs of these services, like flash memory in the router and a bit more CPU power. Harder problems are associated with the commercial requirements. For example, trustworthiness, license management and revocation\cite{rightmanphone}. At the moment, the SOHO routers or modems are not equipped with the necessary devices to provide safe endpoint services for commercial suppliers.

The problem of trustworthiness needs a trusted device in the home network. Since routers don't carry such a device, an other device has to be selected. In our scenario, this task is given to the SIM in the mobile phone. It provides secure storage, revocation capabilities and user identity management. Certainly, the trusted device needs to add an internal DRM to the content delivered to local devices in order to keep rightful usage. This means, that the original DRM has to be removed and a new has to be installed on the content. The legal aspects of DRM handling are out of scope of this paper.

Possibility of manipulating DRM protected content depends on the content providers, this problem does not arise in the second scenario, where the user plays both the role of a content provider and a consumer.

16.5.2 User Content

This scenario shows the upcoming situation of a user, who is sharing a piece of his own content to a friend. He is selecting the appropriate users on the user interface, where he has all of his contacts from the local address book, FaceBook or other online services, of the home right management service and the access keys are delivered to each of them. Content will be accessible through the user's own broadband internet connection or uploaded to a third party provider [7].

The actors in this scenario are more like a Peer-to-Peer (P2P) system, where every user can be a provider and vica-versa. In an explicit situation, the following actors are present:

- User
 - Content consumer: Consumer of the content
 - *Content provider:* Sharing his content to the Content consumer. He made the content accessible and selected the appropriate users.
- *Home Gateway:* A device, which is connected to both the LAN of the Content provider and to the internet. Has the possibility of granting access to content stored in the home network.

This minimal system is capable of sending out an access key to a remote user and letting access to local content [8]. But, there are several problem points: both the user's and the content provider's identities are only *assumed* [9]. Key delivery is done over the local internet connection and as such, possibly eavesdropped. The home gateway is getting access information on the LAN, which can be a problem source in certain situations (e.g. WEP secured Wireless LAN).

In order to create a more secure system, additional measures are required. New entities are introduced to ensure user identity and alternative delivery methods are included (Fig. 16.3).

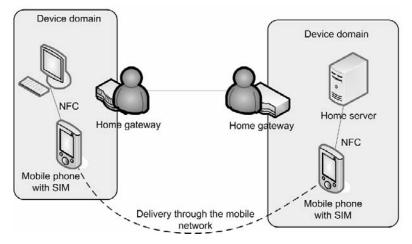


Fig. 16.3 Content sharing between users

- SIM: The Subscriber Identity Module of the user's mobile phone [10]
- *Mobile phone:* Provides a user interface for access key management and alternate delivery network
- *NFC:* Near Field Communication (NFC) interface for very short range transmissions, for example WLAN access setup, access control data exchange between the mobile phone and the home gateway

With the inclusion of the mobile phone, our network is now capable of delivering keys through the mobile network (the keys won't pass through the local internet uplink) and has a trusted device, which is can hold access information in safe storage. As an addition, this phone is capable of generating access control objects (keys, certificates) with internal, tamper resistant routines and is capable of delivering them to local devices and remote devices as well.

The last major addition is, that the mobile phone can represent the user in a limited, but better-than-nothing way. The mobile phone is a personal device, and it can be assumed with a good probability, that if the user knows the PIN of the phone and the access key was delivered to the good number, the actual human being reading the information is the one, who the owner wanted to select.

Out-of-band key delivery would be possible with other solutions, for example using Bluetooth for short range exchange or a USB stick, which can be carried around. The mobile phone provides a more pleasant alternative, since it has a separate network connection, a user interface, which is common for the user and the possibility of revocation [11]. So, for example the owner can select his girlfriend's FaceBook user and grant access to a flickr picture with the key delivered via the mobile network and the access information delivered via the FaceBook messaging service. The mobile phone will generate a key, then it will send it to the girlfriend's mobile number, then the user puts the phone close to his laptop, where the right management client is sending the access information (e.g. URL) to the designated user account on FaceBook and uploads the DRM protected picture to flickr, which is only accessible for the girlfriend.

16.6 Authentication and Encryption

In both scenarios, security is playing a key role. The content needs to be protected against eavesdroppers or other attacks.

Most devices do not have extensive encryption capabilities and a secure infrastructure, they may rely on external units, like a smartcard. A smartcard is tamper resistant, which can support complex encryption functions and provide them to compatible devices.

In Pujolle et al. [3] a smartcard is shown, which implements the Extensible Authentication Protocol (EAP) stack in hardware thus providing high security on a widespread protocol family for WLAN authentication.

While these hardware elements provide good security capabilities, it can be problematic to add those to all the devices in the home network. Besides the costs to equip every single node with a smartcard reader, compatibility issues and additional battery powered devices for certain hardware will make the smartcard solution difficult.

To keep the advantage of a tamper resistant cryptography device and keep costs low, we propose to use the mobile phone's SIM to calculate and the phone hardware to distribute keys for devices.

The phone is becoming a permanent part of the user's personal area. In many cases the handset is already part of the user's identity, because of it's communication services, look and important the role in social connections. Users are taking care of it, since a phone holds a great deal of social and personal information.

According to ETSI [12] it could be possible to use the SIM as a fully featured smartcard as the SIM is capable of storing keys and providing cryptographic functions for third party services, not only for mobile providers.

While the phone is capable of generating a key, the problem of key delivery still remains. If the user has to connect the phone via USB or Bluetooth, it can be problematic, since Bluetooth needs pairing and USB is not supported by a considerable amount of devices. To solve this problem, we propose to use NFC technology to transmit encryption keys between devices. NFC is a short range communication technology based on RFID, but with more limited range and the possibility of using active devices on both sides. An NFC reader adds only a small cost overhead to devices, does not need to be powered continuously and provides contactless transfers for very limited ranges.

Through the mobile phone, the user has full control over the identification process either based on the location e.g. putting the phone close to the reader or on knowledge e.g. typing in a PIN when requested by the remote service.

A key problem is the correct selection of the identifier to be used in a transaction. This can be done either by profiles or by asking the user to allow access to the data, requested by the service.

The public key of the phone represents the root trust in the system. The key pair can be placed to the SIM either by the mobile provider or other, verifiable source, to ensure correct user identity association.

If the private key of the SIM gets compromised, the identifier can be revoked by the identity provider and the user can get a new key without losing access to the services. The remote revocation and user control makes the SIM an ideal device for making payments and gaining access to services.

16.7 Service Architecture

We propose to incorporate the device domain management capabilities and the EAP capable smartcard functions. The EAP family is used for easier cooperation with current network authentication technologies.

With using the SIM's cryptographic functions (Pujolle et al., 2003), we build a device domain, and distribute these keys through the NFC interface.

The constraints, the system has to face are

- continuous network connectivity cannot be assumed between the members of the domain,
- there are no secure clocks in the system,
- no cryptographic hardware is available in the devices,
- key management must be efficient even for large number of devices.

The CPU power of current smartphones makes possible the use of public key operations and so act as a proxy between the provider and the user devices. The provider can be either the DRM broker on the home network or external content providers.

The DRM broker handles rights associated to local and user created content outside the home network. This entity certifies approved devices and revokes expired or compromised ones. No global device identification key is proposed because the phone can deal with the domain's internal right management issues.

This lowers the resource needs at commercial right management providers and also keeps user privacy on a higher level, because he does not have to disclose, what kind of devices he is using. With a DRM broker and an always online phone in the system, we can also extend the proposed systems functionality to physical media, like DVD-s since the networked media played is connected to the broker, which is accessible for example through any mobile IP service.

If a new device is added to the domain, a request is shown on the display of the phone and requires response from the user. This ensures, that access is only granted, if the remote party gets a correct key and in addition, the user confirms his will to permit access. This can be requested once or any other period, based on user preferences.

We recommend the use of NFC interface for distributing keys out of band. With this short range transfer method it is possible to allow the phone to negotiate or generate an authentication and encryption key for the user device, and send it to the mobile device, where no expensive cryptographic methods are needed.

The loss of the mobile phone does not compromise the system's security, since the SIM can be disabled remotely (if the intruder wants to generate a new key, they have to connect to the network). After getting a replacement, the existing keys of the domain will be revoked and the user has to distribute them again.

Usability of the proposed system depends mainly on the easiness and security of key distribution. In the demo system we use either NFC technology to deliver keys to local devices or the mobile network for remote users.

Local key delivery can be accomplished with NFC, because it has very limited range and is convenient for the users, just to put the phone close to the device they want to exchange a key with.

To enable remote access to home content, it is possible to send the access key out of band, via the mobile network to the remote user's phone, where he can use the NFC interface to download the key to the terminal, he wants to use for content access.

One of the key factors of a user centric system is to enable the safe and easy delivery of keys to other users. In order to demonstrate the usefulness of the mobile phone in this task, we created a prototype system, which enables key exchange via NFC or delivery via SMS to an other phone. The other phone can also forward the key for the remote user's terminal. In both ways, the key is delivered out-of-band, which provides better security, as access control information is not transmitted through the network, where the content is accessible.

16.8 Future Work

The current prototype is using the SmartMX chip instead of the SIM for key storage, which limits the possible range of devices. This is a technological limit, which is in the process of being resolved. Nodes need to be equipped with NFC readers to enable key transfer with this technology.

NFC readers are not usual in the home environment. The security of the system depends on the tamper resistance of cryptographic functions on user devices.

Our proposal shows an improvement over the original idea of Popescu et al. [4] by using the possibilities of the mobile phone and the inclusion of a DRM broker, which act as a gateway between different DRM solutions and acts like a home agent for the user's right entities. A possible drawback of using the SIM is that the mobile providers usually do not allow access to the SIM in order to ensure correct functionality of the network.

The authors want to point out, that by using the SIM as secure storage and executing signature and session key generation routines over the SIM (which would be included by the operator on the EAP capable SIM) does not interfere with any networking function of the phone while keeping the advantage of being a widespread device which lowers the introduction costs.

Storage may be also limited, but since an encryption key (for example the master key for adding domain members) can be quite short, well under one kilobyte, even current SIM capacities seem to be enough, but also, high capacity SIMs are already on the horizon [13].

NFC technology is just entering the contactless market, so additional tests are required to test its security against various attacks.

By default, it is a hard problem to ensure a user's identity. As shown by Bhargav-Spantzel et al. [9], a typical Identity Management system has three types of trust models:

- Pairwise: the two entities have direct connection,
- *Brokered:* the two entities are reachable via a network of direct connections,
- Community: for common agreements.

The system shown in the second scenario is fully operational with the minimal entity implementation, where a PGP-like web of trust solution could provide limited identity management (community model) and better-then-nothing security for personal content. The inclusion of the mobile phone offers an easy solution for providing a more secure and easy way for user content management and the possibility of the more exact identity management models (pairwise and brokered).

Future work will focus on implementing a home right management system, where the mobile phone will play the role of a trusted cryptographic device.

16.9 Conclusion

This paper provides an architecture of rights management for home content. While current solutions are device centric, our solution supports both an I-centric and a community centric approach. Two user scenarios are shown, where commercial and personal right management problems are elaborated. With the focus on the user created content, problems and possible solutions for various security requirements are shown.

We have shown that the mobile phone with the SIM card has the potential to provide strong encryption services, being applicable for securing home content. Key generation and distribution are the main functions of the phone, supported by the capability to interconnect devices in the home network. It may also be used to enable access to guests and store device profiles for content adaptation.

Because the phone is practically always online, update and revocation of profiles or keys can be done remotely and nearly instantly. The SIM is trusted by mobile providers and can be the tamper resistant device, which the user needs for building an I-centric rights management infrastructure.

In an always online environment, with networks holding more and more personal information, the user has to be able to control access to his own content.

References

- 1. Chowdhury, M. M. Rahman and Noll J.: Service interaction through role based identity. In: Proceedings of WWRF 17, Heidelberg (2006)
- 2. IST ePerSpace: Towards the era of personal services at home and everywhere. http://www.ist-eperspace.org/ Cited 10 Oct. 2007

- Pujolle, G., Urien, P., Loutrel M.: A smartcard for authentication in WLANs. In: Proceedings of the 2003 IFIP/ACM Latin America conference on Towards a Latin American agenda for network research, Available via ACM. http:// portal.acm.org/citation.cfm?id=1035662.1035673 Cited 10 Oct. 2007
- Bogdan C. Popescu et al.: A DRM Security architecture for home networks. In: Proceedings of the 4th ACM workshop on Digital rights management (2006)
- 5. Heikkila, F.: Encryption: Security considerations for portable media devices. IEEE Security and Privacy Magazine. July/August, 22–27 (2007)
- WAP Forum: UAProf specification. Available online http://www. openmobilealliance.org/tech/affiliates/wap/wap-248-uaprof-20011020-a.pdf Cited 10 Oct. 2007
- Dusparoc, I., Dahlem, D., Dowling, J.: Flexible application rights management in a pervasive environment. In: Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, 2005., Available via ACM. http://portal.acm.org/citation.cfm?id=1048928.1049593 Cited 10 Oct. 2007
- Geambasu, R. et al: Homeviews: peer-to-peer middleware for personal data sharing applications. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Available via ACM. http:// portal.acm.org/citation.cfm?id=1247508 Cited 10 Oct. 2007
- Bhargav-Spantzel, A., Squicciarini, A., Bertino, E.: Trust Negotiation in Identity Management. IEEE Security and Privacy Magazine. March/April, 55–63 (2007)
- Tarkoma, S., Prehofer, C.: Composable mediation for security-aware mobile services. IEEE Communications Magazine. July, 58–65 (2007)
- 11. Merabti, M., Llewellyin-Jones, D.: Digital rights management in ubiquitous computing. In: IEEE MultiMedia Magazine, Available via ACM. http://portal.acm.org/citation.cfm?id=1130720.1130816 Cited 10 Oct. 2007
- ETSI: TS 102 350 V7.0.0 Smart cards, identity files and procedures on a UIC. In: ETSI Technical Specification, ETSI (2005)
- M-Systems: M-Systems and microelectronica announce plan for 1 gigabyte SIM cards by end of 2006. In: Press Release, 3GSM World Congress, Barcelona, Feb. ~15 (2006)
- Abbadi, I. M., Mitchell, C J.: Digital rights management using a mobile phone. In: ACM International Conference Proceeding Series; Vol. 258, Available via ACM. http://portal.acm.org/citation.cfm?id=1282138 Cited 10 Oct. 2007
- Broy, M.: Software engineering from auxiliary to key technologies. In: Broy, M., Dener, E. (eds.) Software Pioneers, pp. 10–13. Springer, Heidelberg (2002)

Part VI Systems

17 Distributed Cross-Layer Approaches for VoIP Rate Control over DVB-S2/RCS

M. Ángeles Vázquez Castro, David Pradas Fernández, and Catherine Morlet

17.1 Introduction

The new standard for Digital Video Broadcasting for satellite applications (DVB-S2) optimizes the performance and the flexibility of the transmission. It improves the capacity gain in about 30% over DVB-S. This gain is mostly thanks to the introduction of an adaptive physical layer, which is implemented using Adaptive Coding and Modulations (ACM). It should be realised that increase of system capacity owed to ACM shall only be used in practise by an actual system if physical layer dynamics are appropriately translated to higher layers of the protocol stack so that the overall system can take advantage of the time-variant capacity.

We will focus on DVB-S2/RCS systems offering unicast services, thus transmitting simultaneously various types of traffic, e.g. web browsing, VoIP, real-time video, video streaming, etc. As we will see, each type of traffic has different requirements on average delay, delay jitter, average rate, peak rate, burst size, possibility of retransmissions, adaptability of the application layer, etc. Since VoIP is gaining popularity due to its cost benefits, our aim is optimising VoIP transmission taking into account the adaptability of the physical layer to the system and channel dynamics. This is evident from the thrust exhibited by enterprises and service providers to migrate towards a converged IP based network

For a communications network to comply with the stringent QoS requirements of VoIP traffic, a number of cross-layer optimization techniques have been recently proposed for improving the overall performance [1,2]. However, we focus on a different approach tailored to the satellite scenario. In particular, we analyze two cross-layer codec rate adaptations, both performing in an end-user-centric scenario. First, we propose a bank of codecs be available at both transmission ends and a codec switch is performed driven by the RTCP report. In this case a transport-to-application layer cross-layer information flow is required. Second, we propose the use of adaptive wideband codecs (inherently cross-layer across the network).

The chapter is organized as follows. Section 17.2 describes the satellite system model (topology and air interface). In Sect. 17.3, it is defined the QoS model and the traffic classification architecture. Section 17.4 introduces our first proposed cross-layer rate control, based on RTCP reports. Section 17.5 introduces our second proposed cross-layer approach, based on adaptive codecs, inherently cross-layer. Section 17.6 presents the delay budget model as well as the E-model for narrowband codecs, which is used in Sect. 17.7 where numerical results are shown. Finally, conclusions are drawn showing the potentiality of both approaches for a satellite scenario.

17.2 System Model

In this chapter, we consider a DVB-S2/RCS air interface (see Fig. 17.1), where network is controlled by the NCC (Network Control Center) entity, which has associated a Hub in order to transmit to the satellite. The satellite is assumed to be transparent, only for the sake of simplicity, and it follows a Geostationary Earth Orbit (GEO).

In the forward link, we assume DVB-S2 [3]. It allows ACM (Adaptive Coding and Modulation) down to a per-time slot basis adaptation depending on the SNIR (Signal-to-Noise-plus-Interference Ratio) at the destination terminal. Therefore, TDM/FDM (hybrid Time and Frequency Division Multiplexing) is implemented, and each time slot transmits a DVB-S2 physical layer packet, which has a constant amount of coded symbols but a time-location dependant number of information bits and symbols with consequent variable transmission time. The ACM mode to be used in each time slot depends on the SNIR (Signal-to-Noise-plus-Interference Ratio) at the destination terminal. We assume transmission in the K_a band (20–30) GHz) and therefore rain is the most affecting atmospheric event. As the figure shows, the satellite transmits several beams to the Earth, covering non-overlapping geographical zones. Without loss of generalization, we consider constant transmit power and fixed beam coverage and bandwidth. Each beam is divided into areas where channel characteristics are assumed to be correlated (i.e. similar weather conditions). We differentiate between two types of areas depending on channel conditions. Areas in good bad channel conditions; where users are usually in clear sky. And areas in bad channel conditions, where users are affected by rain attenuation. We consider the adaptive physical layer model presented in [4].

DVB-RCS implements MF-TDMA and adaptive coding only is allowed [5]. It enables to have bidimensional framing: every time-frequency window is portioned into carriers, superframes, frames, and slots. The superframe is the basis for the timing of all resource control processes, since the Terminal Burst Time Plan (TBTP) is transmitted every superframe. We assume VoIP packets are encapsulated onto ATM packets.

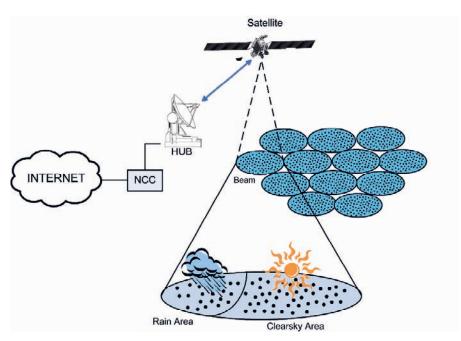


Fig. 17.1 DVB-S2/RCS system model

17.2.1 Centralized vs. Distributed Approaches

Codec selection algorithms can be classified into two different types: network-centric or centralized approach and end-user or distributed approach. In the first case, the network is responsible of controlling the delivery of the best possible speech quality to all the mobile terminals at any instant. In the latter case, the terminal has complete responsibility on the link quality measurements and the control of the speech coding mode. In this case, the terminal monitors the link quality measuring different statistics (such as delay, jitter...). When it realizes the Channel State Information (CSI) changes due to weather conditions, it sends a message to the other end terminal specifying the codec to be used to adapt to channel variations. In a satellite system, as shown in Fig. 17.1, a centralized approach seems to be more appropriate since the system is controlled by the NCC. However, in our work, we have chosen a distributed approach, which is more adequate to the Internet backbone network connected to the satellite system.

The drawback of the distributed approach is that the full benefit can only be obtained if all the terminals implement the advanced mode selection algorithms. However, the objective of this study is to quantify the performance of a distributed approach.

17.3 QoS Model

DVB-S2/RCS does not standardize QoS provision. Instead a number of tools are given at access layer which can be used to provide the desired QoS guarantees. We assume a DiffServ model for our satellite subnetwork thus enabling easy networking with IP external networks. Moreover, we assume that VoIP is primarily transmitted via the highest priority Class of Service (CoS); EF (Expedited Forward) and we provide the delay guarantees of our cross-layer design. EF is implemented to provide premium service to those applications that require a robust network treatment. The main goal of EF is providing low loss rate, guaranteed end-to-end delay, low-latency, low-jitter and an assured bandwidth service.

A part from EF, which we have mentioned above, we propose two more traffic types for our DiffServ architecture. Assured Forward (AF), used in applications that need low loss rates but no guaranteed delays and jitters, such as Audio/Video Streaming. And Best Effort (BE), corresponding to other IP packet flows without priority.

We use the packet scheduler architecture proposed in [6]. This chapter presents a cross-layer approach for the design of the forward link packet scheduler that introduces fairness as a tunable parameter. This approach makes it possible to dynamically adapt the scheduler behaviour in order to guarantee fairness depending on the channel conditions. It allows choosing different levels of fairness between users. The proposed algorithm also supports differentiation of services that comply with the requirements for implementing QoS. Therefore, it is possible to define different scheduling policies depending on the system necessities.

17.4 RTCP-Driven Cross-Layer Distributed VoIP Rate Control

Our first proposal of cross-layer distributed rate control for VoIP flows relies on the RTCP reports [7]. We propose a bank of codecs be available at both transmission ends and a codec switch is performed driven by the RTCP report. In this case a transport-to-application layer cross-layer information flow is required. The architecture is sketched in Fig. 17.2, where the required information flow is also shown.

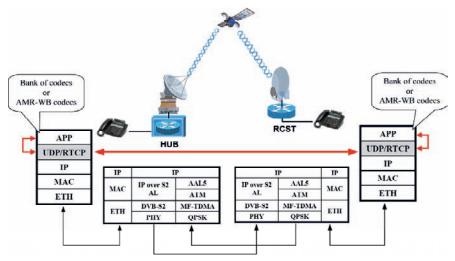


Fig. 17.2 RTCP-driven and AMR-WB architectures

17.4.1 RTCP Reports

RTP receivers provide reception quality feedback using RTCP report packets which may take one of two forms depending upon whether or not the receiver is also a sender. The only difference between the sender report (SR) and receiver report (RR) forms, besides the packet type code, is that the sender report includes a 20-byte sender information section for use by active senders. The SR is issued if a site has sent any data packets during the interval since issuing the last report or the previous one, otherwise the RR is issued.

We assume that the following information is extracted from these reports at each end of the transmission and sent up to the application layer: jitter, delay and packet loss. Jitter is measured by RTCP and included in the RR messages sent by the receiver. As this value is measured in sampling units, in order to convert to time units, one must divide by the sampling rate of the media codec. Delay between two peers can be calculated with the difference among three times, such as, Delay since Last Sender Report (DLSR), Time of Last Sender Report (TLSR) fields in RTCP receiver report packets and report receiving timestamp. Inter-arrival delay jitter and packet loss rate are obtained from the inter-arrival jitter field and cumulative number of packet lost fields in RTCP receiver report packet, respectively. Inter-destination delay jitter can be calculated with the delay values received from all the other multicast group members.

ITU-T codec	Modulation type	Coding bit rate (Kbps)	Reason to be selected
G. 711	Companded PCM	64	Narrowband, most commonly used, "Tool quality"
G. 722.1	Transform Coding	24/32	Wideband
G. 729	CS-ACELP ^a	8	Narrowband, most commonly used after G. 711

Table 17.1 Codecs considered for RTCP-driven scenario

^aConjugate-Structure Algebraic Code Excited Linear Prediction.

17.4.2 Bank of Narrowband Non-Adaptive Codecs

We assume a bank of codecs is available at each transmission end, which is a realistic assumption for many of currently available VoIP software packages. We also assume the application can switch to a different codec according to the information extracted from the RTCP reports. In particular, the switching to a lower (higher) bit-rate codec takes place whenever the delay or jitter value reported by the RTCP is above (below) than the required ones. We consider the codecs shown in Table 17.1, where the key technical characteristics along with the reason to be selected are presented.

17.5 AMR-WB-Based Cross-Layer VoIP Distributed Rate Control

Our second cross-layer mechanism is based on the RTCP reports interpreted along with the signaling of AMR-WB (Adaptive Multi-Rate Wide Band).

17.5.1 AMR-WB Codec

AMR-WB coding algorithms are based on an Algebraic Code Excitation Linear Prediction (ACELP) technology. This same technology has been utilized in various speech codec standards, such as GSM enhanced full rate (GSM-EFR) (3GPP TS 06.51) and narrowband GSM-AMR (3GPP TS 26.071). The main novelty in AMR-WB is the sub-band structure, which enables significant savings in complexity and memory consumption. The audio band is split into two frequency bands so that the internal sampling frequency of the core is 12.8 KHz, having an audio bandwidth of 50–6400 Hz. Separate processing is performed for the frequency range from 6400 to 7000 Hz: more bits can be allocated to the perceptually important lower band.

The multi-rate approach provides the flexibility to achieve an optimal balance between error protection and source coding within a fixed bit-rate budget. When applying a lower source coding rate, the system can allocate more bits to channel coding; thus, enhancing the speech quality by improving robustness.

It should be stressed that the utilization of wideband speech in an IP network does not introduce additional system complexity compared to narrowband. Within the IP network the operation is by definition transcoder free, since compressed speech is transmitted in IP packets end-to-end.

17.5.2 Cross-layer VoIP Rate Control

We propose to use the cross-layer field of the AMR-WB codec called CMR (Codec Mode Request) for the VoIP flow rate adaptation. The CMR is included as header in the RTP packets, which are used to transmit the VoIP frames, and therefore, no signaling is needed in this case. The CMR indicates to the other end the desired coding mode. The CMR must be both computed (to be written in the IP voice payload) and extracted from the IP payload received (to select the appropriate codec for transmission). The CMR is computed based on quality measurements, which is assumed to be based on both the link state and the RTCP reports. The speech coder has nine different encoding rates (23.85 Kbps to 6.60 Kbps) in addition to low background noise encoding rate

Two are the major differences between the adaptive RTCP-based scenario and the AMR-WB-based scenario:

1. The frequency of the reports sent by the RTCP may not be synchronized with the speed of adaptation required by the PHY. 2. The highest bit rate of the AMR-WB codecs is less than half of the G.711 and hence, the same system load allows for a different number of VoIP connections.

Note that the inbound cross-layer signalling of both AMR-WB and RTCP allows for a fully distributed cross-layer VoIP bit rate adaptation. Note also that the algorithms to "measure quality" are open and undefined. In our case, signal quality is measured according to [4], while delay is computed as explained below. Finally, it is interesting also to note that this cross-layer solution is fully scalable and has no impact on the DVB-S2 and DVB-RCS standard and architecture.

17.6 Delay Budget Model and Performance Model

We propose the following delay budget model for our scenarios:

$$T_{tot} = T_{codec} + T_{MAC} + T_{trans} + T_{prop} + T_{playout}$$
(17.1)

$$T_{tot} = T_{codec} + T_{network} + T_{tplayout}$$
(17.2)

 T_{codec} is the delay introduced by the codec, which we model as follows:

$$T_{codec} = T_{fr} + T_{la} + T_{proc} + T_{pack}$$
(17.3)

where T_{pack} is the delay introduced when encapsulating more than one voice packet per IP packet, T_{fr} is the framing delay, T_{la} is the look-ahead delay (for prediction purposes) and T_{proc} is the processing delay.

The rest of the values are: the delay introduced in the MAC queuing and scheduling (T_{MAC}), the transmission delay (T_{trans}), the propagation time (T_{prop}) and the playout delay to smooth out the jitter ($T_{playout}$).

Relevant time delays for the codecs of our cross-layer scenarios are showed in Table 17.2.

Voice quality (clarity) can be measured by subjective methods such as Mean Opinion Score (MOS) mandated in ITU-T Recommendation P.800 [7] and parametric estimation (objective) methods like PSQM (ITU-T Q.861 [8]), PSQM+ and PAMS. The subjective methods are time-consuming and expensive to use while parametric estimation can be done quickly and inexpensively on the voice codecs.

Time (ms)	G. 711	G. 722.1	G. 729	AMR
Tla	0	20	5	5
Tplayout	10	10	10	<=10
Tpack (N=1)	0	0	0	0
Tpack (N=2)	20	20	10	20
Tpack (N=3)	40	40	20	40

Table 17.2 Relevant Codec Delays

We will use the ITU-T "E-Model" [9, 10] for estimating the voice quality. The application of E-Model results in the Transmission Rating or the R-factor. The resultant value of the R-factor is known as the R-value. Table 17.3 summarizes the relationship between the R-value and MOS.

Table 17.3 Relationship between R-value and MOS

	R-value (lower limit)	MOS (lower limit)	User satisfaction	Quality
	90	4.34	Very satisfied	Toll
	80	4.03	Satisfied	Toll
	70	3.60	Some users dissatisfied	Below Toll
	60	3.10	Many users dissatisfied	Unacceptable ^a
_	50	2.58	Nearly all users dissatisfied	Unacceptable ^b

^aUsed only in very special conditions

^bConnections with *R*-value below 50 are not recommended

The E-Model contemplates the case of situations where the user is being provided certain "advantage". In this case, a downward adjustment in the R-value is provided with a factor "A". For instance, when providing access to hard-to-reach locations via multi-hop satellite connections [11], we will use this degree of freedom to adjust our results for both the transparent and regenerative scenarios. It should be stressed that the extension of the E-model to wideband codecs is still under development.

17.7 Numerical Results

We assume a system with the following parameters. As introduced in the QoS model, each class of service has specific requirements. Cross-layer rate control is configured to guarantee a maximum delay of 270 ms (i.e. 20 ms excess w.r.t. propagation delay). On the other hand, delay requirements for AF and BE are 500 ms and 750 ms respectively.

We have chosen a users distribution where 75% of users are in good weather conditions (clearsky users), while 25% of users are in bad weather conditions (rain users).

Since channel conditions can vary depending on geographical areas, we have modeled a set of typical rain events (Fig. 17.3) to observe the attenuation dependence, and specially its effects on VoIP requirements. Deep attenuations (18 dB) have been also considered to study the cross-layer mechanisms in the worst case scenario.

Only 23 out of the 27 ACM proposed in the standard modes have been considered to fix a granularity between modes of approx. 1 dB. System is loaded until 85% of its total capacity.

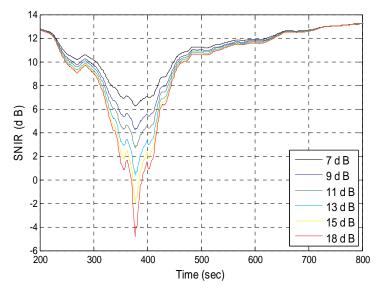


Fig. 17.3 Different rain attenuation models

We present first results for the reference scenario in the aggregated EF type of traffic in the forward link, where no cross-layer mechanisms are enabled, and VoIP is using the codec G. 711. Figure 17.4 shows the system working when a 3 dB rain event is used. It is observed that channel

attenuation is not noticeable since system can manage the load. Time axis of throughput corresponds to seconds. On the other hand, Fig. 17.5 shows the effect on the aggregated throughput when the channel undergoes a deeper attenuation (12 dB). It can be observed that the delay is above the guaranteed delay. Note that there are two curves for the delay, the red refers to the satellite terminals undergoing bad weather conditions and the blue one refers to those in clear sky conditions, showing that the whole beam (not only rain users) is affected by rain in terms of bandwidth reduction. It should be noted that this effect is due to the scheduling policy, as explained in [6]. However, the packet scheduler used could allocate alternative policies in case of necessity, e.g. isolating the users in clear sky conditions, avoiding the effect of the rain in these users.

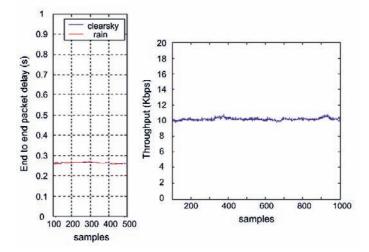


Fig. 17.4 Expedited forward delay for channel attenuation of 3 dB, for the G. 711 codec

Once we have checked the performance of the system without crosslayer design, we will observe the results using the two proposed approaches. Figure 17.6 (right) shows the RTCP-driven adaptive scenario that counteracts such a load peak. Code types are: G.711 / G.722.1 / G.729 VAD. Note the rate adaptation during the rain events (approx. five minutes long). The RTCP-based adaptive scenario can be therefore used to control VoIP bit rate to fulfil delay requirements when for some reason there is peak of VoIP traffic. A side effect is that VoIP adaptation can be seen as a system load control.

Figure 17.6 (left) shows the AMR-WB-based adaptation for the same system and channel conditions. It can be observed that this case outperforms the RTCP-driven based due to the following reasons:

- 1. The cross-layer scenario does not add signalling overhead (RTCP reports) on the system.
- 2. The AMR-WB codec has inherent higher voice quality (Wide Band).
- 3. The AMR-WB highest bit rate is lower than the G.711 bit rate and therefore the system can admit a higher number of connections.
- 4. Like in the RTCP-based adaptive baseline, VoIP adaptation can be seen as a system load control.

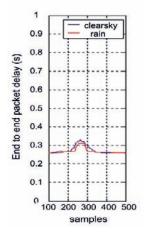


Fig. 17.5 Expedited forward aggregated throughput for a channel attenuation of 12 dB, for the G. 711 codec

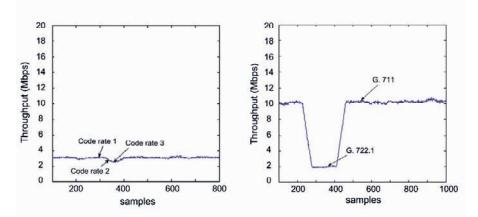


Fig. 17.6 Expedited forward aggregated throughput for a channel attenuation of 12 dB for the two cross-layer approaches. AMR-WB (*left*) and RTCP-driven rate control (*right*)

A comparison between the two approaches has been carried out in Fig. 17.7. We have used as reference scenario the case where only a single VoIP codec is used (G.711). Figure shows that Bank of codecs performs better than the single 64 Kbps code, but it adapts slowly, and therefore probability of exceeding the requirement (270 ms) is already high. For the AMR-WB scenario, we get good results; probability is zero for attenuation lower than 13 dB, only when attenuation is 18 dB the codecs can not avoid exceeding delay threshold. Notice that 18 dB is considered deep attenuation (the probability of this kind of rain events is very low). In that case, other mechanisms (such as Connection Admission Control, weight adaptation, etc.) would be necessaries to avoid congestion effects.

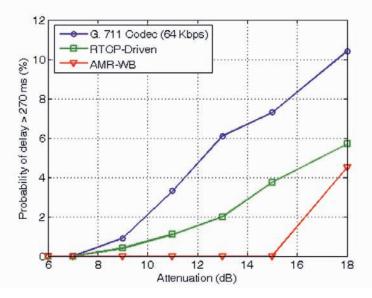


Fig. 17.7 Probability of exceeding delay requirements (270 ms) for Expedited Forward depending on rain attenuation. Comparison between reference scenario and cross-layer approaches

Up to here, we have not considered packetization (i.e. more than one voice frame into the same VoIP packet). We present now the effect of packetization (N=2) on the RTCP-based adaptive scenario. Even if the bandwidth efficiency increases, the premium class (EF) fails to meet the delay requirements for the regular delay threshold of 270 ms. Similar results are obtained for AMR-WB. Extending the threshold to 290 ms (Fig. 17.8), requirements are reached most of time using RTCP driven. For AMR-WB, it can be observed that its performance is also better than the RTCP driven adaptive scenario in terms of meeting the delay requirements

for N=2. The improvement of AMR-WB respect to RTCP is around 5%. Note that in this case, rain users have more probability of exceeding the delay requirement. This is due to the fact that we have chosen the scheduling policy in order to isolate clearsky users.

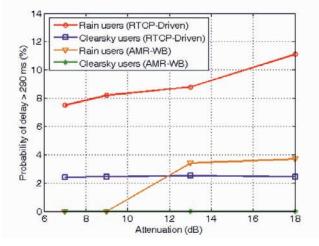


Fig. 17.8 Probability of exceeding the maximum delay (290 ms) for the two approaches for EF with 85% of system load, and N=2

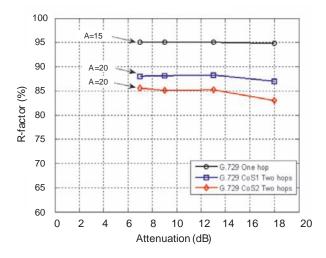


Fig. 17.9 R-factor for G. 729 (with packetization, N=1) for different delay guarantees and number of satellite hops

Finally, Fig. 17.9 shows the R factor (see Table 17.3) assuming two CoSs defined by two different delay guarantees, namely CoS1 guarantees a

maximum delay of 270 ms and CoS2 guarantees a maximum delay of 290 ms. It is observed not only the effect of the number of hops in the quality regardless rate adaptation, but also the difference between class of service guarantees.

17.8 Conclusion

In this chapter we have shown a preliminary study on a cross-layer design of VoIP transmission over a DVB-S2/RCS system. We have proposed and compared two different approaches, both of them distributed and based on rate control. Our results indicate that cross-layer design is needed for providing QoS guarantees when channel attenuation decreases capacity.

We have shown that AMR-WB approach outperforms the RTCP-driven approach in terms of signaling overhead. Moreover, AMR-WB admits higher number of VoIP users with similar voice quality. Furthermore, a side effect has been observed; the rate control on aggregated traffic provides a powerful tool that can be used for load control. This effect should be further investigated. Finally, although packetization improves bandwidth efficiency, it is demonstrated that delay requirements are not fulfilled, and therefore it is not advisable for its use through this satellite link.

References

- 1. Wong R, Sai Shankar N, Van der Schaar M (2005) Integrated application MAC modelling for cross-layer optimized wireless, ICC vol 2, pp 1271–1275
- Van der Schaar M, Krishnamachari S, Choi S, Xu X (2003) Adaptive crosslayer protection strategies for robust scalable video transmission over 802.11 WLANs, IEEE Journal on Selected Areas of Communications, vol 21, n 10
- ETSI: EN 302 207 (2006) Digital Video Broadcasting (DVB); Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications (DVB-S2)
- 4. Vázquez Castro MA, Seco Granados G (2006) Cross-layer packet scheduler design of a multibeam broadband satellite system with adaptive coding and modulation, IEEE Trans. on Wireless Comm., vol 6, n 1, pp 248–258
- Skinnemoen H, Vermesan A, Iuoras A, Adams G, Lobao X (2004) Interactive IP-network via satellite DVB-RCS, IEEE Journal on Selected Areas In Communications, vol 22, n 3
- F. Vieira, M.A. Vázquez Castro, G. Seco Granados, "A tunable-fairness crosslayer scheduler for DVB-S2", International Journal of Satellite Communications and Networking, vol 24, no 5, Sep–Oct, 2006

- 7. Schulzrinne H, Casner S, Frederick R, Jacobson V (2003) RTP: A transport protocol for real-time applications, RFC 3550
- 8. ITU-T Recommendation P.800 (1996) Methods for subjective determination of transmission quality
- 9. ITU-T Recommendation P.861 (1998) Objective quality measurement of telephone band (300–3400 Hz) speech codecs
- 10. ITU-T Recommendation G.107 (2001) The E-model, computational model for use in transmission planning
- 11. ITU-T Recommendations G.108 (1999) Application of the E-Model A planning guide

18 Optical Satellite Downlinks to Optical Ground Stations and High-Altitude Platforms

Dirk Giggenbach, Bernhard Epple, Joachim Horwath, and Florian Moll

18.1 Introduction

Recent years have seen an immense increase in the capability of earth observation (EO) sensors flown on satellites. State-of-the-art payloads like high resolution optical or infrared cameras or SAR systems produce data at a speed of gigabits per second. This is why the conventional RF-downlink has become the bottleneck in EO-systems, as it is limited to some hundred megabits per second. The data acquired by the sensor can only be sent to the ground when it is in the reach of an according RF ground station antenna, which happens a few times per day with each downlink session lasting only around 9 min at maximum for a LEO (Low Earth Orbit) satellite. This limits the effective operational time fraction of such sensors to only some minutes per day. Due to limitations in the available frequency bands but also in technological feasibility the RF downlink technology is currently reaching its limits.

To solve this communications bottleneck we now have the opportunity of using free-space optical (FSO) high speed links. This technology would instantly multiply the downlink data rate by a factor of ten, while even faster links would be feasible in the near future. At the same time the mass, size, and power consumption of the satellite terminal would be cut to only a fraction of the values of conventional RF-antennas, making high speed downlinks an option even for compact- or micro-satellites (a typical Txaperture diameter for a LEO-downlink would be approx. 3 cm). Furthermore, also the size of the according Optical Ground receiving Station (OGS) remains quite compact, with only some decimeters of telescope diameter, enabling transportable or even mobile stations. This is an important benefit compared to RF ground stations which have typically antenna diameters of 5 m and more. The inherent tap-proofness of directed optical beams due to a minimized optical signal spot beam on ground (typically between 10 and 100 m) is also very appreciable in security applications.

The Institute of Communications and Navigation of the German Aerospace Center (DLR) has demonstrated together with its partner JAXA (Japan Aerospace Exploration Agency) the feasibility of direct optical LEO-Downlinks in the Project KIODO (Kirari Optical Downlink to Oberpfaffenhofen). KIODO showed a very good performance with measured bit error rates down to 10^{-6} with a transportable and inexpensive OGS [1].

18.2 Solving the Challenge of Cloud-Blockage

Reliable optical downlinks are of course limited to geographical OGS-sites with a minimum fraction of cloud coverage, as the optical signal is blocked by most clouds. Therefore, these stations should preferably be situated on mountain tops (like the classical astronomical observation site) or in countries with low occurrence of clouds, like the Mediterranean or sub tropic latitudes. In the future, also polar-located OGS will be an option with an immense downlink time fraction for polar orbiting satellites. However, this practice might still not be acceptable for applications with secure near realtime requirements. Also, for non-EO applications like communications or broadcast, a nearly hundred percent availability is required for the satellite link. Therefore the "ground" station has to be positioned above the clouds. Aircraft or aerostatic High Altitude Platforms (HAPs) provide the suitable bases, with the later having the advantage of stationarity together with lesser vibrations and position uncertainty. The final "last mile" to the ground can then be bridged by standard short-range RF point-to-point links as used today in terrestrial applications. With a buffering strategy onboard the HAP even optical downlinks from the HAP to a terrestrial miniature OGS could be used for the HAP-downlink, storing the data during total cloud blockage. This concept here is called StORe for Stratospheric Optical Relays. In a future scenario - with a network of HAPs in range of sight - linking the data from one HAP to an other without cloud blockage underneath would allow a purely optical downlink system [2, 3]. In a future HAP communications network, this networking functionality would be available inherently at no extra expenses.

Other benefits of the StORe-concept compared with RF and terrestrial optical ground stations are the extended visibility time of the LEO satellite (the link can already start at negative elevation angles as long as the line-of-sight stays above the maximum cloud altitude of about 13 km, the top limit for European latitudes) and negligible attenuation by the atmosphere.

Also the challenge of fading (caused by atmospheric index-of-refraction scintillations) is much reduced at stratospheric altitudes. Further, such HAP-relays could provide downlink capability at any place where a HAP is placed. Even reliable nomadic downlink services can thus be provided to temporary end user sites.

18.3 System Comparison

18.3.1 Earth Observation Scenario

The system used for example calculations in this chapter consists of a satellite with a mean orbit height of 500 km and a typical near-polar orbit inclination. The satellite is equipped with a high resolution camera which during operation produces data at a rate of 6.7 Gbit/s (equivalent to 26411 TByte per year at 100% usage). This high data rate illustrates that the down link bandwidth is a limiting factor for the operational time of the satellite. For simplicity in calculations and to focus on the concept of the StORe, there is only one receiving ground station for the data downlink. Here DLR's ground station at Neustrelitz, Germany, is used for calculations.

18.3.2 State of the Art RF Downlink

Currently used RF downlinks from LEO Satellites have effective user data rates of up to 262 Mbit/s (e.g. TerraSAR-X). At Neustrelitz it is possible to start data transmission at an elevation angle of 5° of the satellite. This results in a mean daily contact time of 2360 s or in a maximum transferable data volume of about 77 GByte per day. Since the ground station can be assumed to be available for downlink 100% of time, the transferable data volume per year is 28 TByte and the camera can be used during 0.1% of the operational time of the satellite.

18.3.3 Proposed RF Downlink

Using modern modulation and coding (ModCod) schemes that are adaptive to the elevation angle between ground station and satellite, the downlink capacity can be strongly increased. Targeting a downlink availability of 99.9% and using three different ModCod schemes, simulations show that in the daily mean it is possible to transfer data for 898 s with 295 Mbit/s, for 546 s with 540 Mbit/s and for 927 s with 648 Mbit/s to the ground station in Neustrelitz. This results in a maximum daily transferable data volume of 145 GByte. Since the downlink is available for 99.9% of time this results in a downlink capacity of 53 TByte per year and a utilization of the camera of 0.2%.

18.3.4 Proposed Optical Downlink

Due to atmospheric effects, the availability of an optical downlink is limited to an elevation angle of 10° and more. This reduced Field-of-View results in a mean daily contact time of 1499 s. Nevertheless, using an optical downlink with a data rate of 5 Gbit/s, the resulting transferable data volume per day is 937 GByte, when neglecting cloud blockage. These 5 Gbit/s will be used in the optical link sections in further calculations.

The downlink station at Neustrelitz is in a non-optimum place for an optical ground station because it has a mean availability for optical LEO downlinks (limited by cloud cover) over the year of around 32% [4]. Even with this low availability the transferable data volume per year is at least 109 TByte and the camera could be used during 0.4% of the operational time. A limiting factor in this case is the available data storage on the satellite for times when no data downlink is possible due to bad weather conditions.

An important note at this point is that these values are just for *one* OGS at a site far from the earth's poles and with a relatively high cloud cover probability (Neustrelitz). When using four OGSs distributed over Germany (this concept is called ground station diversity), the combined availability can be boosted to 73% during the winter half year (October to March) and 91% during the summer half year (April to September). Availability approaches 99% when using two or more ground stations in advantageous areas (outside Germany). All OGSs should be separated by several hundred km from each other to ensure uncorrelated cloud cover statistics. Ground station diversity leads not only to an increased downlink availability, but also to more downlink time slots and therefore to an increased average downlink capacity.

18.3.5 Proposed Combined RF-Optical Downlink

For overcoming the cases where an optical downlink is not possible due to cloud cover, one can combine an RF downlink terminal with an optical terminal on the satellite. This approach adds complexity to the data downlink management, e.g. data has to be prioritized and scheduled for downlink via RF or optical channel, but it greatly extends the downlink availability and thus the possible sensor usage. Using the values from above (proposed RF and proposed optical downlink to one ground receiving site only), the combined downlink volume per year can be increases to 162 TByte and the camera can be used during 0.6% of the operational time.

18.3.6 Proposed GEO Relay

A common concept for increasing the available downlink time for a LEO satellite is the use of a GEO satellite as relay station. An example for this concept is ESA's geostationary satellite ARTEMIS. Because a GEO satellite is always visible at the ground station at a high elevation angle, an RF downlink with the most efficient ModCod scheme from the previous section can be implemented. This ModCod scheme gives an available bandwidth of 648 Mbit/s with an availability of 99.9% between GEO and ground. The daily available downlink time from a GEO satellite is one day, or 86400 s. The daily downlink volume is then 6998 GByte. Because of its long distance from earth (~40,000 km) a GEO satellite can communicate with a LEO satellite during about half of its orbit or about 43200 s per day. If the communication is done via FSO communications at 5Gbit/s the LEO satellite can transmit about 27 TByte per day to the GEO satellite. The availability of this link is 100% of geometrical visibility time because cloud blocking does not occur in space. The amount of optically transferable data is much more than the GEO satellite can relay down to earth by its RF-link, so the limiting factor in this scenario is the GEO-downlink. Buffering of the optically received data onboard the GEO is necessary to allow a constant data-flow from GEO to ground also when the LEO is not visible. With this constellation 2552 TByte per year can be transmitted from LEO to GEO to ground station and the utilization of the camera can be raised to 9.7%.

18.3.7 Proposed HAP Relay

The concept of a relay station above the downlink station can also be realized by using a HAP. The use of a HAP has some advantages. First, HAPs are easier to replace than GEO satellites if something fails. Second the environmental constraints for the payload are not as tough as for a GEO payload and third HAPs should be cheaper to build and launch. For the following calculations we assume a HAP placed above the downlink station in Neustrelitz at an altitude of 20 km. For the downlink from HAP to ground station conventional point-to-point RF technology with a steered antenna is used (this "last-mile" link from HAP to ground has 100% availability). Due to the relatively short link distance between HAP and ground station and due to the constant high elevation angel between them, the RF downlink bandwidth can be raised to 1336 Mbit/s with an appropriate ModCod scheme and an availability of 99.9%.

The HAP holds an optical receiver terminal for data downlinks from EO-satellites, again at 5 Gbit/s. Since the HAP is located 20 km above the earth surface, the data link between HAP and LEO can already start at an elevation angle of -2.7° . This results in a mean daily contact time of 4759 s or in a daily transferable data volume from LEO to HAP of 2974 GByte. The link availability between HAP and satellite is 100%, because of the HAP's position above the cloud layer. The HAP can downlink 14429 GByte per day, so the limiting factor in this scenario is the link time between LEO and HAP. With this system the transferable data volume per year can be increased to 1086 TByte and the camera can be used during 4.1% of the operational time. This is less than in the GEO relay scenario, but the HAP in this scenario can theoretically serve about four LEO satellite missions in parallel (limited by contact time between LEO and HAP) while the GEO satellite can serve only one mission at a time (limited by RF-downlink capacity from GEO to ground).

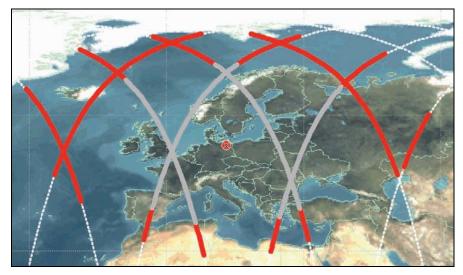


Fig. 18.1 Example for the improved access time by a StORe at 20 km with -2.7° minimum elevation (*red* and *grey lines*) over an RF-ground station with $+5^{\circ}$ minimum elevation (*grey lines* only), both stations situated at *Neustrelitz, Germany* (marker in the image center). The given orbits are the satellite passes of one day. LEO circular orbit height is 500 km with a typical near-polar inclination

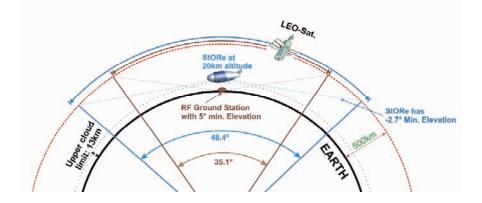


Fig. 18.2 Geometrical visibility constraints of a StORe with -2.7° minimum elevation and a ground station (GS) with 5° minimum elevation for a LEO in 500 km orbit altitude. The StORe has a field-of-view cone with 48.4° planar angle, while the ground station has only 35.1° (angles not to scale)

18.4 Comparison of Downlink Scenarios

Table 18.1 compares the afore discussed downlink scenarios using only one ground receiving station. The table only presents the transferable amount of data and neglects the cost per byte for the different downlink scenarios. The cost per byte is an economical factor that should be considered if the downlink scenario for a space mission is planned.

Downlink Scenario	RF	Proposed RF		Proposed RF-optical	-	Proposed HAP-relay (StORE)
Effective Downlink Rate (Gbit/s)	0.262	up to 0.648	5.0	up to 5.648	0.648	1.336
Mean daily downlink time in s	2360	2360	1499	2360	86400	86400
Availability in %	100	99.9	32–45	100	99.9	99.9
Mean daily trans- ferable data volume from LEO to ground (GByte/day)	77	145	300	377	6998	2974
Mean daily Camera Utilization in % of operational time	0.1	0.2	0.4	0.6	9.7	4.1

Table 18.1 Comparison of the presented downlink scenarios for a ground station in Neustrelitz, Germany

18.5 Cloud Cover Statistics and OGS-Diversity

The biggest problem for direct Free-Space Optical (FSO) downlinks from satellites or HAPs to an OGS is the changing cloud cover above the OGS which determines the availability of the downlink. For selecting favorable locations for an OGS meteorological data is needed for evaluating the cloud cover statistics. Such data is for example available from the International Satellite Cloud Climatology Project (ISCCP), World Data Center for Remote Sensing of the Atmosphere (WDC-RSAT), European Cloud Climatology (ECC), several synoptic observation sites and others. Figure 18.3 gives the mean annual cloud coverage for the entire earth. Concluding from this data, favorable locations for ground stations would be in North and South of Africa, western part of the USA, Australia, the Middle East and parts of the Antarctica.

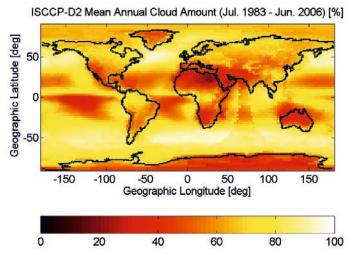


Fig. 18.3 Mean annual cloud coverage for the time from July 1983 to June 2006 based on data from the ISCCP. The color indicates the amount of mean cloud coverage from black = 0% over red, orange and yellow to white = 100%

In most cases the choice of the location for an OGS is not only made on base of cloud cover statistics, but influenced by factors like national interests, political stability in the region, or existing infrastructure. So sometimes it is necessary to set up an OGS in a place with a non-optimum availability. A concept that can be used to boost downlink availability up to 99% in cases where locations with low availability should be used is the concept of multiple ground station diversity. If more than one OGS is used for the downlink and the distance between the OGSs is large enough (> 1000 km) the cloud cover above these stations will be uncorrelated and while one OGS is blocked by clouds another ground station might be available. This system design will also lead to an increased downlink capacity in cases when more than one OGS is cloud free. If the distance between the OGSs is too small, the locations will all have similar weather conditions and the availability will not increase significantly. Also the effect of an increased downlink capacity is not given as the satellite will see all the OGSs at the same time and can only downlink to one of them. Another fact that has to be kept in mind is that the cloud coverage is seasonal dependant as shown by Fig. 18.4. This requires that the data used for calculating the availability of an OGS needs enough resolution in time to reflect the change during the course of the year.

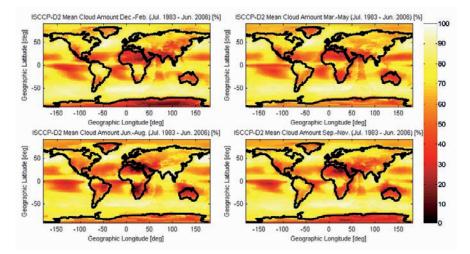


Fig. 18.4 Seasonal dependency of the mean cloud coverage. Based on data from the ISCCP from 7/1983–6/2006. The colors indicate the mean cloud coverage in the selected periods, favorable locations have darker colors. A good example for the seasonal dependency of the mean cloud coverage is Brazil which has a good availability in summer (*lower left* plot, around 80%) but also a bad availability during winter (*upper left*, around 20%)

18.6 Availability of OGS-Networks

In the following calculations of availability of OGS networks are presented. Base of these calculations is either data from ECC or ISCPP and in some cases data from synoptic observations. Because of its low resolution in time and image quality the used data can only be used for approximating of the availability. Since the cloud coverage analysis requires a lot of processing time for each desired location, a pre-selection of suitable locations has been done. The used selection criteria have been the following:

- the locations should be in political stable regions
- infrastructure should already be available
- the yearly mean availability of a location should not be below 60% (not valid within Germany)

18.6.1 OGS Network within Germany

Germany is not really a good place for doing direct optical LEO downlinks, because there is no location that has more than 45% mean annual availability. Since all regions have nearly the same mean cloud coverage, places have been selected to be as far away from each other as possible. If four OGSs are placed within Germany (Fig. 18.5), the system would have an availability of about 82% in the yearly average. If the system is extended to ten OGSs the availability raises only slightly up to 87%. This slight increase in availability despite a heavy increase in the number of used OGSs shows that the weather within Germany is heavily correlated. An OGS network in Germany is also a good example for the seasonal dependency of the mean cloud coverage: During winter half the selected four-OGS-system has an availability of 73% but during summer half the availability reaches 91%.

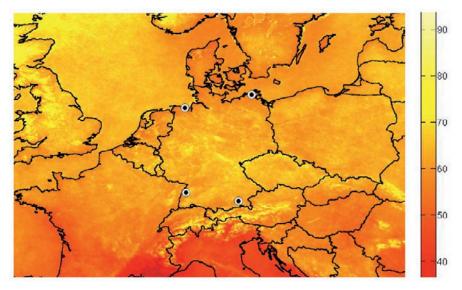


Fig. 18.5 An OGS network with four OGSs (*black* and *white* markers) within Germany. Although the OGSs are positioned as far away from each other as possible, there is still strong correlation in the cloud coverage of the four locations. The legend denotes the yearly mean cloud cover (*darker* color means less cloud probability)

18.6.2 OGS Network within Europe

Regions in Europe that offer mean annual availabilities higher than 45% can only be found in the southern areas, so an OGS network within Europe should be set up across the Mediterranean Sea. The preselected locations from this region are Calar Alto in the south of Spain, Marseille in the south of France, Catania on Sicily (Italy) and Skinakas on Crete (Greece). The selected sites are shown in Fig. 18.6. The fact that there are already astronomical observatories in the chosen regions strengthens the assumption that these locations are suited for FSO. It also guarantees that there is already some existing infrastructure that can be used for setting up an OGS. Unfortunately cloud cover statistics have not been available for all of these locations. In these cases data from observation sites within a few kilometers range have been used.

OGS Location(s)	Availability in %
Skinakas, Marseille, Catania, Calar Alto	98
Skinakas, Marseille, Catania	96
Skinakas, Marseille	92
Skinakas	74
Marseille	72
Catania	69
Calar Alto	64

18.6.3 World Wide OGS Network

For setting up a global network also sites already containing observatories have been chosen. The selected sites are Observatorio del Teide on Tenerife (Spain), Paranal Observatory in the Atacama desert (Chile), Perth in Western Australia and again Skinakas on Crete.

OGS Location(s)	Availability in %
Paranal, Observatorio del Teide, Perth, Skinakas	99
Paranal, Observatorio del Teide, Skinakas	98
Paranal, Observatorio del Teide, Perth	98
Paranal, Observatorio del Teide	95
Paranal	84
Skinakas	74
Observatorio del Teide	71
Perth	60

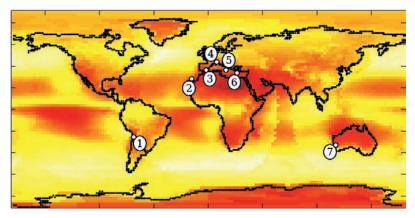


Fig. 18.6 Map of selected locations for a European (3, 4, 5, 6) and a world wide (1, 2, 6, 7) OGS network. 1 = Paranal Observatory, 2 = Observatorio del Teide, 3 = Calar Alto, 4 = Marseille, 5 = Catania, 6 = Skinakas, 7 = Perth

The given networks show that it is possible to reach a direct LEO downlink availability of 99% or more when the concept of ground station diversity is used for the system setup. From the given calculations it can be seen that the number of used OGSs should be at least four OGSs in the network to reach a reasonable availability. Calculating the availability of the networks in this chapter is only done as an average statistical approximation. For a detailed evaluation of OGS locations also small scale local effects like the cloud distribution seen from the OGS and the individual satellite paths need to be taken into account. This detailed information is not available from meteorological satellite data.

18.7 Wavelength Selection and Terminal Architecture

Due to the high altitude of HAPs the effects of the atmosphere on optical beams coming from LEO satellites are much smaller compared to scenarios where the beam is received at an OGS on the ground. Due to the long horizontal propagation distance especially at low elevation angles, the wavelength has to be selected carefully because of the large variance of the absorption coefficient over the wavelength.

Generally there are two attenuation effects inside the atmosphere (besides the free-space loss which is determined by the beam divergence angle): optical absorption and scattering. These effects lead to specific transmission windows which are suitable for optical communications in the atmosphere. Beside these transmission windows it is also important that laser sources, modulators and detectors are available. The three designated wavelength regions for FSO systems are around 850 nm, 1064 nm and 1550 nm, as shown in Fig. 18.7.

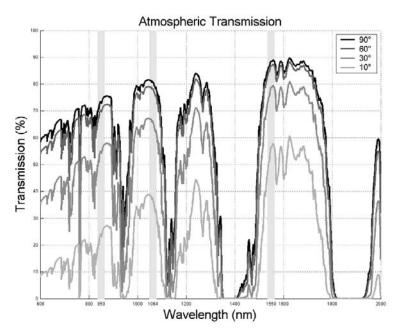


Fig. 18.7 Atmospheric transmission for LEO downlinks to Neustrelitz at selected elevation angles. Clear-sky atmospheric transmission windows that determine the wavelength selection for FSO and at which suitable components are available are marked by grey background coloring. The used atmosphere model is the Midlatitude Summer model combined with a rural aerosol model for the last 2 km and a moderate volcanic activity model for the rest of the atmosphere

In Fig. 18.8 the elevation-dependence of the transmission for these three wavelengths is shown.

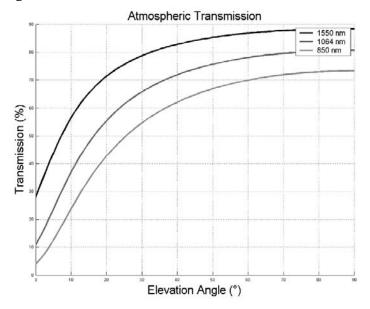


Fig. 18.8 Illustration of the elevation dependency of the atmospheric transmission for an optical LEO downlink to Neustrelitz at the selected wavelengths of 850 nm, 1064 nm and 1550 nm. The atmospheric model is the same as for Fig. 18.7

While sometimes also 10.6 μ m is considered due to its improved cloud-transmissivity, it can be shown that this improvement is more than compensated by the higher free-space loss. This wavelength lacks also available components for communications application.

800 nm technology has some disadvantages: The presence of strong background light from the sun and the higher Rayleigh-scattering compared to 1064 and 1550 nm. For 1064 nm and 1550 nm technology one of the clear advantages is the availability of high power optical fiber amplifiers to boost the transmission signal.

The wavelength of 1064 nm is used for coherent systems with highly stable Nd:YAG oscillators, a laser source with very good coherence and therefore suitable for homodyne systems. This enables the implementation of homodyne binary phase-shift keying (BPSK) modulation. The advantage of these systems is the high sensitivity which leads to small aperture diameters for the optical receivers. Due to the homodyne detection scheme the communication signal is recovered at baseband, which considerably simplifies the communications electronics design compared with heterodyne or intradyne reception, but this reception technology requires diffraction limited super-positioning with the local oscillator, which is a demanding task under atmospheric index-of-refraction turbulence. Therefore, adaptive optics technologies to correct the distorted wave front might be required for its application with terrestrial OGSs. The effect of background radiation can be neglected due to the extremely small noise bandwidth of the homodyne receiver which is in the order of the databandwidth (e.g. 1 GHz signal bandwidth corresponds to only about 3.5 pm optical wavelength at 1064 nm wavelength).

Optical C-Band technology around 1550 nm with on/off-keying and direct detection is widely used in terrestrial fiber-optical transmission systems and has already been tested successfully in a stratospheric test-bed [3]. Current systems are not as sensitive as coherent systems but the use of fast wave-front correction systems (adaptive optics, as mentioned above) to mitigate atmospheric index of refraction turbulence would allow coupling of the received signal into a mono-mode fiber at the receiver. This is the requirement for using optical fiber pre-amplifiers. With optical preamplification at the receiver, the sensitivity is then comparable with current coherent systems [5].

An additional advantage of receiver concepts with wave-front correction systems for coupling into mono-mode fibers would be the enabling of Dense Wavelength Division Multiplexing (DWDM) technology. DWDM a core technology in terrestrial fiber optical transport networks – increases the number of wavelength or channels combined onto a single fiber. Enabling this approach for free-space optics would allow the use of integrated fiber optics off the shelf components. Preference is therefore given to optical C- and L-Band technology due to the low atmospheric attenuation within this wavelength region between 1550.52 nm and 1600.17 nm (more than 120 channels according to the ITU grid specification with 50 GHz channel spacing)

A possible architecture of a unidirectional DWDM-FSO system is sketched in Fig. 18.9.

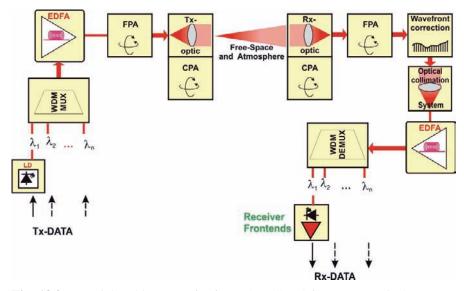


Fig. 18.9 Potential architecture of a future broadband free-space optical DWDM system for simplex EO-Sat downlinks

In this architecture, on the Tx-side each channel or wavelength is coupled into a single fiber by a wavelength-multiplexer and then amplified. Therefore amplifiers with erbium-doped fibers as common to terrestrial fiber communications are used (*EDFA*, Erbium Doped Fiber Amplifier). The output of the coupler is then delivered to the fine pointing assembly (FPA). The FPA is a stabilization and tracking system that removes the high frequency vibrations of the satellite in order to guarantee precise pointing. The coarse pointing assembly (CPA) is a tracking system with a wide angular range (e.g. hemispherical).

On the receiver side there exist major differences between optical terminals for satellites and optical terminals for the StORe. The spatial orientation of a satellite is usually known very precisely and the attitude changes can be controlled with similar precision (e.g. 300 µrad depending on star sensors and reaction wheels). HAPs do not operate in a stable orbit but in the atmosphere. Therefore station keeping maneuvers are necessary to keep the position against the stratospheric wind. The wind can also generate oscillating movements of the payload equipment depending on the vehicle's center of gravity. Also vibrations are much stronger on HAPs compared to satellites' base motion disturbances. The high pointing and tracking requirements for optical terminals require systems that can cope with all these effects on the HAP vehicle. Therefore the control loop of the tracking system on the StORe (mainly FPA-performance) needs higher performance compared to the satellite pointing system. The FPA corrects the angle of arrival of the incoming wave front (tip/tilt correction). Changes in the angle of arrival are caused by vibrations and atmospheric turbulence. Finally the wave front correction building block corrects the higher order wave front distortions in order to reconstruct a plane wave before the signal can be coupled into the mono-mode fiber with the collimation system (this is only necessary with large Rx-apertures). In the fiber the DWDM signal is then amplified by an optical pre-amplifier. Finally the optical signal is demultiplexed and each channel is detected by a single receiver-frontend. Data rates of terminals can be 10 Gbit/s in near future and n-times this data rate by the use of the DWDM technology.

18.8 Conclusion

We have calculated the practical advantage of optical downlinks from earth observation (EO) satellites over conventional RF-downlinks in different downlink scenarios. The usability of the EO-sensor could be boosted by nearly a factor of forty with a future StORe-System (HAP-Relays) or by a factor four with simple direct downlinks to optical ground stations without RF-backup (and an OGS in a non-optimum location). This performance is offered by low-power transmit terminals with very small apertures in the range of few centimeters and according low mass.

Optical data return channels for LEO satellites using GEO relays satellites with RF-downlinks (e.g. the SILEX-system) have the advantage of higher link availability as they reliably cover nearly half of the LEO-orbit, but the system-complexity is also high. The terminal size, power consumption, and weight of optical LEO-GEO link terminals is high due to the high free-space loss and therefore this technology offers lower effective data rates and can not be carried onboard small LEO satellites. When a Ka-Band downlink from the GEO is used, this again causes a bottleneck for the data throughput. Also, the financial effort for setting up a GEO-relay scenario is much higher than for direct LEO downlinks.

It seems favorable to establish a global OGS-network (and later a StORe-Network) based on compatible technology for downlinks from LEO as well as GEO and later possibly for links from deep-space probes.

References

- 1. Perlot N et al. (2007) Results of the Optical Downlink Experiment KIODO from OICETS Satellite to Optical Ground Station Oberpfaffenhofen (OGS-OP). In: Proc. of SPIE, Vol. 6457
- 2. Giggenbach D et al. (2002) Stratospheric Optical Inter-Platform Links for High Altitude Platforms. In: Proceedings of the 20th International Communications Satellite Systems Conference, Montreal
- 3. Horwath J et al. (2006) Backhaul communications for High Altitude Platforms, Results of STROPEX, the Stratospheric Optical Payload Experiment. In: 15th IST Mobile & Wireless Communications Summit Myconos, Greece
- 4. Moll F (2006) Einfluss der Wolkenbedeckung und atmosphärischen Dämpfung von verschiedenen Wellenlängen im MIR auf die Verfügbarkeit optischer Bodenstationen. Diploma thesis, DLR
- Pfennigbauer M, Leeb WR (2003) Optical satellite communications with Erbium doped fiber amplifiers. In: Space Communications, Vol. 19 pp 59–67

19 Wireless Applications in Healthcare and Welfare

Matti Hämäläinen, Pekka Pirinen, Zach Shelby, and Jari Iinatti

19.1 Introduction

Globally recognized problem at hospitals is inefficiency in the healthcare process management. Bottlenecks, such as long waiting and searching times, double writing, lost instruments and a lack of all information needed in situ can easily be pointed at. All these listed issues are far too common problems, and unfortunately they are not limited only to the healthcare sector. So, new innovative solutions and improved procedures developed for hospital are applicable in other environments, such as warehouses, etc. due to the common and generic technological requirements.

The key point in the research area to be discussed next is to define such processes at hospital, healthcare and welfare sectors, which can be rationalized by adopting new wireless technologies and new practices. More benefit can be achieved if the novel actions could be directed to daily working processes. It is also desirable that the promoted techniques are invisible to the user, which makes it much easier to adopt these techniques and processes.

In a public hospital in Finland, it has been estimated that ordering and handling of deliveries in a surgical department may take up to 167 and 147 hours per week, respectively. In a regular ward having 26 beds, 13 hours are spent for ordering and 21.5 hours for handling of deliveries per week [1]. Another study shows that by using an electronic patient record (EPR) and wireless communication, medical doctors can save about 15% of their working time for medical treatment. For nurses, the saving in the work load could be more than 10% if modern technologies are obtained [2].

It is clear that if all this released workload can be redirected to patient care activities by decreasing the double writing etc. subsidiary activities, the financial savings in the whole healthcare sector would be extremely large. In addition to the financial savings, this improves the overall quality of care. It is a well-known problem that there is a lack of labour in the healthcare sector, and it is in everybody's interest to use the labour available as efficiently as possible in nursing activities. In addition, in the future the average age of population is expected to increase. Therefore, there will be more patients per nurse to be taken care. To guarantee the sufficient level of treatment, new methods are needed to assist the daily nursing duties.

The following discussion is based on the ideas and work carried out mostly by the *WILHO Consortium*¹ in Oulu region in Finland. The WILHO Consortium has been formed to improve the existing working processes and utilization of wireless technologies at hospitals, and to promote the concept globally. The general goals of WILHO consortium are introduced in [1,3,4]. The ideas from side projects will complete the discussion.

19.2 Wireless Hospital Concept

There are several areas in the healthcare segment where wireless technologies can effectively be utilized. Some of these are summarized in Fig. 19.1 [1]. The WILHO Consortium has defined the *Wireless Hospital concept* that has been promoted world wide. Because the consortium includes two hospitals (both private and public), the use-case scenarios are adopted from real operational needs. Due to the hospital construction projects both hospitals at the moment have, there will be possibilities to demonstrate and pilot new ideas in practise in those new premises.

The following sections discuss the possibilities and scenarios of applying wireless technologies at hospital environment. Firstly, the basic idea of a wireless hospital concept is shortly illustrated. The key issue in the concept is to integrate wireless support to hospital top level management. In addition, the use of wireless technologies makes it possible to keep track of personnel and goods in real-time inside a hospital.

Wireless integration enables easy combination of all medical, diagnostic and clinical data together whenever needed. Data fusion from several information sources is not location related either. From the patients' point of

¹ WILHO Consortium includes Centre for Wireless Communications (CWC), Intelligent Systems Group at the Computer Engineering Laboratory (ISG) and Optoelectronics and Measurement Techniques Laboratory (OEM), all at the University of Oulu, Oulu University Hospital (OUH). In addition to the academic players, the consortium has two SME's: ODL Health Ltd. (a private hospital) and Medanets Ltd.

view, the benefits are coming from simpler entrance processes, minimized queuing times and more efficient patient centric care.

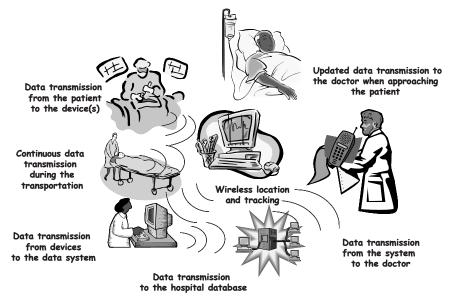


Fig. 19.1 Example of the wireless playground at the hospital

As a consequence, electronic patient records need to support new wireless access methods. EPRs could store all the care history related to a patient, and all the new medical information collected during the period a patient is staying at hospital. After-care and self-care data can also be merged to an EPR. The latter processes can be done, for example, by using light weight wearable sensors and tele-health approaches.

Ubiquitous, distributed and centralized services related to the patient's problems are the most efficient tools to improve hospital transit speed of a patient. However, it goes without saying that the process improvement needs to be done without reflection to the patients' safety or data security.

The work carried out within the WILHO Consortium was realised also in a WILHO Roadmap, which introduces the path towards the final goal, a wireless hospital. Typically the transition plan is made for 5–7 years. As a background information for WILHO works, a comprehensive inquiry within the global healthcare operators has been accomplished and analysed [5].

19.2.1 Network Topologies

The core idea of the WILHO concept is a wireless hospital area network that has an open interface with the hospital information system and electronic patient records, and also with various applications related to different hospital processes [3].

The use of a flexible radio interfaces, e.g. software defined radio (SDR) [6], can support several radio standards. By replacing the cables with wireless connections, the maintenance of a network infrastructure becomes much easier and cheaper. During the installation of new services or communication standards, there is no need for constructional maneuvers due to wireless technology from an infrastructure viewpoint.

19.2.1.1 Wireless Hospital Area Network

One of the most important features that a wireless hospital area network needs to support is heterogeneity. Several standardized radio protocols could be used to access the core network, which can be either wireless or based on a wired solution. A flexible radio interface improves the future enlargement ability of the system. Therefore, the system is not limited only to the radio technologies that are available during the initial system installation. SDR allows easy radio configuration between different standards without a change in the hardware.

Because all patients' real time information is stored in electronic databases, which can be accessed everywhere inside the coverage area of the wireless network, the care taking processes will improve. New diagnostic information coming from the laboratory or monitoring device can be directly taken into account in the decision making process by any doctor involved in the patient's nursing. This eliminates the delays that are coming from the manual data booking processes.

Using the wireless network, patient information can also be automatically transferred into the personal digital assistant (PDA) device of the staff when they are approaching the patient. Due to the automated mechanism, all the latest information related to the patient is always available. Similarly, the vital signs can be measured and included to EPR during the patient's transportation, as is illustrated in Fig. 19.1. Secure autenthication and encrypted data transmission will quarantee the data security.

19.2.1.2 Wireless Body Area Network

On the patient's side, monitoring of vital parameters in real-time is possible using wireless sensors and non-invasive measuring instruments. Removing the cables between the monitoring devices and the patient significantly improves the movement ability. Simple tasks, such as visiting canteen or toilet, might be impossible if a patient is connected to the medical equipment with a cable. Instead of just lying in bed, a patient could walk. And still, the nursing staff has a control to patient's vital functions. The ability to move would raise a patient's spirit and stimulate the overall healing process.

Data from non-invasive sensors or monitoring nodes can be directly transferred to the access point, which is passing the message to the hospital's core network. The other option is to collect data in a centralized manner from all the nodes controlled by a portable base station (PBS) carried by nursing staff. The latter option utilizes point to point links between the sensor nodes and a PBS. The connection between the individual sensor and the fixed infrastructure is therefore created via PBS. Heterogeneous thinking could also be valid inside the wireless body area network (WBAN). This allows the utilization of different radio protocols in the system, as discussed above. The general idea and the main interfaces of a WBAN concept are depicted in Fig. 19.2 [7].

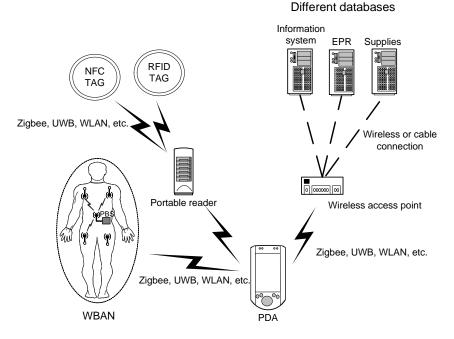


Fig. 19.2 Wireless body area network and its interfaces for health monitoring

For the use of ultra low power active medical implants, the Electronic Communications Committee (ECC) has already defined 1 MHz frequency

bands between 401–402 MHz and 405–406 MHz 8. Similar work has been carried out within an IEEE 802.15.6 study group [9] to define a standard for medical body area networks.

The same procedures discussed earlier can also be exploited in in-human monitoring. WBANs and PBSs can be used, for example, to route signals coming from devoured sensors to the core network.

Not only restricted to controlled medical treatment, these approaches can be used in long term, real-time and out of hospital patient monitoring. Afterwards, all the latest patient data is available for diagnostic use by any doctor or nursing staff, anywhere. In addition, using advanced networking techniques, the data can be delivered to specialists located even in other countries. By establishing a real-time tele-health link, the patient care process could be optimized by selecting the best specialists for treatment process.

19.2.2 Equipment

The key elements when establishing a wireless healthcare network are equipment going to be used. The applicable devices should be small and easy to wear. A typical aspect of human behaviour is a resistance against changes. That is why the tools used in the wireless healthcare sector should be familiar to the user. Any new device that has to be carried with, in addition to the existing gadget, will typically face resistance. This requires that new features should be implemented in the existing hand-held devices and their short term evolutions. If this is not possible, a new device should support all the existing healthcare features that are obtained daily. Such a light personal device can be a PDA, laptop computer or a cellular phone.

The patient is also requiring more advanced services in the patient room. Wishes to Internet access or personal emails even from the ward are not so unusual nowadays. In institutions, personal amusement services can be supported by computers, which have been integrated in the patients' beds. Such a peripheral device can also be used by the nursing staff during their control circulation. In addition, this kind of dual use reduces system implementation costs.

Patient monitoring includes various vital functions, such as electrocardiogram (ECG), saturation of a peripheral oxygen, blood pressure, body temperature, respiratory rate, etc. Non-invasive and wireless technology is widely applicable in this context. For example, as a part of the WILHO research, a sensor belt was developed and tested for respiratory rate measurement [10,11]. High resolution accelerometers and pressure sensors were also utilised successfully in the healthcare experiments. The source signals from various sensors could have different power levels, data rates and update frequency requirements. These issues should be carefully addressed at the access point where the signals are combined and packetized for wireless transfer. Examples of the biomedical measurement data parameters are shown, e.g., in [12]. Furthermore, multi-access protocols are analysed and compared, especially with respect to the power requirements.

19.2.3 Technology

The radio technology used in the peripheral is not the main issue because several features to be utilized can be implemented using the most common radio protocols. On the other hand, different radio standards can give additional value to the application due to their own characteristics. For example, ultra wideband (UWB) technology is superior in the localization accuracy point of view [13]. An UWB based radio interface can also be used when the amount of transferred data is high, or very fast wireless links need to be established. UWB signals can also be used in transmission power limited environments. On the other hand, radio techniques such as Zigbee [14], Wibree [15] and Bluetooth [16] can also be used when data rate requirements are not so high. Near field communication (NFC) [17,18] and radio frequency identification (RFID) [19,20] tags support services that can also be utilized under the wireless hospital concept and as well as in other applications.

Different IEEE 802.11 based radios support both data communication and localization purposes. The technology is also very mature, and there are lots of applications and devices already commercially available. Nowadays, the IEEE 802.11 standard is also a universally utilized wireless technology in hospital wireless communication networks.

In Finland, the use of personal mobile phones is allowed at most of the wards. In addition, IEEE802.11 based wireless local area networks are commonly used. For all that Finland is quite liberal in the technology adaptation, the situation differs globally. Local radio regulations are defining if the wireless system is allowed to be used inside a hospital or not. When increasing the number of transceivers in a space, special care needs to be taken for co-existence issues. None of the new technologies are allowed to disturb existing systems operation, and vice versa. Typically, the regulations for hospital instruments are much tighter than the corresponding ones for commercial devices.

Patients should be capable of taking care of themselves despite different treatment methods, technologies and tools, perhaps under the supervision of an automated healthcare system or care providers. Seamless operation is then the key issue. On the other hand, different technologies and processes in healthcare should support the care providers' tasks, not the other way around. This may mean, for example, context sensitive, location sensitive or problem-sensitive wireless applications. The connection to the service providers' data base does not necessary finish when the patient is sent out of the hospital. Patients might need a connection to the healthcare system even after they have left the hospital. Using wireless technology, this link could be maintained. Again, during remote access, the data security has a key role.

19.3 Application Areas

19.3.1 Wireless Hospital

In the next chapters, some ideas toward wireless hospital concept and utilization of wireless technologies at hospital are illustrated.

19.3.1.1 Hospital Registration and Patient Tracking

When the patient first time arrives to the hospital, or what ever institute adopting the proposed wireless concept, there is only one check-in desk where the registration needs to be done. During the check-in process, all the information related to the patient and his or her visit is asked and recorded. Next, the patient is provided by a light weight wireless tag, which contains all the necessary information concerning the visit to the institution in question. The information stored in a tag could be very generic and it is defined by the requirements of particular treatment processes and hospital practices.

Later on, the intelligent hospital infrastructure can guide the patient's progress inside the building. Patient specific information will be transferred to the corresponding patient by the screens located at public areas and corridors. Monitors could show expected waiting times, correct location information and guidance to proceed if being in a wrong place. The shown information at the screen should not include any private information about the patient, just the tag ID-number to maintain patient's privacy because the screens are visible to all the people around. Of course, there could be dedicated monitors available to read more detailed personal data. Access to personal information can be done easily by using the ID-tag. RFID is a mature technology for this kind of applications.

Both active and passive realizations of the ID-tag could be utilized. In the former case, a patient touches the tag to the reader to get more information. In the latter case, the system automatically detects the presence of a tag, similarly than the burglar alarming system works at shops. When the patient is approaching, e.g., a certain hall or investigation room, a notification can automatically be sent to the nursing stuff after the tag is identified.

19.3.1.2 Real-Time Phase Information

Real-time tracking could be exploited in several hospital logistic actions. For example, an RFID tag embedded into an identification card or a mobile phone, or corresponding device, allows easy access to the data network. A simple illustration is when a cleaner enters an operating-room, the RFID event will be stored to the hospital data base, and the following RFID event indicates that the work process has been completed. Next, immediate alarm could be directed to the nurses and doctors, such as the surgeon or anaesthesia doctor. In a ward, the patient can be prepared in time because the statuses of the other ongoing processes related to the coming medical action are known. The previous case describes one part of the hospital's very important phase information, which links consecutive processes and facility management together.

Optimization of the process flow can be seen as a major source for efficiency improvement at hospitals. Automated processes can then save time, and make the utilization of the operating-room or X-ray, etc. more beneficial by shortening the transfer times between the consecutive patients. At the end, this improvement has a great influence on the hospital's cost structure which is mainly consisting of labour costs.

19.3.1.3 Positioning

The location and tracking (LT) applications inside a hospital can be based on largely adopted wireless local area network (WLAN) standards, such as IEEE 802.11. The adoption of global positioning system (GPS) indoors is not a feasible solution due to the very low received signal power at the ground level. Obstacles, such as roofs, will easily block the signal and make GPS useless indoors.

On the other hand, IEEE 802.11 based LT is already utilized in several environments and the technology is mature. Typically, hospitals and other public premises have WLAN already installed, at least in Finland this is the case. The technology could offer area based LT services if very high accuracy is not needed. For example, a coarse positioning could be enough to locate people or instrument inside a building. If the accuracy demands are higher, there are also more advanced techniques available, such as ultra wideband, which could offer centimetre level accuracy due to the extremely large inherent signal bandwidth [13].

It is also possible to merge a positioning service to a security application. Under physical threat, a staff member could send an alarm signal, which carries also the location information. The need for such services is nowadays becoming more popular in public departments, including the healthcare sector and social services.

19.3.1.4 Real-Time Material Tracking and Monitoring

Interesting applications where wireless technologies can be utilized are, e.g., monitoring storehouses' contents or material ordering process. This part is easily exploitable from the healthcare sector to every storehouse. Automated processes in logistic chain will reduce costs in a long run.

At hospitals, the tracking of medicine, implants and other tools and instruments is typically done manually. In the worst case, the process requires double writing; once in the warehouse and the other time in the office, where information is added to the electronic database or electrical subscription system.

Keeping real time track on medicine and other supplies at hospital is one more challenging task. At the same time, material tracking is linked to the safety processes. By using hospital level stock monitoring, for example the amount of drugs can be maintained as low as possible. Simultaneously, low storage minimizes the potential losses caused by a house-breaking.

Integrating all hospital's electrical databases to planning tools ensures that all the needed supplies and personnel are available during the operation. Real time material tracking relates to stock monitoring but in addition it can be extended to invoicing processes. If all the material flow inside the hospital can be followed piece by piece, real care costs can be easily directed to the corresponding patient. For example, implants and other expensive materials such as screws can be charged based on the real consumption. Material tracking could also be extended to maintenance services.

19.3.1.5 After-Care

After a patient has been treated at a hospital and disbanded, e.g., after the surgery, an after-care monitoring can be carried out using new tele-health techniques. A patient could be equipped with wireless sensors, which are collecting vital parameter information. Data could then be transferred to personal computer and routed to hospital electronic patient record. The

established link can also be two directional, so the patient can easily get feedback and guidance from the hospital.

Using the advanced body area sensor network it is also possible to monitor the patient's feelings. This information can be helpful when designing the medication and training exercises for the patient. If the feelings can be recognized remotely, it also improves the patient's safety sensation. Merging wireless sensor networks and mood recognition has been studied, e.g., in European Union funded FP6-project e-Sense [21].

19.3.1.6 Pain Meter

As a side project of WILHO activities in Oulu, a wireless pain meter has been developed jointly by the ISG and the University of Lapland, Rovaniemi, Finland. This device can be used to collect patient's subjective pain feelings, and to transfer the momentary information immediately to the nursing staff, or a database for after-care monitoring. The meter scales ranges from zero (no pain) to ten (unbearable pain), and the patient can select the most appropriate indication of his/her pain level. If the level exceeds the pre-specified threshold, an immediate alarm to the nurse's terminal will be sent [22].

The pain meter could be used at hospitals but also it gives added value to home diagnostics and after-care.

19.3.2 Wireless Sensors

Embedding wireless sensor networks into the environment increases the utilization of WBANs in healthcare or welfare sector. The sensors are ideal for sensing, e.g., toxic chemicals in close vicinity of patient, or in the neighbourhood. Sensor nodes are possible to implement in other measuring devices, such as heart rate monitors, cars, weather stations, watches etc. The remote monitoring could be carried out through the cellular network, for example. A mobile terminal (a phone or PDA) could also be the final device to be used with the application. Remote links can also be obtained via tele-health services or satellite communication.

Wireless sensors could also be applied in animal care like discussed in [23]. Real time health information from livestock could be collected in cow-houses, etc. environments to get added value to food production chain. Technical solutions for animal monitoring systems are following the ones developed for human control. The attention has to be paid for devices that have low power consumption and are small. Individual animal

identification throughout its life could be based on the same monitoring tag or sensor that is used to monitor its health condition.

19.3.3 Sport Training

The healthcare sector is not the only one where wireless technologies in the human well-being improvement can be utilized. As an example, a wireless positioning service that can be used to deliver simultaneously human vital parameters in real-time is implemented in a skiing tunnel in Vuokatti, Finland. Using the in-tunnel wireless sensor infrastructure, the performance of the skier can be remotely monitored in real-time. It is also possible to evaluate the efficiency of the training session afterward with the trainer. [24] Both professional and amateur athletes can benefit from this kind of exercise environment infrastructure.

19.3.4 Enterprise Resource Planning System

To support the adaptation of wireless technologies in hospital environments, an enterprise resource planning system (ERP) was created within the WILHO project. Using the ERP it is possible to quantitatively measure the cost benefit, which is coming due to the new process models.

To make the decision of purchasing and adopting new technologies in hospitals easier, all the processes within nursing actions, including the equipment expenses, are modelled as consecutive process phases. The calculation outputs the total expenditure advantage from new technology installation taking also into account labour costs. The final evaluation criteria are improved efficiency and quality of nursing, reduced costs and better staff well-being. To get all the information out from the ERP, similar calculations should be done before and after the new technology investments. It should be noted that the final cost improvement can be seen only in a long run.

19.4 Conclusions and Future Visions

This article reviews the recent research activities and visions that are mainly based on the WILHO Consortium's work in Oulu region, Finland, towards the wireless hospital concept. The outlined hospital concept will utilize advanced wireless technologies to improve cost efficiency and quality of care, as well as hospital processes and logistics. The project findings will shift nurse's work load from supporting tasks to nursing and real patient care activities. The findings and ideas from the WILHO work will be piloted in two new buildings by the hospitals involved in the project.

In the future, tele-health will play a big role in the global healthcare business. There are not enough specialists for every hospital, so the utilization of the common knowledge base will increase. Tele-health might be an answer for this increasing problem. Advanced data communication systems are allowing real time video conferences but also remote operations are possible. These technologies are already used in developing countries. Tele-health can also be adopted in sparsely populated regions where nearest health centre or hospital is far away. However, the technology could be made available for all people.

Acknowledgment

The authors would like to thank all the WILHO Institutes for fruitful cooperation. The Finnish Funding Agency for Technology and Innovation (Tekes) through their FinnWell technology program, Nokia Oyj and Sensinode Ltd. are acknowledged for the funding.

References

- Rissanen H, Isokanniainen K, Pirinen P, Alasaarela E (2006) Wilho a new concept of wireless management of healthcare processes. 6th Nordic Conference on eHealth and Telemedicine (NCeHT), Helsinki, Finland, pp 229–230
- 2. C-Quest Partners Oy (2006) Comparative study of a Finnish and a UK district hospital's doctors' time spending patterns. Unpublished research report.
- 3. http://www.wilho.net/
- 4. Hämäläinen M, Pirinen P, Shelby Z (2007) Advanced wireless ICT healthcare research. The 16th IST Mobile and Wireless Communication Summit, Budapest, Hungary.
- Alasaarela E (2006) Strategy of developing wireless solutions in healthcare, unpublished report. Available from http://akseli.tekes.fi/opencms/opencms/ OhjelmaPortaali/ohjelmat/FinnWell/en/system/uutinen.html?id=2631&nav= News
- 6. http://www.sdrforum.org/
- Ashraf MI, Härkönen M, Hämäläinen M, Riekki J (2007) Health care process management supported by wireless technology. Finnish Signal Processing Symposium 2007, Oulu, Finland

- ECC (2006) Coexistence between ultra low power active medical implants devices (Ulp-Ami) and existing radiocommunication systems and services in the frequency bands 401–402 MHz and 405–406 MHz. Lübeck, Germany
- 9. http://www.ieee802.org/15/pub/SGmban.html
- Reinvuo T, Hannula M, Sorvoja H, Alasaarela E, Myllylä R (2006) Measurement of respiratory rate with high-resolution accelerometer and EMFit pressure sensor. Proc. IEEE Sensors Applications Symposium (SAS 2006), pp 192–195
- 11. Alamäki H, Sliz R, Ashraf MI, Sorvoja H, Alasaarela E, Hämäläinen M (2007) Measurement of ECG, respiratory rate, tilt and temperature of a patient and wireless Zigbee data transmission. The 2nd International Symposium on Medical Information and Communication Technology, Oulu, Finland
- Arnon S, Bhastekar D, Kedar D, Tauber A (2003) A comparative study of wireless communication network configurations for medical applications. IEEE Wireless Communications 10: 51–56
- Montillet JP, de Abreu GTF, Saarnisaari H, Oppermann I (2005) Comparison of UWB and WCDMA positioning accuracies. 2005 IEEE 61st Vehicular Technology Conference
- 14. http://www.zigbee.org
- 15. http://www.wibree.com
- 16. http://www.bluetooth.com
- 17. ECMA International (2004) Near field communication, White paper. Ecma/TC32-TG19/2004/1, 9 p
- NFC Forum. Near field communication and the NFC Forum: The keys to truly interoperable communications. Available from http://www.nfcforum.org/resources/white_papers/ nfc_forum_marketing_white_paper.pdf
- 19. Finkenzeller K (2003) RFID handbook, 2nd ed. Wiley, England
- 20. Riekki J, Salminen T, Alakärppä I (2006) Requesting pervasive services by touching RFID tags. IEEE Pervasive Computing 5(1): 40–46
- 21. http://www.ist-esense.org/
- 22. Riekki J, Alakärppä I, Koukkula R, Angeria J, Brockman M, Saloranta T (2007) Wireless pain monitoring. The 2nd International Symposium on Medical Information and Communication Technology, Oulu, Finland
- 23. Hämäläinen M, Kohno R (2007) Prospects for Wireless Technology in Remote Care Processes. The 2nd International Symposium on Medical Information and Communication Technology, Oulu, Finland
- Oppermann I, Stoica L, Rabbachin A, Shelby Z, Haapola J (2004) UWB wireless sensor networks: UWEN – a practical example. IEEE Communications Magazine 42(12): S27–S32

20 Analytical Analysis of the Performance Overheads of IPsec in MIPv6 Scenarios

Zoltán Faigl, Péter Fazekas, Stefan Lindskog, and Anna Brunstrom

20.1 Introduction

The next generation network (NGN) connects different access networks, such as xDSL, 3G, WiFi, and WiMAX to an IPv6-based core network. One of the requirements of NGN is to support the mobility of services, users, and terminal equipments [6]. The mobile IPv6 protocol (MIPv6) [12] and its extensions, such as hierarchical MIPv6 [20], fast handovers for MIPv6 [15], and network mobility protocol [4], provides one major possible mobility service solution. Other solutions also exist, and a discussion and comparison of the main mobility protocols can be found in [16].

In systems supporting mobility a wide variety of threats exist [2, 18]. Reliable and secure communication of mobility signaling protocols, such as MIPv6 and its extensions, is therefore critical. An important challenge is therefore how to integrate security solutions into the signaling protocols of the NGN.

One of the possible choices for security solutions is the use of IPsec [14] with the Internet key exchange protocol version 2 (IKEv2) [13]. MIPv6 recommends the use of IPsec and IKEv2 for the protection of signaling messages, between the mobile node (MN) and the home agent (HA) [1]. However, IPsec and IKEv2 enable a very wide range of configuration possibilities. It is thus an important question for network designers to determine which configuration to apply in a specific situation. An informed decision requires that the security levels and costs of the performance overheads of the possible security configurations are known. The decision for the best security configuration can then be made by specifying a trade-off between security and performance.

This chapter aims to demonstrate how the performance overheads of the different protection policies of IPsec can be analyzed in a MIPv6 scenario.

We highlight two main performance measures in our analysis: the overall utilization of the HA by the protection of MIPv6 signaling processes and the total mean response time for a mobility process in the network. The results can serve as input for network and node dimensioning, and for designing an appropriate trade-off between security and performance. How to dynamically tune IPsec to achieve an appropriate security-performance trade-off has previously been considered in other contexts [21,23].

Our analysis is based on the theory of queuing networks. We use a simplified version of the BCMP¹ theorem, for multiple-class open queuing networks with load-independent arrival rate and service times [3]. A similar approach, based on a closed queuing network, has previously been applied to analyze different types of the Kerberos authentication protocol [11]. However, we choose an open queuing network model since it gives us the possibility to directly influence the arrival rate of mobility processes in the analysis.

Our analysis characterizes generally the overheads imposed by IPsec in protecting MIPv6 signaling, since we use an abstract expression for the processing overheads induced by the different cryptographic algorithms. The performance cost of an algorithm is expressed primarily by the total number of operations required per block for a processor when applying the specific algorithm. These parameters were given in [5] for DES, MD5, and SHA-1, and in [9] for AES. Xenakis et al. [22] gathered them in their work related to a generic characterization of the space and processing overheads of different IPsec configurations.

The structure of the rest of the chapter is as follows. Section 20.2 presents the security configurations considered in the chapter. Section 20.3 describes the reference scenario and the network model used for the analysis. Section 20.4 presents the calculation method, the input parameters and the results of the analysis. Finally, Sect. 20.5 concludes the chapter.

20.2 Security Configurations

The available recommendations for the protection of MIPv6 signaling between a MN and a HA with IPsec suggest a wide range of configurations [1]. The main signaling protocol steps triggered by a movement of a MN to a new IP-domain are illustrated in Fig. 20.1.

¹ BCMP is an abbreviation for Baskett, Chandy, Muntz, and Palacios, which are the inventors of the theorem.

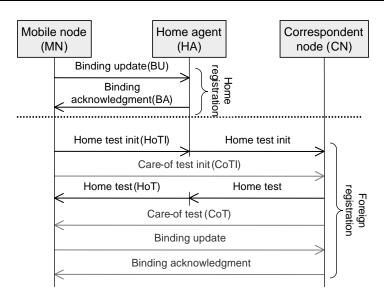


Fig. 20.1 MIPv6 signaling when the MN moves to a foreign IP-domain

In the present work we have analyzed two main policies that are different in granularity and ten different algorithm selection possibilities for each of them. The policy having the coarsest granularity is referred to as Policy 1. It specifies that all MIPv6 signaling traffic should be protected by encapsulating security payload (ESP) in tunnel mode with a non-null encryption and data origin authentication algorithm. The second policy, referred to as Policy 2, has a finer granularity. It specifies that the binding update (BU) and binding acknowledgment (BA) messages should be protected by ESP in transport mode with a non-null authentication algorithm. The home test init (HoTi) and home test (HoT) messages should be protected in the same way as in Policy 1. The other signaling messages related to MIPv6 are not considered in this analysis, since they are not triggered by every movement of the MN. At the algorithm selection level, we consider five types of encryption algorithms, i.e., DES, 3DES, AES with 128 bit key (AES128), AES192, and AES256, and two keyed-hash message authentication code (HMAC) algorithms, i.e., HMAC with MD5 and HMAC with SHA-1. Both Policy 1 and Policy 2 can be configured with any of the encryption and HMAC algorithms. As a reference for the calculations of performance overheads, we also consider a policy, referred to as Policy 3 that does not apply any protection at all.

20.3 Reference Scenario and Network Model

We chose our reference scenario based on current recommendations in standards and projects that suggest MIPv6 as a mobility service [6, 10, 19]. Here, MIPv6 is put forward as a candidate for handling the mobility of terminals performing vertical handovers between different access networks in the NGN. Typically, each access network represents different IPdomains. A MIPv6 signaling process is initiated when a MN arrives to a new IP-domain and gets a new care-of address. As illustrated in Fig. 20.1, the MIPv6 process involves a home registration phase between the MN and the HA, and a foreign registration phase between the MN and the correspondent node (CN). One part of the latter goes through the HA. Figure 20.2 illustrates our reference scenario. By moving, a MN may reach different access networks, such as WiMAX, 3G, and WiFi. WiMAX is supposed to provide internet access to the users with dedicated data rates. In a new access network the MN gets a new IP-address by stateless autoconfiguration.² The new care-of address must then be registered at the mobility service provider of the MN, i.e., the HA. We suppose one mobility service provider with one HA situated somewhere in the IPv6 backbone. Furthermore, we suppose that a MN communicates with one CN while moving, and CNs are located on the internet.

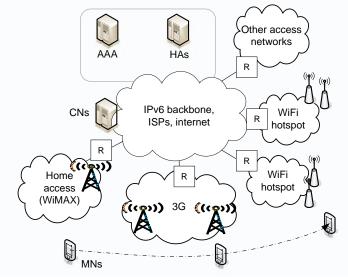


Fig. 20.2 Reference scenario

² Multi-homing questions are not considered in this work. We suppose that a movement always causes alteration of the primary care-of address of the MN.

Based on the scenario, we established the network model presented in Fig. 20.3. In order to fulfill the requirements of the BCMP theorem, which allows basically four node types, we chose the following three node types for the queuing network: M/M/1-FCFS for the UMTS, WiMAX, and WiFi access networks (with separated uplink (UL) and downlink (DL) directions in case of UMTS and WiMAX access); M/G/1-PS for the MN and the HA; and M/G/∞-IS for the modeling of the internet and internetwork delays.³ This choice is motivated as follows. Jobs are served in parallel at PS nodes, resembling the concurrency present in a single node, and in series at FCFS nodes, reflecting the sequential nature of access networks. The IS model reflects high-speed network connections where queue formation is negligible and the packet transfer time can be modeled as a linear function of the distance and the packet size. We did not apply more detailed models for the different parts of the networks, since our aim was to compare the different security configurations from a high-level perspective in a largescale scenario.

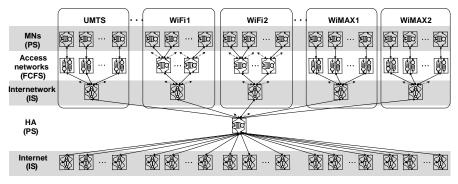


Fig. 20.3 Network model

20.4 Performance Analysis

In the following, Sect. 20.4.1 describes the analytical method that we applied to calculate the performance overheads of IPsec when protecting MIPv6 signaling. Section 20.4.2 gives the input parameters for the calculation. Finally, Sect. 20.4.3 presents our research questions and discusses the results given by the analysis.

³ FCFS, PS and IS are abbreviations for first-come-first-served, processor sharing and infinite server, respectively.

20.4.1 Calculation Method

An analytical solution method has been adopted using Matlab. We applied a simplified version of the BCMP theorem for multi-class switching, open queuing networks that have load-independent arrival rates and service times [3]. It enables studying the steady-state behavior of the network and the nodes without resorting to a system of equations for the global statespace. However, all nodes must fulfill certain assumptions concerning the distributions of inter-arrival and service times, and queuing disciplines. The choice of an open queuing network model (instead of closed) allows the arrival rate of mobility processes to be directly influenced during the analysis. The simplified BCMP method supposes load-independent service times. This limitation primarily influences the latency modeling of the networks. The use of M/M/1-FCFS nodes implies that packets with different sizes are served with exponential service times with equal means. Another limitation is that the arrival process of job chains, i.e., the movement events of MNs, are assumed to be Poisson. The main steps of the method are as follows:

- 1. Calculate the number of visits e_{ir} for the network, where $r \in R$ (the set of job classes) and $i \in I$ (the set of nodes). The definition of job classes and the visit ratios of the classes are derived from the MIPv6 message flow. Each message type, i.e., BU, BA, HoTi, and HoT, was categorized into different job classes. Thus, we had four job classes within one mobility process. In order to keep up the routing information for the mobility processes, we formed separate class groups for the mobility process of each MN.
- 2. Determine the job chains C_u where $u \in U$ (the set of all mobility processes). One chain contains the job classes that form a closed group in terms of routability, i.e., one chain contains those four job classes that belong to the same mobility process. We have one mobility process chain for each MN in our network model. The mobility process chains are triggered by the arrival rate of the MNs in new IP-domains. The illustration of a mobility process chain can be seen in Table 20.1.
- 3. Compute the number of visits e_{iu}^* for each chain:

$$e_{iu}^* = \sum_{r \in C_u} e_{ir} \ . \tag{20.1}$$

Table 20.2 shows the number of visits e_{ir} and e_{iu} of the mobility process chain C_i containing the job classes $r = \{1, 2, 3, 4\}$. The table presents the case of separate UL and DL access channels.

4. Determine the scale factors α_{ir} , i.e., the relative weight of the number of visits of jobs from class *r* in node *i* compared to the number of visits of all the classes from the same chain visiting node *i*.

$$\alpha_{ir} = \frac{e_{ir}}{\sum_{p \in C_u} e_{ip}}, r \in C_u$$
(20.2)

Table 20.3 shows the values for α_{ir} for mobility process chain one (C_1) which contains the job classes $r = \{1, 2, 3, 4\}$ It presents the case of separate UL and DL access channels.

5. Calculate the service times s_{iu}^* for each chain, where s_{ir} is the service time of a job from class *r* in node *i* The calculation of s_{ir} is described in Sect. 20.4.2 and illustrated in Table 20.1.

$$s_{iu}^* = \sum_{r \in C_u} \alpha_{ir} s_{ir}$$
(20.3)

- 6. Derive the performance measures per chain, using a simplified version of the BCMP theorem for open queuing networks with load-independent arrival and service rates.
 - Utilization of node *i* with respect to jobs of the chain C_u can be given as in (20.4).

$$\rho_{iu}^{*} = \begin{cases} \lambda_{u} e_{iu}^{*} s_{i}, & \text{Type-1: M/M/1-FCFS} \\ \lambda_{u} e_{iu}^{*} s_{iu}^{*}, & \text{Type-2,3:M/G/1-PS,M/G/\infty-IS}, \end{cases}$$
(20.4)

where λ_u is the arrival rate of the job chain C_u , i.e., the arrival rate of a mobility process at MN_u In case of M/M/1-FCFS nodes, s_i means that the service times s_{iu} in Type-1 nodes must be equal for all jobs of chains.

- Overall utilization of node *i* is

$$\rho_i = \sum_{u \in U} \rho_{iu}^* \,. \tag{20.5}$$

- The mean response time of all the jobs of a chain C_u in node *i* is

$$T_{iu}^* = \frac{S_{iu}}{1 - \rho_i} \,. \tag{20.6}$$

- The overall mean response time for a chain C_u in the network is

$$T_u^* = \sum_{i \in I} T_{iu}^* .$$
 (20.7)

Table 20.1 Mobility process chain one C_1 triggered between MN₁ and the HA that is caused by a movement of MN₁

Node name &	Jobs related to IPsec Protection: description	Service times (s_{ir})
Index (i)	& job class (<i>r</i>)	
MN (1)	Protect BU (1)	pr ^a _{BU,1} /speed _{MN}
LAN (2)	Transfer BU in the access network (1)	s_2^{b}
$WAN_{1}(4)$	Transfer BU in the internetwork (1)	size _{BU} /data rate _{WAN1}
HA (5)	Unprotect BU (1)	pr _{BU,2} /speed _{HA}
HA (5)	Protect BA (2)	pr _{BA,1} /speed _{HA}
$WAN_{1}(4)$	Transfer BA in the internetwork (2)	size _{BA} /data rate _{WAN1}
LAN (2/3°)	Transfer BA in the access network (2)	s_2 or s_3
MN (1)	Unprotect BA (2)	pr _{BA,2} /speed _{MN}
MN (1)	Protect HoTi (3)	pr _{HoTi,1} /speed _{MN}
LAN (2)	Transfer HoTi in the access network (3)	<i>S</i> ₂
$WAN_{1}(4)$	Transfer HoTi in the internetwork (3)	size _{HoTi} /data rate _{WAN1}
HA (5)	Unprotect HoTi (3)	pr _{HoTi,2} /speed _{HA}
$WAN_2(6)$	Transfer HoTi to CN (3)	size _{HoTi} /data rate _{WAN2}
CN	CN processes HoTi	0^{d}
$WAN_2(6)$	Transfer HoT to HA (4)	size _{HoT} /data rate _{WAN2}
HA (5)	Protect HoT (4)	pr _{HoT,1} /speed _{HA}
$WAN_{1}(4)$	Transfer HoT in the internetwork (4)	size _{HoT} /data rate _{WAN1}
LAN (2/3)	Transfer HoT in the access network (4)	s_2 or s_3
MN (1)	Unprotect HoT (4)	pr _{HoT,2} /speed _{MN}

 ${}^{a}pr$ means processing requirement given in number of instructions. Subscripts 1 and 2 for *pr* were introduced in order to differentiate between the encryption and decryption processes of the same message. This is needed due to the AES encryption algorithm which differs in the performance costs for encryption and decryption (see Table 20.5).

 ${}^{b}s_{2}$ represents a constant service time of M/M/1-FCFS nodes. s_{2} can be calculated as the weighted sum of the service times of job classes visiting this node. The weights are determined by the number of visits of jobs from the given class relative to the sum of the number of visits from all chains.

 $^{c}2/3$: if the DL and UL channel are common in the access network (see WiFi access in Fig. 20.3), then the node index of UL is the same as for the DL direction (i.e., 2), since it is the same node. If the UL and DL channels are independent (see 3G and WiMAX access in Fig. 20.3), then the UL node index is different from the DL node index (i.e., 3).

^dThe processes at the CN were not considered in our calculations.

		$r^{a} =$				
	e_{ir}	1	2	3	4	e_{iu}
MN	$i^{a} = 1$	1	1	1	1	4
LAN _{UL}	2	1	0	1	0	2
LAN _{DL}	3	0	1	0	1	2
WAN_1	4	1	1	1	1	4
HA	5	1	1	1	1	4
WAN ₂	6	0	0	1	1	2

Table 20.2 Number of visits of the jobs of classes $r = \{1, 2, 3, 4\}$

^aTable 20.1 shows the meaning of i and r.

Table 20.3 Scale factors α_{ir}

	<i>r</i> =			
α_{ir}	1	2	3	4
i = 1	0.25	0.25	0.25	0.25
2	0.5	0	0.5	0
3	0	0.5	0	0.5
4	0.25	0.25	0.25	0.25
5	0.25	0.25	0.25	0.25
6	0	0	0.5	0.5

20.4.2 Input Parameters

We divide the presentation of the input parameters into three parts: specification of mobility models, parameters of the network model, and calculation of the service times of different job classes in each node. These distinguished parts were needed in order to calculate (20.4), (20.5), (20.6), (20.7). The input parameters introduced in this subsection were used in the analysis unless otherwise indicated. In some parts of the analysis the processing speed of the MN, the UMTS UL and DL data rates, the number of users in the system, and the mobility rate between different IP-domains were varied.

20.4.2.1 Mobility Scenarios

We used two mobility scenarios for the mobility model of a user: a pedestrian and a car scenario. There were 10^6 users in the pedestrian mobility scenario and 10^5 users in the car mobility scenario. We supposed that in a given scenario every user followed the same mobility model. The mobility of a user is illustrated by the state diagrams in Fig. 20.4(a),(b) for the pedestrian and car scenario, respectively. Note that the WiMAX, UMTS and WiFi states may not represent one instance but an aggregation of access networks of the same type. Moreover, in Fig. 20.4(a) the UMTS states represent one UMTS access network that covers the area globally. A MIPv6 movement to a UMTS access network is performed when other access networks disappear from the range of the MN. Note also that for a given access network type the mean mobility process rate of a MN, used in (20.4), can be calculated as the sum of the outgoing intensities of the states representing the given access network type, multiplied by the steady-state probabilities of the states.

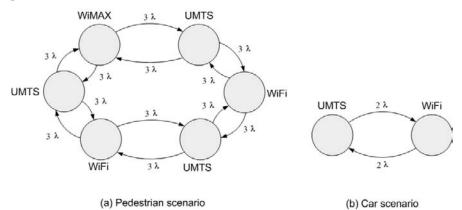


Fig. 20.4 Mobility state diagram of a user

The pedestrian mobility scenario can be illustrated with a campus where the user walks around, and accesses the UMTS, WiFi, and WiMAX networks. The user is supposed to spend 20 min on average in a given access network, with an exponential distribution. On average the user spends half of the time in UMTS, 1/3 of the time in WiFi, and 1/6 of the time in WiMAX. Thus, the mean mobility process rates are 3λ , 2λ , and λ , for UMTS, WiFi and WiMAX, respectively, where $\lambda = 1/7200$ [sec⁻¹].

In case of the car scenario, the user is supposed to spend on average 15 s in a given access network. The scenario can be illustrated with a car moving on a road which is partially covered with WiFi hotspots and where global UMTS network coverage is accessible. The user spends half of the time under WiFi coverage, otherwise the UMTS access network is used. Thus, the mean mobility process rate of the user is $\lambda = 1/30$ [sec⁻¹] for both UMTS and WiFi access network types.

Note that the exact details of the mobility models are not so important. Other mobility models with the same mobility process rates would give the same results. Hence the described mobility models should be seen as representative examples.

20.4.2.2 Parameters of the Network Model

The network model was fixed, having one UMTS, 10^2 WiMAX, and 10^4 WiFi access networks, representing separate IP-domains. The processing speed of the HA was fixed to 3000 million instructions per second (MIPS). When not varied during the analysis, the processing speed of the MNs was set to 600 MIPS. We used the following data rates: 64 kpbs for UMTS UL and DL, 17.6 Mbps for the WiFi, 512 kbps and 1024 kbps for WiMAX UL and DL [8, 17]. The data rates of the internetwork and internet nodes were fixed to 2 Mbps and 700 kbps, respectively.

20.4.2.3 Calculation of Service Times

To calculate the service times we first determined the processing and space requirements of the considered message types for each security configuration. Table 20.4 describes the meaning of the parameters used in the following equations.

Name	Meaning of the parameter
BS	Block size of an encryption algorithm
IP_h	Size of IPv6 header
M_h	Size of the first generic part of the Mobility header (Mh)
BU	Size of Binding Update part of the Mh
BA	Size of Binding Acknowledgment part of the Mh
HoTi	Size of Home Test Init part of the Mh
HoT	Size of Home Test part of the Mh
ACoA	Size of Alternate Care-of Address field of the Mh
DO_h	Size of Destination Options Header
$RT2_h$	Size of Type 2 Routing Header
ESP_h	Size of ESP header
ESP _{trailer}	Size of ESP trailer
ESP _{auth}	Size of ESP authentication data field
T _{encr/decr}	Processing requirement to encrypt
Thash	Processing requirement to hash a data block

Table 20.4 Meaning of the input variables

The service time (s_{ir}) to process a message is calculated as the size of the message (MS) divided by the data rate of the network, or, in case of hosts, the processing requirement (PR) divided by the processor speed given in MIPS. This is illustrated in the last column of Table 20.1.

As a first step to calculate the processing requirements, we calculated the length given in bytes of the parts to be encrypted (L_{encr} , see (20.8) and (20.9) and integrity checked (L_{hmac} , see (20.10). The four separate right-hand sides in the upcoming equations represent values for the four considered

message types, i.e., BU, BA, HoTi, and HoT. Policy 1 and Policy 2 cause different lengths of the parts to be protected in case of BU and BA messages.

$$L_{\text{encr,Policy1}} = \begin{cases} \left[(IP_h + M_h + BU + ACoA + ESP_{\text{trailer}}) / BS \right] \times BS \\ \left[(IP_h + M_h + BA + ESP_{\text{trailer}}) / BS \right] \times BS \\ \left[(IP_h + M_h + HoTi + ESP_{\text{trailer}}) / BS \right] \times BS \end{cases}$$

$$L_{\text{encr,Policy2}} = \begin{cases} \left[(M_h + BU + ACoA + ESP_{\text{trailer}}) / A \right] \times 4 \\ \left[(M_h + BA + ESP_{\text{trailer}}) / 4 \right] \times 4 \\ \left[(M_h + BA + ESP_{\text{trailer}}) / 4 \right] \times 4 \\ \left[(IP_h + M_h + HoTi + ESP_{\text{trailer}}) / BS \right] \times BS \end{cases}$$

$$L_{\text{hmac}} = L_{\text{encr}} + ESP_h \qquad (20.10)$$

As a second step to calculate processing requirements, we need to know the block sizes of the encryption and integrity algorithms and the costs of algorithms to process one block of data. Table 20.5 presents these values for the considered cryptographic algorithms.

Table 20.5 Processing requirement to encrypt, decrypt, or hash one block of data
for specific cryptographic algorithms ^a

Name	Block size BS [bytes]	Encrypt/hash/sign T _{encr} , T _{hash} [# of instructions]	Decrypt/verify T _{decr} [# of instructions]
DES	8	2697	2697
3DES	8	8091	8091
AES128	16	6168	10992
AES192	16	7512	13408
AES256	16	8856	15824
MD5	64	774	_
SHA-1	64	1110	-

^aThe processing requirement values for one block of data were retrieved from [5, 9].

Processing requirements to encrypt or decrypt $(PR_{encr/decr})$ and integrity check (PR_{hmac}) a message were calculated as in (20.11) and (20.12). The resulting values are given in number of instructions.

$$PR_{\text{encr/decr}} = L_{\text{encr}} / BS \times T_{\text{encr/decr}}$$
(20.11)

$$PR_{\text{hmac}} = (n_{k,\text{inner}} + 2) \times T_{\text{hash}} + 32$$
(20.12)

where $n_{k,inner} = \left[(L_{hmac} \times 8 + 512 + 64) / 512 \right]$ is the number of inner hashing operations in the HMAC calculation.

The constants in PR_{hmac} and $n_{k,inner}$ come from [22], and represent parts of the calculation that are independent of the size of the message to authenticate, such as the size of the length field (64 bits) or the length of the mandatory padding block (512 bits) in the number of inner hashing operations.

The space requirements of the messages (MS), given in bytes, were calculated based on the standards, Ethereal traces, and formulas in [22]. See equations (20.13), (20.14), and (20.15) for how to calculate the size of messages in case of the considered policies. Again the four right-hand sides in the equations represent the size of the BU, BA, HoTi, and HoT messages.

$$MS_{\text{Policy1}} = \begin{cases} IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + BU + ACoA + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + BA + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoTi + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{\text{trailer}} / 4 \rceil \times 4 + ESP_{\text{auth}} \\ IP_{h} + RT2_{h} + ESP_{h} + \lceil (M_{h} + BA + ESP_{\text{trailer}}) / 4 \rceil \times 4 + ESP_{\text{auth}} \\ IP_{h} + ESP_{\text{trailer}} / 4 \rceil \times 4 + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoTi + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoTi + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoTi + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoTi + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{auth}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{trailer}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{trailer}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{trailer}} \\ IP_{h} + ESP_{h} + \lceil (IP_{h} + M_{h} + HoT + ESP_{\text{trailer}}) / BS \rceil \times BS + ESP_{\text{trailer}} \\ IP_{h} + ESP_{h} + [OP_{h} + OP_{h} + O$$

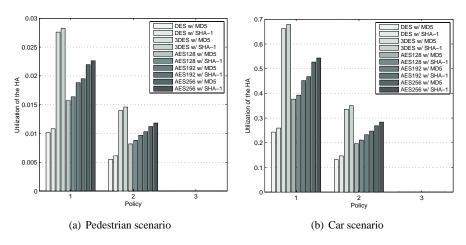
$$MS_{\text{Policy3}} = \begin{cases} IP_{h} + DO_{h} + M_{h} + BU \\ IP_{h} + RT2_{h} + M_{h} + BA \\ IP_{h} + M_{h} + HoTi \\ IP_{h} + M_{h} + HoT \end{cases}$$
(20.15)

20.4.3 Results

The first aim of the analysis was to define the performance overheads of the different security configurations in the HA, and to analyze the utilization of the HA as a function of the mobility rate and the number of users. The second aim was to analyze the effect of altering the data rate of the access network and the processing speed of the MN, and to find the boundary where the MN or the access network creates the dominating cost in the overall mean response time of a mobility process.

20.4.3.1 Utilization of the HA

The utilization of the HA for the considered security configurations is shown in Fig. 20.5. For each policy, the left most bar represents the security configuration using DES encryption with a HMAC using MD5, the right most one represents AES256 with a HMAC using SHA-1. Significant differences between the security configurations are only observed in the car scenario. (Note that the scales on the y axis differ between the graphs.) This is due to the fact that the inter-arrival rate of mobility processes to be handled by the HA is considerably higher in case of the car scenario even if the number of users is one order less than in the pedestrian case. Policy 3 has zero values, since we calculated only with the process time of cryptographic functions, and in Policy 3 no protection is applied. The differences between algorithm selections are due to the encryption and decryption processes with different encryption algorithms, and depend less on the HMAC type. DES performs always the best and 3DES the worst. The processing requirement of the AES cipher depends on the key size, and it utilizes the HA more lightly than 3DES.



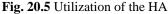


Figure 20.6 shows the overall mean response time for a mobility process in the UMTS access network in relation to Policy 3, where 100% stands for 18.5 ms. Both mobility scenarios resulted in the same values. Thus, Fig. 20.6 represents both the pedestrian and the car scenarios. This is due to the fact that none of the nodes are heavily utilized, and network latencies cause the dominating part in the response time (see (20.6) and (20.7)). Hence, the utilization of the HA does not have a great influence on the mean response time. Instead, the results show that the network latencies cause the dominating part in the response times of the mobility processes.

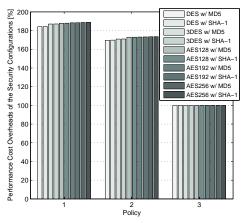


Fig. 20.6 Performance overheads of the security configurations (100% is 18.5 ms)

Figure 20.7 shows at which mobility rates and number of users the utilization of the HA reaches 100%. This represents the maximum values, for the mobility rate and number of users, that can be handled by the HA. From the two curves of the same style (i.e., using the same encryption algorithm), the higher one represents the configuration using MD5 and the lower represents the configuration using SHA-1. As we can see, Policy 2 is more lightweight than Policy 1, and allows more users or a higher mobility rate before reaching full utilization.

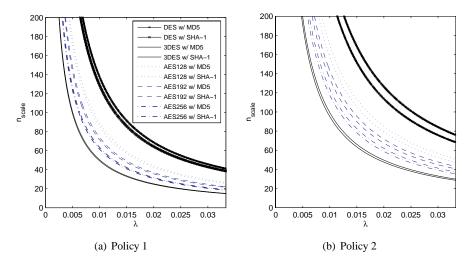


Fig. 20.7 Full utilization of the HA at different security configurations. λ represents the mobility rate, the number of users is $n_{scale} \times 10^4$

The utilization of the HA, as a function of the mobility rate, is shown in Fig. 20.8. The number of users was fixed to 10^5 . From the two curves of the same style, the lower one always represents the configuration with the MD5 algorithm. The increase of the mobility rate causes less overhead in the HA in case of Policy 2.

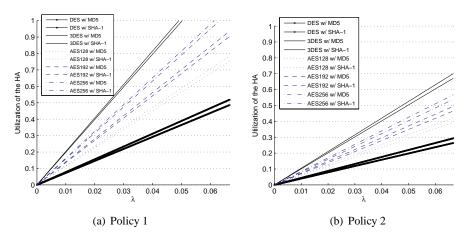


Fig. 20.8 Dependency of the utilization of the HA on the mobility rate (λ). (The number of users is fixed to 10^5)

20.4.3.2 Impact of the Processing Speed of the MN and the Access Network

The aim of the second analysis was to analyze how the alteration of the processing speed of the MN and the data rate of the UL and DL channel of the UMTS access network influences the overall mean response time of the mobility process. When are any of these components becoming the dominating factor in relative terms? The number of users was fixed to 10^5 , and the MIPv6 process rate per user to 1/30 [sec⁻¹], as in the car scenario. In Fig. 20.7, we can see that at these values the utilization of the HA is normal, i.e., the HA is not a bottleneck. The security configuration was fixed to Policy 1 with AES256 encryption and HMAC-SHA-1 authentication. Figure 20.9 shows that a decrease in the processing speed of the access network or the MN can significantly increase the overall mean response time of a MIPv6 process.

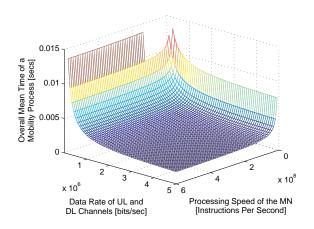


Fig. 20.9 Mean response time of a mobility process

The increase happens when the utilization, p, of the MN or the access network reaches high values (i.e., > 0.95). The bottleneck processing rate of the MN and the access network can be expressed with the general formula given in (20.16).

bottleneck speed =
$$\frac{\text{job arrival rate} \times \text{jobsize}}{\rho}$$
 (20.16)

Figure 20.10 shows the relative weight of the MN and the access network in the overall mean response time of a mobility process. From Fig. 20.10 we can see that when either the MN or the access network becomes a bottleneck, it dominates the response time of a mobility process.

When there are differences between the space or processing requirements of the different security configurations, it is worthwhile to think about designing a security service that provides an appropriate tradeoff between security and performance. Otherwise, the most secure security configuration also provides similar performance and should be selected.

For the decision making of the security configurations one may require to transfer the performance measures into one common measure. One possible solution is to apply the weighted sum of the two (or more) measures. The weights may, however, depend on who makes the decision—HA utilization is more important for the mobility service provider, while end users may be more interested in the mean response time of the mobility process.

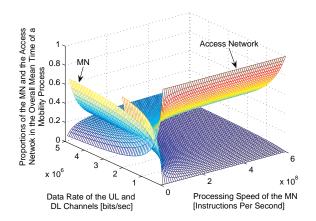


Fig. 20.10 Relative effect of the processing speed of the MN and the data rate of the access channel

20.5 Concluding Remarks and Future Work

Based on queuing analysis, this chapter has evaluated the performance overheads caused by different IPsec protection policies and algorithms when applied to MIPv6 signaling. Significant differences in the utilization of the HA between the studied security configurations could only be seen in case of the car mobility scenario. In this case it might be important to carefully tune the security configuration to meet both the performance and security requirements of the system. In the pedestrian mobility scenario, the most secure configuration also provided similar performance and would be the natural selection. In both scenarios, the overall mean response time for a mobility process was small also for the most secure configuration.

The presented analytical method can be easily adjusted to different scenarios. As long as the scenario can be properly defined, the bottleneck processing rate for each component in the network model can be found. This can be used as a support tool for network dimensioning.

In our analysis, we did not include the security association management processes that are running in parallel with the mobility signaling processes. However, we have already made a performance analysis of the costs of IKEv2 negotiations separately in [7]. We also plan to validate the analytical results with real experiments. The presented analytical method is general in nature and can be applied also to other mobility protocols or any communication flow. Something we intend to also explore in the future.

Acknowledgments

This research was initiated as part of the work on dynamic security services in heterogeneous wireless networks within the EU IST network of excellence in wireless communications (NEWCOM). The first author has been co-funded by the advanced next generation mobile and open network (ANEMONE) project, which is an EU IST project. Part of the work conducted by the third and fourth authors has been supported by grants from the Knowledge Foundations of Sweden with TietoEnator and Ericsson as industrial partners. Finally, the work conducted by the third author has also partly been financially supported by the research council of Norway.

References

- 1. J. Arkko, V. Devarapalli, and F. Dupont. RFC 3776: Using IPsec to protect mobile IPv6 signaling between mobile nodes and home agents, June 2004.
- T. Aura, P. Nikander, and G. Camarillo. Effects of mobility and multihoming on transport-protocol security. In Proceedings of the 2004 IEEE Symposium on Security and Privacy, pp. 12–26, Oakland, CA, USA, May 9–12, 2004. IEEE Computer Society.
- 3. G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. Wiley-Interscience, New York, NY, USA, 1998.
- 4. V. Devarapalli, R.Wakikawa, A. Petrescu, and P. Thubert. RFC 3963: Network Mobility (NEMO) Basic Support Protocol, January 2005.
- O. Elkeelany, M. M. Matalgah, K. P. Sheikh, M. Thaker, G. Chaudhry, D. Medhi, and J. Qaddour. Performance analysis of IPsec protocol: encryption and authentication. In IEEE International Conference on Communications (ICC 2002), pp. 1164–1168, New York, NY, USA, April 28–May 2, 2002.
- ETSI. Telecommunications and Internet converged services and protocols for advanced networking (TISPAN), NGN release 1, release definition. TR 180 001 v1.1.1 (2006-03), ETSI, 2006.
- Z. Faigl, S. Lindskog, and A. Brunstrom. Analyzing IKEv2 performance when protecting mobile IPv6 Signaling. In Proceedings of the IEEE International Symposium on Wireless Communication Systems 2007, pp. 390–395, Trondheim, Norway, October 17–19, 2007.

- 8. L. Ferreira. Definition of reference scenarios for the evaluation of radio resource allocation algorithms. Internal document, NEWCOM, June 2006.
- F. Granelli and G. Boato. A novel methodology for analysis of the computational complexity of block ciphers: Rijndael, Camellia and Shacal-2 compared. In Proceedings of the 3rd Conference on Security and Network Architectures (SAR'04), La Londe, France, June 2004.
- I. Guardini. State of the art on IP mobility and related applicability scenarios: vision of the IST ENABLE project. In Proceedings of the B3G Cluster Workshop on Mobility Technologies in the Internet, Brussels, Belgium, October 2006.
- 11. A. Harbitter and D. A. Menascé. A methodology for analyzing the performance of authentication protocols. ACM Transaction on Information System Security, 5(4):458–491, November 2002.
- 12. D. Johnson, C. Perkins, and J. Arkko. RFC 3775: Mobility support in IPv6, June 2004.
- 13. C. Kaufman. RFC 4306: Internet key exchange (IKEv2) protocol, December 2005.
- 14. S. Kent and K. Seo. RFC 4301: Security architecture for the Internet protocol, December 2005.
- 15. R. Koodli. RFC 4068: Fast Handovers for Mobile IPv6, July 2005.
- D. Le, X. Fu, and D. Hogrefe. A review of mobility support paradigms for the Internet. IEEE Communications Surveys and Tutorials, 8(1):38–51, January 2006.
- 17. Z. Nemeth and Cs. A. Szabo. Measurements to assist access network design with fixed WiMAX in urban environment. In Proceedings of the 1st international conference on Access networks (AccessNets'06), Athens, Greece, September 4–6, 2006. ACM Press.
- P. Nikkander, J. Arkko, T. Aura, G. Montenegro, and E. Nordmark. RFC 4225: Mobile IP version 6 route optimization security design background, December 2005.
- 19. Nokia. Introducing mobile IPv6 in 2G and 3G mobile networks. White Paper, 2001.
- 20. H. Soliman, C. Castelluccia, K. El Malki, and L. Bellier. RFC 4140: Hierarchical Mobile IPv6 Mobility Management (HMIPv6), August 2005.
- E. Spyropoulou, C. Ager, T. E. Levin, and C. E. Irvine. IPsec modulation for quality of security service. In Proceedings of the Third Annual International Systems Security Engineering Association Conference (2002 ISSEA Conference), Orlando, FL, USA, March 13–15, 2002.
- 22. C. Xenakis, N. Laoutaris, L. Merakos, and I. Stavrakakis. A generic characterization of the overheads imposed by IPsec and associated cryptographic algorithms. Computer Networks, 50(17):3225–3241, December 2006.
- P. K. Yogender and H. H. Ali. Impacts of employing different security levels on QoS parameters in virtual private networks. In Proceedings of the 24th IASTED International Multi-Conference on Parallel and Distributed Computing and Networks (PDCN), Innsbruck, Austria, February 14–16, 2006.

Index

AB interleavers advantages and disadvantages, 15 column permutations, 14 construction of, 14-16 error floor region, good performance in, 17 I/2 interleavers, 15, 16 randomization, 15 worst case patterns, avoiding, 14 Access resource areas (ARA), 248 Access resource metrics, combination of. 253 Access selection process access priorities, 257 algorithms, 239-241 dynamic, 254-257 link-performance based, 255 policy-based, 254 resource based, 255–257 Acquisition mode, 195 Adaptive Coding and Modulations (ACM), 315 Adaptive Multi-Rate Wide Band (AMR-WB), 320 A/D conversion and IF-sampling, 196 implementation of I&Q-sampling on FPGA, 197 Additive white Gaussian noise (AWGN), 137 Alamouti scheme, 56 effective channel matrix, 56 vs. SM schemes, 55-58 Allocation efficiency of, 107 and pricing scheme, 101 rule A, 106 Ambient Control Space, 240

AMR-WB-based cross-layer VoIP distributed rate control codec, 321 cross-layer VoIP rate control, 321, 322 higher voice quality and higher number of VoIPs, 326 load control, 326 outperforms RTCP-driven approach, 325, 326 RTCP-based vs. AMR-WB-based scenario, 321, 322 as system load control, 326 Angles of arrival (AoA), 183 Angular power spectrum (APS), 183 Antenna and user subsets, optimal selection of, 25 Antenna design, 190-192 anechoic chamber measurement of single PIFA element, 191 array design, 187–189 DBF algorithms, 186, 187, 193 DBF platform selection, 195 DBF techniques, 195 environmental conditions, analysis of, 182, 183 mechanical subsystem, 192 RF front end of single radiating element, 192 single radiating element on ground plane, 191 spherical section vs. double cone option, 190 Antenna selection, 40 hardware complexity, reducing, 25 with power saving modes, 25 ARTEMIS, 335

Attenuation process, stationarity investigations of, 164-167 structure function of LMS fading amplitude, 166 Attenuation threshold dependent fade duration model, 168-171 CCDF of fade duration. calculating, 168, 170 Markov modeled and measured CCDF of fade duration. 171 5-state Markov-chain, 169 state transition probabilities, determining, 170 transition matrix, 169 Attenuation time series, 163, 170, 173, 176, 177 long term, 163, 173 Authentication and encryption, 307, 308 NFC technology, 308 SIM as smart card, 307 smart card for security, 307 Automatic medical treatment/dosing (closed loop control process), 227, 228 automatic insulin injection for diabetes patients, 228 pacemaker, 228 Available bit rate (ABR), 208 Average sum rate performance, 44–48 versus K with symmetric channels and block diagonalization, 45 and successive optimization, 46 versus number of active users K for block diagonalization, 48 with scheduling and antenna selection algorithm block diagonalization, 47 successive optimization, 48

BAN applications and usage models for assisting persons with disabilities, 229, 230 fall down detection, 230 muscle tension sensors, 230 speech-impaired person, 230

with visual disability, 230 entertainment applications, 231 conceptual block diagram of BAN. 231 medical and healthcare applications, 227–229 automatic medical treatment/ dosing (closed loop control process), 227, 228 with ID-TAG for reducing the risk of wrong treatment/ medicine, 229 increased efficiency, 229 monitoring for inpatients, 228 transmission of vital and healthcare data, 227, 228 vital and health data, examples of, 229 BAN see Body area network (BAN) Base Station (BS), 115, 118 Beamforming matrices, 76, 82-85 iterative algorithm, 84 power-constrained optimization problems (PCO), 84 steps of proposed optimization approach, 83 Beamforming or transmit preprocessing, 23, 41 multiuser diversity, improves, 41 Bell Labs Layered Space-Time (BLAST), 55 BER performance, 64 Beyond 3G (B3G) systems, 261 Bit error rate (BER), 53 Block diagonalization (BD), 24, 25, 27.28 Bluetooth. 357 using IEEE 802.15.1, 233 BM-SC see Broadcast/ Multicast Service Centre (BM-SC) Body area network and standardization at IEEE 802.15.BAN applications and usage models for assisting persons with disabilities, 229, 230

entertainment applications, 231 medical and healthcare applications, 227-229 discussion and future work frequency regulations for WMTs and MICS, 236 wearable BAN and implant BAN, 235 SG-BAN and BAN definition, 224-227 short-range technologies and prototype BAN BAN requirements and, 232 Body area network (BAN), 224 requirements, 232 low emission power, 232 wearable BAN vs. implant BAN, 235 Broadcast/ Multicast Service Centre (BM-SC), 203 Business models for local mobile services basic types Content provider (CP), 282 Network Operator (NO), 282 Service provider (SP), 282 charging and billing systems, extending design of online charging interfaces, 292, 293 system design and benefits, 293, 294 definition of, 282 LOMS role model, 282-284 mobile services and different charge types, categories of, 284-286 news publishing scenario charging of service usage and revenue sharing, 286-288 convergent online charging, requirements for, 290 mobile user inside event area, 289 mobile user outside event area. 288

Carousel services, scheduling of basic algorithm, 209, 210 basic principles, 209 deliberations, 209 consideration of congestion, 211, 212 mapping of congestion and priority, 212 extension with priority scheme, 210.211 initialization of R(i), 210 purposes, 210 interaction for service/content scheduling, 209 Car scenario, 374 CCDF estimation, 152-154, 159, 160 CCDF see Complement cumulative distribution function (CCDF) Cellular system, 72, 73 reuse factor, 73 Channel-aware scheduling, objective of, 24 Channel estimation and quantization errors, impact of, 53, 54, 68 errors at receiver, 59 error model, 59, 60 simulation results, 61, 62 multi-user downlink MIMO transmission. 58 channel estimation errors at receiver, impact of, 59 system model, 59 ZF precoding for multi-user downlink transmission, 59 single link MIMO transmission, 54 system model, 54 transmission schemes, 55-58 at transmitter, 62, 63 random vector quantization, 63, 64 simulation results, 64-67 Channel estimation errors at receiver, impact of error model, 59, 60 iid complex Gaussian error model, 68

simulation results Alamouti scheme and spatial multiplexing, impact on, 61 hybrid transmission scheme, impact on, 61, 62 source of, 53, 54, 68 Channel quantization errors at transmitter, impact of hybrid transmission scheme vs. SM schemes, 63 random vector quantization, 63, 64 simulation results dominant eigenmode transmission, impact on, 64,65 multi-user downlink transmission using ZF precoding, impact on, 64-67 Channel state information (CSI), 27 and Base Station (BS), 115 imperfection of, 53 Charging and billing systems, extending Account & Balance Management Function (ABMF), 291 Charging Gateway Function (CGF), 291 customer care function, 291 event-based charging, 291 online charging interfaces, design of. 292, 293 context-aware charging web service (CWS), 292 functional view of the interfaces, 292 Online Charging System (OCS), 291 rating function (RF), 291 session based charging, 291 system design and benefits, 293, 294 charging zones, 294 Chunk. see Cluster Cloud coverage and downlink time mean annual from (1983-2006), 339 seasonal dependency of, 340

Cluster, 118 CN see Core Network (CN) Coarse pointing assembly (CPA), 347 Codec Mode Request (CMR), 321 Codec selection algorithms, 317 Code rate, 4, 5 Complement cumulative distribution function (CCDF), 152 Computer/network aided clinical record management system, 223 management efficiency, improves, 223 Concatenated zigzag codes, interleaving strategies for multidimensional design of, 5 hamming weight for, 5 multiple interleavers congruential interleavers, 8 cyclic shifted, 9 problem, 6-8 UMTS based interleaver, 9-11 for zigzag codes, 11, 12 simulation results, 16-18 zigzag codes encoding of, 3-5 Conformal array simulations, 187-189 array elements arranged on cone, 188 comparison of gain patterns, 189 comparison of LHCP gain patterns, 189 isolated element pattern, 187 method of moments, 187 Congestion control algorithms closed loop, 204 open loop, 204 Congruential interleavers design of, 6, 7 permutation rule, 8 s-random, 8 suboptimum performance, 8 Constant Bit Rate (CBR), 208

Content-centric services, 285 closed community, 285 open user community, 285 Content consumer, 305 Content Provider (CP), 203, 303, 305 Context-aware charging Web Service (CWS) functions Advice of Charge (AoC), 293 diameter credit control application, 293 online charging indication, 293 Continua Health Alliance, 224 Control entity, 304 Convergent online charging, requirements for, 290 Coordinated Access Band (CAB), 102 Core network (CN), 202 Correlation index, computation of, 140, 141 for different modulation schemes and roll-offs, 141 Correlation threshold, 42-44 average sum rate with block diagonalization versus threshold, 43 effect of threshold selection, 42, 43 objective, 43 Cost-optimised active receive array antenna antenna design, 190-192 digital beamforming implementation aspects, 195-199 simulation assessments conformal array simulations, 187-189 directional land mobile satellite channel model, 183-186 environmental conditions. 182.183 system scenario, 181, 182 "Cost-Optimised High Performance Active Receive Phase Array Antenna For Mobile Terminals (CORPA)", 181 CP see Content Provider (CP)

Cyclic shifted multiple interleavers advantage, 9 asynchronism between users, 9 in interleave division multiple access (IDMA), 9 low Hamming weight, 9 $\Pi \kappa$ from mother interleaver Π , 9 require D=3 and self interleaving operations for good performance, 16 require low memory for storing of interleaving patterns, 9 DBF algorithms, 187 DBF see Digital Beamforming (DBF) DBF weight vector update, 198, 199 radiation pattern, 199 Decentralized RAT selection algorithm enabled by IEEE P1900.4 case study: interference reduction through decentralized RAT selection, 265-268 multi-service scenario, results in, 272 - 275appropriate setting of PLth, 275 throughput improvement, 275 RAT selection enablers defined by IEEE P1900.4, 263-265 simulation model, 269, 270 RRM parameters, 269 single service scenario, results in, 270-272 downlink throughput for different values, 272 uplink load in UTRAN and **GERAN**, 272 Delay budget model and performance model, 322, 323 relationship between R-value and MOS, 323 relevant codec delays, 323 Dense Wavelength Division Multiplexing (DWDM) technology, 346 potential architecture, 347

Device domain, 304 Differentiated Services (DiffServ), 204 DiffServ see Differentiated Services (DiffServ) Digital Beamforming (DBF), 193 advantages and disadvantages, 194 block diagram of DBF processing, 194 implementation aspects acquisition phase in the DBF processing, 195 A/D conversion and IF-sampling, 196, 197 DBF weight vector update, 198.199 pilot despreading and selection of illuminated elements, 197, 198 tracking phase in DBF processing, 196 Digital System E, 182 Digital Video Broadcasting for satellite applications (DVB-S2), 315 Digital Video Broadcasting -Handheld (DVB-H), 201 Directional land mobile satellite channel model performance analysis, 183-186 Dirty paper coding (DPC), 23 Distributed cross-layer approaches for VoIP rate control AMR-WB-based cross-layer VoIP distributed rate control codec. 321 cross-layer VoIP rate control, 321, 322 delay budget model and performance model, 322, 323 numerical results, 324-329 different rain attenuation models, 324 expedited forward aggregated throughput for channel attenuation. 326

expedited forward delay for channel attenuation, 325 probability of exceeding delay requirements, 327 probability of exceeding maximum delay, 328 R-factor for G. 729, 328 QoS model, 318 RTCP-driven cross-layer distributed VoIP rate control bank of narrowband nonadaptive codecs, 320 RTCP reports, 319, 320 system model centralized vs. distributed approaches, 317, 318 DVB-S2/RCS, 317 Dominant eigenmode transmission, 64 schemes, 65 Downlink, 331 increased, 339 optical, see Optical downlink RF, see RF-downlink RF-optical, 335 scenarios, comparison, 338 DRM broker, 304 provider, 303 DSA scheme, 101 DVB-S2/RCS systems, 315 Dynamic access selection access priorities, 257 link-performance based, 255 resource based, 255–257 Dynamic MBMS resource scheduler consideration of channel state information, 206 dynamic resource scheduling, 206-208 performance evaluation, 217 power scheduling, 208 Dynamic Resource Scheduling (DRS), 206 power control technique, 207

Dynamic spectrum allocation (DSA), 99 Earth observation (EO) sensors limitation, 331 Sat downlinks, 347 satellites. 331 Earth observation scenario, 333 Electronic Communications Committee (ECC) ultra low power active medical implants, 355 Electronic patient record (EPR), 351.353 End user, 303 Enterprise resource planning system (ERP) and wireless technology, 362 Equivalent baseband model, 137, 138 second- and fourth order moments, 138 SNR estimation of oversampled narrowband signals, 138 Erbium Doped Fiber Amplifier (EDFA), 347 Errors at receiver, 59 error model, 59, 60 simulation results, 61 Alamouti Scheme and spatial multiplexing, impact on, 61 hybrid transmission scheme, impact on, 61, 62 European Cloud Climatology (ECC), 339 Event modeling with two-state Markov chain, 167, 168 fade/non-fade events, 167 state transition matrix S. 167 Exclusion-compensation principle, 107 FACH see Forward Access Channel

(FACH) (FACH) Fade duration, 168 Fade duration Markov model, investigation of applicability of attenuation threshold dependent fade duration model, 168 - 171event modeling with two-state Markov chain, 167, 168 from fade duration model to twostate fade/non-fade model. 171 - 173measured data, description of, 164 modeling scintillation, 176 simulate single fading event, 173-175 stationarity investigations of attenuation process, 164-167 synthesized time series, evaluation of, 176-178 Fade duration model to two-state fade/non-fade model. 171 - 173with desired duration, 173 fade duration CCDF, 172 simulated fading/non-fading event series, 173 Fair scheduling and antenna selection algorithms, 29, 30 Fine pointing assembly (FPA), 347 Float Point Multiplications (FPMs), 124 Forward Access Channel (FACH), 203, 205 Free-space optical (FSO) high speed links, 331 designated wavelength regions, 344 problem, 338 wavelengths, 344 10.6 µm, 345 800 nm, 345 1064 nm, 345 Frequency/time assignment algorithms, 117

Frequency/time-space resource allocation (FT-S RA) strategy, 116, 124 frequency/time assignment algorithm: user-to-resource assignment, 125 resource allocation strategy, 126 successive projections algorithm, 124, 125 Gateway GPRS Support Node (GGSN), 203 General Packet Radio Service (GPRS), 202 Generic abstraction of access performance and resources access performance abstraction, 246, 247 access resource abstraction metrics, 248-251 structures and combined access resource metrics, 251-253 access selection process dynamic, 254-257 policy-based, 254 multi-radio access architecture, 240 - 242service requirements delay requirements of applications, 244 rate requirements of applications, 243, 244 reliability requirements of applications, 243 service specification, 244, 245 Generic access performance abstraction, 246, 247 link performance abstraction, 246 Generic access resource abstraction metrics. 248-251 resource levels, computing, 249 resource measures, computing, 250.251 structures and combined access resource metrics. 251-253

channelization resource, 252 physical resource, 252 Generic Link Layer (GLL), 240 Geometry-based stochastical model (GSCM), 183 GEO satellite as relay station, 335 vs. HAP relay, 336 GLL see Generic Link Layer (GLL) Global positioning system (GPS), 359 3GPP see 3rd Generation Partnership Project (3GPP) GPRS see General Packet Radio Service (GPRS) Greedy regularized correlation-based algorithm (GRCBA), 116 vs. SPA, 125 Ground station diversity, 334

Hamming weight for concatenated zigzag codes, 5 Healthcare process management, 351 Hidden Markov model (HMM), 176 High Altitude Platforms (HAPs), 332 advantage of receiver concepts with wave-front correction systems, 346 attenuation effects, 344 designated wavelength for FSO, 344 relay, 336, 337 advantages, 336 downlink, 336 vs. GEO relay, 336 wavelength selection and terminal architecture, 344-348 1064 nm, 345, 346 800 nm technology, disadvantages, 345 Higher-order statistics correlation index, 143 even-order moments, 142 sixth-order statistics, 142 High resolution optical/infrared cameras, 331 HMM see Hidden Markov model (HMM)

Home gateway, 305 Home network content adaptation, 301 secure access, 300, 301 Homodyne binary phase-shift keying (BPSK) modulation, 345 Hospital, wireless see Wireless applications in healthcare and welfare, wireless hospital concept Hybrid transmission scheme, 56-58 dominant eigenmode transmission. 58 effective MIMO channel matrix, 57 by shuffling antenna, improves, 57, 58, 61, 62

IEEE 802 802.15.1 Bluetooth, 224 IEEE 802.15.BAN (SGBAN), 224 802.11 WLAN, 224 IEEE P1900.4, 263 IEEE 802.11 radios, 357 I/2 interleavers, 15, 16 IMS trust models, 310 Inphase and quadrature (I&Q), 196 Input-output relationship for Alamouti scheme, 56 hybrid transmission scheme, 57 narrow band system, 54 Institute of Communications and Navigation of the German Aerospace Center (DLR), 332 Integrated Services (IntServ), 204 Intelligent mobile telecommunication (IMT), 223 Inter-cell interference, 71, 73, 80, 101 Interference reduction through decentralized RAT selection (case study), 265 - 268decentralized RAT selection strategy, 267 information transmitted through P1900.4, 268

intercell interference in TDMA and CDMA systems, 266 Interleave division multiple access (IDMA), 9 Interleaver design approaches, 3 criteria, 7-12 memory for storing interleaver patterns, 3, 8, 19 memory requirements, minimizing, 3, 6, 8 of multidimensional parallel concatenated convolutional codes, 7 mutually independent, 13 permutation rule, 4, 8-11, 15 permutations from common mother interleaver, 7, 9, 16 turbo codes in UMTS, 8 Interleavers for zigzag codes, 11, 12 AB Interleavers, 14–16 code words with Hamming weight wH=2, 12wH=3, 12 codewords with low weight, avoiding, 11 interleaver design criteria, 12 optimizing for performance in error floor region, 8, 9 restricted random interleaver. 12 - 14restriction, 12 Intermediate UMTS interleavers, modified, 18 International commission on nonionizing radiation protection (ICNIRP), 232 International Satellite Cloud Climatology Project (ISCCP), 338 cloud coverage data, 339 Internet key exchange protocol version 2 (IKEv2), 365 Internet Protocol (IP), 204 Internet services, 281, 284 Inter-symbol interference (ISI), 73

IP Multimedia Subsystem (IMS), 204.285 IP see Internet Protocol (IP) I&Q see Inphase and quadrature (I&Q) "Islands geography," 101 Japan Aerospace Exploration Agency (JAXA), 332 Joint Radio Resource Management (JRRM), 261 Joint user and antenna selection for block diagonalization (BD), 30-34 first proportionally fair (PF) scheduling metric, 31, 32 *i*th iteration of algorithm, 31 low complexity algorithms, 31 scheduling metrics, 31 selection metric design, 33 simplified proportionally fair (PF), 32 for successive optimization (SO), 34-38 capacity based algorithm, 36 objective, 34, 35 scheduling metrics, 35, 36 simplified Frobenius norm based algorithm, 36 user scheduling, 35 JRRM see Joint Radio Resource Management (JRRM) K-1 interleavers, 6 KIODO (Project), 332 Land mobile satellite (LMS), 183

Leo (Low Earth Orbit) satellite, 331, 332 atmospheric transmission at selected elevation angles (Neustrelitz), 344 elevation dependency at selected wavelengths, 345 downlink availability, 343

increased downlink time, 335 see also GEO satellite optical downlinks, 334 transmission, 331 Liberalization, 99 Line-of-sight (LOS), 183 LMS see Land mobile satellite (LMS) Local Mobile Services (LOMS), 281 FindAFriend, 286 local search, 286 news service, 286 role model charges types, 283, 284 LOMS see Local Mobile Services (LOMS) Low complexity algorithms, 31 Low level radio, 232 Low noise amplifier (LNA), 191 Market-driven dynamic spectrum allocation (DSA) framework, 111 Markov model, 163-165, 167, 168, 170-174, 178 of fade duration, 163, 168, 170-172 Fade duration CCDF, calculating with, 171, 172

method for time series generation, 164, 165 partitioned Fritchman's, 173, 174 two state fade/non-fade, 174, 178 MBMS content distribution carousel service scheduling approach, 202 dynamic MBMS resource scheduler in RAN, 202 multicast streaming services, 202 MBMS see Multimedia Broadcast Multicast Service (MBMS) Measured data, description of, 164 MIMO systems see Multiuser MIMO system Minimum mean square error receiver (MMSE), 55

Minimum Variance Beamforming (MVB), 193 Mobile broadcast and multicast services, scheduling techniques for concepts and algorithms carousel services, 209-212 dynamic MBMS resource scheduler, 205, 206 streaming services, 212-215 and congestion control in context of MBMS, 204, 205 fundamentals, 203, 204 performance evaluation dynamic MBMS resource scheduler. 217 scheduling for carousel services, 215, 216 scheduling for streaming services, 216 problem analysis introduction of MBMS, 202, 203 Mobile IPv6 protocol (MIPv6), 365 Mobile phone, 306 Mobile services and different charge types, categories of content centric service, 284 internet access, 284 local mobile service (LOMS), 284 messaging, 284 Push to talk over Cellular (PoC), 285 voice calls, 284 Model parameterization, 154-156 Gaussian fade slope model, experimental parameters of, 154-156 state transition probabilities, determination of, 156 Modern modulation and coding (ModCod) schemes, 334.335 Moment-based estimation of signal-to-noise ratio computation of correlation index, 140.141

equivalent baseband model, 137, 138 higher-order statistics, 142, 143 correlation index, 143 narrowband signals, 138, 139 simulation results, 144-146 SNR estimation, 138–140 MRRM see Multi-Radio Resource Management (MRRM) Multi-bid profile, 106 allocation rule, 107 allocation's efficiency, 107 exclusion-compensation principle, 107 see also Opponents' profile, 106 Multicast mode vs. broadcast mode, 202 Multi-cell signal processing exchange of data, backhaul required for, 71, 74, 77,78 notation, 72 optimization approach beamforming matrices, 82-85 power allocation, 85, 86 user grouping, 80-82 virtual MIMO configuration, 82 simulation results, 89-93 subsystem partitioning, 86-89 system model cellular system setup, 73 optimization framework, 74-78 optimization problem, 78, 79 Multimedia Broadcast Multicast Service (MBMS), 201 BM-SC as functional entity, 203 broadcast mode, 202 carousel services, 202 dynamic system, 205 efficient distribution of multicast and broadcast services, 201.202 file download services, 202 multicast mode, 202 number of services/content and users, increases, 201

offers interactive multimedia services. 201 reference architecture, 203 requires single transmission, 202 scheduling and congestion control in, 203–205 scheduling techniques in, 201, 204, 205 streaming services, 202 Multiple interleavers condition, 5 congruential interleavers, 8 cyclic shifted, 9 design criteria of minimizing intersection, 7 minimizing multiuser interference impact, 6, 7 K-1, 6 in multidimensional parallel concatenated convolutional codes, 7 multiuser detection in CDMA and IDMA, design of, 6 problem, 6-8 design criteria, 6, 7 memory requirements, 6 of saving memory, 6 UMTS based interleaver, 9-11 for zigzag codes, 11, 12 Multi-radio access architecture, 240 - 242abstraction model, 242 in ambient networks, 241 Multi-radio resource management (MRRM) access connectivity abstraction model and architecture, 242 access selection, 241 Multi-user downlink MIMO transmission channel estimation errors at receiver, impact of, 59 system model, 59 ZF precoding for multi-user downlink transmission, 59

Multi-user joint detection or transmission, 71, 73 see also Virtual MIMO or network MIMO Multiuser MIMO system block diagram of, 26 Muscle tension sensors, 230 MVB see Minimum Variance Beamforming (MVB) Near field communication (NFC), 306, 308, 357 Network model, 369 Neustrelitz, 333 News publishing scenario charging of service usage and revenue sharing, 286-288 charge types, 286 example for context-aware charging and revenue sharing, 287 convergent online charging, requirements for, 290 mobile user inside event area, 289 receiving a discount for traffic and service, 289 mobile user outside event area, 288 user outside film festival area fully charged, 288 Next generation network (NGN), 365 NFC technology, 310 NF see Noise figure (NF) Noise figure (NF), 191 Non-linear vertical BLAST (V-BLAST), 55 Non-polynomial time hard (NP-H), 115 defined, 116 and optimum RA, 115 Non-uniform scatterer cross section (NSCS), 183 N-state Markov chain model, 149, 151-154 basic parameters, 152

schematic representation, 152 state separation, 153 Null space projected channels, 27, 41 OGS-networks availability of, 340 cloud cover statistics and OGSdiversity, 338-340 within Europe, 342 four OGSs within Germany, 341 within Germany, 341 locations for European and world wide. 343 selection criteria, 340 world wide, 342, 343 One-shot multi-bid auction and pricing in DSA networks average and maximum interference, 110 example, 108-111 geographic coupling parameters, 110 interference, 100 multibids, 110 optimal allocation and costs of providers, 111 pricing scheme in DSA model allocation and pricing rules, 106-108 inputs, 105, 106 radio technology coupling parameters, 109 related works, 101, 102 one-shot scheme, Delenda, 101 Progressive Second Price (PSP) Mechanism, 101 rights and obligations, 100 spectrum usage rights, 100 usage rights, parameters, 100 scenario, 109 spatio-temporal DSA model, based on feasible allocation, 104 interference and spectrum efficiency, 102, 103 interference tolerance, 105

trading and liberalization, 99, 100 efficient use of spectrum, 99 market for frequencies, 99 One-shot multibid auction method, 111 One-shot scheme (Delenda), 101 **Open Systems Interconnection** (OSI), 204 Opponents' profile, 108 Optical C-Band technology, 346 Optical downlink, 332, 334 LEO, elevation dependency of atmospheric transmission, 345 limitations. 334 Optical Ground receiving Station (OGS), 331 adaptive optics technologies, 346 factors deciding location of, 339 Optical satellite downlinks to optical ground stations and HAPs, 331, 332, 348 availability of OGS-networks, 340 within Europe, 342 within Germany, 341 world wide, 342, 343 cloud cover statistics and OGS-diversity, 338–340 comparison of downlink scenarios, 338 solving challenge of cloudblockage, 332, 333 system comparison combined RF-optical downlink, 335 Earth observation scenario, 333 GEO relay, 335 HAP relay, 336, 337 optical downlink, 334 RF downlink, 334 state of art RF downlink, 333 wavelength selection and terminal architecture, 344-348 Optimal feasible allocation, 105, 110 Optimal threshold, 43 Optimistic capacity bounds, 71

Optimization (multi-cell signal processing) approach, 79 framework, 74-78 problem, 78, 79 Orthogonal Frequency Division Multiple Access (OFDMA), 115 and BS, 118 Orthogonal space division multiplexing (OSDM) techniques, 24 Orthogonal space-time block codes (OSTBCs), 53 OSDM techniques see Orthogonal space division multiplexing (OSDM) techniques Pacemaker, 228 Partitioned Fritchman model, 149 Patient tracking in wireless hospital, hospital registration and, 358, 359 Pedestrian scenario, 374 Performance analysis of IPsec calculation method, 370-373 mobility process chain, 372 number of visits of jobs of classes, 373 scale factors, 373 input parameters, 373-378 calculation of service times, 375 meaning of input variables, 375 mobility scenarios, 373, 374 mobility state diagram of user, 374 parameters of network model, 375 processing requirement to encrypt, decrypt, or hash one block of data, 376 results, 378-384 dependency of utilization of HA on mobility rate, 381 full utilization of HA at different security configurations, 379

impact of processing speed of MN and access network, 381-383 mean response time of mobility process, 382 performance overheads, 379 relative effect of processing speed of MN and data rate. 383 utilization of HA, 378-381 Performance overheads of IPsec in MIPv6 scenarios, analytical analysis of, 365, 366 performance analysis, 369 calculation method, 370-373 input parameters, 373-378 results, 378-384 reference scenario and network model, 368, 369 security configurations, 366, 367 Personal digital assistant (PDA), 354 Pilot despreading and selection of illuminated elements despreading of single antenna element signal, 198 Planar inverted-F antenna (PIFA), 190 Portable base station (PBS) wireless data transfer in hospitals, 355 Power allocation, 29, 76, 85, 86 Power delay profiles (PDP), 183 Power scheduling, 208 Pricing scheme C, 106 Pricing scheme in DSA model allocation and pricing rules, 106 - 108inputs, 105, 106 Progressive second price (PSP) mechanism drawbacks, 101 Proportional fairness based scheduling, 25, 30 effect of antenna selection, evaluating, 30 Frobenius norm based, 31, 36, 41,49

Prototype BAN system UWB transceiver for wearable BAN, 233 QoS see Quality of Service (QoS) Quality of Service (QoS), 203, 261 model. 318 Quasi-linear utility functions, 101 Radio access networks (RANs), 202, 261 see also Beyond 3G (B3G) systems, 261 Radio access technologies (RATs), 239, 261 Radio frequency identification (RFID), 357, 359 Radio network service providers (NSPs), 111 Radio resource, 118 Radio resource management (RRM), 262 Radio spectrum, 99 Rain attenuation distribution on terrestrial microwave links. estimation of applying proposed model for designated link, 157, 158 principle of parameter transformation, 157 model parameterization, 154-156 N-state Markov model, 151-154 results, 158-160 fade duration statistics prediction, 160 parameters of proposed link, 158 rain attenuation CCDF prediction, 159 stationary examination of rain attenuation process, 150, 151 Rain attenuation process, stationary examination of, 150, 151 generated time series, 150, 153, 158

probability density of measured data, 151 vs. fade duration CCDF values, 153, 159, 160 Rain attenuation/ rain rate modeling, 149 first and second order statistics of microwave link, estimating, 149-151, 155 Random interleaver, restricted, 12 - 14construction of k-th interleaver $\Pi \kappa$. 13 multiple, with restrictions, 13 higher input index pairs, 14 input indices sequence, 13 output indices sequence, 13 Random vector quantization, 63, 64 RANs see Radio access networks (RANs) RAS-II see Receive Antenna Selection Algorithm 2 (RAS-II) RAS-I see Receive Antenna Selection Algorithm 1 (RAS-I) RAT selection algorithms, 263, 266, 269-271 for CDMA/TDMA scenarios, 265 - 268RAT selection enablers defined by IEEE P1900.4, 263-265 IEEE P1900.4 heterogeneous system vision, 264 Network Reconfiguration Manager (NRM) functionalities, 264 RAT selection policy, 268 RATs see Radio access technologies (RATs) 3rd Generation Partnership Project (3GPP), 202, 290 Receive Antenna Selection Algorithm 1 (RAS-I), 41 Receive Antenna Selection Algorithm 2 (RAS-II), 41, 42

Receive antenna selection (RAS), 36,40 advantages, 40 impact of, 40-42 improves multiuser diversity, 41 RAS-I, 41 RAS-II, 41, 42 Reference scenario, 368 Resource allocation strategies for SDMA/OFDMA systems, 115-117, 132 analysis and simulation results, 126-131 frequency/time-space resource allocation, 124 frequency/time assignment algorithm: user-to-resource assignment, 125 resource allocation strategy, 126 SDMA algorithm: successive projections algorithm, 124, 125 performing optimum, complexity, 115, 116 simplify, 115, 117 space-frequency/time resource allocation, 119 frequency/time assignment algorithm: group-to-resource assignment, 121–123 resource allocation strategy, 123, 124 SDMA algorithm: greedy regularized correlationbased algorithm, 120, 121 suboptimal, 116 system model, 117-119 two suboptimal, 116 Resource blocks, 73 Resource matrix, 214 RF-downlink, 331 GEO-downlink, limiting factor. 335 limitations, 331 proposed, 334 RF-optical downlink, 335

Rights management, 301, 302 use of group authentication, 302 Rights management for user content authentication and encryption, 307, 308 background, 298-300 home network with access control and out-of-band key distribution, 299 interconnecting service, 299 devices in home network, 300, 301 future work, 310, 311 high capacity SIMs, 310 home right management system, 311 IMS trust models, 310 NFC readers, 310 NFC technology, 310 SmartMX chip, 310 service architecture, 308–310 constraints, 308 usage scenarios commercial content, 303-305 user content, 305-307 RRM see Radio Resource Management (RRM) RTCP-driven cross-layer distributed VoIP rate control and AMR-WB architectures, 319 bank of narrowband non-adaptive codecs, 320 codecs for RTCP-driven scenario, 320 reports, 319, 320

Sample Matrix Inversion (SMI), 194 SAR systems, 331 Satellite Digital Multimedia Broadcasting (SDMB), 181 Scheduling algorithms, 29, 30 Scheduling and congestion control in MBMS in context of MBMS, 204, 205 fundamentals, 203, 204 Scheduling metrics, 25, 35, 36 average value, 38 Metric 1, 32

Metric 2, 34 Metric 3, 36 Metric 4, 36 performance, 46 Scintillation defined, 176 filter parameters, 176 with HMM method, 176 modeling, 176 SDMA, FT-S RA strategy, 126 greedy regularized correlationbased algorithm, 120, 121 group-to-resource assignment, 121-123 see also Frequency/time-space resource allocation (FT-S RA) strategy SDMA, S-FT RA strategy, 123, 124 successive projections algorithm, 124, 125 user-to-resource assignment, 125 see also Space-frequency/time resource allocation (S-FT RA) strategy SDMA/OFDMA systems, 115 system model, 117–119 see also Resource allocation strategies for SDMA/OFDMA systems Security configurations, 366, 367 MIPv6 signaling, 367 Selection metrics design, 33 squared Frobenius norm, 34 Selective virtual MIMO, 71 Service requirements delay requirements of applications insensitive, 244 sensitive, 244 rate requirements of applications discrete-rate, 244 elastic, 244 reliability requirements of applications error-sensitive, 243 error-tolerant, 243

Service specification, 244, 245 S-FT RA and FT-S RA strategies, performance, 126-131 comparison, for different allocation objectives, 130 simulation parameters, 127 SG-BAN and BAN, 224-227 BAN vs. other IEEE 802.15 standards, 226 current status of IEEE 802.15 working group, 225 definition and scope of BAN, 225 Short-range wireless technologies bluetooth using IEEE 802.15.1, 233 low rate UWB of IEEE 802.15.4a, 233 prototype BAN system, 233, 234 UWB transceiver for wearable BAN, 233 specific low level radio, 232 zigbee using IEEE 802.15.4, 232, 233 Signal Interference Ratios (SIRs), 207 Signal-to-interference-plus-noise ratio (SINR), 193 Signal-to-leakage ratios (SLR), 89 Signal-to-noise ratio (SNR), 137, 138 Simplified channel-aware greedy scheduling and antenna selection fair scheduling and antenna selection algorithms, 29, 30 joint user and antenna selection for block diagonalization (BD), 30-34 joint user and antenna selection for successive optimization (SO), 34-38 subspace correlation based user grouping, 38, 39 receive antenna selection (RAS), impact of RAS-I, 41 RAS-II, 41, 42

simulation results average sum rate performance, 44-48 correlation threshold, 42-44 system model block diagonalization (BD), 27,28 successive optimization (SO), 28.29 Simulation assessments directional land mobile satellite channel model, 183-186 normalised squared absolute value, 185 PDF of power of sum receive signal, 186 PDP for near scatterers with NSCS, 186 time series azimuth angles of LOS and multipath signals, 185 time series of elevation angles of LOS and multipath signals, 184 environmental conditions, 182, 183 Simulation results channel scheduling and antenna selection average sum rate performance, 44 - 48correlation threshold, 42-44 average sum rate with block diagonalization versus threshold, 43 mobile broadcast and multicast services average difference in Eb/N0, 219 discard rate in dependence of standard deviation, 218 number of supplied users in dependence of standard deviation, 218 multi-cell cooperative signal processing centralized / decentralized performance (UL), 91

comparison of different precoding schemes, 92 different backhaul scenarios in DL, impact of, 90 suitable number of quantization bits. 90 multidimensional concatenated zigzag codes, 16-18 signal-to-noise ratio, 144-146 evolution of E[], 144evolution of success rate, 146 medium SNR range, 145 normalized mean square error, 145, 146 Single fading event, simulate, 173-175 elementary fading event generation process, 174 generating single elementary fading event, 175 maximal attenuation, 174 synthesized fading event, 175 Single link MIMO transmission system model, 54 Single Sign On (SSO), 298 SIRs see Signal Interference Ratios (SIRs) Smart card, 307 SmartMX chip, 310 SNR see Signal-to-noise ratio (SNR) Software defined radio (SDR), 354 Space Division Multiple Access (SDMA), 115 algorithms, 116, 117 channel response, 118 S-FT RA and FT-S RA strategies, performance of, 126-131 simplifying tasks for, 118 spatial correlation between two vector channels, 119 Space-frequency/time resource allocation (S-FT RA) strategy, 116, 119 considering nonunique and unique SDMA groups, 128

frequency/time assignment algorithm: group-toresource assignment, 121 - 123and number of FPMs, 124 RCBA and GRCBA in unique SDMA groups, comparison between, 129 resource allocation strategy and SDMA, 123-124 SDMA algorithm: greedy regularized correlationbased algorithm, 120, 121 Spatial multiplexing, 55 minimum mean square error receiver (MMSE), 55 V-BLAST receiver, 55 zero-forcing receiver, 55 Spatio-temporal DSA model feasible allocation, 104 interference and spectrum efficiency, 102, 103 interference tolerance, 105 Spatio-temporal transmission schemes, 53 requirement of CSI, 53 Spectrum allocation, 107 distribution, 99 efficiency, 99, 102, 103 interference, 100 quality, 104 rights and obligations, 100 Spectrum broker entity (SB), 105 Sport training, wireless, 362 S-random congruential interleaver, 8 with s=2J+1. 16 StORe (Stratospheric Optical Relays), 332, 347 geometrical visibility constraints of, 337 improved access time (example), 337 Streaming services, scheduling of concept overview, 213 basic phases, 213

performance evaluation scheduling for streaming services, 216 spatial user distribution and RAN resources, 214 resource matrix, 214 splitting of reception groups, 215 weighting, 213, 214 Strong codes, 3 Subscriber Identity Module (SIM), 300, 306 Subspace correlation based user grouping, 38, 39 Subsystem partitioning, 86-89 aim of, 88 gains of different cooperation sizes, 88 increased performance, 89 interference cancellation, 87 **SINR**, 86 Successive optimization (SO), 24, 25, 28.29transmit power minimization, 28 Successive projections algorithm (SPA), 116, 123-125 vs. GRCBA, 125 Synthesized time series, evaluation of. 176-178 CCDF of measured and generated, 177 fade duration of measured and generated, 178 realization of, 177 System model block diagonalization (BD), 27, 28 successive optimization (SO), 28, 29 System scenario antenna specifications, 182 **Temporal Reference Beamforming** (TRB), 193 Time division duplex (TDD), 27 Tracking mode, 196

Trading, 99

Transmission schemes, 55-58 Alamouti scheme, 56 Alamouti vs. SM schemes, 55–58 hybrid transmission scheme, 56-58 dominant eigenmode transmission, 58 hybrid transmission scheme vs. SM schemes, 55-58 spatial multiplexing, 55 minimum mean square error receiver (MMSE), 55 V-BLAST receiver, 55 zero-forcing receiver, 55 Transmit power minimization block diagonalization, 24 successive optimization, 24 TRB see Temporal Reference Beamforming (TRB) UE see User Equipment (UE) Ultra-wideband (UWB), 223, 357 UMTS based interleaver, 9-11 construction of, 10, 11 high performance, 17 improved interleavers, obtaining, 11 methods of, 9 performance results, 17 randomization method, 10, 11 transpose or modified read out order, 11 turbo code. 9 using intermediate steps by reading out permutation rules, 9, 10 see also Intermediate UMTS interleavers, modified Unspecified Bit Rate (UBR), 208 Usage rights of spectrum, 100 Usage scenarios commercial content, 303-305 entities, 303, 304 limitations, 303, 304 user content, 305–307 content sharing between users, 306 User equipment (UE), 202

User grouping, 80–82 algorithm, 39 origin of main interference in downlink, 81 partitioning cellular network into subsystems, 80 proposed resource block assignment, 81 search complexity of users, reducing, 25, 39, 42, 49 selective virtual MIMO schemes, effective, 80 strategies, 33, 38 UWB of IEEE 802.15.4a, 233 UWB see Ultra-wideband (UWB) Variable Bit Rate (VBR), 208 V-BLAST receiver, 55 Virtual MIMO or network MIMO, 75,82 Wibree, 357 WILHO Consortium, 352, 353 and ERP, 362 patient monitoring gadgets, 356 and wireless pain meter, 361 WINNER C1 channel model, 61 multi-user downlink transmission. 66 LOS scenario, 66 WIPL-D simulation, 187, 188 Wireless Application Protocol (WAP), 298 Wireless applications in healthcare and welfare, 351, 352 application areas enterprise resource planning system, 362 sport training, 362 wireless hospital, 358-361 wireless sensors, 361, 362 wireless hospital concept, 352, 353 equipment, 356, 357 network topologies, 354-358 technology, 357, 358

Wireless body area network (WBAN), 355 Wireless broadband systems, advantage, 53 Wireless communication network technologies role in supporting medical and healthcare services, 223 Wireless data transfer (hospital), 357 see also Bluetooth; Ultra-wideband (UWB); Wibree; Zigbee Wireless hospitals, 352, 353 advantages, 352, 353 after-care, 360, 361 data transfer modes, 355 equipment, 356, 357 example, 356 hospital registration and patient tracking, 358, 359 location and tracking (LT) applications, 359 using WLAN vs. GPS, 359 network topologies, 354-358 wireless body area network, 354-356 wireless hospital area network, 354 pain meter, 361 patient monitoring, 356 positioning, 359, 360 process flow optimization, 359 improved cost structure of, 359 real-time material tracking and monitoring, 360 real-time phase information, 359 technology, 357, 358 data communication and localization purposes, see IEEE 802.11 radios see also Wireless data transfer (hospital) use of NFC and RFID, 357

wireless playground, example of, 353 Wireless local area networks (WLAN), 223, 359 Wireless metropolitan area networks (WMAN), 223 Wireless personal area networks (WPAN), 223 Bluetooth, 223, 233 increase quality and efficiency, 223 Ultra-wideband (UWB), 223, 233 ZigBee, 223, 232, 233 Wireless sensors, 361 in animals, 361, 362 World Data Center for Remote Sensing of Atmosphere (WDC-RSAT), 338, 339 WPAN see Wireless personal area networks (WPAN) Zero-forcing receiver, 55 Zigbee, 357 using IEEE 802.15.4, 232, 233 Zigzag codes, 3 code performance, 4 and concatenated zigzag codes, 4 concatenation, 3 with different interleavers, 17, 18 encoding of, 4 high code rate, 3 interleavers for, 11, 12 iterative decoding, 3, 5 low complexity for high performance, 3 minimum Hamming weight, 4 multiple interleavers, use of, 5 mutually independent for good performance, 5 performance results, 16, 17 principle, 4 single (weak performance), 3