Universidade de Aveiro
2009

Departamento de Electrónica e Telecomunicações

Alberto de Jesus Nascimento

Multilayer Optimization in Radio Resource Allocation for the Packet Transmission in Wireless Networks

Optimização Multicamada de Atribuição de Recursos Rádio para Transmissão de Pacotes em Redes em Fios

Universidade de Aveiro
2010

Departamento de Electrónica, Telecomunicações e Informática

Alberto de Jesus Nascimento

Optimização Multicamada de Atribuição de Recursos Rádio para Transmissão de Pacotes em Redes em Fios

Multilayer Optimization in Radio Resource Allocation for the Packet Transmission in Wireless Networks

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Electrónica e Telecomunicações, realizada sob a orientação científica do Dr. Atílio Manuel da Silva Gameiro, Professor Associado do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Apoio financeiro do CITMA
Centro de Ciência e Tecnologia da Madeira
Programa Operacional Plurifundos da Região Autónoma da Madeira do Fundo Social Europeu.

Apoio financeiro da FCT e do FSE no âmbito do III Quadro Comunitário de Apoio.

À minha esposa Olívia

Pelo seu apoio constante nos bons e maus momentos, pela sua infinita paciência e pelo espírito de sacrifício demonstrado perante a nossa separação, decorrente da minha estadia em Aveiro durante cerca de dois anos e meio, para que este trabalho pudesse ter sido realizado.

Aos meus filhos:

Tiago José Câmara do Nascimento
Ana Catarina Câmara do Nascimento

Para que possam um dia entender que, apesar do que parece, a vida, e o alcançar de objectivos, implicam sempre trabalho, dedicação, lealdade e espírito de sacrifício.

o júri

presidente

Prof. Dr. Carlos de Pascoal Neto
Prof. Catedrático da Universidade de Aveiro

Prof. Dr. Paulo da Fonseca Pinto
Professor Associado com Agregação da Universidade de Lisboa

Prof. Dr. Fernando José da Silva Velez
Professor Auxiliar da Faculdade de Engenharia da Beira Interior

Prof. Dr. Manuel Alberto Pereira Ricardo
Professor Associado da Faculdade de Engenharia da Universidade do Porto

Prof. Dr. Amaro Fernandes de Sousa
Professor Auxiliar da Universidade de Aveiro

Prof. Dr. Atílio Manuel da Silva Gameiro
Professor Associado da Universidade de Aveiro

agradecimentos

Ao finalizar este trabalho gostaria de endereçar algumas palavras de agradecimento a todas às pessoas que me ajudaram a que ele pudesse ter sido realizado.

Ao meu orientador, Prof. Dr. Atílio Gameiro, por ter aceite a tarefa de me orientar neste trabalho, pela sugestão do tema e pelo empenho e colaboração manifestados, sem os quais este trabalho não teria sido possível.

Ao Shahid do pólo de Aveiro pela realização das simulações que levaram à obtenção das *Look Up Tables* utilizadas nas simulações. E pela camaradagem demonstrada.

Um agradecimento especial aos colegas do IT Aveiro: Jonathan, Valdemar e Joaquim, pelos comentários e sugestões surgidos ao longo das diversas conversas decorridas no bar do IT e mesmo nos seus corredores. Muitas vezes ajudaram a esclarecer e contribuíram para o enriquecimento do trabalho desenvolvido.

Aos demais colegas do Instituto de Telecomunicações, pólo de Aveiro, pelo apoio prestado durante a minha estadia em Aveiro.

Ao Instituto de Telecomunicações, pólo de Aveiro, pela cedência de instalações e equipamento para a realização do trabalho. Ao Pedro Silva pelo apoio prestado ao nível da logística.

Aos meus colegas do Departamento de Matemática e Engenharias (DME) da Universidade da Madeira, Tiago, Eduardo, Karolina, Dionísio, Lina, Maurício e Luís Lopes pela amizade e salutar convívio existente no departamento, os quais me ajudaram, e muito, nas boas e más horas, especialmente nos longos momentos de ausência.

Ao Amândio pelos conselhos antes, durante e depois da minha despensa lectiva e ao Sr. Nélson pelo apoio e camaradagem prestados.

A todos os membros do DME em geral.

Ao CITMA e a FCT pela atribuição das respectivas bolsas.

Finalmente, à minha esposa pela sua dedicação, carinho e, acima de tudo, infinita paciência para suportar todas as dificuldades, variações de humor e a minha ausência prolongada, ao longo dos últimos 4 anos, dispendidos na realização deste trabalho.

O autor gostaria ainda de agradecer aos membros do júri a elegância de ter facultado, por intermédio do orientador, a identificação de algumas gralhas bem como sugestões para a melhoria de alguns aspectos de redacção, o que permitiu a incorporação e consequente melhoria na versão final da tese.

palavras-chave

Redes móveis sem fios de 3ª e 4ª gerações; projecto de redes móveis sem fios em modo *cross-layer*; qualidade de serviço em redes móveis sem fios; algoritmos de escalonamento de pacotes em redes móveis sem fios; algoritmos de gestão de recursos rádio; atribuição dinâmica de recursos; muliplexagem de recursos por divisão espacial; algoritmos de escalonamento de pacotes através da função utilidade; simulações a nível de sistema do protocolo de comunicação das redes móveis sem fios.

resumo

Na última década tem-se assistido a um crescimento exponencial das redes de comunicações sem fios, nomeadamente no que se refere a taxa de penetração do serviço prestado e na implementação de novas infra-estruturas em todo o globo. É ponto assente neste momento que esta tendência irá não só continuar como se fortalecer devido à convergência que é esperada entre as redes móveis sem fio e a disponibilização de serviços de banda larga para a rede Internet fixa, numa evolução para um paradigma de uma arquitectura integrada e baseada em serviços e aplicações IP. Por este motivo, as comunicações móveis sem fios irão ter um papel fundamental no desenvolvimento da sociedade de informação a médio e longo prazos.

A estratégia seguida no projecto e implementação das redes móveis celulares da actual geração (2G e 3G) foi a da estratificação da sua arquitectura protocolar numa estrutura modular em camadas estanques, onde cada camada do modelo é responsável pela implementação de um conjunto de funcionalidades. Neste modelo a comunicação dá-se apenas entre camadas adjacentes através de primitivas de comunicação pré-estabelecidas. Este modelo de arquitectura resulta numa mais fácil implementação e introdução de novas funcionalidades na rede. Entretanto, o facto das camadas inferiores do modelo protocolar não utilizarem informação disponibilizada pelas camadas superiores, e vice-versa acarreta uma degradação no desempenho do sistema. Este paradigma é particularmente importante quando sistemas de antenas múltiplas são implementados (sistemas MIMO). Sistemas de antenas múltiplas introduzem um grau adicional de liberdade no que respeita a atribuição de recursos rádio: o domínio espacial. Contrariamente a atribuição de recursos no domínio do tempo e da frequência, no domínio espacial os recursos rádio mapeados no domínio espacial não podem ser assumidos como sendo completamente ortogonais, devido a interferência resultante do facto de vários terminais transmitirem no mesmo canal e/ou slots temporais mas em feixes espaciais diferentes. Sendo assim, a disponibilidade de informação relativa ao estado dos recursos rádio às camadas superiores do modelo protocolar é de fundamental importância na satisfação dos critérios de qualidade de serviço exigidos.

Uma forma eficiente de gestão dos recursos rádio exige a implementação de algoritmos de agendamento de pacotes de baixo grau de complexidade, que definem os níveis de prioridade no acesso a esses recursos por base dos utilizadores com base na informação disponibilizada quer pelas camadas inferiores quer pelas camadas superiores do modelo. Este novo paradigma de comunicação, designado por *cross-layer* resulta na maximização da capacidade de transporte de dados por parte do canal rádio móvel, bem como a satisfação dos requisitos de qualidade de serviço derivados a partir da camada de aplicação do modelo.

Na sua elaboração, procurou-se que o standard IEEE 802.16e, conhecido por *Mobile WiMAX* respeitasse as especificações associadas aos sistemas móveis celulares de quarta geração. A arquitectura escalonável, o baixo custo de implementação e as elevadas taxas de transmissão de dados resultam num processo de multiplexagem de dados e valores baixos no atraso decorrente da transmissão de pacotes, os quais são atributos fundamentais para a disponibilização de serviços de banda larga. Da mesma forma a comunicação orientada à comutação de pacotes, inenente na camada de acesso ao meio, é totalmente compatível com as exigências em termos da qualidade de serviço dessas aplicações. Sendo assim, o Mobile WiMAX parece satisfazer os requisitos exigentes das redes móveis de quarta geração.

Nesta tese procede-se à investigação, projecto e implementação de algoritmos de encaminhamento de pacotes tendo em vista a eficiente gestão do conjunto de recursos rádio nos domínios do tempo, frequência e espacial das redes móveis celulares, tendo como caso prático as redes móveis celulares suportadas no standard IEEE802.16e. Os algoritmos propostos combinam métricas provenientes da camada física bem como os requisitos de qualidade de serviço das camadas superiores, de acordo com a arquitectura de redes baseadas no paradigma do cross-layer. O desempenho desses algoritmos é analisado a partir de simulações efectuadas por um simulador de sistema, numa plataforma que implementa as camadas física e de acesso ao meio do standard IEEE802.16e

Keywords

Beyond 3G and 4G mobile wireless networks; cross-layer design framework; quality of service in mobile wireless networks; packet scheduling; radio resource management; dynamic resource allocation architecture; space division muitple access; utility-based packet scheduling; system level simulations; dynamic rersource allocation architecture.

Abstract

In the last decade mobile wireless communications have witnessed an explosive growth in the user's penetration rate and their widespread deployment around the globe. It is expected that this tendency will continue to increase with the convergence of fixed Internet wired networks with mobile ones and with the evolution to the full IP architecture paradigm. Therefore mobile wireless communications will be of paramount importance on the development of the information society of the near future.

In particular a research topic of particular relevance in telecommunications nowadays is related to the design and implementation of mobile communication systems of 4th generation. 4G networks will be characterized by the support of multiple radio access technologies in a core network fully compliant with the Internet Protocol (all IP paradigm). Such networks will sustain the stringent quality of service (QoS) requirements and the expected high data rates from the type of multimedia applications to be available in the near future.

The approach followed in the design and implementation of the mobile wireless networks of current generation (2G and 3G) has been the stratification of the architecture into a communication protocol model composed by a set of layers, in which each one encompasses some set of functionalities. In such protocol layered model, communications is only allowed between adjacent layers and through specific interface service points. This modular concept eases the implementation of new functionalities as the behaviour of each layer in the protocol stack is not affected by the others. However, the fact that lower layers in the protocol stack model do not utilize information available from upper layers, and vice versa, downgrades the performance achieved. This is particularly relevant if multiple antenna systems, in a MIMO (Multiple Input Multiple Output) configuration, are implemented. MIMO schemes introduce another degree of freedom for radio resource allocation: the space domain. Contrary to the time and frequency domains, radio resources mapped into the spatial domain cannot be assumed as completely orthogonal, due to the amount of interference resulting from users transmitting in the same frequency sub-channel and/or time slots but in different spatial beams. Therefore, the availability of information regarding the state of radio resources, from lower to upper layers, is of fundamental importance in the prosecution of the levels of QoS expected from those multimedia applications.

In order to match applications requirements and the constraints of the mobile radio channel, in the last few years researches have proposed a new paradigm for the layered architecture for communications: the *cross-layer design framework*. In a general way, the cross-layer design paradigm refers to a protocol design in which the dependence between protocol layers is actively exploited, by breaking out the stringent rules which restrict the communication only between adjacent layers in the original reference model, and allowing direct interaction among different layers of the stack.

An efficient management of the set of available radio resources demand for the implementation of efficient and low complexity packet schedulers which prioritize user's transmissions according to inputs provided from lower as well as upper layers in the protocol stack, fully compliant with the cross-layer design paradigm. Specifically, efficiently designed packet schedulers for 4G networks should result in the maximization of the capacity available, through the consideration of the limitations imposed by the mobile radio channel and comply with the set of QoS requirements from the application layer.

IEEE 802.16e standard, also named as Mobile WiMAX, seems to comply with the specifications of 4G mobile networks. The scalable architecture, low cost implementation and high data throughput, enable efficient data multiplexing and low data latency, which are attributes essential to enable broadband data services. Also, the connection oriented approach of Its medium access layer is fully compliant with the quality of service demands from such applications. Therefore, Mobile WiMAX seems to be a promising 4G mobile wireless networks candidate.

In this thesis it is proposed the investigation, design and implementation of packet scheduling algorithms for the efficient management of the set of available radio resources, in time, frequency and spatial domains of the Mobile WiMAX networks. The proposed algorithms combine input metrics from physical layer and QoS requirements from upper layers, according to the cross-layer design paradigm. Proposed schedulers are evaluated by means of system level simulations, conducted in a system level simulation platform implementing the physical and medium access control layers of the IEEE802.16e standard.

Table of Contents

Contents

TABLE OF CONTENTS		I
ABBREVIATIONS		Vi
CHAPTER 1 Introduction		1
1.1	Preliminaries	1
1.2	Broadband Wireless Technologies	2
1.2.1	3G Cellular Systems	2
1.2.2	Long Term Evolution (LTE)	3
1.2.3	Wireless Fidelity Systems (Wi-Fi)	3
1.2.4	Wireless Interoperability for Microwave Access (WiMAX)	4
1.3	Technical Challenges for Broadband Wireless	4
1.4	Cross-Layer Design	5
1.5	Packet Scheduler	6
1.6	Objectives	7
1.7	Outline of the Dissertation	7
1.8	Publications	9
CHAPTER 2 Cross-Layer Design		11
2.1	Introduction	11
2.2	ISO/OSI Reference Model	13
2.3	Motivation for Cross-Layer Design in Mobile Communications	14
2.3.1	Link Adaptation in Single User Point-to-Point Communication	15
2.3.2	Multiuser Diversity Gain in Multi-User Point-to-Multi-Point Communications	15
2.3.3	TCP over Wireless Links	16
2.3.4	Multi-User Gain with Quality-of-Service	16
2.3.5	Application's Adaptability to Changes in Physical and Network Layers	17
2.4	A Model for the Coordination of the Cross-Layer Entities	17
2.5	Cross-Layer Design Concepts	19
2.5.1	Types of Information Flow across Layers	19
2.5.2	Types of Cross-Layer Management Entities	20
2.6	Disadvantages of the Cross-Layer Design Reference Model	20
2.7	Related Work	21
2.8	Conclusion	25
CHAPTER 3 Mobile WiMAX		26
3.1	Introduction	26
3.2	WiMAX Evolution	28
3.3	WiMAX Physical Layer (PHY)	31
3.3.1	Orthogonal Frequency Division Multiplexing (OFDMA) Basics	31
3.3.2	OFDMA Sub-Channelization in WiMAX	32
3.3.2.1	Diversity Permutation Sub-Carrier Sub-Channelization	32
3.3.2.2	Band Adjacent Multi-Carrier (AMC) Sub-Channelization	33
3.3.3	Frame Structure of Mobile WiMAX	33
3.4	WiMAX Medium Access Control Layer (MAC)	37
3.4.1	MAC Layer Functional Blocks	37
3.4.2	Mechanisms for Quality of Service Support	38
3.4.3	Service Classes in Mobile WiMAX	39
3.4.4	Bandwidth Request and Assignment in Mobile WiMAX	41
3.4.5	Advanced Features for Performance Enhancement	42
3.4.6	Adaptive Modulation and Coding (AMC)	43
3.4.7	Advanced Antenna Systems (AAS)	43
3.4.8	Hybrid Automated Repeat Request (HARQ)	44
3.4.9	Fractional Frequency Reuse	44

3.5	Cross-Layer Implementation in Mobile WiMAX	45
3.5.1	Introduction	45
3.5.2	Cross-Layer Design for Capacity Improvement	46
3.5.2.1	Cross-Layer Design for Effective Resource Allocation	46
3.5.2.2	Cross-Layer Design for Advanced Antenna Techniques	46
3.5.2.3	Cross-Layer Design for Detection and Error Recovery	47
3.5.3	Cross-Layer Design for Quality of Service Support	47
3.6	Related Work	47
3.7	Conclusion	49
CHAPTER 4 System Level Simulator for Mobile WiMAX System		51
4.1.	Introduction	51
4.2.	System Level Simulation Methodology	54
4.2.1	Simulation Execution Flow	55
4.2.2	Simulation of Packet Decoding Process	57
4.3	Network Scenario and Layout	58
4.4	Propagation Channels	59
4.4.1	Path-Loss Model	60
4.4.2	Shadowing (Slow Fading) Model	60
4.4.3	Fast Fading Model	61
4.5	Signal to Interference plus Noise Ratio (SINR) Modelling	62
4.6	Link Level Interface Modelling – General Concepts	65
4.6.1	Link Level Interface Modelling for Mobile WiMAX System	67
4.6.1.1	Effective SIR Mapping Functions	69
4.6.1.2	Exponential Effective SINR Mapping (EESM)	70
4.7	MIMO Channel Modelling in System Level Simulations	70
4.7.1	3GPP Spatial Channel Model	71
4.7.2	Modelling the SINR at the Mobile Station	75
4.7.2.1	Modelling the Desired User Signal at the Mobile Station	75
4.7.2.2	Modelling Inter-Cell Signal Interference at the Mobile Station	76
4.7.3	Link to System Interface for MIMO channel	77
4.8	Traffic Models	78
4.9	Performance Metrics	79
4.10	Related Work	79
4.11	Conclusion	81
CHAPTER 5 Dynamic Resource Allocation Architecture for Mobile WiMAX		82
5.1	Introduction	82
5.2	System Profile for Mobile WiMAX DRA	84
5.3	Implementation of the Map of Resources	87
5.4	DRA Architecture	90
5.4.1	Introduction	90
5.4.2	Link Adaptation	91
5.4.3	Asynchronous Hybrid Automatic Repeat Request (HARQ)	93
5.4.4	Scheduler	93
5.4.5	Resource Manager	94
5.4.6	Resource Allocation Procedure	97
5.5	Related Work	100
5.6	Conclusion	103
CHAPTER 6 System Validation		104
6.1	Introduction	104
6.2	Validation of the Basic System Level Simulation Platform	105
6.2.1	Fast Fading Channel Model Implementation	105
6.2.1.1	Multipath Power Profile	106
6.2.1.2	Rayleigh Distribution of the Path Amplitude	106
6.2.1.3	Time Correlation	107

6.2.1.4	Fast Fading Channel Correlation over Each Frame Period	108
6.2.2	Shadowing	109
6.2.2.1	Log-Normal Law	109
6.2.2.2	Space Correlation	109
6.2.2.3	Inter-Site Correlation	110
6.2.3	User Distribution over the Network	111
6.2.4	Validation of the MIMO Channel Model Used in the System Level Simulations	112
6.2.4.1	Root Mean Square (rms) Delay Spread	112
6.2.4.2	Root Mean Square (rms) of the Angle Spread at the Base Station	113
6.2.4.3	Root Mean Square (rms) of the Angle Spread at the Mobile Station	114
6.2.4.4	CDF Distribution of the Powers from all Paths of the Multi-Path Channel	114
6.2.4.5	Dynamic Range of Variation of the Powers of all Paths in the Multi-Path Channel Model	115
6.3	Validation of the Dynamic Resource Allocation Module	115
6.3.1	User Geometry	117
6.3.2	User Residual Frame Erasure Rate (FER)	118
6.3.3	Average Number of Transmissions Attempts per Packet	119
6.3.4	Number of Times a User has been Scheduled	120
6.3.5	Average Received SINR per Packet	121
6.3.6	User Service Throughput	122
6.3.7	Average Packet Delay	122
6.4	Scheduling Algorithms	124
6.4.1	Round Robin (RR)	124
6.4.2	Maximum C/I (CI)	125
6.4.3	Max C/I over Average C/I (AvgCI)	125
6.4.4	Proportional Fairness (PF)	126
6.4.5	Modified Largest Weighted Delay First (M-LWDF)	127
6.4.6	Exponential (EXP)	128
6.5	Performance Evaluation of the Dynamic Resource Allocation Module	128
6.5.1	Performance Evaluation for the Full Queue Traffic Model	131
6.5.2	Performance Evaluation for the Near Real Time Video (NRTV) Traffic Model	135
6.5.3	Performance Evaluation for the World Wide Web (WWW) Traffic Model	138
6.5.4	Performance Evaluation for Full Queue Traffic Model with MIMO Channel	140
6.6	Conclusion	143
CHAPTER 7 Utility-Based Packet Scheduling under Mobile WiMAX Network Scenario		145
7.1	Introduction	145
7.2	Packet Scheduling in Broadband Wireless Access (BWA) Networks	147
7.3	Utility-Based Packet Scheduling	149
7.3.1	Introduction	149
7.3.2	Utility-Based Scheduling Principle and Cross-Layer Design	150
7.3.3	Packet Utility and Utility Function	152
7.3.4	Scheduling Algorithm	154
7.3.5	Guidelines for Utility Function Definition	158
7.4	Multi-Class Utility-Based Packet Scheduling	159
7.4.1	Proposed Algorithm	159
7.4.2	Proposed Utility Functions	161
7.4.2.1	Proposal of a Utility Function for Voice over IP (VoIP) Traffic Users	162

7.4.2.2	Utility Function for the World Wide Web (WWW) Traffic Users	162
7.4.3	Packer Bundling for VoIP Scheduling Efficiency	163
7.4.4	Algorithm Implementation	165
7.4.4.1	Scheduler	165
7.4.4.2	Resource Allocation	165
7.4.5	Simulation Scenario	167
7.4.6	Results	168
7.5	Joint Utility-Token Bucket Based Packet Scheduler	172
7.5.1	Scheduling Principle	173
7.5.1.1	Non Real Time (NRT) Users	173
7.5.1.2	Best Effort (BE) Users	176
7.5.2	Utility Functions Definition	177
7.5.2.1	Real Time Service Flow – VoIP	178
7.5.2.2	Real Time Service Flow – NRTV	179
7.5.2.3	Non Real Time Service Flow – WWW	179
7.5.2.4	Best Effort Service Flow – File Transfer Protocol (FTP)	179
7.5.3	Simulation Scenario	180
7.5.4	Results	180
7.5.4.1	Users Satisfaction Ratio	180
7.5.4.2	Average Packet Delay per User	182
7.5.4.3	Average Packet Drop Rate per User	184
7.5.4.4	Average Throughput per User	185
7.5.5	Performance Distributions for Maximum System Load	186
7.5.5.1	Performance for VoIP Users	186
7.5.5.2	Performance for NRTV Users	187
7.5.5.3	Performance for WWW Users	188
7.5.5.4	Performance for FTP Users	189
7.6	Related Work	190
7.7	Conclusion	194
CHAPTER 8 Space Division Multiple Access with Utility Based Packet Schedulers for Mobile WiMAX		196
8.1	Introduction	196
8.2	Spatial Beamforming	198
8.3	SDMA Scheme for IEEE802.16e Mobile WiMAX	199
8.3.1	User Spatial Separability	200
8.3.2	SINR Estimation after Performing Beamforming	202
8.4	Proposed SDMA-Enabled DRA Architecture	203
8.4.1	Computation of Users Correlation	204
8.4.2	Computation of SINR for Resources in SDMA and in non-SDMA Zones	205
8.5	Joint Scheduling and Resource Allocation Algorithm Using SDMA Multiple Access	207
8.6	Results	210
8.6.1	Performance for Full Queue Traffic Users	210
8.6.2	Performance for VoIP and WWW Traffic Users	211
8.6.2.1	Voice over IP (VoIP) Traffic Model	211
8.6.2.2	World Wide Web (WWW) Traffic Model	216
8.7	Related Work	219
8.8	Conclusion	220
CHAPTER 9 Joint Time and Frequency Domains Packet Scheduler for Mobile WiMAX		223
9.1	Introduction	223
9.2	Resource Space for Time-Frequency Domain Scheduler	225
9.3	Proposed Scheduler	226
9.3.1	Time Domain Packet Scheduler	226

9.3.2	Frequency Domain Packet Scheduler	228
9.3.3	Computation of Final Priority	229
9.4	Results	229
9.5	Related Work	236
9.6	Conclusion	238
CHAPTER 10 Conclusions		240
10.1	Preliminaries	240
10.2	Cross-Layer Design	242
10.3	Mobile WiMAX Networks	242
10.4	System Level Simulations for Mobile WiMAX	242
10.5	Dynamic Resource Allocation Module Architecture	243
10.6	Validation of the System Level Simulator and Dynamic Resource Allocation Module	244
10.7	Utility-Based Packet Scheduling for Mobile WiMAX	244
10.8	Space Division Multiple Access with Utility-Based Packet Schedulers for Mobile WiMAX	250
10.9	Joint Time and Frequency Domains Packet Scheduler for Mobile WiMAX	253
10.10	Future Research	256
REFERENCES		255
ANNEX A Methods for Effective SINR Mapping		274
A.1	Introduction	274
A.1.1	One Dimensional Data Compression and Mapping	274
A.1.1.1	Mean Instantaneous Capacity Mapping Method (MIC)	274
A.1.1.2	Exponential Effective SINR Mapping (EESM)	275
A.1.1.3	Effective SINR Mapping Based on Mutual Information (MI-ESM)	275
A.1.2	Two Dimensional Data Compression and Mapping	276
A.2	Calibration of the Link to System Level Interface Model	277
ANNEX B Traffic Models		278
B.1	Introduction	278
B.2	Voice Over IP (VoIP) Traffic Model	278
B.3	3GPP Near Real Time Video (NRTV) Traffic Model	280
B.4	3GPP World Wide Web (WWW) Browsing Traffic Model	281
B.5	3GPP File Transfer Protocol (FTP) Traffic Model	282
ANNEX C Performance Metrics		284
C.1	Introduction	284
C.2	Throughput Performance Metrics	284
C.3	Performance Metrics for Delay Sensitive Applications	288
C.4	Fairness Criteria	289

Abbreviations

Abbreviations

16QAM	16 Quadrature Amplitude Modulation
1xEV-DO	1 x Evolution Data Only
2G	Second Generation
3G	Third Generation
3GPP	Third Generation Partnership Project
3GPP2	Third Generation Partnership Project 2
4G	Fourth Generation
64QAM	64 Quadrature Amplitude Modulation
AAS	Advanced Antenna Systems
ACK	Acknowledge
ADSL	Asynchronous Digital Subscriber Line
AES	Advanced Encryption Standard
AMC	Adaptive Modulation and Coding
AoA	Angle of Arrival
AoD	Angle of Departure
ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode
AvgCI	Average C/I
AWGN	Additive White Gaussian Noise
B3G	Beyond Third Generation
BE	Best Effort
BER	Bit Error Rate
BLER	Block Error Rate
BPSK	Binary Phase Shift Keying
BWA	Broadband Wireless Access
CAC	Connection Admission Control
CBR	Constant Bit Rate
CC	Chase Combiner
CC	Convolutional Encoder
CDMA	Code Division Multiple Access
CDMA 2000	Code Division Multiple Access 2000
CI	Maximum CI scheduler
CID	Connection Identifier
COA	Compatibility Optimization Algorithm
CP	Cyclic Prefix

CPS	Common Part Sub-layer
CQI	Channel Quality Information
CQICH	Channel Quality Information Channel
CS	Convergence Sub-layer
CSMA	Carrier Sense Multiple Access
CSTD	Shift Transmit Diversity
CTC	Convolutional Turbo Encoder
DCA	Dynamic Channel Allocation
DL	Downlink
DLFP	Downlink Frame Prefix
DoA	Direction of Arrival
DOCSIS	Data Over Cable Service Interface Specification
DPS	Doppler Power Spectrum
DRA	Dynamic Resource Allocation
DSL	Digital Subscriber Line
EAP	Extensible Authentication Protocol
ECN	Explicit Congestion Notification
EESM	Exponential Effective SINR Mapping
ELN	Explicit Loss Notification
ertPS	Enhanced Real Time Pooling Service
ETSI	European Telecommunications Standard Institute
EXP	Exponential scheduler
FCH	Frame Control Header
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
FDPS	Frequency Domain based Packet Scheduler
FEC	Forward Error Control
FER	Frame Erasure Rate
FFT	Fast Fourier Transform
FIFO	First In First Out
FTP	File Transfer Protocol
FUSC	Full Usage Sub-Carrier sub-channelization
GSM	Global System for Mobile communications
HARQ	Hybrid Automatic Repeat Request
HOL	Head of Line
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSUPA	High Speed Uplink Packet Access
ICI	Inter-carrier interference
IE	Information Element
IETF	Internet Engineering Task Force
IFFT	Inverse Fast Fourier Transform
IP	Internet Protocol
IS-95	International Standard 95

ISI	Inter Symbol Interference
ISO	International Standards Organization
ITU	International Telecommunications Union
LA	Link Adaptation
LLC	Logical Link Control
LTE	Long Term Evolution
LUT	Look-Up Table
MAC	Medium Access Control
MAP	Mobile Application Part
MCM	Multi-carrier modulation
MCS	Modulation and Coding Scheme
MIMO	Multiple-Input Multiple-Output
MISO	Multiple-Input Single Output
MLD	Maximum Likelihood Decoder
M-LWDF	Modified Largest Weighted Delay First scheduler
MMSE	Minimum Mean Square Error
MPDU	Medium Access Control layer Protocol Data Unit
MRC	Maximum Ratio Combiner
NACK	Negative Acknowledge
NGN	Next Generation Network
NLOS	Non Line of Sight
NRT	Non Real Time
nrtPS	Non Real Time Pooling Service
NRTV	Near Real Time Video
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	OFDM Multiple Access
OS	Opportunistic Scheduling
OSI	Open Systems Interconnection
PDU	Packet Data Unit
PF	Proportional Fairness scheduler
PHY	Physical Layer
PMT	Point to Multi-Point
PUSC	Partial Usage of Sub-Carrier channelization
QoS	Quality of Service
QPSK	Quaternary Phase Shift Keying
RAM	Resource Allocation Map
RAU	Resource Allocation Unit
RB	Resource Block
RF	Radio Frequency
RR	Round Robin scheduler
RRM	Radio Resource Management
RT	Real Time
RTG	Receive to Transmit Gap
rtPS	Real Time Polling Service

RU	Resource Unit
SAP	Service Access Point
SCM	Spatial Channel Model
SCS	Scheduling Candidate Set
SDMA	Space Division Multiple Access
SDU	Service Data Unit
SER	Symbol Error Rate
SFID	Service Flow Identifier
SIMO	Single-Input Multiple-Output
SINR	Signal to Interference plus Noise Ratio
SISO	Single-Input Single-Output
SM	Spatial Multiplexing
SMTP	Simple Mail Transfer Protocol
SNR	Signal to Noise Ratio
SOFDMA	Scalable OFDMA
SOS	Sum of Sinusoids
SPDU	Service Packet Data Unit
SS	Subscriber Station
STBC	Space Time Block Code
TCP	Transmission Control Protocol
TDD	Time Division Duplexing
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
TD-SCDMA	Time Division Synchronous Code Division Multiple Access
TTG	Transition to Transmit Gap
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UGS	Unsolicited Grant Service
UL	Uplink connection
UMTS	Universal Mobile Telecommunications System
UTRAN	UMTS Terrestrial Radio Access Network
VoIP	Voice over IP
WCDMA	Wideband Code Division Multiple Access
WiFi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
WWW	World Wide Web
ZF	Zero Forcing

Chapter 1

Introduction

1.1 Preliminaries

In the last decade mobile communications have witnessed an explosive increase in the amount of users and in the amount of data transferred over the air interface. High bandwidth demanding multimedia applications, originally intended for the fixed Internet, are currently being envisioned for the mobile wireless environment due to technology advances, such as the implementation of Multiple-Input Multiple-output (MIMO) antennas [1-3] and Adaptive Modulation and Coding (AMC) transmission schemes [4], which result in a more efficient utilization of the scarce bandwidth and the increase in the available capacity for data.

First (analog) (1G) and second (digital) generation mobile networks were originally designed for the provision of voice services in a circuit switched transmission mode. Although some data applications are also provided by 2G networks, such as Global System for Mobile Communications (GSM) or Code Division Multiple Access (CDMA) based standards, high bandwidth demanding multimedia applications and services are not envisioned. Third generation (3G) mobile communication systems, based on wideband code division multiple access (WCDMA) and CDMA 2000 radio access technologies, have seen widespread

deployment. As more packet-based applications are implemented the need for more bandwidth, quality of service provision and higher spectrum efficiency increases. Therefore, the research community has been actively involved in the design of a new generation of mobile communication networks: the fourth generation (4G). With this new concept of mobile communications a completely new air interface is justified by the need to serve current and future multimedia and high bandwidth consuming type of applications. Typically, a reasonable approach would be to aim at 100 Mbps full-mobility wide area coverage and 1 Gbps low-mobility local area coverage in about 2010 [5].

Next-generation wireless involves the concept of a major move toward ubiquitous wireless communications systems and high-quality wireless services. 4G mobile communications involve a mix of concepts and technologies in the making of this vision. Some of these can be recognized as derived from 3G, such as High Speed Packet Access (HSPA), which is an evolution from Universal Mobile Telecommunications Standard (UMTS) WCDMA; whereas others involve new approaches to wireless mobile, like Orthogonal Frequency Division Multiple Access (OFDMA) scheme. 4G mobile networks evolution [6] are relevant in the movement toward a new wireless world that is a total convergence of wireless mobile and wireless access communications.

1.2 Broadband Wireless Technologies

Broadband wireless networks attempt to provide broadband type of applications, initially envisioned for fixed networks, into the mobile environment. There are two different types of broadband wireless services: the first type attempts to provide a set of services similar to that of the traditional fixed-line broadband, using wireless as the transmission medium. This is called *fixed wireless broadband*. The second type of broadband wireless is called *mobile wireless broadband* and it differs from the fixed one by offering the additional functionality of portability, nomadicity, and mobility. In this section different proposed wireless systems intended for broadband wireless access (BWA) are enumerated.

1.2.1 3G Cellular Systems

3G networks are currently being implemented by cellular network operators to deliver broadband applications over wireless. GSM operators evolved their second generation (2G) networks by deploying UMTS High Speed Downlink Packet Access (HSDPA) and High Speed Uplink Packet Access (HSUPA), as part of the UMTS evolution. Traditional CDMA operators are deploying 1xEV-DO (1x Evolution Data Only) as part of their 3G solution. In China and some Asian countries TD-SCDMA (Time Division-Synchronous CDMA) standard is also being implemented as the 3G solution for broadband wireless access networks.

High Speed Packet Access (HSPA)

HSDPA [7] is a downlink evolution of the UMTS standard, defined in the Third Generation Partnership Project (3GPP) UMTS Release 5. It is capable of providing a peak user data rate of 14.4 Mbps, using a 5 MHz channel, provided that all 15 sub-channelization codes available in the radio frame are assigned to one user only. HSDPA is based on a Time Division Duplex (TDD) frame with 15 slots and duration of 5 ms. Some enhancements such as: spatial processing, diversity reception and multiuser detection are used to provide higher performance over basic systems. An uplink version, HSUPA, supports peak data rates up to 5.8 Mbps and is standardized as part of the 3GPP release 6 specifications.

1xEvolution Data Only (1xEV-DO)

1xEV-DO [8] is a high-speed standard defined as an evolution to second-generation IS-95 CDMA systems, by the 3GPP2. It supports a peak downlink data rate of 2.4 Mbps in a 1.25 MHz channel. The Revision A of this standard supports 3.1 Mbps and Revision B will support 4.9 Mbps. 3G systems are also evolving to support multimedia services.

1xEvolution Data Only Revision C (1xEV-DO Rev. C)

This is the major revision of the 1xEV-DO standard from 3GPP2 for longer-term evolution (LTE) to offer data rates of 70 Mbps to 200 Mbps in the downlink and 30 Mbps to 45 Mbs in the uplink, using 20 MHz channel bandwidth.

1.2.2 Long Term Evolution (LTE)

3GPP is developing the next revision of the 3G standards, named Long Term Evolution [9] (LTE). The objective of this new standard is to provide peak data rates up to 100 Mbps in the downlink and 50 Mbps in the uplink, with an average spectral efficiency of three to four times that of Release 6 HSPA. This will be achieved with the introduction of a completely new air interface based on OFDM/OFDMA (Orthogonal Frequency Division Multiplexing; OFDM Multiple Access), and MIMO.

1.2.3 Wireless Fidelity Systems (Wi-Fi)

Wi-Fi systems are based on the IEEE 802.11 family of standards from IEEE. It is primarily a local area networking (LAN) technology. Current Wi-Fi systems, based on IEEE 802.11a/g support a peak data rate of 54 Mbps with indoor coverage less than 100 m. Wi-Fi provides higher peak data rates than do 3G systems thanks to the larger bandwidth used (over 20 MHz). But capacity for outdoor scenarios is mainly reduced due to the inefficient multiple access scheme used, CSMA (Carrier Sense Multiple Access), along with the interference constraints of operating in the license-exempt band. Another limitation of Wi-Fi standard is that it was not designed to support user's mobility. New technologies are emerging with the revised IEEE 802.11n standard: multiple-antenna, spatial multiplexing and transmit diversity, which are envisioned to support a peak data rate of at least 100 Mbps.

1.2.4 Wireless Interoperability for Microwave Access (WiMAX)

According to the WiMAX Forum, Mobile WiMAX is a broadband wireless solution that enables convergence of mobile and fixed broadband networks through a common wide area radio access technology and flexible network architecture [10-11]. In order to achieve high peak data rate transmissions in Non-Line of Sight (NLOS) scenarios, its air interface is based on the OFDMA multiple access for both the downlink as well as for the uplink connection. In order to address different channel bandwidths, ranging from 1.25 MHz to 20 MHz, according to regulator specifications from different markets, Mobile WiMAX implements scalable OFDMA (SOFDMA) [12]. This technique results in the modification of the size of the Fast Fourier Transform (FFT) used in the OFDM modulation, according to channel bandwidth, while keeping inter-carrier frequency separation constant, to provide orthogonality. WiMAX Medium Access Control (MAC) layer is connection oriented for Quality of Service (QoS) support [13]. Some of the key features provided by Mobile WiMAX are enumerated as follows:

- OFDMA as multiple access technology for both downlink and uplink connections.
- Scalable OFDMA.
- Adaptive Modulation and Coding (AMC).
- Very high peak bit rates.
- Use of retransmission schemes at the link layer for fast and robust retransmissions.
- Use of flexible and dynamic per user resource allocation by means of OFDMA multiple access and sub-channelization schemes.
- Support for advanced antenna techniques such as MIMO and beamforming.
- Implementation of a connection oriented MAC layer for the provision of QoS.
- Robust security schemes and protocols.
- Support of mobility and nomadicity.
- All IP-based architecture for cost efficiency and convergence with fixed IP networks.

1.3 Technical Challenges for Broadband Wireless

Broadband wireless systems must deliver high peak data rates per each cell to end users, while providing QoS requirements from a set of services such as voice over IP, real time video, web browsing, games, etc. These networks must be able to support IP-based applications in order to provide the convergence which is expected to happen between broadband wireless and Internet [14-15].

In order to be successfully implemented as a viable competitor to the solutions provided by Asynchronous Digital Subscriber Line (ADSL) or Cable Modems, WiMAX, a particular implementation of BWA networks, must deliver significantly better performance than current

alternatives such as 3G and Wi-Fi systems. In order to meet these stringent requirements some key technical challenges must be addressed.

- To develop reliable transmission chains which are able to overcome the limitations resulting from such an aggressive transmission medium such as the mobile radio channel, namely for the kind of broadband data envisioned for transportation.
- To achieve high spectral efficiency and coverage in order to deliver broadband services to a large number of users, with limited spectrum available.
- To multiplex multimedia services with a variety of QoS requirements.
- To support mobility through seamless handover and roaming.
- To achieve low power consumption on hand-held devices.
- To provide robust security.
- To adapt IP-based protocols and architectures for the wireless environment to achieve low cost and convergence with wired networks.
- To provide a low cost implementation solution.

1.4 Cross-Layer Design

Traditional communication networks are implemented based on a hierarchical protocol layering architecture composed of seven layers, according to the Open Systems Interconnection (OSI) model from International Standards Organization (ISO). This communication parading resulted in the rapid proliferation of services and applications and the widespread deployment of the Internet. With the strict layering approach followed in this architecture, protocols can be deployed inside each layer, without taking into account other layers functionality and implementation details in the protocol stack. Adjacent layers communicate by means of standardized signalling messages sent through standardized interfaces named Service Access Points (SAPs).

Although appropriate for fixed communication networks, in the last years a new communication paradigm has evolved into a new architecture model, in which the strict layering and modular approach barriers are erased, allowing different layers in the protocol, be they adjacent or not, to interact by means of the exchange of signalling information. In particular, it was realized, that this c*ross-layer design* approach is appropriate for the scenario of wireless networks, as it can potentially result in performance enhancement [16-17], provided the breaking of the layers barriers is performed with care [18].

One of the main objectives of this work is the design and implementation of scheduling algorithms based on the cross-layer design paradigm. Proposed schedulers are based on the interaction and exchange of messages conveying the state of different layers in the protocol stack, in order to improve system capacity, maximize resource efficiency and provide

application's QoS requirements. It is worth mentioning that the IEEE 802.16e standard for Mobile WiMAX networks was specifically designed with control and signalling channels which can be used in the exchange of information, according to the cross-layer design approach. The architecture of the Dynamic Resource Allocation (DRA) module implemented in this work for Mobile WiMAX system level simulations is based on the control channels available in the Mobile WiMAX radio frame.

1.5 Packet Scheduler

Making efficient use of radio resources is a challenging task for 3G/4G wireless communication systems as the scarcity of radio resources, diverse QoS requirements and wireless channel conditions pose difficulties to the design of the scheduling and radio resource management [19-21].

The design of scheduling algorithms for wireless networks is constrained by the intrinsic characteristics of the mobile radio channel, which is a hazard means for conveying information. Fast fading and slow fading shadowing, as well as path-loss, result in a time-varying wireless link, both in time and space. This random behaviour of the signal impinging on the receiver is also location dependent and is influenced by the user's mobility across the cell.

Another issue that needs to be considered in the design of a packet scheduler is the limitation in the maximum level of the power available for data from batteries, as well as the desire to prolong batteries life as long as possible in order to maximize mobile's autonomy.

The scheduler, together with the DRA, the Connection Admission Controller (CAC), the Congestion Controller (CC) and the Link Adaptation (LA) modules, compose the Radio Resource Manager (RRM) and the functional behaviour of these modules depends on the scheduling policies implemented. The main function of the scheduler is to intelligently allocate radio resources to achieve high system performance in terms of efficiency and fairness in radio resource allocation. Depending on the type of multiple access scheme implemented in the air interface, resources are defined in time, frequency and space domains. Schedulers operate across different sessions (connections or service flows) in order to ensure that requested QoS parameters, such as packet delay and delay jitter, packet loss rate, minimum guaranteed throughput and maximum sustained throughput, are provided to each session.

Scheduling decisions require some information to be inputted into the implemented scheduling algorithm, such as: the number of active sessions, QoS constrains, link state and the state of the queues with new packets and with packets in retransmission. This information must be provided by signalling channels implemented in the radio frame for both downlink as well as the uplink connection. A well designed scheduler should be able to provide the following features:

Efficient link utilization: radio bandwidth is a scarce and costly resource. As so, it must be efficiently used, which means that resources should not be provided to users with bad channel

quality and, therefore, with a high probability of errors to occur in packet transmission. Also, under some scenarios of application, bandwidth should be assigned to users who can maximize data transmission over it.

Fairness: resources should be distributed fairly across sessions, in order to avoid starvation of some users in data transmission. However, fairness can result in a decrease in the level of efficiency achieved, which means that there is a trade-off in the scheduler design for the provision of fairness and optimization of the link utilization.

Complexity: the scheduler should be of low complexity in terms of system implementation. The objective is twofold: save battery power (a high complexity scheduler would require too much computations and therefore would be too much power consuming); and minimize the time required to perform scheduling decisions according to the scheduling algorithm.

Isolation: the scheduling algorithm should provide isolation among the different sessions. This means that a given session performance should not be affected by a misbehaving session (typically requiring more bandwidth than the one that should be initially attributed).

Energy consumption: the algorithm should take into account the need to save battery power.

Scalability: the scheduling algorithm should operate efficiently no matter the number of sessions attempting to access radio resources.

1.6 Objectives

The objective of this thesis is the investigation, design and implementation of packet scheduling algorithms for the management of radio resources in broadband wireless networks of future generation. These algorithms combine the metrics coming from the physical layer as well as the quality of service requirements from service applications in the application layers, according to a cross-layer design paradigm framework.

1.7 Outline of the Dissertation

This dissertation is organized in 10 chapters and 3 appendixes.

Chapter 1: gives a short introduction to the main issues covered in this research and outlines the objectives to be achieved with this thesis

Chapter 2: gives a detailed description about the Cross-Layer Paradigm and the details to be followed in the implementation of a cross-layer based architecture into a mobile broadband wireless network. The different types of cross-layer architectures are explained, some examples of algorithms implementing this concept are given and some cautions to be followed with such implementation are reinforced.

Chapter 3: describes the IEEE 802.16 standards for both Fixed WiMAX networks and Mobile WiMAX networks. The issues which are of particular importance for the implementation of the DRA and packet schedulers elaborated in this research are presented and the peculiarities of the

physical (PHY) and medium access (MAC) layers, which are considered in the implementation of the simulator used for system level simulations, are provided.

Chapter 4: this chapter provides all details followed into the design, modelling and implementation of the System Level Platform used in the realization of system level simulations for the Mobile WiMAX standard. Used traffic models for data generation as well as the channel models, for Single-Input Single-Output (SISO) and MIMO channels, are provided. The cellular layout architecture used into the simulator is described. Of particular interest in the realization of system level simulations are: the definition of link to system level interfaces, and the generation of a set of look up tables, to be used in the physical layer performance abstraction. The procedures followed in the derivation of the Signal to Interference plus Noise Ratio (SINR) and the function used to map the vector of SINRs into a single scalar, which can be inputted into the look-up tables, is detailed.

Chapter 5: together with chapters 7, 8 and 9, this chapter is the core of this thesis. It is about the design and implementation of the Dynamic Resource Allocation (DRA) module in the system-level platform. The proposed schedulers are plugged into this DRA architecture. All details regarding the DRA functionalities for the exchange of data and signalling messages are provided and the functionalities of the cross-layer design are enforced. The resources available for data allocation and the protocol followed into the signalling of used resources attributed to active and scheduled mobiles are given.

Chapter 6: After implementation of the DRA it must be validated by means of commonly benchmark packet schedulers under simplistic scenarios and/or traffic models. This is the purpose of this chapter. In particular some basic packet schedulers are used in this validation, ranging from full queue to World Wide Web and voice over IP traffic models. Simulations are conducted for different channel models, for SISO and MIMO.

Chapter 7: elaborates on the implementation of packets schedulers for QoS provision, fairness satisfaction and resource utilization efficiency, based on the notion of Utility Functions, a concept derived from economics. The principle behind the implementation of a utility-based packet scheduler is presented and the aspects which must be followed in the definition of the shape and set of parameters, which characterize the type of utility function to be assigned to each type of service class, are given. A modification of the original utility-based packet scheduler is explained, which is based on the jointly implementation of a token bucket packet scheduler with the original utility algorithm. This scheduler is shown to provide QoS to three different types of traffic: real time with maximum packet delay, non real time with minimum throughput guarantees and best effort with no QoS requirements. Commonly used schedulers such as the maximum C/I, the proportional fairness and the modified-largest weighed delay first, available in the literature, are used as benchmark to compare the performance against.

Chapter 8: One of the advanced features for the Mobile WiMAX standard is the implementation of advanced antenna systems (AAS), which can be used for beamforming generation in the implementation of Spatial Division Multiple Access (SDMA). In this chapter the utility-based scheduler is extended to be implemented in conjunction with a resource allocation algorithm implementing an SDMA access scheme. Therefore, besides the original time domain, another degree of freedom is achieved with SDMA: the space domain. The performance enhancements resulting from such DRA architecture are provided.

Chapter 9: In this chapter a variation of the original utility-based packet scheduler is also described. This new packet scheduler is based on the Adjacent Multi-Carrier (AMC) sub-channelization mode of Mobile WiMAX. This mode results in a multi-user frequency diversity gain if sub-channels are assigned to users according to the highest level of the channel quality which is achieved. Results are provided for the capacity with two basic types of sub-channelization modes: Partial Usage Sub-Carrier, PUSC (which is the sub-channelization implemented in all previous work) and AMC.

Chapter 10: draws the main conclusions of this thesis and discusses future research topics.

1.8 Publications

The following articles have been published during the PhD study:

- A. Nascimento, J. Rodriguez and A. Gameiro, "Utility Based Scheduling for Wireless Systems", Proc. of IST Summit2006, Mykonos, June 2006.
- A. Nascimento, J. Rodriguez and A. Gameiro, "Utility-Based Scheduling for MC-CDMA", Proc. of WPMC2006, San Diego, USA, Sep. 2006.
- A. Nascimento, J. Rodriguez and A. Gameiro, "**Dynamic resource allocation for IEEE802.16e, Proc. of the 3rd international conference on Mobile multimedia communications,** Nafpaktos, Greece, 2007.
- A. Nascimento, J. Rodriguez and A. Gameiro, "Dynamic Resource Allocation Architecture for IEEE802.16e: Design and Performance Analysis , Journal on Mobile Networks and Applications, Springer, vol. 13, no. 3-4, Aug 2008, pp. 385-397.
- V. Monteiro, A. Nascimento, J. Rodriguez, A. Gameiro, "Packet based system level simulator for cellular wireless B3G networks", Proc. Simtools, 2008.
- A. Nascimento, J. Rodriguez and A. Gameiro, "Dynamic Resource Allocation Architecture for IEEE802.16e: Design and Performance Analysis", Proc. ICT Mobile Summit, 2008.
- A. Nascimento and A. Gameiro, "Jointly Cross-Layer Scheduling and Dynamic Resource Allocation for RT and NRT Traffic Types for IEEE802.16e" accepted for IEEE VTC conference, Spring 2009, Barcelona.

- A. Nascimento, J. Rodriguez, "Dynamic Resource Allocation for IEEE 802.16e" The 1st International Conference on Mobile Lighweight Wireless Systems, Athens, Greece, May 18-20, 2009.
- A. Nascimento, J. Rodriguez, A. Gameiro, "A Joint Utility-Token Bucket Packet Scheduling Algorithm for IEEE 802.16e WiMAX", The Sixth International Symposium on Wireless Communication Systems (ISWCS 09), Sienna, Italy, September, 7-10, 2009.
- A. Nascimento, J. Rodriguez, A. Gameiro, "*A New Cross-Layer based Dynamic Resource Allocator for IEEE 802.16e Networks"* , The Second International Workshop on Cross-Layer Design (IWCLD 2009), Palma Mallorca, Spain, June, 11-12, 2009.
- A. Nascimento, J. Rodriguez, "Dynamic Resource Allocation for IEEE 802.16e", Mobile Lightweight Wireless Systems, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Volume 13, Springer Berlin Heidelberg.

Chapter 2

Cross-Layer Design

2.1 Introduction

A layered and hierarchical protocol reference model was the original proposal for the development of communication protocols intended for fixed wired networks. The enormous success of this model for the deployment of communication networks and the widespread use of the Internet resulted in the proliferation of many new applications. In the last decade, the exponential increase in the penetration rate of mobile users, resulting from the maturation of wireless networks of second generation (2G) and also from the development of new wireless standards, brought out the natural idea of coupling both worlds: multimedia applications, originally designed for fixed Internet and the availability of these applications for the mobile and wireless scenario.

On future wireless networks, traffic is expected to be a mixture of real-time applications, such as voice, multimedia teleconferencing or games, together with more delay insensitive applications, in which users do not necessarily have any minimum requirements in terms of time for the delivery of the information, such as web browsing, messaging and file transfers. These applications pose big challenges to the wireless medium as they require diverse levels of quality of service (QoS) and higher amounts of bandwidth for data transfer in restricted periods of time.

With the development of new high-consuming bandwidth multimedia applications it is normal to expect an increasing demand for wireless data services [22]. However, differently from the wired fixed channel, the mobile radio channel is characterized as being an aggressive medium to convey information. Multipath propagation and signal attenuation due to path-loss and shadowing pose significant challenges in the provision of QoS requirements and on the big amounts of bandwidth associated to those applications. Also, radio spectrum is scarce and must be efficiently utilized. Very early these aspects evidenced that although layered architecture has served well for wired networks they do not suit the requirements and constraints posed by wireless ones.

In order to match applications requirements and the constraints of the mobile radio channel, in the last few years researches have proposed a new paradigm for the layered architecture for communications: the *cross-layer design framework*. In a general way, cross-layer design refers to protocol design in which the dependence between protocol layers is actively exploited, by breaking out the stringent rules which restrict the communication only between adjacent layers in the original reference model, and allowing direct interaction among different layers of the stack, in order to obtain performance gains [16].

This chapter details the principles and motivations behind the cross layer design paradigm and is organized as follows. Section 2 is an introduction to the ISO/OSI protocol layering stack used in the implementation of general communication networks. The motivations for the design of such a modular protocol stack are presented. Section 3 stresses out the limitations resulting from the use of such strict protocol layered architecture in wireless networks. The fundamental properties which differentiate wireless networks, both in concept and behavior, from fixed communication ones are presented and the benefits resulting from the implementation of such cross layer design framework are pointed out. Section 4 elaborates on a model proposed for the cross layer design protocol architecture. This model specifies the set of functionalities provided by each entity in the model. Four planes, extending vertically across the main layers of the OSI reference model are proposed, namely: Physical/Link, Network, Transport and Application. Section 5 describes the different approaches proposed in the literature for the enabling of information exchange among the different layers of the protocol stack, in a cross layer design framework. The concept of cross-layer design can result in network malfunctioning and instability may arise if the interfaces among layers and the type of information to be exchanged are not properly designed. As such, some caution must be followed in implementing cross layer-based wireless network protocol stacks. This is the issue of section 6. Section 7 presents the related work in cross layer design and section 8 concludes the chapter.

2.2 ISO/OSI Reference Model

The Open Systems Interconnection Basic Reference Model (OSI Reference Model or OSI Model) [23] is an abstract description for the design of communication systems whose reference architecture is stratified into modular sub-layers, and in which each sub-layer performs individual and self-contained tasks for the whole system. This reference model was proposed and developed as part of OSI initiative. The goal of this model is the sub-division of complex tasks, encompassing the transmission of data over the network, into smaller and simpler ones, performed by each layer of the protocol stack. In this modular approach each layer is seen as a black box from the remaining ones and they communicate only through well defined primitives sent over interfaces among adjacent layers, named Service Access Points (SAP). By means of these primitives each layer provides services to the layer above and receives services from the layer below. This modular approach also makes it easier the upgrade and portability of protocols designed for specific layers of the model across different systems.

The OSI reference model is sub-divided into seven layers. From top to bottom: Application, Presentation, Session, Transport, Network, Data-Link and Physical. However, the most important ones, from the point of view of the transportation media architecture, are: the Application, Transport, Network, Data-Link and Physical. It is important to mention that this architecture is more theoretical than practical, regarding implementation in communication networks, as most of the protocols in use on the Internet were designed as part of the TCP/IP model, and may not fit cleanly into the OSI model.

Application layer: this is the layer closest to the end user as it interacts directly with the user's software application. Typical functions of the application layer are the identification of the partners involved in the communication process, the determination of resource availability and the synchronization of the communication. Some examples of applications layer protocols are: Telnet, File Transfer Protocol (FTP) and Simple Mail Transfer Protocol (SMTP).

Presentation layer: this layer has to do with the fact that there exist different ways to present the information to application layer protocols. Protocols on this layer convert data to the standard ASCII pattern and perform data compression and data encryption. Presentation service data units are then encapsulated into session Protocol Data Units PDUs and moved down the stack.

Session layer: this layer establishes, manages and terminates the connections (sessions) between computers which are used in the transportation of data and control information over the network.

Transport layer: this layer provides the transport of data between end users in a transparent way, i.e., hiding the details of the whole communication process across the physical entities of the network to users' applications. It implements data segmentation, flow and error controls for the reliable transport of data, no matter what type of physical system is used. In the Internet the

Transport Control Protocol (TCP) and User Datagram Protocol (UDP) are the best examples of transport layer protocols in the OSI model.

Network layer: this layer has built-in functionalities for the transfer of data from the source to the destination nodes of the communication process. The data coming from the transport layer is segmented into smaller pieces of information, with variable lengths, and is routed through different paths into the destination. These paths are established among physical entities (routers) which can be inside or in different networks. The routing of these pieces of data must comply with the quality of service (QoS) required from the transport layer. The best known example of network protocol is the Internet Protocol (IP).

Data Link layer: this layer provides services to the physical layer for the reliable transfer of data. It performs functions for error detection and correction. It is sub-divided into two sub-layers: the **Medium Access Control (MAC)** sub-layer and the **Logical Link Control (LLC)** sub-layer. The MAC sub-layer determines and controls the access to the resources in the physical layer and the LLC sub-layer is responsible for the control of the connection.

Physical layer: this layer generates the physical signals for the modulation of the bits of information. Among its tasks are: the specification of the physical parameters used in the representation of the logical bits (voltages, currents and frequencies) and the coding and modulation of these bits into the format appropriate to the characteristics of the physical medium.

In the OSI reference model the lower three layers are related and specific to the type of media being used in the transportation of the information. The upper three layers closely depend on the type of the host running the end-user application. Figure 1 shows the OSI reference model.

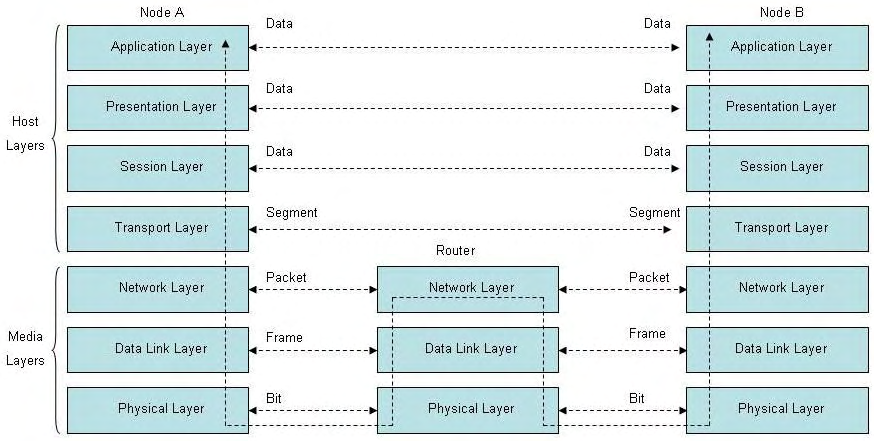


Figure 1 - The seven layers of the ISO/OSI Reference Model

2.3 Motivation for Cross-Layer Design in Mobile Communications

By nature, the wireless channel is a very aggressive medium to convey information: large-scale propagation phenomena accrued from path-loss signal attenuation and shadowing, coupled with

fast-scale multipath propagation and co-channel interference in both time, frequency and space domains, pose challenges in the transmission of information over the mobile channel [24, 25].

- From a pessimistic point of view, these small and large scale signal variations affect the system performance, either in terms of the QoS provided to the different users, or in the efficient use of the scarce radio resources available, and create several new problems for protocol design that cannot be well handled in the framework of the layered architectures.
- From an optimistic point of view, however, wireless networks offer new opportunities for communications which cannot be addressed in a strictly layered protocol design. For instance, the variations in link quality in time, frequency and space domains enables the opportunistic use of the channel if the transmission parameters can be dynamically adapted, according to these variations.

It seems logical to admit that the strict modular and hierarchical layering principles of the OSI reference model, in which the functionality of each layer is not influenced by the performance of others, do not address such constraints and requirements associated to the mobile wireless channel, and do not provide the multiuser gains which can be achieved whenever link adaptation is performed, according to variations in radio channel quality along time and space. It is straightforward to see that these issues can be more efficiently addressed if the design of the original protocol stack is modified according to the new cross-layer design paradigm.

The following sections illustrate some examples of scenarios, where the direct communication among different layers in the protocol stack results in significant improvements in the performance of the network.

2.3.1 Link Adaptation in Single User Point-to-Point Communication

The modulation and coding schemes (MCS) used in the transmission of each symbol can be modified dynamically over time, according to the state of the channel reported by the end user [26]. For example: mobile users report the state of the downlink (from base to mobile station) connection to the base station in pre-defined periods of time. This information is used in the base station to estimate the downlink channel and to select the most appropriate MCS scheme, for the maximization of the data rate over the channel, provided the estimated ratio of packets received with error to the total amount of packets transmitted is under a given upper bound.

2.3.2 Multiuser Diversity Gain in Multi-User Point-to-Multi-Point Communications

Consider a conventional cellular system with a fixed base station and a number of mobile users. Packets arrive from the wired Internet and are queued temporality at the base station. Consider also a simple Time Division Multiple Access (TDMA) system. A packet scheduler selects, in the beginning of each time-slot, the user which should transmit. As the mobile radio channels vary independently for each user (assuming they are not collocated), and assuming the base

station has perfect information regarding the state of the channel to each user, the user to be selected could be the one with best channel quality (measured as the signal to noise and interference ratio – SINR) in the beginning of the time slot. This simple strategy was proposed in [26] and results in the maximization of the system throughput conveyed in the cell. Its drawbacks have mainly to do with the channel starvation sensed by other users as they can remain without channel access for a long period of time if their channel has bad quality and/or varies slowly. This gain in throughput due to channel-dependent scheduling is called multiuser diversity gain [26, 27].

2.3.3 TCP over Wireless Links

TCP is a connection-oriented end-to-end data transfer protocol for the reliable end-to-end transmission of information over the Internet. It has mechanisms for error detection and correction and for congestion control. TCP protocol was designed for wired networks. Whenever congestion occurs in a router, packets are dropped and this causes the source to decrease the transmission rate by decreasing the size of the congestion window. The problem with TCP, when implemented over a wireless channel, is that it has no means to identify the cause for the dropping of packets. However, most drops are due to errors in the wireless radio channel. As a consequence, the TCP protocol interprets all drops as being related to congestion and it reacts to such drops by decreasing the packet transmission rate, which results in losses in the achievable throughput. New versions of the TCP protocol include the Explicit Congestion Notification (ECN) mechanism [28] used to notify the receiver whenever such congestions occur. This allows the TCP to differentiate drops due to congestion from ones due to errors and appropriate counter-measures may be taken [29]. This is an example where direct notifications from the physical layer can be used at the network layer to improve network-layer throughput performance.

It is important to mention that both the TCP and the MAC sub-layer present mechanisms for error detection and correction. Avoiding both layers to undertake the same actions for losses due to errors result in an improvement in the performance of the network. These can be conducted more efficiently and faster by the MAC sub-layer.

2.3.4 Multi-User Gain with Quality-of-Service

Future beyond third generation (B3G) and fourth generation (4G) wireless networks are expected to support a mixture of services with different requirements in terms of QoS requests. The voice over IP (VoIP) application, for example, is time-sensitive as it requests the timely transport of voice packets in order to not affect the quality of the speech as perceived by the end user. The web browsing application is more relaxed in terms of delay, but more demanding in terms of the achieved block error rate (BLER) in the transmission of packets over the wireless channel. A packet scheduler, located in the MAC sub-layer, must be aware of these service

constraints, to decide which packets to follow over the channel at each scheduling instant. This information must be directly provided from the application layer to the MAC sub-layer. Also, for the efficient use of the available set of radio resources, the scheduler must have a good estimate of the channel quality state for the whole set of radio channels, from every user in the cell. This information is provided by the physical layer to the MAC sub-layer. In each scheduling period the scheduler decides the set of packets to transmit and assign the set of resources which results in the maximization of the cell's throughput while it accomplishes the QoS requirements for the service. This application example encompasses information coming from both the application layers and from the physical layers to the MAC sub-layer [30].

2.3.5 Application's Adaptability to Changes in Physical and Network Layers

Application transmission adaptation refers to an application's capability to adjust its behavior to changing network and channel characteristics. Adaptive applications that are network and channel aware can cope with the adversities imposed by the physical layer and to congestions occurring in the network layer, by adapting the transmission parameters accordingly. For example: different source coding algorithms and different coding rates can be used depending on the state of the channel; unequal error protection can be also employed to distinguish high priority frames from less important ones. Depending on the degree of congestion in the network layer, the frame loss rate can change with time. A media server in the application layer can track packet losses and adjust the media source rate accordingly. To reduce information loss, the medium server can employ packet forward error correction (FEC) coding and unequal error protections. Some examples of proposals for application's layer adaptability according to the cross-layer design paradigm are given in [31-34].

2.4 A Model for the Coordination of the Cross-Layer Entities

In the reference OSI layered protocol architecture the functionalities of the protocols inside a layer are independent from the functionalities embedded within protocols from other layers of the same architecture model. Inside the protocol stack, exchange of control and data information may take place only between adjacent protocol layers through well defined interfaces designated by Service Access Points (SAP). A SAP provides access to a selected subset of the functionalities, inside the set of protocols designed for each layer, via a defined set of primitive operations. In this respect, any attempt to violate this OSI reference model and the way communication is exchanged among the different protocols in each layer of the stack is considered a cross-layer design.

In the last years numerous proposals for cross-layer design in the research literature have been made to improve the performance of wireless communication systems. Generally the design objective with cross-layer design for wireless networks is the maximization of the achieved

system throughput per service area (cell), by optimizing the access to radio resources, and the satisfaction of the QoS requirements, perceived by as many users as possible, from the set of users subscribing to the services provided by the network. The ability to support cross-layer interactions across the layers of the protocol stack for network operation is a fundamental characteristic of B3G wireless networks.

Cross-layer design allows communication to take place between non-adjacent layers of the protocol stack through additional entities introduced into the system's architecture. In [35] a model is proposed to specify the set of functionalities each entity must provide in the cross-layer architecture. This model encompasses four vertical planes extending across the main layers of the OSI reference model, namely: Physical/Link, Network, Transport and Application, as illustrated in figure 2.

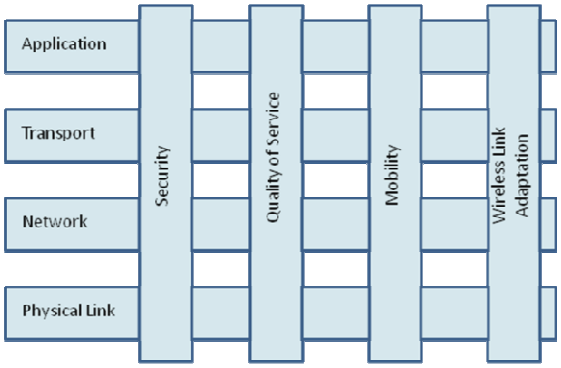


Figure 2 - Cross-layer coordination model from [35]

Each one of these planes should encapsulate the functionalities provided by each type of cross-layer design algorithm with a particular objective in mind. In wireless networks these objectives are: security, mobility, quality of service and adaptation of the wireless link.

The security plane: the security plane coordinates security technologies and encryption protocols across different layers.

The QoS plane: QoS solutions such as integrated services (IntServ) and differentiated services (DiffServ) have been proposed by the IETF to support QoS in the TCP/IP protocol stack. However, as these services have been designed with the fixed layered architecture of the OSI protocol reference model in mind, they are not appropriate for QoS provisioning over cross-layer architectures and, as a consequence, QoS requirement from application layer cannot be sent along the protocol stack to lower layers. The satisfaction of QoS in wireless environments calls for the joint processing of their requirements and the state of the wireless channel at well defined instants of time. The QoS plane must facilitate and coordinate the exchange of QoS information across multiple layers of the protocol stack.

The Mobility plane: the Mobility plane handles all cross-layer design functionalities designed for the mobility support, associated to the movement of mobile terminals across the coverage area of base stations, from the same or different access networks. When the handover occurs

among base stations of different access networks it is designated as vertical, while for handovers across base stations of different access networks it is termed as horizontal. The purpose of cross-layer design algorithms is here to make the transition as smooth as possible, resulting in connectivity to the IP core network, independent of the type of technology used. This caters for the adaptation of the services provided in the application layer to the underlying wireless technology.

The wireless link adaptation plane: Cross-layer design proposals are an efficient strategy in wireless networks because they make it possible the implementation of link adaptation algorithms in upper layers, according to the channel state. In general these algorithms adapt the rate of symbols transmission over the channel, by adapting the MCS scheme used according to the state of the channel, for the satisfaction of a given packet error rate (PER) design threshold. Power adaptation is also conducted to optimize interference patterns, according to desired levels of SINR and to save battery. These design objectives are encapsulated under the wireless link adaptation plane.

2.5 Cross-Layer Design Concepts

In the new architecture proposed for the cross-layer design each layer remains as a self contained entity regarding the execution of specific functions in the network. However, differently from the strict fixed layout, in the cross-layer design paradigm, the different layers in the stack are allowed to cooperate, by exchanging information regarding their functionalities and states. This information sharing affects the decisions and behavior of the different layers with time. A crucial point in the cross-layer design is the clear specification on the amount and type of information to be shared, which is implementation-specific. Also, for the exchange of information among layers, new interfaces must be created, and new modules specifically designed and included in the protocol stack, to manage and coordinate the steps followed in the exchange and sharing of information. This section provides some hints about the different possibilities for information sharing and on some proposals in the literature for the coordination modules.

2.5.1 Types of Information Flow across Layers

According to [16], the information can be exchanged across layers in one of three different ways:

- **Upward – (from lower layers to a higher layer):** a higher-layer protocol, which is based on the lower layers functioning require direct interfaces for the exchange of the information provided by these lower layers. For example, an interface between the TCP and lower layers can be used in the explicit congestion notifications (ECN), from the router to the TCP layer at the sender, to differentiate packet losses due to congestion from

packet losses due to errors in the wireless channel. Opportunistic scheduling algorithms based on the channel state are also examples of this type of information flow [36].

- **Downward (from higher layers to a lower layer)**: some cross layer design proposals rely on setting parameters on the lower layers of the stack using a direct interface from some higher layer. For example: applications can inform the link layer about their delay requirements in order for the link-layer to prioritize packets from delay-sensitive applications in detriment of packets from non-delay sensitive applications [37].
- **Back and forth (iterative flow between two layers)**: two layers performing different tasks can collaborate with each other at runtime, with information flowing back and forth between them. For example, in [38] the authors present an architecture where both PHY and MAC layers cooperate in the avoidance of collision resolutions in the uplink of a wireless local area network (WLAN) system.

2.5.2 Types of Cross-Layer Management Entities

Another important cross-layer design principle is how to manage cross-layer interactions among layers in such a way that system fluctuations can be smoothed and instability, arising from continuous jumps across transient states, avoided. To this sense, an interlayer manager must be implemented for the coordination of the cross-layer processes occurring in the network. In [39] different types of layouts for the coordination manager are considered:

- The manager reside inside the protocol stack: in this case the manager is an internal entity to the protocol stack and it may be either an inter-layer entity that coordinates the operation of all protocol layers or a set of intra-layer entities, each of which is collocated with a protocol layer.
- The manager is an external entity to the protocol stack: in this case the manger may be centralized and hosted by a specific network element or it may be distributed across several network nodes.

2.6 Disadvantages of the Cross-Layer Design Reference Model

There can be some disadvantages or dangers associated to a network architecture based in the cross-layer design framework. In [18] the authors elaborate on some aspects of the cross-layer design paradigm which should be carefully considered in order to avoid some drawbacks to this model. Basically, some side effects which could result from a protocol stack reference model, in which the different layers interact and influence the functional behavior of each other, are presented. Authors base their reasoning on some sound examples, where the use of the static protocol layering model was fundamental and they argue that some caution should be exercised while engaging into cross-layer design approaches:

1. A modular architectural design has proven itself time and time over. Modularity provides the very essence of abstraction needed for researches and engineers to fully understand the overall system. The layered protocol model is used as an example in which the Internet, and its tremendous success, is loosely based on. A modular design can also accelerate the development since different designers can focus their efforts on different subsystems with the assurance that the entire system will interoperate once it is brought together.
2. By nature cross-layer design creates interactions among processes and layers. Some interactions are intentional while others might be unintentionally created. Loops can be created and instability can arise.
3. Standardization allows subsystems to be used across many applications, thus resulting in lower development costs and time-to-market which in turn increase usage. In contrast, a cross-layer design system will need to be adapted to every application and this will increase cost and time-to-market greatly.
4. Once the layering is broken, the possibility to review and redesign parts of the system is lost since everything is interconnected one way or the other. Protocols have to be redesigned in a cross-layer fashion, by taking into account several layers as opposed to earlier strict modular lay-out, where a protocol could be developed in isolation.
5. A system-wide cross-layer design could lead to “spaghetti” implementation, which in turn hamper further innovation and be difficult to maintain. Future design improvements may become impossible, because it will be difficult to foresee exactly how a new modification will affect the overall system operation.

The main point raised in [18] is that while cross-layer design gives a short term performance gain, good architectures are usually based on longer term consideration. While this claim is impossible to contradict or prove wrong or right, since the cross-layer design concept is rather new, it might be wise to use it with caution in the evolutionary approach and perhaps avoid the more system-wide cross-layer design. It seems, however, that the examples presented above are applied mainly to very complex cross-layer design architectures. Nevertheless they bring out interesting aspects that must be taken into consideration when designing a system based in the cross-layer design principle.

Some aspects associated to the cross-layer design can be pointed:

- The increase in the signaling load in the whole system needed to convey the interactions among layers.
- The optimization of the interactions among layers.

2.7 Related Work

In the last few years there has been much interest in the research for cross-layer design. This growing interest is mainly coupled with the proposals for the new wireless architectures envisioned for both 3G and B3G scenarios. Cellular, ad-hoc, mesh and sensor networks benefit from the potentialities accrued from the implementation of a cross-layer design protocol framework. In this section some review of the state-of-art for cross-layer design is presented. Basically, the work conducted can be divided into two branches:

- Analytical work, envisioning the optimization of the cross-layer design protocol stack, with or without the use of utility functions, and using convex optimization tools. The scenarios for this analysis are rather simplistic: single cell, downlink connection, simplistic traffic models (full queue for example) and simple analytical channel models. Invariably the objective is the maximization of some utility metric as a function of the data rate, under some constraints, such as the maximum power allowed in the base station for data transmission.
- Experimental work conducted under simple simulation scenarios, or more rather complex scenarios with real channel and traffic models and for a realistic network layout. Invariably there are many proposals for implementing wireless standards such as High Speed Downlink Packet Access (HSDPA), Worldwide Interoperability for Microwave Access (WiMAX), Wireless Fidelity (WiFi), Long Term Evolution (LTE) and ad-hoc, as well as mesh networks.

In [40] a cross-layer design framework for the Mobile WiMAX network is proposed. Higher system performance is achieved by exploiting cross-layer interactions between the MAC and PHY layers. The proposed scheduler and resource allocator, residing in the MAC layer, determine the number of packets which can be transmitted to each user and allocates the available frequency bands by using a channel-aware subcarrier allocation algorithm, respectively. Information regarding channel quality is passed to the MAC layer, which, with the help of an ARQ protocol decides the MCS scheme to use in the transmission.

A scheduling strategy is implemented in [41] for Wideband Code Division Multiple Access (WCDMA) networks. The architecture enables the exploitation of cross-layer information to improve system performance in terms of capacity and delay, by assigning priorities to users based on short-term channel variations. The scheduling scheme extends the concept of opportunistic scheduling: the priority function considers not only the channel state (as done in opportunistic scheduling) but also the past and predicted evolution in the channel fluctuations for the near future, experienced by each user, in the computation of the priority attributed to the user. According to simulation results, achievable gains as up to 30% in capacity and 35% percent reduction in average channel access times are claimed.

Many cross-layer design solutions were proposed in order to improve TCP's performance over wireless links. As mentioned before, the inability of the TCP protocol to correctly identify the cause of packet losses results in degradation of the performance achieved over wireless links. One possible way to deal with such problems arising from this misunderstanding is the use of Explicit Loss Notification (ELN) schemes for indication of transmission errors over the wireless medium, as done in [42].

Adaptive Modulation and Coding (AMC) is a widely known technique to match the transmission rate to time-varying channel conditions. It can realize several benefits for TCP's performance over wireless links. For example, in [43] it is advocated that a cross layer design approach that effectively conjugates AMC with TCP can maximize TCP throughput, while sustaining a prescribed PER. By selecting the channel-dependent parameters such as the average of the received signal-to-noise ratio, the mobility-induced Doppler spread, the fading parameter and the number of packets the link layer' queue can serve, the TCP throughput is improved for a prescribed PER.

Some proposals for cross-layer design, envision the increase in the capacity of served users. For example, in [44] a proposal is done for a cross-layer design between physical and data link layers in order to achieve high spectral efficiency. The proposal encompasses an AMC scheme and ARQ mechanism at the data link layer. An optimal design for the AMC scheme at the PHY layer is based on the restriction on the maximum number of retransmissions allowed per packet and the probability of packet loss after this number is achieved. Simulations indicate that a small number of retransmissions in conjunction with the chosen modulation-coding pair can improve spectral efficiency in term of bits per transmitted symbol.

In much the same concept [44] combines AMC with an HARQ scheme against the truncated ARQ scheme used in [43]. But the scheme proposed uses an HARQ type I mechanism for packet retransmission, aiming for maximum optimization under a prescribed delay constraint. The proposal is based on a fixed packet size and adjusts the size of the packet, together with the regions for link adaptation to improve spectral efficiency.

In [45] another cross-layer design proposal based on AMC and a scheduler in the MAC layer is presented. Under a TDM/TDMA multiple access scheme the cross-layer encompasses information regarding the queue length at the link layer and the channel state at the physical layer to decide the number of time-slots that can be actually scheduled and assigned to each user.

Some proposals in the literature are also performed for the adaptation and optimization at the application layer in the mobile. In [46] an unequal error protection scheme is implemented through a cross-layer approach that protects important information from impairments caused by channel errors. In [47] cross-layer is accomplished by exchanging information between the source and channel codec in the application and physical layers respectively. In [48] an

evolutionary approach to the joint source and channel coding is presented, that conveys joint control between the source coding and power control in terms of the source rate at the video codec and the average signal-to-noise ratio at the physical layer. It attempts to achieve an end user's QoS-level by adjusting the combination of both parameters.

In [1] a survey of cross-layer scheduling algorithms for Multiple-Input Multiple Outputs (MIMO) systems is conducted, under an information theoretic framework.

2.8 Conclusion

Cross layer revealed itself as a new strong design paradigm to be considered in the implementation of the communication architectures of new wireless networks of 3G and B3G generations. Cross layer design is basically a deviation from the strict layering modular architecture proposed by ISO. In cross layer design different layers in the protocol stack can exchange information through well defined interfaces and influence the behavior of other layers in the stack, depending on the their functional state at a given point in time. This layer state exchange is particularly important in the scenario of wireless networks, which are characterized by the transfer of information through such a very aggressive medium as the mobile radio channel.

In this chapter the importance of cross layer design paradigm for wireless networks of 3G and B3G was enforced, by pointing out the benefits acquired from the implementation of this protocol architecture on top of the radio transmission channel. Details regarding principles and motivations behind the cross layer design paradigm were provided. A model specifying the set of functionalities which can be provided by each entity in any type of cross layer architecture is given. This model comprises four planes, extending vertically across the main layers of the OSI reference model: Physical/Link, Network, Transport and Application.

If not properly analyzed, designed and implemented, the cross-layer design can result in network malfunctioning and/or instability. As such, some caution must be followed while implementing cross layer-based wireless network protocol stacks.

A revision of some proposals encompassing the design and implementation of cross layer architectures in the literature, and which reflect the many variants for the application of the cross layer principle was also conducted.

Mobile WiMAX standard is the subject of the next chapter. Its functionality perfectly suits the cross layer design-based network architecture, as it is designed with specific control channels and signaling messages in mind, which can be used for the exchange of information regarding the state and functioning of the different layers in the protocol stack. Cross-layer design turns out to be a strong argument in the proposal of WiMAX as a strong potential candidate for the 4G evolution. A WIMAX network, designed according to the principles of cross-layer design, results in the maximization of resource usage efficiency, fairness in resource allocation and

satisfaction of the QoS parameters associated to the type of multimedia applications envisioned for such scenario of 4G multimedia wireless networks.

Chapter 3

Mobile WiMAX

3.1 Introduction

The last decade has witnessed the explosive growth in the penetration rate of mobile cellular wireless networks of second generation (2G), from which Global System for Mobile Communications (GSM) and Interim Standard 95 (IS-95) are particular examples. In the mean time, widespread deployment of wired Internet is a fact now and has ignited the development of new types of applications characterized by stringent QoS requirements, such as: bandwidth, latency and delay jitter. The convergence of both fixed Internet and wireless networks was then envisioned as the next step to follow, and as a consequence, in the last few years, the implementation of third generation cellular networks (3G), such as High Speed Downlink Packet (HSDPA) or Code Division Multiple Access (CDMA 2000) as well as ad-hoc networks, such as Wireless Fidelity (WiFi), corroborated the vision shared by researches, system designers and manufacturers, of providing those applications over the air and on the move, while at the same time guaranteeing the same set of QoS requirements posed to fixed Internet networks. Another key factor is the need for cost-efficient solutions, which could ease the implementation of a wireless infrastructure to replace costly solutions, such as those available with cable or

Digital Subscriber Line (DSL) operators, and to provide fast wireless Internet in undeveloped countries.

IEEE 802.16e, also known as Mobile WiMAX [49] (Worldwide Interoperability for Microwave Access) is a broadband wireless solution that enables the convergence of mobile and fixed broadband networks, through a common wide area broadband radio access technology (Orthogonal Frequency Division Multiple Access - OFDMA) and flexible network architecture. Mobile WiMAX is an extension to the previous IEEE 802.16d standard [50], also known as Fixed WiMAX, which is rapidly proving itself as a technology that will play a key role in the next generation of broadband wireless networks.

Fixed as well as Mobile WiMAX deployments were motivated by the success of WiFi, from which many vendors and operators joined efforts and created the WiMAX Forum [51] and the IEEE 802.16 Working Group [52] to develop a new end-to-end solution to address the new demands and opportunities. The IEEE 802.16 WG was established by the IEEE Standards Board in 1999. Since then, this group has developed and published several versions of air-interface standards for Wireless Metropolitan Area Networks (WMANs), with a focus on the Medium Access Control (MAC) and Physical (PHY) layers. The standards produced by the IEEE 802.16 group are adopted by both the IEEE and ETSI HIPERMAN group.

The Mobile WiMAX standard is a highly promising technology for implementation in wireless cellular networks of fourth-generation (4G). Because cross-layer optimization is the most important concept for next-generation wireless communication systems, Mobile WiMAX protocol architecture supports cross-layer operation. In fact, Mobile WiMAX has inherent mechanisms specifically designed for the vertical coupling and interaction between layers. These mechanisms consist of control channels and signalling messages used in the exchange of information among the PHY, MAC and upper layers in the protocol stack, resulting in an effective cross-layer based architecture design.

This chapter describes the underlying technology for WiMAX standard with a focus on the standardization activities which are being conducted in the IEEE 80216 WG. A detailed description of the physical and medium access control layers functionalities is provided, with a particular emphasis on the mechanisms implemented for the provision of QoS in the support of multimedia applications. Cross-layer schemes for capacity increase and efficient QoS support are presented as examples of the potentialities which can be achieved by using the cross-layer design mechanism available in the Mobile WiMAX standard.

This chapter is organized as follows. Section 2 presents the evolution of the WiMAX standard which resulted in the IEEE 802.16e version. Sections 3 and 4 describe in detail the PHY and MAC layers of IEEE Mobile WiMAX standard, respectively. The advanced features which were included in the new version of the standard are presented in section 5. The WiMAX standard comprises different functionalities which could result in many types of

implementations, for each type of scenario. Section 6 is about the implementation of the cross-layer design in the WiMAX protocol stack. The control and management messages available in the standard for exchange and interaction between MAC, PHY and upper layers are presented in detail. Section 7 is a review of the state-of-the-art regarding the implementation of cross-layer design architectures, with the Mobile WiMAX standard as a case study. Section 8 concludes the chapter.

3.2 WiMAX Evolution

One of the biggest potentials of Mobile WiMAX is that it offers a scalability architecture in both radio access technology and network architecture which provides a great deal of flexibility in network deployment options. According to the predicted data rates and level of quality of service (QoS) to be expected from multimedia applications, and also to the throughput figures achieved under 3G cellular networks, it was realized since the beginning that the new standard should be built around the following key technical aspects [53-54]:

- **Adequate Multiple Access Technology** – Mobile WiMAX supports OFDMA as a multiple access technology, whereby different users can be allocated different subsets of OFDM sub-carriers. OFDMA results in the exploitation of frequency diversity, besides the multiuser diversity achieved with opportunistic scheduling, under the normal OFDM air-interface in Fixed WiMAX. OFDM technology offers good resistance against multipath which suits WiMAX to operate in Non-Line-of-Sight (NLOS) conditions.
- **Scalability for different Scenarios of Application** – Mobile WiMAX has a scalable PHY layer architecture that allows for data rates to scale with available channel bandwidth ranging from 1.25 to 20 MHz [12]. This scalability is supported in the OFDMA mode, where the size of the Fast Fourier Transform (FFT) symbol may be changed according to the spectrum availability, while the inter-carrier frequency separation is kept constant.
- **Availability of High Data Rates -** Mobile WiMAX technology implements fast link adaptation schemes and scheduling over resources in time and frequency. With link adaptation a number of modulations and forward error correction (FEC) coding schemes can be changed dynamically, according to the state of the channel reported on a frame basis. The scheduler, located in the base station, multiplexes users in either frequency and/or time domains. The placement of the scheduler in the base station results in a faster resource allocation process.
- **Advanced antenna techniques** – WiMAX supports the use of multiple antenna techniques to increase overall system capacity and spectrum efficiency, such as space-time coding, spatial multiplexing and beamforming [55].

- **Provision of Quality of Service (QoS) demands** – One of the main achievements of the standard is the provision of QoS mechanisms designed to support a variety of applications including voice and multimedia services. The MAC layer has a connection-oriented architecture, capable of supporting multiple connections per each user terminal and is responsible for the assignment of specific service flow QoS parameters, according to the type of application. Bandwidth request mechanisms were developed for the provision of QoS under WiMAX standard.
- **Advanced Error Protection Mechanisms** – Mobile WiMAX supports a hybrid mechanism composed of automatic retransmission requests (ARQ) and Forward Error Correction (FEC) coding for connections requiring transmission feasibility (Hybrid Automatic Repeat Request - HARQ). Two types of ARQ mechanisms are considered in the standard: Chase Combining (CC) and Increment Redundancy (IC).
- **Robust Security** – Mobile WiMAX supports strong encryption, using Advanced Encryption Standard (AES) and has a robust privacy and key-management protocol. It also offers flexible authentication architecture based on Extensible Authentication Protocol (EAP).
- **Mobility** – Mobile WiMAX has inherent features for the support of seamless handovers of real-time applications with very small latencies.
- **IP-based architecture** – The reference network is based on an all-IP platform which relies on IP-based protocols for end-to-end transport, QoS, session management, security and mobility.

The IEEE 802.16 WG initial focus was the development of a line of sight (LOS)-based point to multipoint (PMT), fixed wireless broadband system for operation in the 10 GHz - 66 GHz milimeter wave band. This standard, designated as IEEE 802.16 was completed in 2001. It was based on a single-carrier PHY layer with a time division multiplexing (TDM) MAC layer and many concepts were adapted from the popular cable modem DOCSIS (Data Over Cable Service Interface Specification) standard. In order to include non-line-of-sight (NLOS) applications in the 2 GHz - 11 GHz band, an amendment to the original standard was produced in 2003 and was designated as IEEE802.16a. In this new version the PHY layer is based on the OFDM technology to favour deployment in urban areas. Some additions to the MAC layer were also included such as the support for multiple OFDMA access.

One of the problems in the earlier draft of IEEE 802.16 is that it covers too many profiles and PHY layers, which can lead to potential interoperability problems [22]. From the initial variants, the IEEE 802.16 standard has undergone several amendments and formed the basis for the first WiMAX solution which was designated as IEEE 802.16d (also known as IEEE 802.16-2004), whose specification was completed on June 2004. IEEE 802.16d is also commonly referred as fixed WiMAX [50]. The standard provides technical specifications for the PHY and MAC

layers for fixed wireless access and addresses the first or last mile connection in wireless metropolitan area networks (WMANs). In December 2005, the IEEE group completed and approved IEEE 802.16e, an amendment to the IEEE 802.16d standard that added mobility support. The IEEE 802.16e (also known as Mobile WiMAX) forms the basis for the WiMAX solution for nomadic and mobile applications [49].

It is important to mention that both Fixed and Mobile WiMAX standards offer a variety of fundamentally different design options. There are multiple PHY layer choices: a single carrier based PHY layer named WirelessMAN-SCa, an OFDM-based PHY layer named WirelessMAN-OFDM, and an OFDMA-based PHY layer named WirelessMAN-OFDMA. Similarly, there are multiple choices for the MAC architecture, duplexing, frequency band of operation, etc. This is because these standards were designed to suit a plethora of applications and scenarios. The basic characteristics of both Fixed and Mobile WiMAX standards are summarized in table 1 from [13].

	802.16	**802.16-2004**	**802.16e-2005**
Status	Completed December 2001	Completed June2004	Completed December 2005
Frequency Band	10GHz-66GHz	2GHz-11GHz	2GHz-11GHz for fixed 2GHz-6GHz for mobile
Application	Fixed LOS	Fixed NLOS	Fixed and Mobile NLOS
MAC architecture	Point-to-multipoint; Mesh	Point-to-multipoint; Mesh	Point-to-multipoint; Mesh
Transmission scheme	Single carrier only	Single carrier, 256 OFDM or 2 048 OFDM	Single carrier, 256 OFDM or scalable OFDM with 128, 512, 1024 or 2048 sub-carriers
Modulation	QPSK, 16QAM, 64QAM	QPSK, 16QAM, 64QAM	QPSK, 16QAM, 64QAM
Data Rate	32Mbps-134.4Mbps	1Mbps-75Mbps	1Mbps-75Mbps
Multiplexing	TDM/TDMA	TDM/TDMA/OFDMA	TDM/TDMA/OFDMA
Duplexing	TDD and FDD	TDD and FDD	TDD and FDD
Channel Bandwidths	20MHz, 25MHz, 28MHz	1.75MHz, 3.5MHz, 7MHz, 14MHz, 1.25MHz, 5MHz, 10MHz, 15MHz, 8.75MHz	1.75MHz, 3.5MHz, 7MHz, 14MHz, 1.25MHz, 5MHz, 10MHz, 15MHz, 8.75MHz
Air-interface	WirelessMAN-SC	WirelessMAN-SCa WirelessMAN-OFDM WirelessMAN-OFDMA WirelessMAN-HUMAN	WirelessMAN-SCa WirelessMAN-OFDM WirelessMAN-OFDMA WirelessMAN-HUMAN
WiMAX implementation	None	256-OFDM as Fixed WiMAX	Scalable OFDMA as Mobile WiMAX

TABLE 1 BASIC DATA ON IEEE 802.16 STANDARDS

All Mobile WiMAX certification profiles use scalable OFDMA as the PHY layer [12]. At least initially, all mobility profiles will use a PMP MAC and be TDD-based. Due to the many different implementation possibilities it is important to define which ones should be mandatory and which ones should be optional, in the early drafts of proposals for equipment implementation and for interoperability and certification testing, among proposals from different equipment and system solutions suppliers. These are performed under the auspices of the WiMAX Forum [51], which is an industry consortium for the promotion of the WiMAX standard. It provides certification and interoperability testing for product manufacturers and

participates in the development of new versions of the standard. For interoperability testing the WiMAX Forum defines a limited number of system and certification profiles. A system profile defines the subset of mandatory and optional PHY and MAC layers features of the WiMAX standards.

3.3 WiMAX Physical Layer (PHY)

This section details the implementation of the OFDMA-based PHY layer in Mobile WiMAX networks.

3.3.1 Orthogonal Frequency Division Multiplexing (OFDMA) Basics

The WiMAX PHY layer is based on the OFDM technology, which is an efficient and low-cost scheme for high data rate transmission in a NLOS or multipath radio environment. OFDM is a spectrally efficient version of multicarrier modulation (MCM) where the sub-carriers are selected such that they are all orthogonal to one another over the symbol duration, thereby avoiding the need to have non-overlapping subcarrier channels to eliminate inter-carrier interference (ICI). OFDM modulators/demodulators are implemented in discrete time by the use of Inverse Fast Fourier Transform (IFFT) and Fast Fourier Transform (FFT) chips, respectively. OFDM is very efficient in eliminating signal distortion due to delay spread arising from multipath propagation because the stream of data is split among the orthogonal sub-carriers, resulting in an increase of the symbol time interval, which makes it more immune to Inter-Symbol-Interference (ISI). Also, ISI can be completely eliminated with the insertion of guard intervals between OFDM symbols, larger than the expected multipath delay spread [54]. This guard interval is called Cyclic Prefix (CP) in OFDM. As long as the CP is longer than the channel delay spread, ISI is completely eliminated. The CP is a repetition of the last samples of the OFDM symbol that is appended to the beginning of the data payload. This mechanism makes the channel circular and enables the use of simple Maximum Ratio Combiners (MRC) as decoders in the receiver, instead of complex multi-user decoders such as the ones used in CDMA for example. Figure 1 illustrates the creation of the CP.

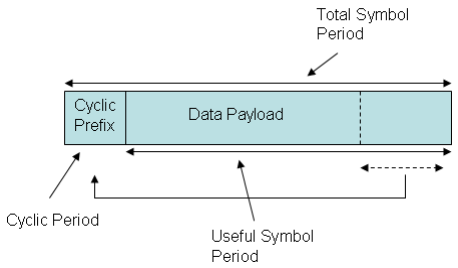


Figure 1 - Data Symbol Structure and creation of cyclic prefix

The OFDMA symbol consists of three types of sub-carriers:

· Data sub-carriers for data transmission.

· Pilot sub-carriers for estimation and synchronization purposes.

· Null sub-carriers for no transmission. These are used for guard bands and DC sub-carrier.

Active (data and pilot) sub-carriers are grouped into subsets of sub-carriers called sub-channels. Figure 2 illustrates the OFDMA sub-carrier structure implemented in Mobile WiMAX standard [54].

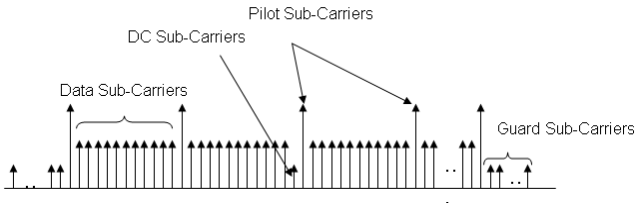


Figure 2 - OFDMA sub-carrier structure from [54]

3.3.2 OFDMA Sub-Channelization in WiMAX

The WiMAX OFDMA-based multiple access supports sub-channelization in both downlink and uplink. The minimum time-frequency unit of sub-channelization is one slot, which comprises 48 data sub-carriers.

Fixed WiMAX PHY layer is based on OFDM technology and it allows a limited form of sub-channelization in the uplink direction: the multiple access scheme is OFDM/TDMA for downlink and OFDMA for uplink. In the downlink only one mobile can transmit over all sub-carriers for the time interval corresponding to one OFDM symbol. In the uplink 16 sub-channels are defined for allocation: 1, 2, 4, 8 or all 16 sub-channels can be assigned to a subscriber station (SS).

Mobile WiMAX physical layer is based on the OFDMA multiple access technology. It allows the allocation of small sets of sub-carriers to different users in downlink and uplink directions. Sub-channels in Mobile WiMAX are comprised of sub-carriers which may be allocated contiguously or distributed according to a pseudo-random pattern, depending on the cell index, over the spectrum.

3.3.2.1 Diversity Permutation Sub-Carrier Sub-Channelization

The distributed sub-carrier permutation mode is a very useful scheme for averaging out inter-cell interference and avoiding deep fading, by allocating in each sub-channel sub-carriers in a pseudo-random way, provided different random sub-carrier distribution patterns are defined and assuming transmission is synchronized among neighbouring cells [49]. This mode is efficient in achieving frequency diversity, which makes it effective for scenarios in which it is difficult to track frequency selective channel variations, in order to allocate resources and/or select transmission parameters adaptively according to these variations. Therefore, it is expected to be of particular interest for users with high velocity and/or low signal-to-interference-plus-noise

ratio (SINR). In this mode basic resource units in frequency domain are called diversity sub-channels and result in one degree of freedom for resource allocation (time domain).

In WiMAX there are several sub-channelization schemes based on the pseudo-random distribution of sub-carriers for both uplink and downlink. Two of the most important ones are Partial Usage of Sub-carriers (PUSC), which is mandatory for all WiMAX implementations, and Full Usage of Sub-carriers (FUSC). PUSC enables the implementation of segmentation in the MAC layer. A segment is an individual instance of the MAC layer. Sub-carriers belonging to sub-channels from different segments do not collide, even if the same pattern is used in adjacent cells. This has to do with the way in which random allocation is performed for data and pilot sub-carriers. In FUSC mode all sub-carriers in each sub-channel are spread over the whole FFT spectrum, which forbids the implementation of segmentation for this channelization mode. For further details please refer to [49, 54].

3.3.2.2 Band Adjacent Multi-Carrier (AMC) Sub-Channelization

In adjacent sub-carrier permutation mode, adjacent sub-carriers are grouped into clusters of contiguous sub-carriers. In this mode the channel response can be seen as a flat fading channel. Thus, the frequency selectivity of the channel cannot be exploited, but the system can make better use of multiuser diversity over frequency domain, as long as the channel state does not change significantly during the scheduling period. This mode is particular interesting for scenarios with high SINR and/or with low mobile speeds, because it is more sensitive to inter-cell interference and to errors in the estimation of the channel quality [54].

The band AMC sub-channelization results in two degrees of freedom for resource allocation, as resources are available in both frequency and time domains. Band AMC allows system designers to exploit multiuser diversity, allocating sub-channels to users based on their frequency response over each sub-channel, assuming the state of each sub-channel is independent among different users (i.e., channel states are not correlated). Multiuser diversity over frequency domain provide significant gains in overall system capacity, provided the system assigns to each user a sub-channel that maximizes its received SINR.

3.3.3 Frame Structure of Mobile WiMAX

Mobile WiMAX systems can support time-division duplex (TDD) or frequency-division duplex (FDD) modes. For both FDD and TDD duplex modes the frame structure is the same, except that both uplink and downlink sub-frames are transmitted simultaneously over different frequency bands for FDD. In the first profile released for Mobile WiMAX the duplex mode to be used is TDD. In TDD mode the frame is subdivided in two sub-frames separated by one guard interval and the downlink-to-uplink-sub-frame ratio may be varied to support different traffic profiles. The frame is composed of several zones that are divided according to sub-carrier

allocation methods or Multiple-Input Multiple-Output (MIMO) modes. Figure 3 illustrates the structure of the TDD frame in Mobile WiMAX standard.

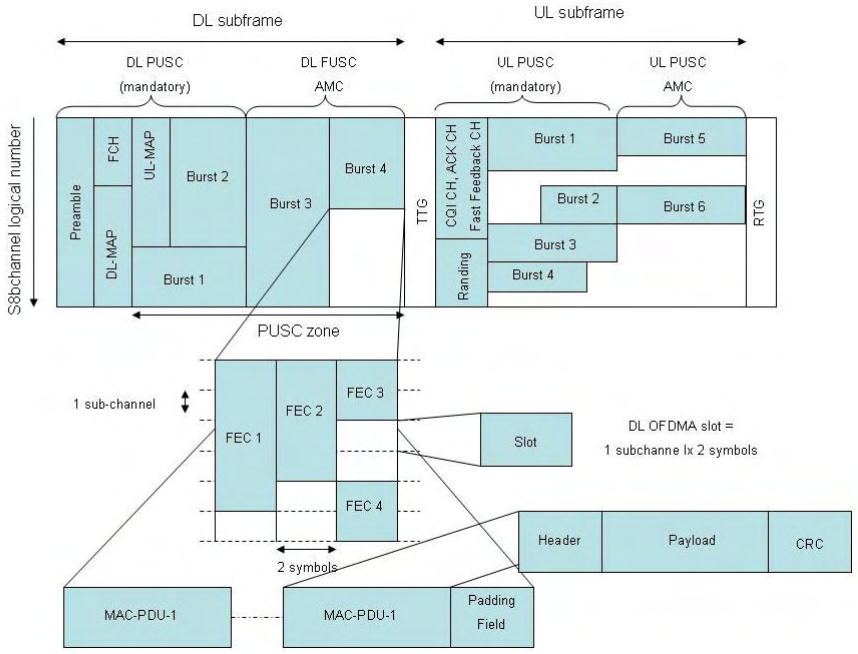


Figure 3 - Frame structure for Mobile WiMAX using TDD duplex mode

The following fields are implemented:

Downlink Channels

- **Preamble** – This is the first OFDM symbol in the frame and is used for time and frequency synchronization as well as for channel estimation.
- **Frame Control Header (FCH)** – This field provides frame configuration information such as the Mobile Application Part (MAP) message length, the modulation and coding scheme (MCS) used in the DL-MAP and the used sub-carriers.
- **DL-MAP and UL-MAP** – These fields convey the MAP messages (named as Information Elements – IEs) which indicate the starting times of bursts. IE messages are used in the indication of the slots which are assigned to each burst in the data region. These fields are broadcast after the FCH in the downlink sub-frame. Since both MAP fields contain critical information that needs to reach all users, they are often transmitted with the most robust modulation and coding scheme (BPSK with rate 1/2 coding and repetition coding). MAP messages could form a significant overhead in the frame, resulting in less efficiency for data allocation, particularly when there are a large number of users with small packets. To mitigate overhead, WiMAX can optionally use multiple sub-map messages where the dedicated control messages to different users are transmitted at higher rates, based on their individual SINR conditions. Each MAP message is

composed of Information Elements (IE) containing information regarding each burst allocation in the TDD frame. Namely each IE contain:

- The description of the burst profile (modulation and coding combination for each burst).
- An optional field with the list of Connection Identifiers (CIDs) with packets mapped on the downlink burst to which the DL-MAP is referred to, including the number of connection identifiers in the list.
- The Burst Allocation information field:
  - OFDMA symbol offset
  - Sub-channel offset
  - Number of sub-channels
  - Number of OFDMA symbols
  - Use of power boosting (+6dB to -9 dB)
  - Indication on the use of Repetition Coding (1/2/4/6)

· **Data Bursts** – The minimum time-frequency resource which can be allocated for a given user is the slot. Each slot consists of one sub-channel over one, two or three OFDM symbols, depending on the sub-channelization scheme used. A contiguous sequence of slots assigned to a given user constitutes a burst. All slots in the same burst must transmit with the same MCS scheme. Bursts are allocated depending on user's traffic demand, QoS requirements and channel conditions.

Uplink Control Channels

Mobile WiMAX provides a number of control channels which can be used to exchange cross-layer information such as channel quality information and Acknowledge/Negative ACK (ACK/NACK) feedback for HARQ.

· **Ranging** – This is a region in the uplink sub-frame for contention-based access, which is used for a variety of purposes: it can be used to perform closed-loop frequency, time and power adjustments during network entry as well as periodically. It can also be used by the mobile station to make uplink bandwidth requests in contention mode.

· **Channel Quality Indication Channel (CQICH)** – This channel is used by the mobile station to feedback channel quality information that can be used in the base station scheduler for link adaptation purposes. This control channel is used to report the downlink SINR for either diversity sub-channels or band AMC sub-channels. This channel occupies one uplink slot in the fast-feedback region in the uplink sub-frame. For diversity sub-channels, the mobile terminal reports the average SINR of the preamble broadcast in the downlink sub-frame, from which the base station is able to determine the downlink MCS scheme level. For band AMC sub-channels a mobile terminal can report the differential of

SINR values of five selected frequency bands after reporting the SINR measurements of the five best bands.

- **Uplink ACK Channel** – This is a region allocated in the uplink sub-frame and is designed for the inclusion of one or more ACK channels for enabling HARQ in data transmission. Each uplink ACK channel occupies one half-slot in the HARQ ACK region and is implicitly assigned to each HARQ-enabled burst, according to the order of the HARQ-enabled downlink bursts in the DL-MAP. Thus the mobile terminal can quickly transmit ACK or NACK feedback for downlink HARQ-enabled packet data using this uplink ACK channel.
- **Uplink Sounding** – Mobile WiMAX defines an uplink sounding zone in the uplink sub-frame for the definition of uplink sounding symbols, which are used in the support of MIMO and smart antenna beamforming. The base station measures the uplink channel response from uplink sounding waveforms transmitted by each mobile station and translates the measured uplink channel response to an estimated downlink channel response, under the assumption of channel reciprocity for the TDD duplex mode.

Support of Advanced Antenna Technologies

Mobile WiMAX supports various multiple antenna technologies which are applied in different zones within a frame. For example, Adaptive Antenna Systems (AAS) is a kind of smart antenna processing which can be used for Space Division Multiple Access (SDMA), and allows the transmission of data bursts concurrently to spatially-separated mobile stations, by applying different beam patterns to them. Figure 4 shows a logical frame structure for AAS support.

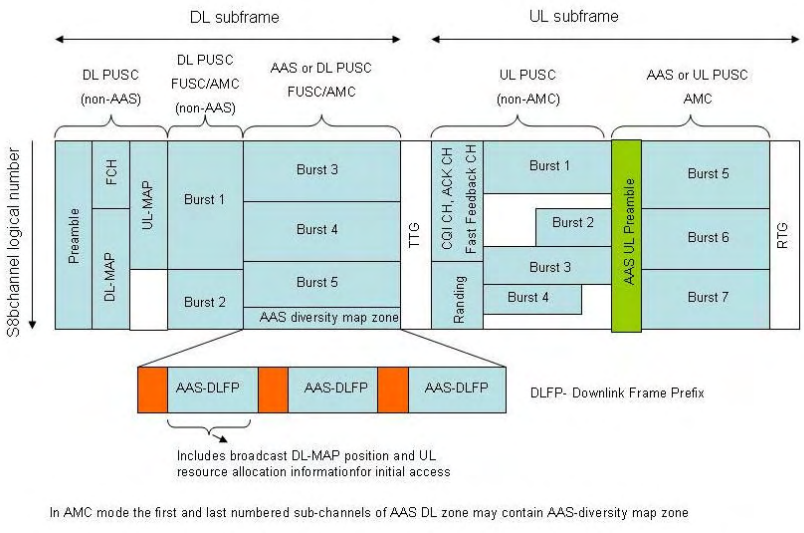


Figure 4 - Frame structure for Mobile WiMAX using TDD duplex mode for AAS support

Downlink and uplink AAS zones are defined by a special broadcast MAP message and, as can be seen from the figure, the downlink AAS zone includes an AAS diversity map zone which occupies two sub-channels. AAS-Downlink Frame Prefixes (AAS-DLPFs) in the DL AAS zone

are transmitted using different beams from each other. Each mobile station, in AAS mode, scans these AAS-DLPFs using known AAS downlink preamble patterns which were previously reported and they choose the one with the best beam. Each AAS-DLFP includes the position of the broadcast DL-MAP which is beamformed. It can also be used to page a specific mobile that cannot receive the normal DL-MAP or be used for resource allocation information for uplink initial access. Once the mobile obtains the information regarding initial resource allocation through a broadcast DL-MAP pointed to by the AAS-DLFP, subsequent allocations can be managed with private DL-MAP and UL-MAP messages that are unicast and beam formed with high MCS levels.

3.4 WiMAX Medium Access Control Layer (MAC)

Mobile WiMAX standard was designed and developed from the outset for the delivery of broadband applications. Its MAC layer, in particular, has inherent features designed for the joint support of those burst data traffic applications with high peak rate demands and streaming and/or delay sensitive ones. The fine granularity and flexibility provided by the MAC layer in resource allocation, according to user's bandwidth needs, and the lower latency incurred in handling user's bandwidth requests and in making scheduling decisions, makes it possible to send data through the air-interface under the stringent QoS requirements of each type of service flow, and the efficient use of radio resources, with the consequent maximization of the achieved spectrum efficiency. Figure 5 illustrates the MAC layer architecture proposed for Mobile WiMAX.

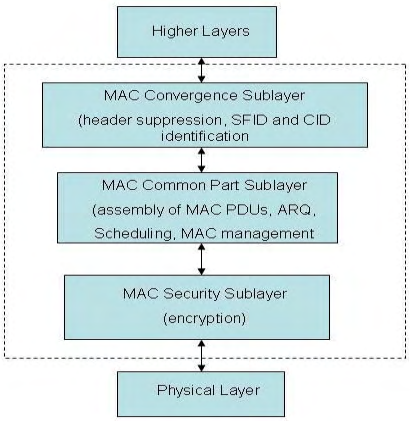


Figure 5 - Frame structure for Mobile WiMAX using TDD duplex mode for AAS support

3.4.1 MAC Layer Functional Blocks

Basically the MAC layer is sub-divided into: Convergence Sub-layer (CS) and Common Part Sub-layer (CPS).

· The CS sub-layer is responsible for the interface between the MAC layer of WiMAX and other backhaul networks such as (Asynchronous Transfer Mode) ATM and IP-based

networks. The CS sub-layer maps the QoS parameters from external networks to the set of QoS parameters used in the WiMAX standard. This sub-layer classifies Service Data Units (SDUs) to an adequate connection in the CPS sub-layer with specific QoS parameters.

- The CPS sub-layer is the most important element in the MAC layer. It is in the CPS sub-layer where the packet scheduler resides. Other MAC layer functionalities such as: packet concatenation and/or fragmentation, packet error control through retransmissions and resource allocation, are also performed inside the CPS sub-layer.

The MAC layer provides the interface between the PHY layer and upper transport layers. It takes advantage of different PHY layer services that address the needs of various mobile environments. The MAC layer performs several functions including: scheduling of bursts, link adaptation and error recovery, network entry procedures, and standard Packet Data Unit (PDU) creation tasks, such as fragmentation or packing, for each active connection. The MAC layer supports five QoS service classes that serve the data transfer needs of different real-time and non-real time media types.

The MAC layer receives SDU packets from higher layers and organizes them into PDUs for transmission over the air. Multiple SDUs of same or different lengths may be aggregated into a single PDU (MPDU) in order to save MAC header overhead. In the same way, large SDUs may be fragmented so that they may be sent across frame boundaries. MPDUs may be of variable length and multiple MPDUs may be concatenated into a single burst to reduce MAP overhead. Each burst is transmitted using a single MCS that is signaled within the MAP message and may include MPDUs intended for one or more users. Each MPDU is segmented into forward error correction (FEC) blocks that are coded and interleaved within the burst. A number of contiguous OFDMA symbols, using the same permutation formula to map sub-carriers to sub-channels, is called a permutation zone.

In Mobile WiMAX the MAC layer is designed for handling applications with different QoS requirements. All services are connection-oriented, i.e., each service is mapped to one or multiple connections and is handled by the CS sub-layer and then the CPS. This is illustrated in figure 6. As can be seen from the figure, the CS classifies the SDUs to a connection with specific QoS parameters and, depending on the QoS requirements from each type of service data flow in the MAC layer, there are different mechanisms available for bandwidth allocation which is the responsibility of the MAC layer.

3.4.2 Mechanisms for Quality of Service Support

Quality of service support is one of the essential features in WiMAX standards [13, 56]. Strong QoS control is achieved by using a connection-oriented MAC architecture where all downlink and uplink connections are controlled by the serving base station.

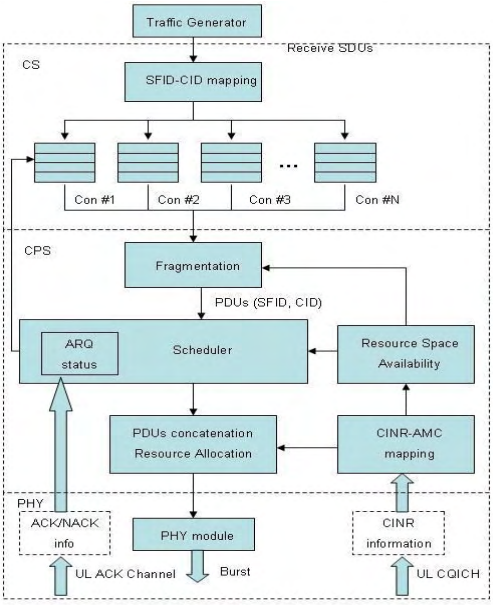


Figure 6 - Mobile WiMAX Protocol layer

Before any data transmission takes place the base and mobile stations must establish a unidirectional logical link called a connection between both two MAC peers. Each connection is identified by a connection identifier (CID) which is a temporary address for data transmissions over the particular link. Fundamental to the provision of QoS in WiMAX is the implementation of service flows. A service flow is defined as a one-way flow of MAC Service Data Units (MSDUs) on a connection associated with specific QoS parameters such as: latency, jitter and throughput. These parameters are inputs to the scheduler. Each service flow is identified by a service flow identifier (SFID). The base station is responsible for issuing the SFID and for mapping it to unique CIDs. The packet scheduler is placed inside the MAC layer in order to be fast enough and reduce the latency incurred by scheduling decisions, which would otherwise increase if it was placed inside the base station controller.

3.4.3 Service Classes in Mobile WiMAX

To support a wide variety of applications WiMAX defines five scheduling services:

Unsolicited Grant Service (UGS) – This is designed to support Real-Time (RT) service flows which periodically generate packets of fixed size. Service flows of type UGS are granted radio resources periodically without the need for the scheduler intervention and bandwidth request from mobiles. This results in a reduction of the associated signalling overhead and latency incurred in bandwidth requests. This type of service flow was designed for the support of applications with a constant bit rate (CBR applications). For uplink, this service offers fixed-size grants for data transport on a real-time periodic basis (implicit request). Connections configured with UGS are not allowed to utilize random access opportunities. Examples of applications implementing this type of scheduling service are T1/E1 and VoIP without silence suppression.

Real-Time Polling Service (rtPS) – This is designed to support delay sensitive real-time service flows with variable-size data packets on a periodic basis. This service offers periodic dedicated request opportunities to meet real-time requirements, which results in more signalling overhead and latency than UGS. It is well suited for connections associated to service flows of type VoIP or video streaming services, such as Near Real Time Video (NRTV). For uplink this service offers periodic unicast request opportunities (piggyback request/unicast polling). Because the mobile station issues explicit requests, the protocol overhead and latency is increased, but this capacity is granted only according to the real need of the connection.

Extended Real-Time Polling Service (ertPS) – This is a mix of both UGS and rtPS service classes. It is designed to support real-time service flows that generate delay sensitive variable sized data packets on a periodic basis, as in rtPS. Also, it performs like the UGS scheduling service, as the base station is allowed to issue unicast grants in an unsolicited manner. For uplink, this service offers a mechanism for periodic allocations, which may be used for requesting the bandwidth as well as for data transfer, considering the traffic characteristics of VoIP with silence suppression (piggyback request/unicast polling).

Non-Real-Time Polling Service (nrtPS) – This is designed to support delay-tolerant service flows consisting of variable-sized data packets for which a minimum data rate is required. The nrtPS particularly addresses Internet type of applications such as File Transfer Protocol (FTP) and web browsing. For uplink, this service offers unicast polls on a regular basis, in an interval on the order of 1 second or less. For this service, connections may utilize random access transmit opportunities for sending bandwidth requests.

Best Effort (BE) – This is designed to support service flows that have no minimum service requirements and which are serviced on a resource-availability basis. BE service flows provide no guarantees either for minimum throughput assurance or maximum packet delay. For uplink, this service may offer contention request opportunities (contention based polling). The mobile sends requests for bandwidth in either random access slots or dedicated transmission opportunities. The occurrence of dedicated opportunities is subject to network load and the mobile cannot rely on their presence. The email application is an example of a BE service flow.

Table 2 illustrates the different types of data delivery services and respective QoS provisioning parameters in Mobile WiMAX systems.

Finally, table 4 describes the classification proposed for each one of the four traffic types implemented in this work, and the QoS parameters which are considered in taking scheduling decisions.

The Maximum Sustained Rate is actually not defined and/or used in the schedulers proposed in this work, because whenever a given user is scheduled for transmission it is assigned the amount of radio resources needed to empty its buffer. However, if a maximum rate would be defined, it would result in no significant modifications in the architecture of the schedulers

proposed and in the advocated performance, when compared to the performance of most commonly packet schedulers in the literature and used as benchmark in this work.

Service Type	Application	QoS Parameters
UGS	T1/E1, VoIP without silence suppression	· Minimum reserved traffic rate · Maximum latency · Tolerated jitter
rtPS	Streaming audio or video	· Maximum reserved traffic rate · Minimum sustained traffic rate · Maximum latency · Traffic priority
ertPS	VoIP with silence suppression	· Maximum reserved traffic rate · Minimum sustained traffic rate · Maximum latency · Tolerated jitter · Traffic priority
nrtPS	File Transfer Protocol (FTP)	· Maximum sustained traffic rate · Minimum reserved traffic rate · Traffic priority
BE	Data Transfer, Web browsing	· Maximum sustained traffic rate · Traffic priority

TABLE 2: TYPES OF DATA DELIVERY SERVICES AND THEIR QOS REQUIREMENTS

Table 3 provides the definitions for the QoS parameters.

Parameter	Definition
Maximum reserved traffic rate	Peak information rate of service
Maximum sustained traffic rate	Minimum amount of data to be transported when averaged over time
Maximum latency	Maximum allowable time between ingress of packet to convergence sub-layer and the forwarding of SDU to air interface
Tolerated jitter	Maximum delay variation for the connection
Traffic priority	Priority assigned to service flow

TABLE 3: QOS DEFINITIONS

Application	QoS Category	QoS Specifications Used
VoIP	rtPS	Maximum Sustained Rate Maximum Latency Tolerance Priority
NRTV	rtPS	Maximum Sustained Rate Maximum Latency Tolerance Priority
FTP	nrtPS	Maximum Sustained Rate Minimum Reserved Rate Priority
WWW	BE	Maximum Sustained Rate Priority

TABLE 4: SPECIFICATIONS FOR THE TRAFFIC TYPES IMPLEMENTED IN SYSTEM LEVEL SIMULATOR

3.4.4 Bandwidth Request and Assignment in Mobile WiMAX

Whenever the mobile has multiple connections with the base station it has some control over bandwidth in order to share the resources among the connections. Scheduling in downlink and uplink is done by the base station. In the downlink the base station is the only one transmitting during the downlink sub-frame as it schedules all downlink connections. It allocates bandwidth according to the needs of the incoming traffic in its buffers, without mobile station intervention. The bandwidth assignments are broadcast to all mobile stations in the DL-MAP and the mobiles

will know exactly when to receive its own packets. In uplink the base station allocates bandwidth based on requests sent from each mobile station wishing to transmit and signals this allocation in the UL-MAP.

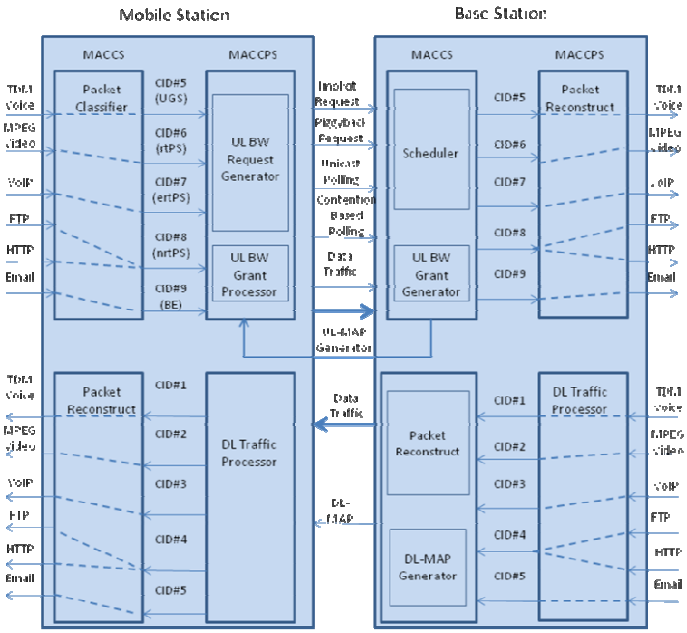


Figure 7 - Example of service flow exchange in Mobile WiMAX

The mobile station has a plethora of ways to request bandwidth. Depending on the particular QoS and traffic parameters associated with a service type, the mobile station can combine the determinism of unicast polling with the responsiveness of contention-based requests and the efficiency of unsolicited bandwidth. A conventional way to request bandwidth from the mobile is for the base station to allocate dedicated or shared resources periodically to each mobile station, which can be used by these to send a bandwidth request PDU consisting of the bandwidth request header and the payload (polling). Polling can be done either individually (unicast) or in groups (multicast). When a mobile station is polled individually, it is allocated bandwidth to send bandwidth request messages. Multicast polling is a contention-based bandwidth request method used whenever there is insufficient bandwidth to individually poll many inactive mobiles. The allocation is multicast or broadcast to a group of mobile stations that have to contend for the opportunity to send bandwidth requests.

Figure 7 illustrates the mapping process from service applications into connection and service flows in the MAC layer and the association to the service classes described above [12].

3.4.5 Advanced Features for Performance Enhancement

Mobile WiMAX includes a number of advanced features for performance improvement. The most important ones are: advanced antenna features for the support of multiple antenna techniques, hybrid automatic repeat request and frequency reuse.

3.4.6 Adaptive Modulation and Coding (AMC)

Mobile WiMAX supports a variety of MCS schemes for data transmission. The MCS scheme is selected according to the channel state and this selection can be performed on a burst-by-burst basis per link. On the downlink the base station uses the channel quality feedback indicator transmitted by the mobile station on the CQICH channel. On the uplink the base station estimate the channel quality based on the received signal. The MCS used is the one which maximizes the throughput while keeping the estimated block error rate (BLER) lower than the pre-defined threshold. This threshold depends on the type of service.

AMC significantly increases the overall system capacity as it allows real-time trade-off between throughput and robustness on each link. The various modulation and coding schemes supported by WiMAX in downlink are: QPSK, 16 QAM and 64 QAM, which are mandatory for both Fixed and Mobile WiMAX. In uplink 64QAM is optional for Fixed WiMAX. FEC using convolutional coding is mandatory. The standard optionally supports turbo codes and low-density parity check (LDPC) codes at a variety of code rates as well.

3.4.7 Advanced Antenna Systems (AAS)

Significant gains in system capacity and spectrum efficiency can be achieved with the implementation of the advanced antenna features [57-59] provided with the Mobile WiMAX. These are the following:

Transmit diversity: Transmit diversity in the downlink connection can be achieved with the use of space-time block coding schemes (STBC). With 2 antennas at the base station emitter and 1 antenna at the mobile station receiver a 2x1 Alamouti STBC can be implemented. This is referred as Matrix-A coding in WiMAX terminology. Other variations of STBC can be implemented with more than 2 antennas at the transmitter and receiver. STBC is appropriate for scenarios whereby the SINR is low and/or the mobile speed is high. It improves the quality of the signal at the mobile receiver, in the same sense as with a maximum ratio combiner (MRC) with a SIMO 1x2 antenna configuration. The advantage of the 2x1 Alamouti STBC is that the complexity of the receiver is transferred to the transmitter at the base station.

Spatial multiplexing (SM): In spatial multiplexing multiple independent streams can be transmitted in parallel across multiple antennas. Spatial multiplexing is implemented with a MIMO channel using multiple antennas at the transmitter and receiver. A linear receiver is used to separate the contributions from the different antennas at the transmitter and to attenuate the interference among the antennas in the array. A typical linear receiver used is the Minimum Mean Square Error (MMSE). Spatial multiplexing increase the capacity of the system provided the received SINR is good enough and/or the mobile moves with low speed. Assuming a rich multipath environment the capacity of the system can be increased linearly with the minimum number of antennas at the transmitter and receiver. In the uplink, spatial multiplexing can be

implemented assuming collaboration is achieved through the coding across multiple users with one single antenna. In WiMAX terminology spatial multiplexing is referred to the Matrix B configuration.

3.4.8 Hybrid Automated Repeat Request (HARQ)

This is a combination of ARQ with FEC at the physical layer and provides for improved link performance over traditional ARQ, at the cost of implementation complexity. With HARQ retransmission is requested if the decoder is unable to correctly decode the received block. When a retransmitted coded block is received it is combined with the previously detected coded blocks and fed back to the input of the FEC decoder. The probability of success in the decoding of the data block is increased by combining the different replicas of the block. Two types of HARQ can be implemented: Chase Combining and Incremental Redundancy. Chase Combining was used in all simulations conducted in this work.

3.4.9 Fractional Frequency Reuse

Mobile WiMAX supports frequency reuse of one, i.e. all cells operate on the same frequency channel to maximize spectral efficiency. However, users on cell edge will suffer heavily degradation in connection quality due to inter-cell interference arising from co-channel users. In Mobile WiMAX users can operate on sub-channels which only occupy a small fraction of the whole channel bandwidth. This is achieved by sub-channel segmentation and the definition of permutation zones. A segment is a subdivision of the available OFDMA sub-channels and each segment is equal to a single instance of the MAC layer. The availability of sub-channelization schemes, such as PUSC, allows the coordination of sub-channel allocation to users at the cell edges in order to limit the inter-cell interference. This is achieved by grouping different sets of sub-channels in segments and by assigning each segment to a particular cell. In fractional frequency reuse, users with good channel conditions (measured as the SINR) have access to the full set of sub-channels, operating under a frequency reuse of 1 and users with bad channel conditions will be allocated non-overlapping sets of sub-channels in order to preserve channel orthogonally, namely at the cell edge. This type of sub-channel allocation leads to the effective reuse factor taking fractional values greater than 1.

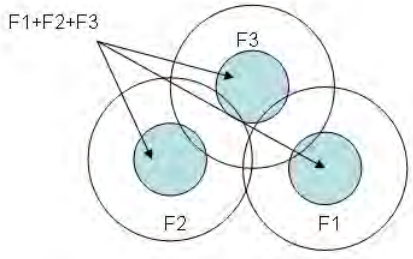


Figure 8 - Fractional Frequency Reuse Implementation in Mobile WiMAX

In figure 8, F1, F2 and F3 represent different sets of sub-channels in the same frequency spectrum. The frequency reuse one is maintained for users in the centre of the cells, to maximize spectrum efficiency and fractional frequency reuse is implemented for users in the edge of the cells to assure edge-user connection quality.

3.5 Cross-Layer Implementation in Mobile WiMAX

Cross-layer optimization is one of the most important concepts for next-generation wireless communication systems, and the protocol architecture of the IEEE 802.16e standard for Mobile WiMAX was designed with mechanisms which efficiently support cross layer operation between layers of its protocol stack [40, 10]. In this section the mechanisms implemented in the Mobile WiMAX standard for cross-layer operation are described and some examples of their use for cross-layer optimization, namely for capacity enhancement and quality of service support, are presented in detail.

3.5.1 Introduction

In the Mobile WiMAX protocol architecture the PHY layer implements strategies such as adaptive modulation and coding schemes, different sub-channelization schemes and multiple antenna technologies, and the MAC layer implement strategies such as hybrid automatic repeat request (HARQ), scheduling and flexible resource allocation in both time and frequency domains. These different strategies can be jointly adapted through both layers, in a cross-layer design approach, according to system constraints such as delay requirements, total power available for data transmission and fairness in resource allocation.

For example, the MAC scheduler selects packets according to channel condition, using information from the PHY layer, and QoS parameters, using information from application layer, in order to maximize system throughput and meet QoS requirements. Table 5 shows examples of feedback information that can be utilized for cross-layer operation.

Layer	Examples of Feedback Information
Physical	SINR, Received Signal Strength (RSSI)
Medium Access Control	Current FEC scheme, number of transmission attempts, medium availability time.
Application	QoS requirements: delay, delay jitter, required throughput, accepted packet loss rate
User	User-perceived QoS

TABLE 5: EXAMPLES OF FEEDBACK INFORMATION FROM DIFFERENT LAYERS IN PROTOCOL STACK

The cross layer architecture framework is accomplished through a number of uplink control channels in the uplink sub-frame. These channels are tailored for the fast exchange of information for cross layer operation and are used in the signalling interaction between the base and the mobile stations. This particular design of the TDD frame makes it possible for the scheduler to rapidly adapt the packet transmission to the dynamic propagation and interference conditions as well as to the traffic demand.

In the following sections two types of cross-layer design for Mobile WiMAX are presented: one for capacity improvement and the other for QoS support.

3.5.2 Cross-Layer Design for Capacity Improvement

Mobile WiMAX has inherent features available for capacity improvement. These are: band AMC sub-channel allocation, adaptation between space diversity and spatial multiplexing with MIMO and SDMA using enhanced AAS and HARQ.

3.5.2.1 Cross-Layer Design for Efficient Resource Allocation

Mobile WiMAX implements two different types of sub-channelization schemes: diversity and band AMC. Diversity sub-channels average out inter-cell interference because each cell or sector has a different sub-carrier permutation pattern, resulting in the implementation of a frequency reuse factor of one. For band AMC, cell edge users are much more subjected to inter-cell interference if neighbouring cells or sectors use the same set of sub-channels. The use of Band AMC results in more capacity than diversity mode, because of the multiuser gain over frequency domain. It was reported in [60] an increase in capacity of up to 30% over diversity mode. But it requires more overhead and is more complex than diversity mode because it is necessary to estimate channel quality with greater accuracy and a lot of feedback is required in order to report the channel quality for each band. The selection of each channel mode requires careful cross-layer design between PHY and MAC layers by exchanging information such as mobile's SINR, mobility and required QoS.

3.5.2.2 Cross Layer Design for Advanced Antenna Techniques

The Mobile WiMAX standard supports variable options for multiple-antenna technologies, such as: STBC based on Alamouti code, SM using vertical encoding, two-user collaborative SM, and an AAS for beam forming. STBC reduces the fade margin by providing spatial diversity and SM improves capacity by transmission of multiple streams over multiple antennas of a mobile with good SINR and low spatial correlation. The standard adopts an adaptive MIMO switch by adaptively selecting between both two MIMO modes according to channel conditions. This information is reported from the PHY layer to MAC layer using cross-layer design primitives.

Mobile WiMAX standard also supports adaptive beam forming by applying a weight to each antenna element. This can be used for multiplexing users over different spatial beams in the spatial domain. Since the spatial resources have different data transmission capabilities, depending on how spatially separable users are clustered, the design of a scheduler and resource allocator that utilizes spatial resources is a very sophisticated cross layer design. The adaptive allocation of sub-channels and power with beam forming improves system performance and is another cross-layer design issue.

The algorithm for SDMA resource allocation operates on the basis of QoS requirements received from the MAC layer and information about spatial channels received from the PHY layer, and provides the scheduler with sets of possible solutions for scheduling and allocation of resources.

3.5.2.3 Cross-Layer Design for Detection and Error Recovery

The use of adaptive link adaptation together with HARQ results in capacity improvement through spectrum efficiency maximization. The choice of the correct MCS scheme depends on channel conditions reported to the MAC layer from the PHY layer and on the estimation of the predicted BLER for the first transmission as a HARQ-enabled connection. The definition on the number of allowable retransmission attempts depends on the type of service and on the overhead resulting from control messages needed for the control of the retransmissions. When HARQ is applied to real-time services it is desirable to design retransmission strategies that consider service delay bound as well as channel quality. This information exchange requires the definition of cross-layer primitives among application, PHY and MAC layers.

3.5.3 Cross-Layer Design for Quality of Service Support

The predefined set of QoS parameters assigned to each one of the service data flow types implemented in Mobile WiMAX standard is used in scheduling and resource allocation. These QoS parameters are managed dynamically through MAC management messages, which are interchanged with upper layers in the protocol stack. These parameters are also used for granting bandwidth requests in uplink.

Downlink packets are received from upper layers and are classified into service flows by a packet classifier within the base station. These packets are transmitted once they are selected and their resources are allocated by the downlink scheduler. The scheduler considers cross-layer information such as the downlink channel quality information reported from the PHY layer and QoS parameters from upper layers, in order to maximize system throughput and meet the requirements for QoS service flows.

Uplink packets from upper layers are classified into service data flows by a packet classifier within the mobile station and the mobile station requests bandwidth according to a number of mechanisms which were implemented for the request of bandwidth in uplink, such as unsolicited grants, multicast/unicast polls and piggyback. From the amount of bandwidth requested the base station estimates the queue status information of each mobile station.

3.6 Related Work

There is some work available in the research literature regarding cross-layer based WiMAX systems.

In [61] the authors investigate the performance of a WiMAX network based on link and system level simulations. Some dynamic resource allocation methods for the basic WiMAX system profile are proposed and their performance investigated. A set of recommendations on design and deployment configurations are proposed and standard scheduling algorithms such as the Maximum C/I and Proportional Fairness, coupled with simple full queue traffic model are used in performance analysis. However, a real simulation scenario, with traffic models for the emulation of real traffic data, is not considered. The assignment of radio resources from the MAC frame is a very simple process. As a consequence not enough performance curves are proposed in the validation of the proposal.

In [40] a cross layer architecture framework for performance improvement of WiMAX compatible WiBro system in Korea is presented and investigated by means of a prototype for a Mobile WiMAX network, under a variety of traffic models for commonly services envisaged for Next Generation Networks (NGN), such as web browsing and FTP. The framework architecture is based on the exchange of signalling information between PHY and MAC layers. However, although the authors provide detailed explanation about the use of the control channels for the exchange and implementation of a cross-layer architecture, the details of the implemented DRA architecture are not presented and no system level simulations in a realistic scenario, with traffic models for emulation of data traffic, are considered. Rather a very simplistic analysis of the main achievements of the proposed cross-layer architecture is elaborated. System and link level simulations were conducted and the performance of the network was evaluated and compared to the performance of 3GPP HSDPA systems, by means of typical metrics such as system and user throughput as well as spectral efficiency. However, the authors do not elaborate on the resource allocation on the provision of real-time data services, nor on the satisfaction of the respective QoS requirements.

In [62-63] the performance of Mobile WiMAX for different scenarios such as the use of MIMO antennas for the downlink and uplink are investigated. System level simulations were used to determine the throughput performance of the proposed WiMAX network architecture for full buffer and web browsing traffic models. In particular the authors investigate on the coverage efficiency of the different scenarios considered for the control channels in the MAC frame. A comprehensive set of simulations is conducted for different kinds of channel models.

In [64] the authors are concerned in particular with the performance of WiMAX networks using MIMO schemes. Again system and link level simulations were conducted to infer about the performance of WiMAX networks under such scenarios via performance metrics such as sector and user throughput, spectral efficiency and network quality, measured as the estimated packet error rate (PER).

In [65] the system performance for the IEEE802.16-2004 standard is given. The influence of configuration parameters such as packet fragmentation, the amount of padding bits and the size

of the payload of each MAC Protocol Data Unit (PDU) packet on the system performance are analysed with recommendations for optimum system parameters.

Most of the approaches available in the literature for investigation on WiMAX system performance do not consider and/or investigate scheduling algorithms specifically designed for the provision of QoS. Although in [40] the authors elaborate on a generic scheduling algorithm for QoS provisioning via the implementation of Utility Functions [66-68], no consideration is given to dynamic resource allocation design for real-time services. Also the work available in the literature does not consider the specifics of the system behaviour when Hybrid Automated Repeat Request (HARQ) protocol is implemented.

Utility function-based scheduling is a typical example of resource management based upon the cross-layer paradigm, for both QoS provision and enhancement of throughput [69]. The central idea of utility-based scheduling is to map resource usage such as bandwidth and power or performance criteria, such as data rate and delay, onto the corresponding utility, in such a way as to take into account the characteristics of the application and channel conditions. In [70] rate-based utility optimization problem is formulated, assuming a fixed allocation of power, for OFDM-based wireless networks. The use of concave utility functions of the achievable user's data rate on a given sub-channel foresees the simplification of the problem into a gradient-based scheduling algorithm, in which the user selection depends on the marginal utility function of the data rate on the given sub-channel and the achievable data rate.

In [71] the initial rate-based resource management scheme was extended for the provision of QoS. The argument used in the utility function is the delay of the head-of-line packet in each user's queue. User selection depends on the combination of the delay and achievable data rate of the user.

In [68] the optimization problem is formulated by defining a decreasing utility function for head-of-line packet delay. It was demonstrated that the long-term optimization objective with respect to average waiting times leads to an optimization problem in which the best user is the one resulting in the best combination of the marginal utility function of delay, queue length and channel state.

3.7 Conclusion

In this chapter a detailed description of the protocol architecture of the Mobile WiMAX standard was performed. The standardization process conducted by IEEE 802.16 working group, which resulted in the IEEE 802.16e as an amendment for the IEEE 802.16-2004 standard, was detailed. Mobile WiMAX is an extension of Fixed WiMAX for mobility scenarios and presents new advanced features such as OFDMA multiple access for both downlink and uplink, different types of sub-channelization, use of advanced antennas systems and HARQ. The ultimate goal is to increase system capacity and support of the stringent QoS

requirements posed by the type of multimedia applications envisioned for fixed Internet networks. These advanced features constitute strong arguments for the consideration of the Mobile WiMAX standard as a potential candidate for wireless networks of fourth generation.

The PHY and MAC layers were described in detail. In particular, the TDD frame structure and the control and data fields in uplink and downlink sub-frames were presented. It was verified how the frame structure addresses the implementation of a cross-layer design framework between PHY, MAC and upper layers in the protocol stack. Specifically designed control channels allow fast and efficient exchange of information in a truly cross-layer design framework. Examples of cross layer design in Mobile WiMAX were given. Finally a resume of the related work available in the literature regarding cross layer design architectures, whose performance is inferred by means of system level simulations, was presented.

Next chapter proposes a dynamic resource allocator (DRA) module which includes Mobile WiMAX functionalities for cross-layer design. Packet schedulers are developed by taking into consideration the intricacies of the implemented DRA.

Chapter 4

System Level Simulator for Mobile WiMAX System

4.1 Introduction

A great deal can be learnt about an air interface technology by analysing its performance in a link level setting, consisting of one base station and one mobile station. This link level analysis is fundamental in the evaluation of the technologies associated to the given air interface, namely for the study of the variation of the Bit Error Rate (BER) with the Signal to Noise Ratio (SNR) per bit sent along the transmission chain, under the influence of such an aggressive medium for signal transmission as the wireless mobile channel. A link level model is used to study the transmission between a base station and one or more mobile stations. Link level models concentrate on physical layer and performance is measured in terms of bit per second throughput. In real-world, multiple base stations are deployed in a service area and operate in the presence of a large number of active mobile users. Therefore, system performance can only be evaluated through a system-level analysis, where the point-to-point radio link communication scenario is replaced by one in which all radio links among the mobile and base

stations must be considered. System level simulations consider a network of base stations. Performance is expressed by user perceived quality of service parameters and physical layer details are abstracted as much as possible.

Typically, network simulations are divided into two parts: link and system level simulations. Although a single simulator approach would be preferred, the complexity of such simulator (including everything from transmitted waveforms to multi-cell network) is far too high with the required simulation resolutions and simulation times. Also the time granularity of both domains of simulation are symmetrically different: at link level bit transmissions are at the order of mili-seconds, while at the system level, traffic and mobility models require time intervals of some tens of seconds to minutes. Therefore, separate link and system level simulations are needed. The link level simulator is able to predict the receiver's Frame Erasure Rate/Bit Error Rate (FER/BER) performance, taking into account channel estimation, interleaving and decoding. Because the simulation is divided into two parts, a method to interconnect them and a proper interface has to be defined.

Link level simulations are conducted for the definition of a model with appropriate interfaces for the interaction with system level simulations. System performance evaluation of a given mobile wireless access technology requires simulations that carefully capture the dynamics of a multipath fading environment, the architecture of the air interface, mobile stations behaviour (in terms of mobility) and the types of applications used (properly defined traffic models). Also, since system level results depend intrinsically on the scenario simulated (propagation and interference environments, number and distribution of users within the cells) it is important that the assumptions and parameters used in the analysis be reported jointly with performance results. This procedure is needed for the correct validation of the results obtained and for benchmarking against other proposed scenarios from different wireless systems.

This chapter details all steps followed in the development of a system level simulation tool for conducting system level simulations of wireless systems in general, and Mobile WiMAX in particular. Section 2 presents the system level simulation methodology which should be followed whenever conducting system level simulations and classifies the different types of simulations that can be performed. The modules which compose the system level simulator are also presented. Section 3 explains the two different types of simulation flows commonly employed when performing system level simulations: combined-snapshot and dynamic modes. The network and cell layout scenario used in the simulations is introduced as well as the model for the antenna pattern used in the simulation of tri sector base stations. Section 4 is about the propagation models used in simulations at system level. The kind of propagation model used depends on the type of channel configuration implemented: Single-Input-Single-Output (SISO) or Multiple-Input-Multiple-Output (MIMO). The propagation models used were developed for simulations performed at the network level, encompassing a base station transmitting/receiving

from a number of mobile stations. The propagated signal is characterized by three components: path-loss, shadowing and fast fading, and corresponding models must be presented for each one of them.

Propagation models are used in the computation of the desired and interfering signal components impinging on the receiver antennas, and these are used in the derivation of the link quality, as measured as the Signal-to-Interference-plus-Noise-Ratio (SINR). Section 5 describes the method used in the computation of the SINR value for each link between the base and mobile stations. This computation depends on the cellular layout implemented, the type of channel used (MIMO/SISO) and the type of multiple access scheme used in the air interface. Section 6 details the link to system level interface needed to interconnect both two domains of simulations. As Mobile WiMAX is based on the Orthogonal Frequency Division Multiple Access (OFDMA) scheme a vector of SINR values, one for each sub-carrier of the sub-channel, is obtained and a method must be applied to convert this vector into a single scalar value. The method used along the system level simulations performed in this work is described: Exponential Effective SINR (EESM). After obtaining the compressed SINR value the corresponding Block Error Rate (BLER) is derived by means of Look-Up Tables (LUT) which reflect physical layer performance. Section 7 describes the channel model used for conducting system level simulations under a MIMO channel scenario. The model implemented for MIMO channel simulations at system level is the Spatial Channel Model (SCM) from 3GPP. SCM is a ray tracing-based propagation channel model where the signal for each path in the receiver results from the superposition of different signal components arriving from different paths around the mobile and base station. Strategies followed in the reduction of the complexity, inherent in modelling every single link between each mobile station, as well as the serving and neighbouring base stations are presented. The link to system level interface and the mapping method between the vector of SINR values and the scalar value is also presented. Section 8 presents the models commonly used in the research literature for user's traffic simulations, according to the type of application used. A detailed description of each traffic model used in this work is included in appendix A. Section 9 is about the performance metrics used in the evaluation of the system level platform and, in next chapters, for Dynamic Resource Allocation (DRA) architecture and scheduling algorithms performance evaluation, at system level. An exhaustive list of the performance metrics used in this work is given in appendix B. Section 10 presents the related work regarding system level simulations available in the research literature for wireless network systems such as HSDPA, LTE and WiMAX. Section 11 concludes the chapter.

4.2 System Level Simulation Methodology

The performance evaluation of a broadband wireless system such as Mobile WiMAX, for scenarios of application as close as possible to reality, must be conducted on system-level simulation platforms, with realistic models for traffic and signal propagation phenomena such as: path-loss, shadowing and fast fading; mobility patterns and traffic generation for supported users; inter-cell influence, etc. [106]. The more accuracy achieved in the implementation of these models in the simulation platform, the closer the simulator outputs are to reality. A properly-designed system-level simulation platform suits the derivation of the performance figures needed in the evaluation of the impact and satisfaction of the standard, in terms of system requirements such as: spectrum efficiency, system capacity, quality of service support, end-user satisfaction and cost-efficiency. If the results obtained from system-level simulations are satisfactory then hardware can be designed and manufactured.

The methodology followed in system level simulations depends on a different set of assumptions regarding: type of wireless system simulated, air interface technology, simulations complexity and time resolution, interface with other layers of the protocol stack, such as the physical layer, network layout, channel and interference modelling and application traffic models.

In particular, the following aspects must be considered with care in developing a system level tool and on performing system level simulations:

Network Scenario

- This is related to the type of environment considered in the simulations: urban, rural, vehicular or indoor.

Network Layout

- Amount of tiers and number of base stations simulated.
- Type of cells: one omnidirectional cell or three, six sectored cells for example.
- Number of mobile stations and their distribution over the network coverage area.

Radio Resource Management

- Enable/disable power control.
- Enable/disable user mobility and handover.
- Definition of the radio resources according to the type of air interface and medium access layer.

Physical Layer Modelling and Abstraction

- Definition of the metrics used to map physical layer performance to higher layers of the protocol stack.
- Definition of the types of interfaces used in the interaction between system and physical layers.

Propagation and channel modelling

- Path loss propagation.
- Slow fading (shadowing) propagation.
- Fast fading channel modelling.

Interference modelling

- Intra-cell, inter-cell and inter-system interference.

Implemented radio access system

- Multiple access to radio resources, circuit switch/packet switch,

Traffic models for application services

- Choice of traffic models: emulation by using pre-defined traffic models or use of real traces from real networks.

Performance metrics

- Metrics for network evaluation performance.
- Metrics for user satisfaction evaluation.

Simulation complexity and time resolution

There is a trade-off between accuracy and simulation execution time. The correct balance must be found. The main simulation components of a complete simulation tool are illustrated in figure 1, according to simulation procedures elaborated in [72, 107].

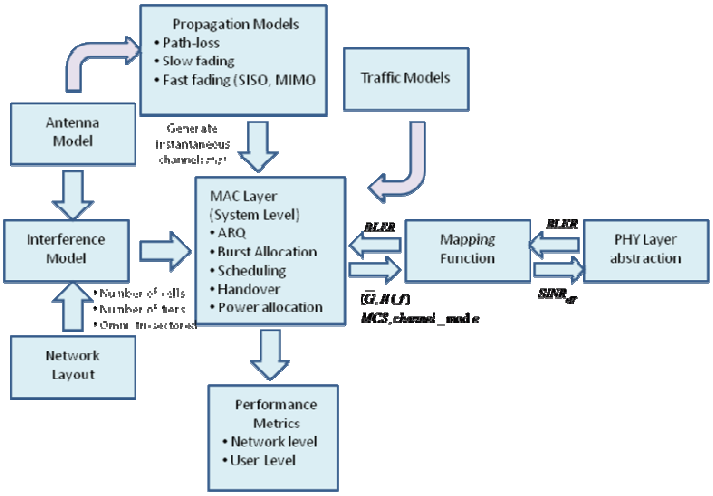


Figure 1 - Simulation Components

4.2.1 Simulation Execution Flow

Two different types of simulations can be performed at system level: using a Combined Snapshot-Dynamic or a Dynamic mode.

- **Dynamic mode:** mobility is enabled as mobiles travel along the network coverage area performing handovers. Mobiles are dropped in the network in the beginning of the simulation run and remain active since the instant of activation, which can be coincident with the beginning of the simulation run or be defined by some random distribution. Only

one simulation run is performed and mobiles are removed at the end of the simulation. Statistics are collected as mobiles travel through the network coverage area. Path-loss, shadowing and fast fading propagation components are re-computed at every transmission time interval. The new position of the mobile station in the next transmission time interval is also computed according to the chosen mobility model.

- **Combined snapshot-dynamic mode**: mobility and handovers are disabled. A given number of simulation runs are performed. Mobile stations are drawn on the network in the beginning of each simulation run and are removed at their end. They remain active since the instant of activation, which can be coincident with the beginning of the simulation run or be defined by some random distribution. In this mode, path-loss and shadowing are computed at the beginning of each simulation run and remain constant until the end of the run. Fast fading is re-computed at every transmission time interval. This mode increases simulation speed as the different simulation runs (snapshots) can be performed in parallel.

In both modes mobiles are randomly uniformly distributed over a hexagonal network of base stations. Each base station can be configured with one sector (omnidirectional antenna pattern) or with three sectors/cells (directional antenna pattern).

In this work, system level simulations were conducted according to the combined snapshot-dynamic mode. A number of simulation runs (snapshots) was performed along each simulation execution. In the beginning of each run mobile stations are dropped in the first tier of cells of the network layout, which can amount to one cell (omni-cell scenario) or three cells (tri-sectored scenario). Neighbouring cells contribute only to interference generation.

Path-loss and shadowing are computed in the beginning of each run and for each pair of mobile to base station radio link (including neighbouring cells) and are kept constant until the end of the run, whilst fast fading is executed for each transmission time interval.

Therefore, the steps followed in the simulation flow of a single run of a general system level simulation tool are as follows:

- Mobile stations are dropped independently, with uniform distribution throughout the system. Each mobile corresponds to an active user session that runs for the whole duration of the run.
- Mobiles are assigned channel models. This can be in support of a channel mix or separate statistical realizations of a single type of channel model.
- Mobiles are assigned a traffic model and packets are generated according to the desired traffic model.
- Cell assignment is based on the received power at the mobile station from all potential serving cells. The cell with the best path to the mobile station, taking into account slow fading, path-loss and antenna gains, is chosen as the serving sector.

- For simulations that do not involve handover performance, evaluation of the location of each mobile station remains unchanged during a drop and the mobile's speed is used only to determine the Doppler effect of fast fading. The mobile station is assumed to remain attached to the same base station for the duration of the drop.
- For a given drop the simulation is run for the pre-defined duration and then the process is repeated with the mobile stations being dropped at new random locations.
- Performance statistics are collected for mobile stations in all cells.

Each run is made up of a number of transmission time intervals or frame periods, as each transmission time interval lasts for the period of time equal to the transmission of a single OFDM frame. Regarding execution flow, the simulator developed for Mobile WiMAX system level simulations is basically a finite machine whose states repeat at each transmission time interval. A transmission time interval is equal to 5 ms.

For each transmission time interval the following events are performed:

- Fast fading is computed for each mobile station in each transmission time interval. Slow fading and path loss are assumed as constant during the whole simulation run.
- Packets are withdrawn from buffers assigned to traffic models. Packets are not blocked, as the queues are assumed as infinite. Start times for each traffic type for each user should be randomized.
- Map of radio resources allocation is updated.
- Packets are scheduled with a packet scheduler using the required metric. Packet decoding errors result in packet retransmissions. In the Dynamic Resource Allocation (DRA) module a Hybrid Automatic Repeat Request (HARQ) process is modelled by explicitly rescheduling a packet as part of the current packet call and after a specified feedback delay period.
- The map of radio resources allocation is computed, according to the implemented scheduler. This is performed by the Dynamic Resource Allocation (DRA) Module.
- Packets are transmitted.
- Packet quality detection is performed and feedback regarding the status of the decoding is reported back.

4.2.2 Simulation of Packet Decoding Process

The Block Error Rate (BLER) resulting from decoding the information transmitted along a single Resource Unit (RU) is denoted by $BLER_{RU}(SINR_{RU})$ and is obtained from the link-to-system interface between the PHY and MAC layers, using as input the Signal to Interference plus Noise Ratio, $SINR_{RU}$. Then a random variable, uniformly distributed between 0 and 1, is drawn. In case the random value is less than $BLER_{RU}(SINR_{RU})$ the block is considered as erroneous and a Negative Acknowledge (NACK) message is sent back to the base station on the

associated signalling channel. Otherwise, the block is deemed as error free and an Acknowledge (ACK) message is transmitted.

The success or failure in the decoding of the transmitted block of information is computed from decoding each individual resource unit into which the data block is mapped. Assuming that a total amount of N_{res} radio resources are used in the transmission and that the decoding is an independent and identically distributed random process, the BLER for the whole radio block is given by equation (1).

$$BLER_{RB} = 1 - \left[1 - BLER_{RU}\left(SINR_{RU}\right)\right]^{N_{res}} \quad (1)$$

The main simulation components of a complete simulation tool are illustrated in figure 1, according to simulation procedures elaborated in [72, 107].

4.3 Network Scenario and Layout

All system level simulations conducted in this work were performed assuming an urban environment scenario. The simulated network is constituted of 57 sectors (19 base stations with 3 sectors each), composing a 3 tier hexagonal cellular network layout, as illustrated in Figure 2.

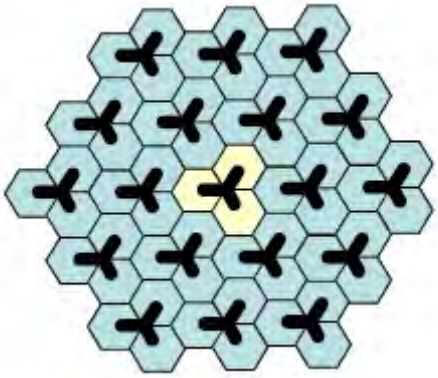


Figure 2 - Network layout deployment

The antenna pattern used in each sector is plotted in figure 3.

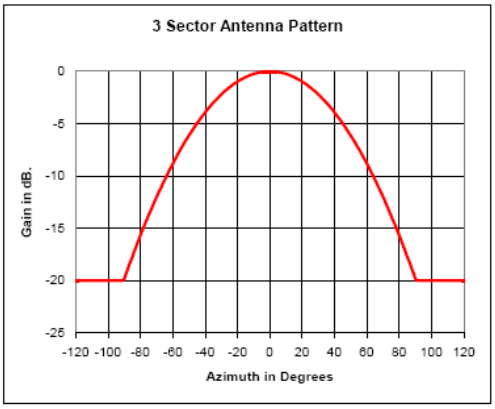


Figure 3- Antenna pattern for 3 sectors

The antenna pattern used for the sectored antenna deployment only considers the horizontal pattern, corresponding to a main sector of 70 degrees. According to the model for the typical antenna pattern proposed in [73], power attenuation is computed as a function of the angle between the antenna pointing direction and the mobile to base station direction, as given by equation (2).

$$A(\theta) = -\min\left[12\left(\frac{\theta}{\theta_{3dB}}\right), A_m\right] \quad (2)$$

Where:

- -180< θ <180 is the angle between the antenna's pointing direction and the mobile to base station line-of-sight direction in degrees.
- $\theta_{3dB} = 70^{o}$ is the beam width at 3dB.
- $A_m = 20^{o}$ dB is the maximum attenuation.

Two types of cell configurations can be defined for simulations: central-cell and non-central cell approach. In the central-cell approach mobiles are dropped along the coverage of the central base station and statistics are collected only for the cells of this base station. Naturally the central cell approach simulation method can be enabled only in conjunction with the combined snapshot-dynamic mode, as mobility modelling is disabled. The cells in the remaining tiers are assumed as fully loaded, i.e., transmitting with maximum power, and contribute to interference only.

4.4 Propagation Channels

The radio propagation is divided into three distinct components, namely path loss, slow fading (shadowing) and fast fading. The decrease of the transmitted radio signal impinging on the receiver antennas is the result of their contribution. Accurate modelling of each one of these three radio propagation components depends on the simulation scenario envisaged for the system-level simulations. Namely the simulation scenario can be described according to the following characteristics:

- Type of environments: indoor, urban, suburban and rural.
- Mobile speed: pedestrian, vehicular, train.
- Type of receiver used in the signal processing at the receiving end.
- Antenna radiation pattern.
- Antenna configuration used in the communication between the transmitter and the receiver (SISO, SIMO, MISO, MIMO).
- Radio transmission parameters: carrier frequency, system bandwidth, etc.

4.4.1 Path-Loss Model

Path loss is defined as the power loss due to the propagation environment. The signal attenuation is directly proportional to a power of the distance between the transmitter and the receiver. The attenuation also depends on the carrier frequency and on the type of environment. The model used for the computation of the attenuation of the radio signal between the transmitter and the receiver is the one proposed in [73] for vehicular environments. This model is suitable for both urban and suburban scenarios, in which the buildings form a relatively homogenous clutter. According to [73] the path loss in dB is given by equation (3).

$$L_{[dB]} = \left[40\left(1 - 4x10^{-3}\,\frac{\Delta h_b}{m}\right)\right]\log_{10}\left(\frac{R}{Km}\right) - 18\log_{10}\left(\frac{\Delta h_b}{m}\right) + 21\log_{10}\left(\frac{f}{MHz}\right) + 80dB \qquad (3)$$

Where:

- R represents the distance in kilometres between the base station and the mobile.
- f is the carrier frequency.
- Δh_b is the base station antenna height from the roof level.

Simulations were performed assuming the carrier frequency proposed in WiMAX profile, which is equal to 2.5 GHz. Also, it was assumed an antenna height, Δh_b, at the base station equal to 15 m. For this setting, the path loss in equation (3) results in the formula presented in equation (4), which is the expression used for the computation of the path loss in the simulations.

$$L_{[dB]} = 130.18 + 37.6\log_{10}\left(\frac{R}{Km}\right) \qquad (4)$$

4.4.2 Shadowing (Slow Fading) Model

Shadowing is the slow variation of the signal power at the receiver. It is given in dB and is modelled by a Gaussian random variable with linear autocorrelation, which is an exponential function of the de-correlation distance, d_{corr}, according to [74] and is given by equation (5).

$$\rho(d) = e^{-\ln(2)\frac{d}{d_{corr}}} \qquad (5)$$

The parameter d_{corr} is the length of the de-correlation distance for which the auto-correlation of the shadowing process, ρ, is equal to 0.5.

Although Gaussian random processes can be modelled as a sum of sinusoids (SOS), conventional one-dimensional channel models (1-D) cannot capture the spatial correlation of shadowing processes. For example, when a given mobile is moving along a closed path around its base station, 1-D models cannot capture the influence of the slow shadowing, in the variation of the signal received at the mobile station, and this affects the performance of the handover algorithm.

According to this, [75] proposes a two-dimensional (2-D) SOS-based channel model to simulate slow fading. The shadowing $SH^{j}_{(x,y)}$ in dB between one mobile station at position (*x,y*) and base station *j* is the sum of two spatial functions, F_0 and F_j, having a Gaussian distribution, with standard deviation mean equal to σ_{SH} (the shadowing standard deviation in dB) and auto-correlation given by (4), using the method described in [75]. It is given by equation (6).

$$SH^{j}_{(x,y)} = \sqrt{0.5}x\left[F_o(x,y) + F_j(x,y)\right] \quad (6)$$

Parameters	Values
Log-Normal Shadowing std σ_{SH}	8dB
De-correlation length d_{corr}	20m

TABLE 1: PARAMETERS FOR SHADOW FADING MODEL

The standard deviation, σ_{SH}, and the de-correlation length, d_{corr}, for the urban scenario used in the system level simulations are the ones listed in table 1.

4.4.3 Fast Fading Model

The fast fading component of the signal is simulated by fast generation of independent Rayleigh faders, according to a modified Jake's model from the method proposed in [76-77]. In order to speed-up simulations, the multi-path channel model is used for the serving cell, while a flat fading channel model (with only one tap) is assumed for neighbouring cells. The mobile speed and carrier frequency are the parameters considered in the generation of the fading statistics. In this context a channel model corresponds to a specific number of paths, a power profile giving the relative powers of these multiple paths and Doppler frequencies to specify the fade rate.

ITU multi-path channel models for narrowband SISO are proposed in [78]. These models are based on a discrete version of the scattering function of the propagation channel and are designated as tapped delay line models. Each tap is characterized by an attenuation, A_i, a corresponding delay, τ_i, a Doppler frequency, f_d, and a Doppler Power Spectrum (DPS), $P_s(f_d, \tau_i)$, at the i^{th} tap. A separate link level simulation must be performed for each specific channel model and mobile stations' velocity combination.

Channel Model	Multi-path Model	Number of Paths	Speed (Km/h)	Fading
Model 1	Ch-100	1	30	Jakes
Model 2	Ch-100	1	120	Jakes
Model 3	Ch-104	6	30	Jakes
Model 4	Ch-104	6	120	Jakes
Model 5	Ch-102	4	3	Jakes
Model 6	Ch-103	6	3	Jakes

TABLE 2: PARAMETERS FOR THE DIFFERENT TYPES OF FAST FADING CHANNEL MODELS FOR SISO

Channel Model		Path 1	Path 2	Path 3	Path 4	Path 5	Path 6
Flat Fading Ch-100	Path Power (dB)	0	-	-	-	-	-
ITU Ped. A	Path Power (dB)	-0.51	-10.21	-19.71	-23.31	-	-

Ch-102	Delay (ns)	0	110	190	410	-	-
ITU Ped. B Ch-103	Path Power (dB)	-3.92	-4.82	-8.82	-11.92	-11.72	-27.82
	Delay (ns)	0	200	800	1200	2300	3700
ITU Veh. A Ch-104	Path Power (dB)	-3.14	-4.14	-12.14	-13.14	-18.14	-23.14
	Delay (ns)	0	310	710	1090	1730	2510

TABLE3: MULTI-PATH CHANNEL MODELS FOR PERFORMANCE SIMULATION

Tables 2 and 3 detail the parameters used in the definition of each type of channel model proposed by ITU [79]. The channel model assigned to a specific user remains fixed over the whole duration of a simulation run.

In [54] system level simulations are conducted to validate Mobile WiMAX standard using the ITU's normalized power profiles for channel models such as ITU Vehicular A: Ch-104 and ITU Pedestrian-B: Ch-103 [79-80]. These models are illustrated in Table 3. The absolute power values are normalized so that they sum to zero dB (unit energy) for each given channel.

4.5 Signal to Interference plus Noise Ratio (SINR) Modelling

In system level simulations mobile stations are randomly dropped along the simulated coverage area. When the mobile station becomes active, its serving cell is selected according to signal strength and the mobile station camps on this cell. As mentioned in the previous sections, the signal coming from the serving cell is modelled as a frequency selective fading channel, whereas the signal coming from neighbouring cells is modelled according to a flat frequency fading channel.

Assume then a given mobile station MS_i is camping in the coverage area of cell $Cell_i$. Assume also a SISO channel. The power received from serving base station $BS_{serving}$ for data sub-carrier $i, i \in [0,..., N_{data} - 1]$ on mobile station MS_i in the n^{th} frame interval is given by equation (7).

$$P_{BS_{serving}}^{(i)}(n) = \frac{P_{data}^{(i)} \cdot \left| H_{BS_{serving}}^{(i)}(n) \right|^2 G_{BS_{serving}} G_{MS_i}}{PL_{MS_i BS_{serbing}} SH_{MS_i B_{serving}} L_{loss}} \quad (7)$$

Where:

- $\left| H_{BS_{serving}}^{(i)}(n) \right|^2$ is the instantaneous power from the serving base station $BS_{serving}$ at the i^{th} data sub-carrier at the n^{th} frame interval.
- G_{MS_i} is the gain of the antenna at the mobile station MS_i.
- $G_{BS_{serving}}$ is the gain of the antenna at the serving base station $BS_{serving}$
- $PL_{MS_i BS_{serving}}$ is the path-loss between serving base station $BS_{serving}$ and mobile station MS_i
- $SH_{MS_i BS_{serving}}$ is the shadowing loss between serving base station $BS_{serving}$ and mobile station MS_i.

- L_{Loss} encompasses the other losses in the transmission (cable losses, body loss,…).

As sub-carriers are mutual exclusively assigned inside each cell there is no intra-cell interference. Therefore, only inter-cell interference must be considered. The interfering power arriving at mobile station MS_i from neighbouring cells is given by equation (8)

$$P_{Inter}^{(i)}(n) = \sum_{BS_j \in \{BS_{Inter}\}} \frac{P_{data,BS_j}^{(i)} \cdot \left|H_{BS_j}^{(i)}(n)\right|^2 G_{BS_j} G_{MS_i}}{PL_{MS_i BS_j} SH_{MS_i BS_j} L_{loss}} \quad (8)$$

Where:

- $\{B_{Inter}\}$ is the set of interfering base stations and $BS_j \in \{B_{Inter}\}$.
- $\left|H_{BS_j}^{(i)}(n)\right|^2$ is the instantaneous power from the interfering base station BS_j at the i^{th} data sub-carrier at the n^{th} frame interval.
- G_{BS_j} is the gain of the antenna at the interfering base station BS_j.
- $PL_{MS_i BS_j}$ is the path-loss between interfering base station BS_j and mobile station MS_j.
- $SH_{MS_i BS_j}$ is the shadowing loss between interfering base station BS_j and mobile station MS_j.
- L_{Loss} encompasses the other losses in the transmission (cable losses, body loss,…).

According to equations (7) and (8), the SINR at sub-carrier *i* and for the n^{th} frame interval is given by equation (9).

$$SINR^{(i)}(n) = \frac{P_{BS_{serving}}^{(i)}(n)}{P^{(i)}{}_{Inter}(n) + N_0 W_i F_{MS_i}} \quad (9)$$

Where:

- N_0 is the received noise spectral density.
- W_i is the sub-carrier bandwidth.
- F_{MS_i} is the noise figure at the mobile station.

The method followed in equations (7-9) for the derivation of the SINR is perfectly general. It was adapted to OFDM system level simulations in [81-82], and adopted in the system level simulator platform for the computation of the SINR in each data sub-carrier *k*, as given by equation (10).

$$SINR^{(k)}(n) = P^{(k)}(n).\overline{G}.\left(\frac{N}{N+N_p}\right).\frac{R_D}{N_{SD}/N_{ST}} \quad (10)$$

Where:

- $P^{(k)}(n)$ is the frequency selective fading power profile for the serving cell (propagation from interfering cells is modelled as flat fading).
- $\overline{G}$ if the Geometric Factor between the mobile station and its serving and interfering cells and is given by equation (11).
- N is the FFT size, including pilot, data and guard sub-carriers.
- N_p is the cyclic prefix length.
- R_D is the percentage of maximum total available transmission power allocated to data sub-carriers.
- N_{SD} is the amount of data sub-carriers per each OFDM symbol.
- N_{ST} is the amount of useful (pilot plus data) sub-carriers per OFDM symbol.

The Geometric Factor is defined by equation (11)

$$\overline{G} = \frac{G(Cell_0, MS)x\dfrac{1}{PL(Cell_0, MS)xSH(Cell_0, MS)}}{\sum_{k=1}^{N} G(Cell_k, MS)x\dfrac{1}{PL(Cell_k, MS)xSH(Cell_k, MS)} + N_0 WF_{MS}} \quad (11)$$

In equation (11) $Cell_0$ is the serving cell and N is the total amount of interfering base stations.

Assuming that multi-path fading magnitudes and phases, respectively $M_p(n)$ and $\theta_p(n)$, are constant over the frame interval for each path p of the tapped delay channel filter, the frequency-selective fading power profile for the k^{th} sub-carrier is given by equation (12) [81].

$$P^{(k)}(n) = \left| \sum_{p=1}^{N_{paths}} M_p A_p \exp(j\theta_p) \exp(-j2\pi f_k T_p) \right|^2 \quad (12)$$

Where:

- p is the tap index (from 1 to 6) of the tapped delay model.
- A_p is the amplitude value of the long-term average power for the p^{th} tap of the tapped delay filter.
- T_p is the relative time delay of the p^{th} tap of the tapped delay filter.
- f_k is the relative frequency offset of the k^{th} sub-carrier within the spectrum of the OFDM symbol.

Parameters A_p and T_p depend on the type of ITU channel used in the modelling of multi-path channel propagation.

Mobile WiMAX standard is an OFDM-based technology. If one designates the set of sub-carriers available for data transmission in each OFDM symbol as N_{data}, the power available at the base station for data transmission (not considering the power boost used in the transmission

of pilot sub-carriers, used in channel estimation) by P_{data}, and if one splits this power uniformly over the set of data sub-carriers, the power assigned for the transmission of data sub-carrier $n, n \in [0,...,N_{data}-1]$ is given by equation (13):

$$P_{data}^{(n)} = \frac{P_{data}}{N_{data}} \tag{13}$$

4.6 Link Level Interface Modelling – General Concepts

The performance evaluation of a practical system by means of simulations has to consider the overall system layers of the whole communication protocol stack: physical layer, link layer (both Logical Link Control – LLC and Medium Access Control – MAC) and upper layers (network, application and transport layers).

Simulations carried out on the radio link level are performed for a point-to-point link between the base and mobile stations, either in a SISO or MIMO propagation channel. The whole transmission chain is simulated at the granularity of the bit level. It includes all modules responsible for the transmission and reception of the signal to either end of the link. The ultimate goal is the generation of a set of curves illustrating the variation of the Bit Error Rate (BER), Block Error Rate (BLER) or Symbol Error Rate (SER) with the Signal to Noise Ratio at the bit level: E_b / N_0.

At the system level, simulations are conducted for a group of base stations, in a typical hexagonal cellular layout, which transmit (downlink connection) and receive (uplink connection) to/from a group of mobile stations attached to its area of coverage (cell). At this level simulations are conducted in a point-to-multi-point configuration, where a group of mobile stations are attached to each cell in the network and the ultimate goal is the generation of a set of metrics, which reflects the performance of the network in terms of: achieved user and cell throughput, packet drop ratio, average packet delay, etc.

Both layers perform simulations under different time-scales: physical layer simulations are performed at the bit level and system level simulations are performed at the frame interval, or transmission time interval.

Although desirable, in terms of the accuracy and validation of the results obtained, it is not practical, in terms of complexity and simulation time, to simulate the whole physical link between the base station and a single mobile station, for all mobile stations in the network. The integration of all layers functionalities would be a huge task for the simulator and it would definitely take a huge amount of time, especially if several scenarios such as: traffic type, environment and cellular layout have to be considered. Therefore, there is a need to adopt a simple model that would be accurate enough to capture the signal statistics and impact on

performance metrics, whilst still maintaining the simulation time frame within an acceptable limit.

For this reason, system performance evaluation is based on a separation among the different layers functionalities:

- On one hand system level performance evaluations do not consider the steps performed on the physical layer for the transmission of each bit of information between both ends of the communication link. System Link level Radio Resource Management (RRM) algorithms performing at the frame interval level of granularity consider the physical layer as a "black box", interacting with it by means of well defined interfaces.
- On the other hand, physical layer simulations are unaware of the algorithms performed on higher layers for RRM, such as Dynamic Channel Allocation (DCA), Power Control, Handover, Connection Admission Control (CAC) and Scheduling.

This separation among layers implies the definition of interfaces which must be properly designed in order to, as accurately as possible, affect system simulation results with the performance of the physical layer. This strategy results in an implicit trade-off: on one hand it allows obtaining results in an acceptable time interval and, on the other, there is the drawback of loosing accuracy in the obtained results and accuracy of the physical layer performance, as seen by upper layers.

The performance of the physical layer is modeled by means of Look-Up Tables (LUT) in which the behavior of the radio link is encapsulated. An example of a metric that could be used in the performance abstraction of the physical layer is the variation of the Frame Error Rate (FER) with the Signal-to-Interference plus Noise Ratio (SINR), averaged over many channel realizations for the specific channel model used.

According to the simulated scenario two types of interfaces can be defined in physical layer abstraction for system level simulations [83-84]:

Average Value Interface – this type of interface reflects the radio link quality for a long time interval. This scenario is typical for mobile speeds corresponding to values of the coherence time smaller than the duration of a single transmission time interval, making it unrealistically to assume the channel constant along one or two radio frames. Only statistical channel behavior is assumed as channel state value is averaged over time. The average value interface is not accurate if there are fast changes in the interference due to, e.g., high bit rate packet users.

Actual Value Interface – this type of interface reflects the instantaneous value of the radio link. It is suitable for scenarios of low mobility, resulting in a slow fading channel profile.

Figure 4 illustrates the processing blocks involved in system level simulations and their dependencies.

These blocks reflect the functionalities implemented in both PHY (left) and MAC (right) layers. It can be observed that:

· The performance of the physical layer, as modeled by the Link to System Interface Module (LSIM), depends on both the channel quality and the interference conditions, and is reflected in the computation of the SINR metric. The SINR depends on the position of each mobile on the network (geometric factor) and on the traffic load due to the total amount of active mobiles in the system.

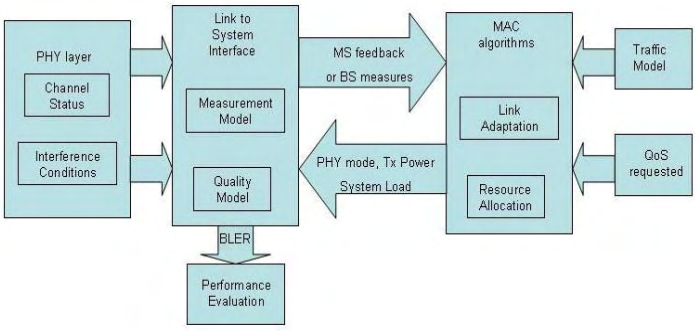


Figure 4 - Overview of Link to System Level Interface

· On the System Level the MAC sub-layer schedules the amount of resources required by each service flow. Its performance depends on the Quality and Measurement Models used. The Quality model is responsible for the estimation of the link performance based on the resource allocation. Link quality estimation is performed by means of the Packet Error Rate (PER) or the Block Error Rate (BLER) and these metrics are the outputs of the Look-Up Tables (LUT). The Measurement Model is the block responsible for the computation of the quality metric used by MAC algorithms such as: channel-dependent scheduling and resource allocation, power control and link adaptation.

4.6.1 Link Level Interface Modelling for Mobile WiMAX System

Figure 5 illustrates the methodology followed in system level simulations and performance evaluations for the OFDM radio technology, in which Mobile WiMAX is based [81]. The following aspects are considered in physical layer modeling and interfacing for OFDM-based air interfaces:

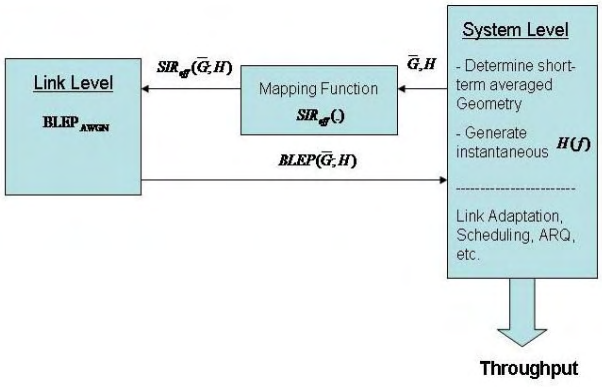


Figure 5 - Schematic view of system level methodology

- On system level a set of mobile stations are randomly dropped over the network deployment area. Depending on its physical location each mobile station is characterized by a Geometric Factor, $\overline{G}$, which depends on the path-loss between the serving and neighbouring cells, shadowing value and thermal noise power. The Geometric Factor is derived by averaging-out the influence of the fast fading component on the propagated signal.
- In each transmission time interval the instantaneous value of the radio channel between each base station and mobile station is derived, assuming only the fast fading component. As the Mobile WiMAX standard is OFDM-based it is preferred to express the channel in the frequency domain, by its frequency response, $H(f)$, which depends on the type of radio scenario being simulated: pedestrian, vehicular, etc. For low/medium Doppler, i.e. low/medium channel coherence time, the channel is assumed constant during a transmission time interval and its instantaneous value is considered in the computation of the SINR used in the LUT (Actual Value Interface).
- Although OFDM uses cyclic prefixes for the limitation of Inter-Symbol Interference (ISI) at each sub-carrier, the performance (SINR) over the whole set of sub-carriers encompassing each OFDM symbol will change due to the influence of the frequency selectivity of the mobile radio channel.
- The output from the System Level Simulation Module is a vector of frequency response channel gains, $H(f)$, and a scalar Geometric Factor, $\overline{G}$. A mapping function $SIR_{eff}(\overline{G};H(f))$ is used in order to map the geometry and the frequency response to an effective SINR value. The interference part of the effective SINR includes both the inter-cell interference and noise, as the intra-cell interference is not considered due to the orthogonality among adjacent sub-carriers.
- The effective SINR is used in the computation of the BLER probability from the basic Additive White Gaussian Noise (AWGN) performance curves which emulate the link level performance.

The only task that remains is the definition of the mapping function, $SIR_{eff}(\overline{G};H(f))$. The mapping function should depend on the exact modulation and coding scheme (MCS) used as well as on other transmission formats (SISO and MIMO), but not on the channel model implemented on the simulations. Any mapping function designed to be used in a system level evaluation must be thoroughly verified by means of link level simulations for different MCS schemes/transmission modes.

4.6.1.1 Effective SIR Mapping Functions

The role of an SINR-based mapping function is to provide the expected BLER/FER of a coded block of information as a function of the vector of SINR values of its data symbols, which can be transmitted through different types of resources (time, frequency, codes or spatial beams), depending on the type of multiple access implemented in the air interface. This calculation may involve the computation of the weights of both pre and post processing matrices, at the transmitter and the receiver, as for example whenever beamforming or spatial multiplexing is used in connection with MIMO.

The vector of instantaneous SINR values, associated to the resources assigned to the transmission of a data block, is mapped into a given set of data sub-carriers and is computed from the corresponding fading amplitudes of the different sub-carriers at the receiver. Due to multi-path propagation the SINR values in this vector are received with different levels of quality and for this reason, at least theoretically, they should be considered separately by the LUT for the generation of the BLER/FER/PER/BER metrics. However, the amount of resource elements in all domains (power, time, frequency, code and space) is far too high for them to be considered individually by the LUT. As a consequence the vector of SINR values must be compressed to a lower dimension order, preferably to a one or two-dimensional vector, before being inputted to the LUT.

Assume an OFDMA based multiple access scheme with SISO, and inter-cell interference, modeled as AWGN, accumulated in the thermal noise component of the SINR. Assume also that at the n^{th} frame interval, the radio block data symbols are mapped into a set of N data sub-carriers of the OFDM symbol. An approach for the derivation of a compressed SINR metric would be the arithmetic mean over the whole set of N sub-carriers, as given by equation (14).

$$SINR_{eff}(n) = \overline{G}\frac{\left(\sum_{k=0}^{N-1}\frac{\left|h_k(n)\right|^2}{\sigma_n}\right)}{N} \tag{14}$$

Where:

- $\overline{G}$ is the actual geometry factor between the user, its serving and neighboring cells, given by equation (11).
- $\left|h_k(n)\right|^2$ is the instantaneous power in the k^{th} data sub-carrier of the sub-carrier set to which the data is mapped into in the n^{th} frame interval.
- N is the size of the sub-carrier set.
- $\sigma_{'n}$ is the noise power which includes interference from other cells (modeled according to Gaussian distribution).

Figure 6 illustrates this compression principle.

The arithmetic mean-based compression scheme is the simplest and less accurate effective SINR mapping. It is not accurate because it averages-out the variations of the channel along the set of sub-carriers, i.e., it underestimates some samples and overestimates others, and the inefficiencies of this approach can be seen from figure 6.

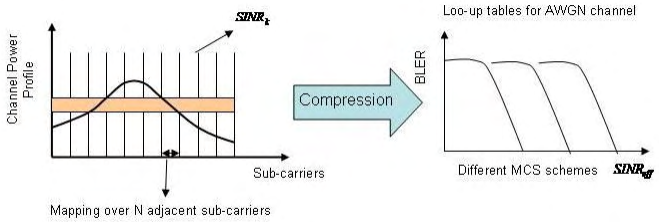


Figure 6 - Illustration of SINR compression and mapping

4.6.1.2 Exponential Effective SINR Mapping (EESM)

Among the different proposals for effective SINR mapping, this is the one which seems to have the most acceptance in the research community, standardization bodies and equipment manufacturers. Its derivation and performance analysis is described in detail in [192-194]. The proposed approach for the mapping/compression of the SINR values, one for each sub-carrier, into a single effective (scalar) SINR value is the Exponential Effective SINR Mapping (EESM). The EESM is used together with link level results for the different MCS schemes on AWGN channels to determine the BLER. The EESM is given by equation (15).

$$SINR_{eff} = -\beta \ln\left(\frac{1}{N}\sum_{k=1}^{N} e^{-\frac{SINR_k}{\beta}}\right) \qquad (15)$$

Where β is a correction parameter used to adapt the formula to the different types of scenarios used in the simulations (different MCS schemes and MIMO techniques) and is independent of the channel model used. It must be optimized from link-level simulations for each type of modulation and coding rate combination used. Also, a subset of the data sub-carriers space can be used to evaluate the effective SINR for reasons of computational efficiency.

Other methods for compressing the vector of SINR values into a single scalar SINR are proposed in the literature. These are described in Annex A.

4.7 MIMO Channel Modelling in System Level Simulations

There are different methods for modelling the MIMO channel at the system level. These methods are grouped into two different categories:

- **Ray-based**: the channel coefficient between each transmit and receive antenna pair is the summation of all rays at each tap of the multi-path filter at each time instant, according to the antenna configuration, gain pattern, angle of arrival (AoA) and angle of departure (AoD) of each ray. The temporal channel variation depends on the travelling speed and direction relative to the AoA/AoD of each ray.

- **Correlation based**: The MIMO channel coefficients at each tap are mathematically generated according to independent and identically distributed Gaussian random variables, according to the antenna correlation and the temporal correlation, corresponding to a particular Doppler spectrum.

All the simulations conducted along this work were performed by means of the 3GPP Spatial Channel Model for MIMO modelling [85]. In this model the channel gain between each pair of antennas, at both ends of the communication link, results from the superposition of the contributions from each individual path of the tapped delay line model.

4.7.1 3GPP Spatial Channel Model

The Spatial Channel Model (SCM) proposed by 3GPP [85-86] is a detailed empirical system level model for simulating urban micro-cell, urban macro-cell and suburban macro-cell fading environments. The MIMO channel is represented as a superposition of clustered constituents, with stochastic powers, angles of departure (AoD) and arrival (AoA), as well as times of arrival. Each cluster emulates the effects of small scale fading mechanisms, which contribute to multi-path propagation. The received signal consists of *N* time-delayed replicas (paths) of the transmitted signal (the number of paths depends on the specific channel model). Each path is associated to a power and delay values, which are randomly generated according to pre-defined random distributions, whose parameters depend on the type of environment used in the simulations. Each path is also associated to clusters of *M* sub paths which are not resolvable at the receiver end.

3GPP SCM channel model has a total of *N=6* paths and each path has *M=20* sub-paths. For the n^{th} path, $\mathbf{H}_n(t)$ denotes the MIMO channel matrix generated by the SCM model. It is a multi-dimensional matrix and is given by equation (16).

$$\mathbf{H}_n(t) = \begin{bmatrix} h_{11}^n & \cdots & h_{1n_T}^n \\ \vdots & \ddots & \vdots \\ h_{n_R1}^n & \cdots & h_{n_Rn_T}^n \end{bmatrix} \tag{16}$$

The entry $h_{n_in_j}^n(t)$ denotes the complex channel gain (complex amplitude) between the n_i receiving antenna and the n_j transmitting antenna for the n^{th} path. It is generated by the superposition of a number of sinusoidal signal components corresponding to each sub-path. The number of rows is equal to the number of receiving antennas, N_R, in the base station linear antenna array and the number of columns is equal to the number of transmitting antennas, N_T, in the mobile station linear antenna array.

The overall procedure for generating the channel matrices consists of three basic steps:

1. Specification of the environment in which the empirical channel model is to be used: suburban macro, urban macro or urban micro.
2. Derivation of the channel parameters to be used in the simulations associated with that environment.
3. Generation of the channel coefficients based on the derived parameters. This step gives the complex channel gains of the channel matrix $\mathbf{H}_n(t)$ for each path.

Once the scenario has been chosen and the location of the base stations (with desired geometry and inter-base distances) has been determined, one may start instantiating users in the area of interest. This entails first randomly generating the user locations and specifying other user-specific quantities such as their velocity vector **v**, with its direction drawn from a uniform [0,360º) distribution. The specifics of the mobile station antenna or antenna array have to be determined, such as the array orientation, Ω_{MS}, also drawn from a uniform [0,360º) distribution. The derivation of the channel parameters in step 2 is based on the execution of random distributions whose configuration parameters were previously derived from measurement campaigns, and which depend on the type of environment defined in step 1.

Figure 7 depicts the 3GPP SCM channel model for MIMO simulations where only one cluster of scatterers is shown. Two uniform linear arrays at both ends of the communication path, with 2 antennas each, are illustrated in this figure.

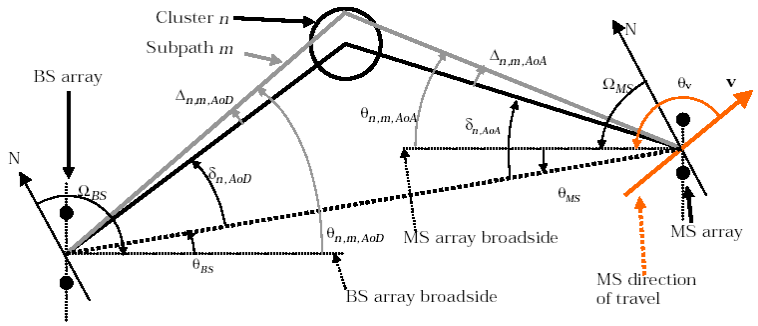


Figure 7 - SCM model for MIMO channel system level simulations

The following parameters are derived directly from the cell layout, mobile positions and input configurations:

- θ_{BS} is the angle between BS-MS line-of-sight and the BS broadside array.
- θ_{MS} is the angle between BS-MS line-of-sight and the MS broadside array.
- d_{n_T} is the distance, in meters, from base station antenna element n_T from the reference $n_T = 1$ antenna (for the reference antenna $n_T = 1$, $d_1 = 0$).
- d_{n_R} is the distance in meters from mobile station antenna element n_R from the reference $n_R = 1$ antenna (for the reference antenna $n_R = 1$, $d_1 = 0$).

The remaining parameters such as the angle of arrival (AoA) and departure (AoD) of each ray are generated according to specific distributions and parameterization which depend on the type of environment and channel model being simulated. These parameters are assumed as constant over the whole simulation (in case of mobile stationarity) or as changing in a very slowly time scale:

- $\theta_{n.m,AoD}$ is the angle of departure (AoD) for the m^{th} sub path of the n^{th} path at the base station with respect to the base station broadside. It is computed as follows: $\theta_{n.m,AoD} = \theta_{BS} + \delta_{n,AoD} + \Delta_{n,m,AoD}$. The parameters $\delta_{n,AoD}$ and $\Delta_{n,m,AoD}$ denote, respectively, the AoD and the offset for the m^{th} sub path of the n^{th} path with respect to $\delta_{n,AoD}$.
- $\theta_{n.m,AoA}$ is the angle of arrival (AoA) for the same m^{th} sub path of the same n^{th} path with respect to the mobile station broadside. It is computed as follows: $\theta_{n.m,AoA} = \theta_{MS} + \delta_{n,AoA} + \Delta_{n,m,AoA}$. The parameters $\Delta_{n,m,AoA}$ and $\delta_{n,AoA}$ denote, respectively, the AoA and the offset for the for the m^{th} sub path n^{th} path with respect to $\delta_{n,AoA}$.

These parameters are used in the computation of the complex channel $h^n_{n_R n_T}(t)$ between receiving antenna n_R and transmitting antenna n_T, according to equation (17).

$$h^n_{n_R,n_T}(t) = \sqrt{\frac{P_n \sigma_{SF}}{M}} \sum_{m=1}^{M} \begin{pmatrix} \sqrt{G_{BS}(\theta_{n,m,AoD})} \exp(j\left[kd_{n_T} \sin(\theta_{n,m,AoD}) + \Phi_{n,m}\right])x \\ \sqrt{G_{MS}(\theta_{n,m,AoA})} \exp(jkd_{n_R} \sin(\theta_{n,m,AoA}))x \\ \exp(jk\|\mathbf{v}\|\cos(\theta_{n,m,AoA}) - \theta_{\mathbf{v}})t) \end{pmatrix} \quad (17)$$

- P_n is the power associated to the n^{th} path, which is modelled as an exponential Power Delay Profile (PDP) by the model.
- σ_{SF} is the lognormal shadow fading applied to the *N* paths for a given mobile drop.
- $G_{BS}(\theta_{n.m,AoD})$ and $G_{MA}(\theta_{n.m,AoA})$ are, respectively, the base station and mobile station antenna gains of each array element.
- $k = 2\pi / \lambda$ is the wave number.
- $\|\mathbf{v}\|$ is the magnitude and θ_v is the angle of the mobile station velocity vector.

For a given scenario, realizations of each user's parameters such as: the path delays, powers and sub-paths, angles of departure and arrival can be derived using the procedures described in detail in [85].

The following steps are required to generate the channel element of one sub-path of a multipath from the link between a desired user receive antenna and a particular transmit antenna at the base station:

- *Step 1. Choose an environment.*
- *Step 2. Determine distances and orientation parameters.*
- *Step 3. Determine the Delay Spread (DS), Angle Spread (AS) and Shadow Fading (SF) components. DS and SF are correlated and log-normally distributed. AS is also log-normally distributed and correlated to DS and SF.*
- *Step 4. Determine random delays* ($\tau_n, n = 1,...,N$) *for each of the* N *multipath components.*
- *Step 5. Determine random average powers* (P_n, $n = 1,...,N$) *for each of the* N *multipath components.*
- *Step 6. Determine AoDs for each one of the* N *multipath components at the base station.*
- *Step 7. Associate the multipath delays with AoDs.*
- *Step 8. Determine the powers, phases and offsets AoDs of the* $M = 20$ *sub paths for each one of the* N *multi-paths at the base station.*
- *Step 9. Determine AoAs for each one of the* N *multipath components at the mobile station.*
- *Step 10. Determine the offset AoAs at the mobile station for each one of the* $M = 20$ *sub paths of each one of the* N *paths.*
- *Step 11. Associate the base station and mobile station paths and sub paths.*
- *Step 12. Determine the antenna gains of the base station and mobile station sub paths as a function of their respective sub path AoDs and AoAs.*
- *Step 13. Apply the path-loss based on the base-station to mobile station distance from Step 2 and the log normal shadow fading determined in step 3 as large-scale parameters to each one of the sub-path powers of the channel model.*
- *Step 14. Generate the channel coefficients.*

The SINR at each receive antenna is computed for each path of the model and they are all combined at the SINR level (using the EESM mapping method). In order to do so interference from all other interferers is explicitly modelled in system-level simulations. It is assumed that due to OFDM cyclic prefix transmitted in each symbol intra-cell interference is neglected, i.e., only inter-cell interference is assumed in the simulations.

Figure 8 below illustrates the scenario assumed (cellular and base station layout, as well as mobile stations and antenna array orientation) for conducting MIMO channel simulations in the system level simulator. In particular, it differs from figure 7 (reprinted from 3GPP

specifications), as it describes the steps followed in the computation of the values assumed for the different parameters, which are used in the MIMO channel model for conducting system level simulations, according to 3GPP specifications.

After deploying the cells and base stations in the network layout, the angles θ_{BS} and Ω_{BS} result automatically defined. The angles θ_{MS} and Ω_{MS}, regarding each mobile station dropped in the network, depend on the orientation of each mobile station antenna array and on the direction followed by each mobile moving along the coverage area pertaining to its serving base station. These parameters are computed and modified dynamically, as the simulation evolves.

As mentioned before, these angles are central to the computation of the complex gains for each pair of transmitting and receiving antennas in the MIMO channel matrix.

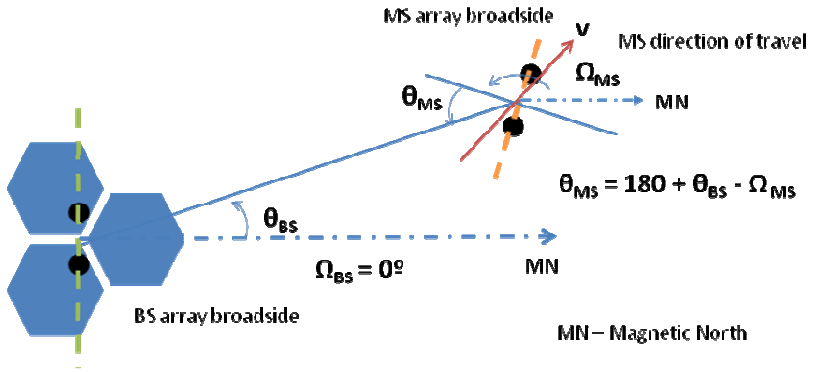


Figure 8 - Network layout for MIMO channel computation

4.7.2 Modelling the SINR at the Mobile Station

In what follows the downlink communication is assumed (although the same explanation is also valid for the uplink communication). Also, the channel can be assumed as symmetrical because all system level simulations are performed for the time division duplex mode (TDD). As it was mentioned in previous sections, for the computation of the performance metrics in system level evaluations, the figure of merit used is the BLER, which is obtained from the look-up tables, according to the SINR value computed at the mobile station. The derivation of SINR for the MIMO channel is performed in two separate parts:

(i) Computation of the desired user radio signal arriving from the mobile station's serving cell.

(ii) Computation of the interfering user radio signal arriving from neighbouring cells.

4.7.2.1 Modelling the Desired User Signal at the Mobile Station

The desired signal from the serving cell is computed according to what is described in the previous section. After the derivation of the channel matrix for each one of the N paths the channel response for each pair of transmitting-receiving antenna is derived from equation (17) by combining the contributions from all paths.

4.7.2.2 Modelling Inter-Cell Signal Interference at the Mobile Station

Although the spatial characteristics of the signals received from the serving as well as from interfering cells can be modelled according to the empirical MIMO channel model, it is very complex and computationally intensive to explicitly model the MIMO channel from all interfering cells, especially those cells from which receiving powers are relatively weak. The performance difference achieved when modelling signals from relatively weak interferers as spatially white (ignoring their spatial characteristics) is negligible.

In a MIMO channel the modelling of other-cell interference is done by considering three types of neighbouring cells: near strong cells whose interference is modelled by a MIMO channel, near weak cells whose interference is modelled as wideband (frequency-selective) channel and far sectors whose interference is modelled as narrowband (flat) channel. According to this approach, and to simplify system performance evaluation, the set of neighbouring cells can be divided into three groups: MIMO (list-A), SISO wideband list (list-B) and SISO narrowband list (list-C). These lists are filled according to the values of path-loss and shadowing to each cell, i.e., the cells are ranked in order of the received power. Strongest A cells are inserted into list A whose size is a trade-off between computational complexity and performance. 3GPP recommends a number of 8 cells for list-A in a 3-sectored cell deployment [85]. The next B cells below list A in the rank are inserted as members of list B and the remaining ones are inserted as members of list C.

Interference from Strong Interferers

The strong interferers are modelled according to the 3GPP SCM MIMO channel model. The interference from one interferer over MIMO coming to each receive antenna is collected from all multi-paths ($N_T x N$). Then the total interference for each receiving antenna is computed as the sum of the interference from all interferers in the list-A, according to equation (18).

$$I_{lis-A} = \sum_{a \in list-A} \sum_{i=1}^{N_T} \sum_{j=1}^{N} P_{a,i,j} \tag{18}$$

Where $p_{a,i,j}$ is the received power over time from the i^{th} transmit antenna for the j^{th}, $j=1,...,N$ path of the wideband channel model of the a^{th} cell in list-A.

Interference from Weak Interferers

The weak interferers are modelled as spatially while Gaussian noise processes whose variances are based on a multi-path Rayleigh fading process (wideband SISO channel), depending on the simulation environment. The fading processes for each cell and receive antenna are independent and equivalent for each mobile receive antenna. The total received noise power from cells in B list, at the n_R^{th} antenna, is given by equation (19)

$$I_{list-B} = \sum_{b \in list-B} \sum_{i=1}^{N} p_{b,n_R,i} \tag{19}$$

Where $p_{b,n_R,i}$ is the received power over time for the n_R^{th} receive antenna, coming from the i^{th}, $i=1,...,N$, path of the wideband channel model of the b^{th} cell in list-B.

Interference from list-C Interferers

Interferers in list C are modelled as spatially white Gaussian noise processes whose variances are based on a flat Rayleigh fading process (single path). The fading processes for each cell and receive antenna are independent and the fading is equivalent for each mobile receive antenna. The total received noise power at the n_R^{th} receive antenna, due to all cells in list C is given by equation (20).

$$I_{list-C} = \sum_{c \in list-B} p_{c,n_R} \tag{20}$$

Where p_{c,n_R} is the received power over time for the n_R^{th} receive antenna, coming from the c^{th} cell in list C.

The total amount of inter-cell interference on the n_R^{th} antenna is given by the contribution from cells in the three lists, according to equation (21):

$$I_{total} = I_{list-A} + I_{list-B} + I_{list-B} \tag{21}$$

4.7.3 Link to System Interface for MIMO channel

The implementation of a MIMO scheme in the radio access of a wireless system is intended to increase transmission diversity in order to improve the BER, increase the transmitted data rate or trade-off both. Transmit diversity is achieved by means of Space Time Block Coding schemes (STBC) of which the Alamouti's STBC [87], with 2 antennas at the base station and one antenna at the mobile station, is an example. The 2x2 Alamouti's STBC scheme turns out to be a receiver simplification because the coding is performed at the transmitter. In the 2x2 Alamouti's STBC scheme a Maximum Ratio Combiner (MRC) receiver increases the diversity of the signal at the mobile station receiver. Vertical 2x2 BLAST (VBLAST) [88-91] is an example of a spatial multiplexing scheme where two independent symbols are transmitted at each antenna in the base station and recovered at the mobile's antenna array with a Maximum Likelihood Decoder (MLD), Minimum Mean Square Error (MMSE) or Zero Forcing (ZF) receiver. The modelling of the transmission chain of the link layer depends highly on the type of MIMO scheme implemented as well as on the performance results obtained.

The WiMAX Forum has selected two different multiple antenna profiles for use on the downlink. In Mobile WiMAX the first multiple antenna profile is the simple 2x1 or 2x2 Alamouti STBC scheme referred to as Matrix A in the specifications. In OFDMA-based WiMAX system this technique is applied subcarrier by subcarrier. The second multiple antenna profile included is the 2x2 SM scheme referred to as Matrix B in the specifications. Only Matrix A MIMO was implemented in the simulator.

Matrix A MIMO Implementation

Matrix A is denoted as $\begin{bmatrix} s_1 & s_2^* \\ s_2 & -s_1^* \end{bmatrix}$. The rows represent the 2 antennas at the transmitter and the columns represent two adjacent transmission OFDM symbols. During the first transmission period transmit antenna 1 transmits symbol s_1 and transmit antenna 2 transmits symbol s_2. During the second transmit period transmit antenna 1 transmits symbol s_2^* and transmit antenna 2 transmits symbol $-s_1^*$. The optimum receiver estimates the transmitted symbols s_1 and s_2 as illustrated in equation (22).

$$\begin{aligned} x_1 &= \left(|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2\right)^2 s_1 + n_1 \\ x_2 &= \left(|h_{11}|^2 + |h_{12}|^2 + |h_{21}|^2 + |h_{22}|^2\right)^2 s_2 + n_1 \end{aligned} \quad (22)$$

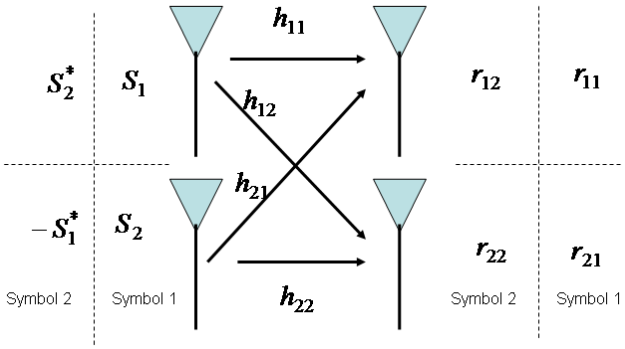


Figure 9 - STBC Alamouti scheme

In equation (22) s_1 and s_2 correspond to the signal at receiving antenna 1 and 2, respectively. These expressions clearly show that the impact of the Alamouti scheme is an enhancement of the channel conditions by a fourth diversity order. This corresponds to improvements in the BER. It is also clear that the two signals can be completely separated at the receiver as they do not interfere to each other. The SINR for the information transmitted along two consecutive symbols and subcarrier k is given by equation (23).

$$SINR_k = \overline{G}\left(\left|h_{11}^k\right|^2 + \left|h_{12}^k\right|^2 + \left|h_{21}^k\right|^2 + \left|h_{22}^k\right|^2\right) \quad (23)$$

Where $\overline{G}$ is the user's geometric factor. This value is used in the computation of the compressed scalar SINR value for the derivation of the BLER.

4.8 Traffic Models

In the simulations widespread traffic models have been used for system validation and performance results under the proposed set of DRA. These are traffic models at the IP packet level:

- Full Queue (FQ) traffic model in which it is assumed that there is an infinite amount of data bits waiting in the queue of each active user in the system. That is, users are designated as backlogged. This traffic model is particularly interesting in accessing the maximum capacity of the network.
- Voice over IP traffic model.
- Near Real Time Video with an average source bit rate of 32 kbps, 2 Mbps and 10 Mbps.
- World Wide Web (WWW) traffic model with a source bit rate of 64 kbps, 2 Mbps and 10 Mbps.
- File Transfer Protocol (FTP) traffic model with a source bits rate of 64 kbps and 384 kbps.

Annex B provides a detailed description of each one of these traffic models.

4.9 Performance Metrics

In the execution of each transmission time interval a number of statistics are collected for the computation of the metrics used in the evaluation of the performance of the system level simulation platform. These performance statistics are generated as outputs from the system level simulations and are used in the performance evaluation of the used scenarios and proposed algorithms. The following parameters are used as inputs for the computation of performance metrics:

- Simulation time per run: T_{sim}.
- Number of simulation runs: D.
- Total number of cells being simulated: N_{cells}.
- Total number of users in cells of interest (cells being simulated): N_{users}.
- Number of packet calls for user u: p_u.
- Number of packets in i^{th} packet call of user u: $q_{i,u}$.

Annex C presents an exhaustive list with all performance metrics used in all system level simulations performed in the scope of this work.

4.10 Related Work

There is some work available in the research literature regarding the simulation of mobile networks for B3G and 4G evolutions such as WiMAX and HSDPA. Some of these proposals are based on the use of proprietary simulation platforms and others are based on open platforms for system simulation such as Network Simulator 2 (NS2).

In [95] the authors propose to estimate the capacity of a WiMAX network using a module developed for NS2 simulation tool. Different scheduling algorithms are considered for the

estimation of the spectrum efficiency of a WiMAX network scenario and the results obtained are compared against theoretical values.

[93-94] presents the performance of a WiMAX OFDMA system, based on system level simulations for multi-user scheduling algorithms in the frequency domain. It is shown that multi-user scheduling in frequency domain can potentially improve OFDMA system efficiency in frequency-selective broadband channels.

In [96-97] a Fixed WiMAX network, based in the IEEE 802.16d standard is simulated at the system level. Detailed performance comparisons of two different scenarios with tight frequency reuse schemes are presented.

In [98] detailed link and system level simulations have been performed in interference limited cellular environments for tight 1x1 and 1x3 frequency reuse schemes. The authors conduct exhaustive system level simulations to infer about the dependency of the achieved service throughput, modulation and coding distribution and channel utilization on the applied system load.

In [99] a similar work to the one described in this chapter is performed. Namely, the authors elaborate on the implementation of a DRA protocol architecture based on cross-layer signalling in MC-CDMA networks. The proposed DRA is intended to support a very large amount of users with inherent flexibility and granularity necessary to support heterogeneous traffic with limited complexity. Extensive system level simulations are performed to evaluate the DRA performance under such network scenario.

In [100] an extensive set of simulations are performed for verifying the effectiveness of Mobile WiMAX QoS mechanism for different types of traffic profiles, in managing traffic generated by data and multimedia sources. The authors concluded that the performance of the system depends on such factors as the frame duration, the mechanisms used for requesting uplink bandwidth and offered load. Uplink and downlink packet scheduling and transmission are simulated.

In [101-102] dynamic system level simulations are performed to infer whether commonly used schedulers available in the literature are able to guarantee QoS requests of VoIP traffic users, on a scenario of mixed VoIP and WWW traffic users. Algorithms are divided into two different sets: QoS-differentiated and non-QoS-differentiated algorithms.

In [103] it is proposed a new scheduling approach based on an strategy where users are temporary removed from the active set of users considered in the scheduler if the channel quality is lower than a given admission threshold. The temporary removal is easily combined with any conventional scheduling technique providing considerable performance benefits. An exhaustive set of system level simulations are conducted to infer about the benefits of such strategy over simpler scheduling algorithms (without the temporarily removal strategy).

[104] presents the results from extensive simulations to evaluate the downlink performance of Mobile WiMAX employing MIMO channel and [105] analyses the performance of the same

Mobile WiMAX networks according to different PHY layer configurations. Handover and link adaptation are jointly used to limit the amount of inter-cell interference and improve cell coverage and service satisfaction.

4.11 Conclusion

The evaluation of the performance and suitability of a new wireless standard, such as Mobile WiMAX, demands that most studies are conducted by means of simulators which effectively model all the key features of the system, according to the communications protocol layers.

Network simulations encompasses two domains of simulations which run at two different time-scales: link level simulations run at the time granularity of each single transmitted bit and system level simulations run at the granularity of each transmitted packet over the network. As explained along the chapter, it is virtually impossible to perform network simulations with both domains working together, either due to the low speed of simulations, or to the extremely high complexity resulting from such an approach. This motivates the development of two separated sets of simulations, one for each level. The implementation of independent simulators for link and system levels demand for the definition of proper interfaces for using the outputs from link level simulators in modelling its behaviour and performance and which can influence the system level simulator functioning behaviour.

This chapter details the different aspects which must be considered in the implementation of a general system level simulator, for conducting system level simulations. The modelling framework which is normally followed in designing a system level simulator is presented. Specifically, a system level simulator must contain models for the different components of signal propagation (path-loss, shadowing and fast fading), signal interference, user mobility and traffic models. This chapter explains in detail the models used for the different components of signal propagation and the methodology followed in the derivation of the vector of SINR values, as well as the methodology used in the compression of this vector to a scalar, suitable for the derivation of the BLER value. The LUT models link level functionally and behaviour.

The definition of properly designed link-to-system-level interfaces, for abstracting physical layer performance to system level, is also presented in detail and examples of different approaches proposed in the literature are mentioned.

The Dynamic Resource Allocation module (DRA), whose architecture is detailed in chapter 5, is now to be plugged into this system level architecture design. Validation of the proposed system level architecture, with the use of a cross-layer based DRA is the subject of chapter 6, in which the different procedures to be followed in this validation and the performance metrics to be derived are presented.

Chapter 5

Dynamic Resource Allocation Architecture for Mobile WiMAX

5.1 Introduction

The Mobile WiMAX standard was designed and developed from the outset for the delivery of broadband applications. In particular, its MAC layer has inherent features designed for the joint support of burst data traffic applications (with high peak data rate demands) and streaming applications (which are delay-sensitive regarding the time instant of packet transmission).

The fine granularity and flexibility provided by the MAC layer in resource allocation, according to user's bandwidth needs, the lower latency incurred in handling user's bandwidth requests, and in making scheduling decisions, makes it possible to send data through the air-interface, under the stringent quality of service (QoS) requirements of each type of service flow, and the

efficient use of radio resources, with the consequent maximization of the achieved spectrum efficiency. These functionalities are exploited by the packet scheduler in the Dynamic Resource Allocation (DRA) architecture, which is able to combine the peculiarities of each type of service being supported in the system. By itself, the design of the architecture and the implementation of the DRA is one of the most fundamental problems in wireless networks. Its architecture encompasses the scheduler, the radio resource manager (RRM), the connection admission controller (CAC) and the link adaptation (LA) modules.

Mobile WiMAX specifications define only the PHY and MAC layers. The DRA architecture design and implementation is based on these specifications. The proposed DRA exploits cross-layer information to provide efficient mapping of data onto radio resources and ensures seamless connectivity for the different service flows at the MAC layer. The cross-layer architecture framework is accomplished through a number of uplink control channels in the uplink sub-frame (in TDD mode). These channels are tailored for the fast exchange of information for cross-layer operation and are used in the signalling interaction between the base and mobile stations. This particular design of the TDD frame makes it possible for the DRA to rapidly adapt the packet transmission to the dynamic propagation and interference conditions, as well as to the traffic demand fluctuations.

In spite of the significant work which has been done by the research community, regarding the performance evaluation of WiMAX networks, pertaining to the scheduling of Real Time (RT) and Non-Real-Time (NRT) service flows, in both fixed (IEEE802.16-2004) [50] and mobile (IEEE802.16e) WiMAX [49], very few of the different approaches consider practical DRA architectures, which can be implemented in practical scenarios. Also, the performance of many of the implemented algorithms is based upon analytical models applied to very simple cellular layouts, whereby a single cell broadcasts, in a point-to-multi-point configuration, to a number of mobile stations. In such scenarios no neighbouring cells are considered and mobility and/or handover are not implemented. Traffic models are not implemented and mobiles are normally assumed as backlogged all the time.

This chapter presents, in great level of detail, all the steps followed in the design of the architecture proposed for the implementation of a DRA module for Mobile WiMAX. This DRA is to be embedded into the system level simulation platform presented in chapter 4 and used in the implementation of the packet schedulers detailed in chapters 7, 8 and 9. It is fully compliant to the standard IEEE 802.16e, defined for the deployment of Mobile WiMAX networks.

This chapter is organized as follows. Section 2 is about the system profile used in system level simulations and, in particular, in the DRA implementation. The system profile is in accordance to the one proposed by the WiMAX Forum, as an attempt to harmonize simulations and results from different partners working in the field. It presents the parameter values chosen from the plethora of options available in the WIMAX standard. This choice will influence the DRA

design. Section 3 explains how the radio resources are modelled in the DRA. Radio resources are formed by grouping the OFDM symbols and sub-channels, which are simulated in the system level tool. The size of each resource unit and the number of resource units available influence the performance of the scheduler. Section 4 is the core of this chapter in a sense that it describes in detail all the steps followed in the design and implementation of the modules for the proposed DRA. Section 5 presents the state-of-the-art available in the research literature, regarding system-level evaluations performed under WiMAX networks. Section 6 concludes the chapter.

5.2 System Profile for Mobile WiMAX DRA

In order to coordinate results and architectural implementations in the telecommunications community, including researches, product developers, and service providers, the Application Working Group of the WiMAX Forum has developed a standard simulation methodology that describes the key parameters and features of the Mobile WiMAX standard [107]. This methodology should be used by anyone performing system level simulations and producing performance figures, for a fair comparison of the performance of Mobile WiMAX against other potential competitors, such as 3GPP Long Term Evolution (LTE), in the investigation about the suitability of Mobile WiMAX as a potential 4G wireless standard.

It is important to understand the differences between WiMAX versions of the IEEE 802.16 standard, in the sense that real networks envision to implement only a subset of the features and parameter values allowed. Therefore, in order to compare performance results from different research studies or vendors, it is important that system-level simulations are conducted on a similar set of features and parameter values, and be representative of real-world equipment, according to the methodology and system profile proposed in [107].

For Mobile WiMAX (IEEE 802.16e) standard the WiMAX Forum has approved several profiles using the orthogonal frequency division multiple access (OFDMA) air-interface, with bandwidths ranging from 1.25 to 20 MHz. The separation between adjacent sub-carriers is kept constant no matter the chosen bandwidth, as the number of sub-carriers is set proportional to the width of the spectrum available. This is designated as scalable OFDMA (SOFDMA).

This work follows the guidelines from the WIMAX Forum system profile for Mobile WiMAX.

In this system profile the duration of the TDD frame is equal to 5 ms (this is the corresponding transmission time interval) and the size of the Fast Fourier Transform (FFT) is 1024 sub-carriers, which corresponds to a channel bandwidth of 10 MHz. Table 1 lists the parameters used in the DRA architecture implementation regarding the physical layer.

In the TDD frame, resources are available in two domains: frequency (sub-channels) and time (OFDM symbols). Each sub-channel comprises a number of sub-carriers according to the type

of sub-channelization used and direction of the connection. For the different types of sub-channelization schemes implemented in the Mobile WiMAX standard please refer to chapter 3.

Parameters	**Values**	
System Channel Bandwidth (MHz)	10	
Sampling Frequency (F_p in MHz)	11.2	
Subcarrier Frequency Spacing (f kHz)	10.94	
FFT Size (N_{FFT})	1024	
	DL	UL
Null Subcarriers	184	184
Pilot Subcarriers	120	280
Data Subcarriers	720	576
Data Subcarriers per Subchannel	24	16
Number of Subchannels (N_S)	30	35
Useful Symbol Time ($T_b = 1/f$) in μs	91.4	
Guard Time ($T_g = T_b/8$) in μs	11.4	
OFDM Symbol Duration ($T_s = T_b + T_g$) in μs	102.9	
Number of OFDMA Symbols per frame (5ms)	48	
Data OFDM Symbols	44[1]	

TABLE 1 – MOBILE WIMAX SYSTEM PROFILE

The system profile determines that DL PUSC (Downlink Partial Usage Sub-Channelization) sub-channelization mode be implemented as a mandatory feature in the system and that DL FUSC (Downlink Full Usage Sub-Channelization) sub-channelization mode may be implemented also, but as an optional feature. In the implementation of the DL-PUSC mode 720 data sub-carriers are organized into three segments of 240 data sub-carriers each and each sub-channel comprises 24 data sub-carriers. Therefore, there are 30 sub-channels available per OFDM symbol in the frame and 10 sub-channels per segment. The sub-carriers in the preamble are divided into 3 segments and the channel quality information (CQI) is estimated from a sample of the sub-carriers inside each segment. Three different measures are used to estimate the CQI, one for each segment, and all sub-channels inside each segment are described by the same CQI associated. This segmentation scheme is particularly useful for interference limitation amongst adjacent cells of the same base station, if each single segment is assigned to each single cell.

The map of resources associated to the PHY radio frame comprises a number of slots. Each slot is the smallest resource granularity which can be allocated for data transmission. A group of slots using the same modulation and coding scheme (MCS) constitutes a burst. A burst can be assigned to more than one user provided the same MCS scheme is followed in the transmission of all packets mapped into the slots of the burst.

[1] 1 OFDM symbol for Preamble, 2 OFDM symbols for TTG and RTG and 1 more for synchronization, guard and alignment issues.

In DL-PUSC configuration each sub-channel comprises 24 data sub-carriers and each slot comprises a resource space equal to two OFDM symbols by one sub-channel. Table 2 lists the system parameters regarding the implementation of the DL-PUSC-based DRA in the basic system level simulator platform described in chapter 4.

Parameter	Value
Number of DC subcarriers	1 (index 1024, counting from 0)
Number of guard subcarriers, left	160
Number of guard subcarriers, right	159
N_{used}, number of used subcarriers (which includes the DC subcarrier)	1729
Total number of subcarriers	2048
Number of pilots	192
Number of data subcarriers	1536
Number of physical bands	48
Number of bins per physical band	4
Number of data subcarriers per slot	48

TABLE 2 - WIMAX PHY INFORMATION – UL/DL PUSC SUB-CHANNELS

Table 3 lists the size of the transport block carried in each slot for each one of the MCS schemes implemented in the DRA using DL-PUSC. These are the basic constituents of each resource addressed in the MAC layer for data transmission.

MCS Level	Modulation	Coding Rate	Symbol size (bits)	FEC block size (in bits
0	QPSK	1/2	24	48
1	QPSK	2/3	32	64
2	QPSK	3/4	36	72
3	16QAM	1/2	48	96
4	16QAM	2/3	64	128
5	16QAM	¾	72	144
6	64QAM	½	72	144
7	64QAM	2/3	96	192
8	64QAM	¾	108	216

TABLE 3 - DATA RATES FOR MCS LEVEL

From a resource managing point of view it is possible to use all sub-channels available in the frame within neighbouring cells, amounting to a 1/1/1 frequency reuse factor from an interference perspective. Alternatively, sub-channel segmentation may be employed to divide the available sub-channels into three segments, each allocated to one of three cells within a base station, amounting to a 1/3/1 reuse factor from an interference perspective.

Although the Adjacent Multi-Carrier sub-channelization (AMC) scheme is not included in the set of mandatory features in the first commercial release of a Mobile WiMAX network (according to the system profile from the WiMAX Forum), the performance of a DRA using AMC as the basic sub-channelization scheme is analyzed in chapter 9 of this thesis. In AMC sub-channelization mode each band comprises 4 bins and each bin comprises 8 data subcarriers. With 48 physical bands there exists a total of 48*4=192 bins with 192*8 = 1536 sub-carriers available for data (8 data sub-carriers per bin). For more details regarding AMC sub-channelization please refer to chapter 3. Table 4 lists the system parameters regarding the implementation of the DL-AMC-based DRA.

Parameter	Value
Number of DC subcarriers	1 (index 1024, counting from 0)
Number of guard subcarriers, left	160
Number of guard subcarriers, right	159
N_{used}, number of used subcarriers (which includes the DC subcarrier)	1729
Total number of subcarriers	2048
Number of pilots	192
Number of data subcarriers	1536
Number of physical bands	48
Number of bins per physical band	4
Number of data subcarriers per slot	48

TABLE 4 - WIMAX PHY INFORMATION DL/UL AMC

5.3 Implementation of the Map of Resources

In Mobile WiMAX the basic resource granularity for data transmission is the slot, whose size depends on the type of sub-channelization and on the direction of the communication: either downlink or uplink. Different slots may be grouped together in the bi-dimensional space of the OFDMA frame to compose a burst of data for transmission. Actually, each burst is sub-divided into a group of Forward Error Correction Blocks (FECs) which are mapped to individual sets of resource units with the same number of slots in the resource space of the OFDMA frame. Protocol Data Units (PDU) arriving from the MAC layer are mapped into these FECs. If necessary they are fragmented and/or concatenated to fit the space available. Padding bits are used in this process.

All slots in the same burst must use the same MCS scheme and may contain information for different users in the cell. Therefore, resource allocation in WiMAX is a bin allocation optimization problem and each user must decode all FECs in the burst to determine its destination. Figure 1 is a schematic description of the Resource Allocation Map (RAM) corresponding to the downlink sub-frame of the TDD mode for the DL-PUSC sub-channelization scheme. As can be seen from the figure, the resource allocation space is a matrix of sub-channels per OFDM symbols.

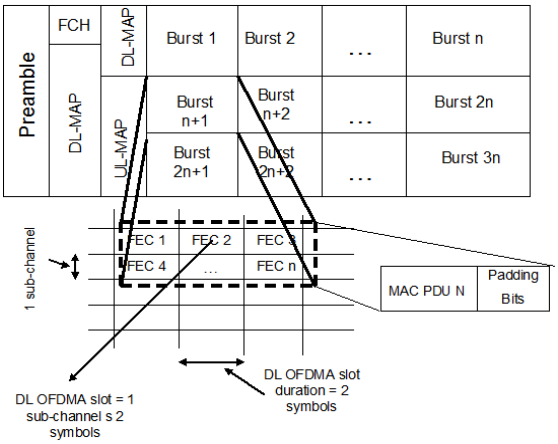


Figure 1 - WiMAX Resource Allocation Map (RAM)

Not all OFDM symbols in the frame are available for data transportation. Some of these symbols are used for conveying signaling and control information and also for channel estimation. This constitutes an overhead whose size depends on the number of users scheduled in the frame and on the size of each burst in slots [108]. In particular, the total amount of users which can be scheduled per frame depends on the type of MCS scheme used in the transmission of the downlink and uplink Mobile Application Part (MAP) control regions, defined after the preamble in the downlink sub-frame. As MAP messages are broadcasted to all mobiles in the cell and they contain vital information for the decoding of the data bursts, they must be properly decoded even for those users in the cell edge, which are more affected from inter-cell interference. As a consequence they must be transmitted with the most robust MCS scheme (rate 1/2 convolution coding (CC) and modulated with Quaternary Phase Shift Keying (QPSK)) and with a higher degree of protection by means of repetition coding (coded symbols are repeated one, two, four or six times so that mobile stations in the cell edge can successfully decode them). Therefore, the number of OFDM symbols available for data transmission decreases as the number of users scheduled in the frame increases and/or more robustness is used in transmission of DL/UP-MAP signaling.

An example of the problematic nature of the overhead creation is presented in [63], in which two scenarios of 5 and 10 users scheduled per frame for each downlink and uplink transmission and 10 MHz of channel bandwidth is assumed. Table 5 enumerates the MAP Information Elements (IE) and respective sizes which comprise the DL and UL MAP fields of the TDD PHY frame in Mobile WiMAX. For the IEs listed in table 5, table 6, also from [63], illustrates the computation of the total MAP overhead, in OFDM symbols, as well as the amount of OFDM symbols available for data transmission.

MAP IEs	**Size (bits)**
Fixed compressed MAP (DL+UL+CRC)	152
Ranging region allocation IE (3 IEs: initial, periodic and bandwidth request IEs)	168
Fast feedback region allocation IE	32
HARQ ACK (CQICH) region allocation IE	56
UL interference and noise level IE	28
Fixed overhead in HARQ DL MAP IE	72
Fixed overhead in HARQ UL MAP IE	64
UL HARQ per scheduler user	40
DL HARQ per scheduler user	44

TABLE 5 – IES AND RESPECTIVE SIZES (IN BITS) IN DL/UL MAP FIELDS

MAP overhead symbols with repetition o0f 6	**5 users scheduled per frame**	**10 users scheduled per frame**
MAP overhead symbols with Rep = 6	10	12
Other overhead symbols (Preamble, guard time, …)	5	5
Symbols for (DL+UL) data transmission for Rep = 6	33	31
MAP overhead symbols with Re = 4	6	8
Symbols for (DL+UL) data transmission for Rep = 4	37	35

TABLE 6 - SIGNALING OVERHEAD AND USEFUL SYMBOLS FOR DATA TRANSMISSION

According to table 6, if 10 users are scheduled per frame (uplink and downlink) the amount of overhead, in bits, is given by: 152+168+32+56+28+72+64+400 (40*10users) + 440 (44*10 users) = 1412 bits. DL plus UL MAP fields are transmitted in the DL-PUSC mode, with the most robust MCS scheme (QPSK, CC 1/2) and with a repetition of 6. Therefore, the 720 data sub-carriers per OFDM symbol convey a total amount of 2*1/2*1/6 = 120 data bits, which amounts to 1412/120 = 12 OFDM symbols used only for MAP signaling. As there are 48 symbols in the frame there remains 48 – 1 – 12 = 35 OFDM symbols for data transportation. If the number of scheduled users is increased then the amount of OFDM symbols for data will decrease. From a practical point of view, in all simulations conducted within the system level simulator tool used in this thesis this was the amount of overhead considered. Thus, the amount of symbols available for data was assumed to be equal to 30.

Figure 2 illustrates the map of radio resources used in the DRA architecture for data transmission and the slots assigned for signaling and control channels in the downlink and uplink sub-frame.

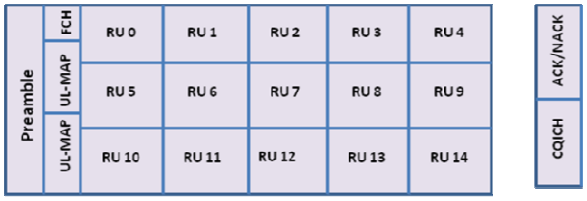


Figure 2 - WiMAX Resource Allocation Map (RAM)

These control sub-channels are used in the implementation of the cross-layer framework for the WiMAX system. The minimum granularity of resource allocation is the Radio Access Unit (RAU). For the DL-PUSC mode each RAU is a container with dimensions 10 sub-channels per 6 symbols, corresponding to 30 slots, in which each slot is equal to one FEC (Forward Error Correction) block. With a focus on downlink TDD operation, out of the 48 OFDM symbols available in the TDD MAC frame (see table 1), 35 symbols are assigned for the DL sub-frame, of which a fixed control overhead of 5 symbols is used in the preamble, FCH, DL-MAP and UL-MAP, and the remaining 30 symbols are available for data transmission.

As each RAU comprises 30 slots, with 30 symbols available for data transmission, there are (30/2)*30 = 450 slots per frame and 450/30 = 15 RAUs per DL MAC sub-frame. Each RAU is large enough to accommodate a MPDU size of 6480 bits using the 3/4 coded 64 QAM MCS scheme. This corresponds to a peak bit rate of 1.296 Mbps per RAU or 19.44 Mbps per frame. It is worth mentioning that a smaller RAU size will result in higher overhead, because more users can be assigned resources in the frame (depending on the offered load per user). Meanwhile, a larger RAU size will result in higher error rates because the BLER must be computed for all RAUs used in the transmission of the MPDU, assuming a large burst of data is transmitted.

The size of the RAU is thus a trade-off between efficiency in data allocation per resource overhead (in terms of DL/UL-MAP size and padding bits to fill each RAU) and the probability of error in the decoding of each resource. The adequate size depends on the type of traffic model being supported. For example: web traffic model packets have much larger size than VoIP ones. It results then that it is more efficient to map web packets into resources of larger size than VoIP packets and vice-versa for VoIP packets.

5.4 DRA Architecture

This section details all the steps followed in the design and implementation of the cross-layer based dynamic resource allocation module (DRA), implemented for system-level simulations of Mobile WiMAX networks.

5.4.1 Introduction

The DRA is a module constituent of the radio resource manager (RRM). The RRM improves the efficiency and reliability of wireless transmissions. At the same time it is responsible for the establishment, maintenance and finalization of each connection. Besides the DRA, the RRM architecture comprises the Connection Admission Control (CAC), Power Control (PC) and Handover (HO) modules.

The modules comprising the DRA architecture are: (i) Packet Scheduler, (ii) Resource Allocator, (iii) Hybrid Automatic Repeat Request (HARQ) and (iv) Link Adaptation (Adaptive Modulation and Coding – AMC). The architecture of a generic DRA is illustrated in Figure 3. This is the architecture of the DRA actually implemented in the system level simulator.

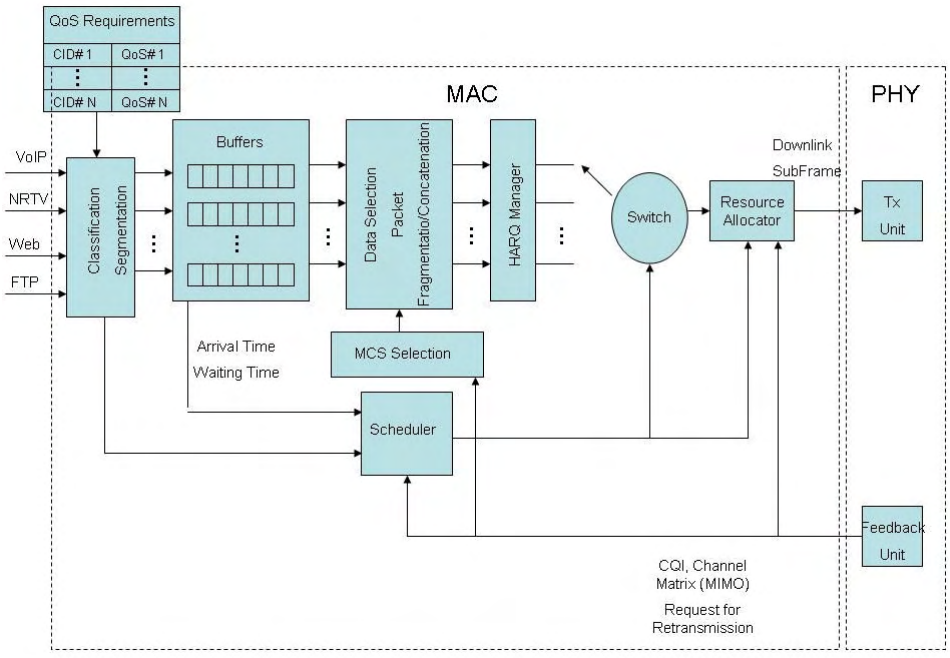


Figure 3 - Dynamic Resource Allocator architecture with constituent modules

At the base station there is one buffer, in a First-In-First-Out (FIFO configuration), dedicated to each type of connection established in the MAC layer, and used for storage of the IP packets arriving from upper layers in the protocol stack. Each connection is assigned a connection identifier (CID) and is mapped into a given set of quality of service parameters (QoS), which depend on the type of connection. In each frame period the DRA allocates radio resources in the TDD frame, according to the priority list metrics outputted from the scheduler. Whenever a user is selected for transmission, the DRA withdraws the required data bits from the respective queue, in order to fill up a data burst in the resource space of the TDD frame. The selected packets constitute a MAC-Protocol Data Unit (MPDU) which is filled with padding bits, if necessary, in order to fill up the assigned resource unit. In the system level simulator implemented under the scope of this work each user is assigned a single type of service only and there exists a single buffer in the base station for each user. Packets generated according to the selected traffic model are stored in this buffer.

In what follows the DRA procedures are described, bearing in mind downlink communication because uplink data transmission is not simulated.

5.4.2 Link Adaptation

At the beginning of each transmission time interval (frame period) the link adaptation module selects the transmission mode (i.e. MCS scheme) to be used if the user is scheduled for transmission. In the simulations 9 MCS schemes, encompassing QPSK, 16 QAM and 64 QAM and the convolution encoder, are used, according to the profiles envisioned by the WiMAX forum, which are listed in table 3. The MCS scheme employed in the transmission is chosen according to the decision rule defined in equation (1):

$$i = \arg \max_{i \in MCS_{set}} \left[R_i (1 - BLER_i) \right] \quad (1)$$

Where MCS_{set} represents the set of modulation and coding schemes available, R_i is the throughput achieved for the selected MCS scheme and $BLER_i$ is the predicted Block Error Rate (BLER), obtained from the look-up table used in the abstraction of the physical layer, for the MCS scheme used. This BLER is a function of the channel quality indicator (CQI) metric, $\gamma_{k,CQI}$, reported from each mobile station, as defined in equation (2):

$$BLER_k^{(i)}(n) = f\left(\gamma_{k,CQI}(n)\right) \quad (2)$$

The CQI metric provides an estimation of the state of the channel in the downlink connection, and gives an indication about the range of values expected for the SINR achieved in the downlink at the mobile receiver. Thus, the probability of success in decoding every piece of information on the mobile's receiver is increased.

The BLER is a threshold which depends on the type of service being processed. For voice services a lower delay per packet is preferred in detriment of a higher BLER. For data services a higher delay can be supported but with a lower BLER. The DRA selects the most spectrally efficient MCS, i.e., the one that maximizes the achievable bit rate and at the same time complies with the expected BLER under the desired threshold level.

The channel quality is tracked by the CQI parameter for each channel used in the mobile station. For the DL-PUSC mode, and according to the inherent frequency diversity achieved with sub-carrier pseudo-random allocation, it is assumed that all sub-channels have the same state. In the simulations the CQI is computed from the state of each data sub-carrier in the OFDM symbol used in the transmission of the preamble. In order to reduce complexity a smaller group of data sub-carriers is sampled from the whole set of data sub-carriers.

The CQI is then updated with a period of T_{CQI} frames, which is a multiple of the frame period, combining past information and information provided by measurements, according to the time-smoothing formula expressed in equation (3):

$$CQI_k^{(i)}(n) = 0.7x\gamma_k^{(i)}(n) + 0.3xCQI_k^{(i)}(n-1) \quad (3)$$

Where:

- $CQI_k^{(i)}(n)$ is the value of the CQI for data sub-carrier k of user i for time period $\left(lT_{CQI}, (l+1)T_{CQI}\right)$.
- $\gamma_k^{(i)}$ is the reported CQI from measurements performed on data sub-carrier k of user i.

The CQI is reported back to the base station via the Channel Quality Indicator control Channel (CQICH) on the uplink sub-frame, on a frame-by-frame basis.

The resulting CQI is a scalar value computed by the Exponential Effective SINR Method (EESM) mapping rule applied to the vector of CQI values according to equation (4):

$$SINR_{eff} = -\beta \ln\left(\frac{1}{N}\sum_{k=1}^{N} e^{-\frac{SINR_k}{\beta}}\right) \quad (4)$$

Where:

- β is the correction parameter used to adapt the formula to the different types of scenarios used in the simulations.
- N is the number of sub-carriers in the vector of CQI values.

In the cell, each user performing a transmission in a given transmission time interval is assigned an amount of Resource Units (RU). These RUs constitute the burst into which the user's MPDU is mapped. The number of RU's used depends on the size of each packet and on the selected MCS scheme.

5.4.3 Asynchronous Hybrid Automatic Repeat Request (HARQ)

In the MAC layer error recovery is implemented by associating each burst of data to an empty HARQ process. HARQ error recovery is carried out by soft combining the information associated with new and previous erroneous transmissions, in an attempt to minimize the amount of redundant information and power transmitted over the air interface. The mobile station combines the current version with previous ones of the same MPDU using Chase Combining [109]. Retransmissions of the same MPDU keep the original MCS scheme used in the first transmission attempt.

In the MAC layer there exist a number of HARQ processes implemented for each single user. This mechanism grants simultaneous transmissions from the same user, in the same frame interval. Each HARQ process is associated to one buffer in the mobile station to store the result of the combination of successive versions of the same MPDU. Therefore, according to this mechanism, each MPDU being transmitted for the first time is then mapped into one of the available HARQ processes, and each HARQ process is in charge of transmission and re-transmissions of a single MPDU until it is successfully received. Once an HARQ process has been selected, the DRA must wait for an ACK/NACK message from the mobile station before selecting the HARQ process again. HARQ buffers are freed when the radio block is successfully received or when the maximum number of allowable transmission attempts, $N_{attempts}$, has been achieved. Due to the time required for signaling feedback and processing of the information at both ends of the transmission chain, the minimum time interval between two successive transmissions of a particular HARQ process is equal to two frame periods.

In the uplink sub-frame there is an HARQ Acknowledge (HARQ – ACK) channel region for the inclusion of one or more ACK channels(s) for HARQ support. This UL-ACK channel is implicitly assigned to each HARQ-enabled DL burst according to its order in the DL-MAP. Thus, the user can quickly transmit ACK or NACK feedback messages for DL HARQ-enabled bursts using this UL ACK channel (see chapter 3).

5.4.4 Scheduler

Packet schedulers must be designed to be reactive to changes in the channel and traffic patterns, in order to respond fast to deviations from the requested QoS of even the most delay sensitive applications. The scheduler is located inside each base station to enable rapid response to traffic requirements and channel conditions. As data packets are associated to service flows with well defined QoS requirements, the scheduler can correctly determine the packet transmission ordering through the air interface. Packets must be given priority according to the set of QoS metrics which have been negotiated between the network service provider and the end user.

Bearing this in mind, the ultimate goal of the most proposed and implemented packet schedulers, in the selection and attribution of resources to each user, is the satisfaction of the

application's QoS requirements. Although the end user is not particularly concerned with the mechanisms followed by the scheduler in the prosecution of this goal, the perceived user's satisfaction for the service is highly influenced by the scheduler's performance.

Dynamic scheduling is based on different metrics reported from other layers in the protocol stack, in agreement with the cross-layer architecture paradigm. This information is combined in the computation of the weight assigned to each user, which defines the priority in the access to radio resources in the MAC resource space. The metrics commonly used are:

- CQI reports sent from users in every uplink sub-frame. Such channel-sensitive scheduling techniques exploit multi-user diversity, resulting in the so-called "Multi-User Diversity Gain" [26, 110].
- The user average data transmission rate.
- Quality of Service (QoS) metrics such as: the delay incurred to each packet waiting in the user's buffer for a transmission opportunity and/or the minimum expected service data rate.
- Priority assigned to each type of application in the attribution of resources.
- Amount of information in each user's buffer (in bits).
- Number of transmissions already attempted.

At each frame period the scheduler provides transmission opportunities to eligible mobiles with data to send, starting with the highest ranked user and then proceeding to lower ranked ones in sequence. The CQI reports are obtained from every user on a frame-by-frame basis and the scheduler re-computes the mobile's access priority at every frame period.

One of the main contributions of this work is the proposal, validation and analysis of packet schedulers for Beyond Third Generation (B3G) networks using the Mobile WiMAX standard as a case study. The proposed schedulers and their interaction with the DRA architecture are elaborated in greater detail in chapters 7, 8 and 9.

5.4.5 Resource Manager

Once the list of scheduled users is composed the resource manager starts with the resource allocation process. The resource manager is responsible for the allocation of sub-channels and power over each sub-carrier in the resource unit. The allocation of the proper sub-channels should be able to exploit physical layer information such: as SINR, MCS level and velocity of the user, in an effort to maximize resource efficiency and be constrained to the power available for data transmission at the base station. The velocity is very important because it determines the adequate type of sub-channelization mode used (adjacent AMC or diversity permutation). The sub-channel allocation algorithm should also satisfy the applications' QoS requirements. But, as the scheduler is normally assigned the task of selecting the users and dimensioning the

amount of resources to be allocated in each frame, the resource manager does not take into account applications requirements.

In the beginning of each TDD frame the DRA assigns free RAUs from the resource map according to the individual priorities assigned to each user by the scheduler. Whenever a user is selected, the DRA withdraws IP packets from the respective queue or from the respective HARQ process in case of retransmission. The number of RAUs requested for transmission depends on the size (in bits) of the amount of packets in the queue. The size is computed by the link adaptation module. Packets are concatenated and/or segmented in order to fill the amount of RAUs available for allocation in the frame. Padding bits are added if necessary. The set of RAUs assigned to the user constitutes a given burst in the frame and each burst is composed of a group of contiguous RAUs. Each burst is individually assigned to a single user and all slots in the same burst are transmitted with the same MCS scheme.

Subsequent to scheduling and mapping of the IP packets to the available RAUs, the base station broadcasts the map with the RAM to the network via DL-MAP control fields in the DL sub-frame. Each mobile station utilizes this signaling information to anticipate packet transmission and to perform the demodulation and decoding of the transmitted data using the appropriate MCS scheme. The Resource Allocation Map is updated every frame by the base station. The whole process comprises four different steps (A to D) as illustrated in figure 4.

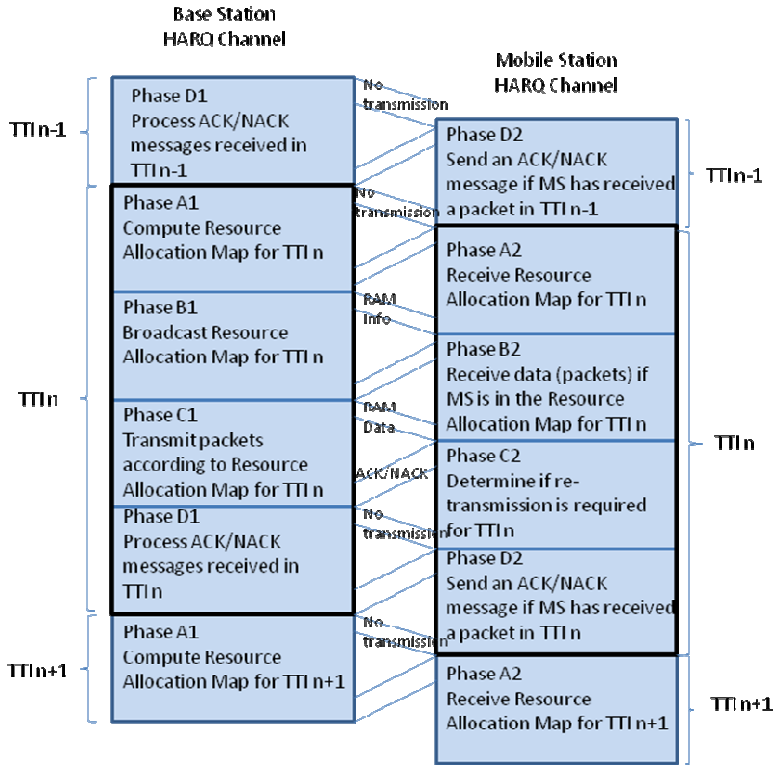


Figure 4 - WiMAX DRA cycle

· **Phase A1:** The base station computes the Resource Allocation Map (RAM) to be used in data transmission during Phase C1 that follows.

- **Phase B1:** The base station broadcasts the Resource Allocation Map (RAM) computed in Phase A1. This map contains the signaling used in the identification and allocation, in the MAC frame, of all resources assigned to each mobile user who is scheduled for transmission in the downlink, for the current transmission time interval. The RAM is transmitted in the DL-MAP sub-field of the MAC sub-frame. Therefore all users in the cell have the information required to demodulate the packets they will receive during Phase C1.
- **Phase C1:** The base station sends data according to the RAM determined in Phase A1.
- **Phase D1:** The base station receives ACK messages from users which have received their packets successfully during Phase C1, and likewise receives NACK messages from users which have received their packets un-successfully during Phase C1. This signaling is sent in the UL-ACK region in the uplink MAC sub-frame.
- **Phase A2:** The mobile station receives the broadcasted RAM from the base station.
- **Phase B2:** The mobile station receives the useful information transmitted from the base station in the downlink. This data comprises a set of MAC Layer Packet Data Units (MPDUs), which are mapped into a set of Radio Alocation Units (RAUs) in the MAC sub-frame. These RAUs were previously reserved and allocated to the mobile during the scheduling phase, according to the RAM received in Phase A2.
- **Phase C2:** The mobile station processes the received MPDUs. If it is only another version of MPDUs transmitted initially (first transmission) the mobile station performs Chase Combining on the different replicas of the same packet.
- **Phase D2:** After decoding the received MPDUs the mobile station sends an ACK/NACK message to the base station if it has received the given MPDU, with or without error respectively, during Phase B2. This is to inform the base station about the status of the decoding process.

It is important to note that the whole cycle of four steps lasts for two frame intervals. This is because broadcasting of signaling information and transmission of data are performed in the same frame interval, while the sending of feedback information on behalf of the mobile user occurs, at least, on the next frame, as there is no time to decode the information in the same frame period. Therefore, the transmission time interval actually encompasses two frame periods, amounting to a time interval equal to 10 ms.

Each HARQ process of each mobile station handles only one MPDU mapped onto a given amount of RAUs of the MAC frame. The same amount of RAUs is used in the retransmission of the MPDU with the same MCS scheme. In each frame period each HARQ process is in one of two states:

- **Active State:** The HARQ process number *n* of mobile station *k* is busy handling a MPDU. Different versions of the same MPDU may be sent over the air-interface if the HARQ process is retransmitting the MPDU. The initial version of the MPDU is stored in the buffer dedicated to the HARQ. The result from the combination of successive received versions of the same MPDU is stored in the buffer dedicated to HARQ process number *n* of mobile station *k*.

  Inactive State: The HARQ process number *n* of mobile station *k* is idle. In the base station the buffer dedicated to the HARQ process number *n* is empty or is being used by another HARQ process number *n* from another mobile station different from mobile station *k*.

5.4.6 Resource Allocation Procedure

The description of the resource allocation procedure encompasses the methodologies for packet scheduling, resource map allocation and link adaptation. The proposed resource allocation procedure is sub-divided into four steps:

Step 1: Determining HARQ processes lists

In the beginning of each scheduling period two lists of HARQ processes are computed: the Low Priority list and the High Priority list.

- **High Priority List** – This list comprises only HARQ processes waiting for another transmission attempt. After the first transmission attempt, and if there is an error, a timer (Timer Priority) is activated. If the HARQ process is not scheduled again before the expiration of this timer, it will be inserted into the High Priority list. At the same time another timer (Timer Discard) is activated. If the HARQ process is not released before the expiration of this timer, the HARQ process is initialized and the corresponding information stored in its buffer is lost. Figure 5 illustrates the flow chart describing the creation of the two types of priority lists and the activation/deactivation of both timers. The delay threshold is service specific: for real-time services such as VoIP or near-real time video (NRTV), the delay threshold is zero, which means that the HARQ process will be inserted into the High Priority list in the next scheduling period. For non-real time services, such as the WWW, the delay threshold can be larger.
- **Low Priority List** – For a mobile with new packets stored in their buffer, an inactive HARQ process, which can be assigned for the first transmission attempt, is searched for in the set of HARQ processes assigned to the mobile. If there is any it is inserted into the low priority list. This process is repeated for all active users in the cell and with new packets. The HARQ processes waiting for another transmission opportunity and with the Priority Timer not set are also inserted into this list.

It is assumed that the base station hosts a waiting queue for each mobile station. Both lists are then sorted in descending order of the priority, computed for each HARQ process according to the scheduling algorithm. The base station selects the HARQ processes with highest priority in each list.

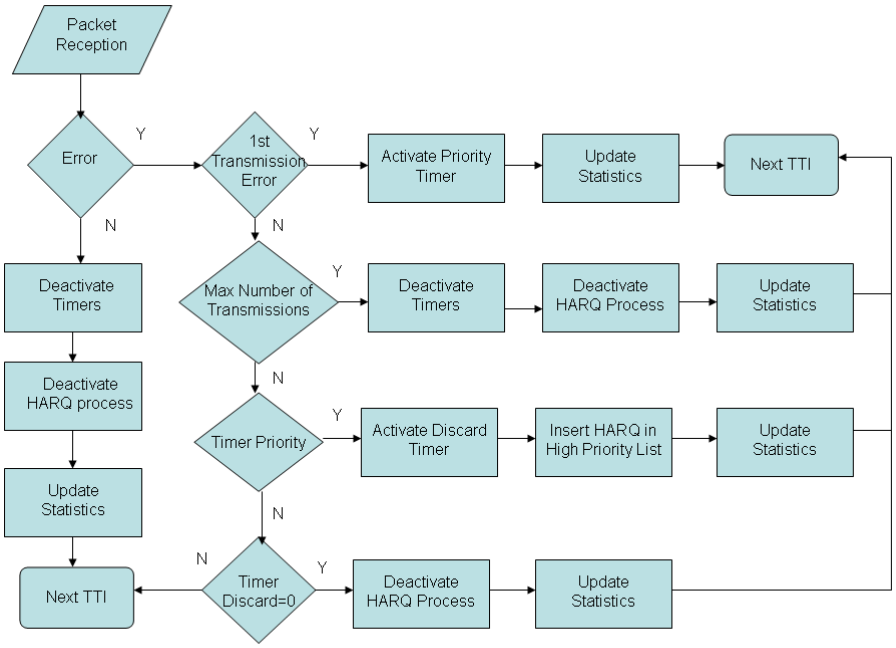


Figure 5 - Flowchart with the creation of priority lists and activation/deactivation of timers

The principle behind the definition of the Priority Timer is to increase the probability of serving a given user with an active HARQ process in retransmission before the delay bound is achieved for the packets stored in the buffer. Of course this mechanism is not fully QoS-compliant because new packets can remain in buffer until they are dropped without service, as they are not inserted into the High Priority List. The Discard Timer avoids an HARQ process being indefinitely in active state, something which could block the access to free HARQ process, from users attempting the transmission of new packets which have arrived to its buffer.

The definition on the adequate number of HARQ process to assign to each user depends on the maximum delay bound of the service, the value assigned to the Discard Timer and on the complexity pretended for the simulation.

Step 2: Scheduling HARQ processes in High Priority list

All HARQ processes in High Priority list are attempting a retransmission. The resource allocation module iterates the High Priority list, starting with the process with highest priority. Then the same RAUs in the map of resources are allocated, with the same power, data and MCS scheme used in previous transmissions. If there are any, RAUs remaining unallocated after all HARQ processes in the list are serviced, are deactivated and put into the state of availability in order to be assigned to those HARQ processes in the Low Priority list.

Step 3: Scheduling HARQ processes in Low Priority list

If there are still RAUs available for data transportation, after the HARQ processes in the High Priority list are serviced, then the resource allocation module iterates the Low Priority list. The same MCS scheme is selected for the transmission of a new version of the MPDU.

HARQ processes attempting retransmissions are scheduled only if there are enough RAUs for allocation, as the MPDU stored in its buffer cannot be fragmented. The power assigned per RAU is computed according to what follows.

For HARQ schemes performing a first transmission attempt, the link adaptation module selects the most appropriate MCS scheme, $SINR_{selected}$ according to the predicted SINR, $SINR_{pred}$, based on the frame's preamble and reported from each mobile. Mobile stations with a CQI value lower then a given admission threshold are not considered by the scheduler even if they have information ready for transmission. Mobile stations with a CQI which do not correspond to a transmission with quality good enough to result in a high probability of success in the decoding of the packet, but within the admission threshold limits, are transmitted with the most robust MCS scheme.

The following steps are performed until one of the following three conditions is not satisfied anymore: (i) the list of priorities is empty, (ii) there is no more information to be transmitted (iii) or there are no more resource units for allocation.

Computation of the power assigned to each RU

The power available for data in the cell is uniformly distributed per each resource unit in the map of resources, according to equation (5).

$$P_{t,RU} = P_{dat} / N_{RU} \tag{5}$$

Computation of size of the MPDU

The number of bits carried in each resource unit is deduced from the selected MCS scheme and the available number of resource units. It is given by equation (6).

$$N_{transmitable} = N_{RU} x Size_{RU} \tag{6}$$

The number of bits that will be transmitted is a function of the number of bits waiting in the base station queue, N_{queue}. It is computed according to equation (7).

$$N_{transmit} = \min(N_{queue}, N_{transmitable}) \tag{7}$$

Determination of the required amount of RUs

The number of resource units to be allocated in the map of resources is computed by equation (8)

$$N_{RU,transmit} = \left\lceil \frac{N_{transmit}}{Size_{RU}} \right\rceil \tag{8}$$

The Resource Allocation Map is updated by allocating $N_{RU,transmit}$ to the given mobile with the selected MCS scheme $MCS_{selected}$ and the computed transmit power per resource unit: $P_{t,RU}$.

The base station then updates the mobiles list as well as the amount of resource units still available for allocation, N_{budget} N_{budget}.

Step 4: Interaction with the physical layer

Finally the Resource Allocation Map is broadcasted by the base station, providing the respective mobiles with data mapped in the resources map with control information regarding the allocation of bursts, the MCS scheme used in the transmission and respective transmission power assigned to each resource unit.

In Summary the main characteristics of the DRA are the following:

- It handles mixed traffic (in particular non-real time services and real-time services).
- HARQ processes are organized onto two priority lists. Higher priority is given for HARQ processes active and waiting for a retransmission after a given period of time defined by the priority timer.
- In the lower priority list the priority of the new transmissions and re-transmissions is defined by the scheduler.
- Re-transmissions are performed with the same amount of resources and the same MCS scheme selected from the first transmission attempt.
- Priority computations are based on the type of service, waiting queue status and channel current conditions.

5.5 Related Work

Although there is a significant amount of research available in the literature regarding packet scheduling using the OFDMA multiple access scheme, most of the work conducted is based on simplistic cell layouts (one serving base station) and no traffic models are considered. In particular, in most of the scenarios implemented a singular base station is assumed as transmitting to a number of mobile stations with no inter-cell interference modelling. Also, users are backlogged or very simple traffic generator models are assumed: typically packets of constant size are generated according to a Poisson distribution and service time is modelled as an Exponential distribution. Most research also follows an analytical approach. Namely scheduling is modelled as an optimization problem with constraints regarding maximum power available for data transmission and some simple quality of service requirements, such as minimum supported data rate or maximum allowable packet delay.

Few contributions deal with realistic scenarios involving the use of system level simulations, and very few involve some sort of DRA implementation.

In [92] the authors implement models for the performance evaluation of signaling resources which are implemented in the PHY layer of the Mobile WiMAX standard, for data and signaling control information. System level simulations are exhaustively performed for definition of the optimum number of symbols used as overhead for control channels, according to system parameters.

In [62] the analysis of the cell coverage achieved with the control MAP channel is performed by means of system level simulations. The improvement in the cell coverage is investigated with the use of Cyclic Shift Transmit Diversity (CSTD) [111] which is a type of space-time coding scheme used to achieve spatial diversity. With CSTD each transmit antenna sends a circularly shifted version of the same OFDM symbol in time-domain samples. Simulations for different types of repetition coding, using 1, 2 and 4 transmit antennas and with half of the users experiencing a speed of 3 Km/h and the other half experiencing a speed of 30 Km/h are conducted. The conclusion is that 95% coverage for a MAP error rate (MER) equal to 1% is achieved using CSTD with 4 antennas at the transmission and 2 antennas at the reception, 1/2 CC code with QPSK modulation and repetition of 4 with a simple MRC receiver.

An example of a simulation platform for performing system level simulations for HSDPA networks is given in [112]. A packet scheduler is proposed to avoid excessive packet delays that may invoke the slow-start on the Transmission Control Protocol (TCP) connections.

In [61], an exhaustive performance analysis of a DRA proposed for the IEEE802.16e standard of Mobile WiMAX is conducted for both link and system levels under SISO channel. Both types of sub-channelization, band AMC and diversity mode are considered. System level simulations are performed for both Max C/I and Proportional Fairness schedulers. Users are selected for transmission as long as their CQI is above a defined threshold. CDF curves for user throughput are obtained for the different scenarios considered.

In [63] an overview of Mobile WiMAX and performance figures are presented for system-level simulations conducted under various configurations, channel and traffic models. An analysis of MAP control channels coverage using robust transmission encoding is conducted and configuration setups are illustrated.

In [113] a radio resource manager is implemented for evaluation of the downlink connection of a Mobile WiMAX system. The resource manager encompasses both the scheduler and a call admission control module.

In [114] different downlink resource allocation strategies for the OFDMA multiple access scheme are conducted by means of system level simulations. The analysis is performed for the inherent trade-off between system capacity maximization due to multi-user diversity and fairness, in terms of packet delay.

In [40] the authors provide a cross-layer design framework based on the functionalities available at the MAC layer of IEEE 802.16e standard. Cross-layer protocols for performance

improvement are presented and a set of primitives for cross-layer operation between both PHY and MAC layers are provided.

In [115] another example of DRA design is conducted for the analysis of some interrelated problems in resource allocation such as dynamic sub-carrier allocation, admission control and capacity planning in OFDMA-based Wireless Metropolitan Area Networks.

In [116] the authors perform the evaluation for the first time of different configuration parameters at MAC layer level, such as: MAC frame size and protocol data unit and number of connections at MAC layer. Through system level simulations they evaluate the trade-offs achieved between the total amount of connections at the MAC level and signalling in overhead.

In [117] an investigation on the performance of the different sub-channelization schemes available in a multi-cell scenario with OFDMA as multiple access technology and with full frequency utilization over all cells is conducted. The number of collisions among the same set of sub-carriers being used in adjacent cells is stored for two types of scenarios: full and partial load, and for diversity sub-carrier sub-channelization (FUSC) and band adjacent sub-carrier sub-channelization. Performance is inferred according to the spectrum efficiency.

In [64] system level simulations are performed for the performance analysis of WiMAX networks using multiple transmit antennas at transmitter and receiver (MIMO) for both downlink and uplink. Different types of MIMO mode are considered in the simulations. The performance of the different MIMO schemes is analysed by means of the spectral efficiency in bits per channel access per SNR at the receiver. Both full queue and web traffic models and different types of receivers (MMSE and MLD) are considered in the simulations.

In [118] a general reference model for conducting system level simulations for evaluation of WiMAX system performance is presented. The approach is similar to the one followed in the analysis of system level performance of other 3G systems such as HSDPA and CDMA2000 EV-DO. The study is conducted for VoIP and full queue traffic models, both downlink and uplink connections.

In [119] a comprehensive performance study of commonly proposed scheduling algorithms in the literature for the scenario of the downlink connection for Mobile WiMAX is performed. The algorithms are evaluated according to their ability to support QoS, fairness amongst service classes and bandwidth utilization.

The authors of [120] design a radio resource management architecture encompassing both the packet scheduler and the connection admission control modules for the support of real-time and non-real time services in an OFDMA-based multiple access network. Different types of non-real time applications (FTP and HTTP) and real-time (VoIP) coexist in the system. The combined effects of the proposed CAC and packet scheduling by using the cross-layer simulation between physical and Internet application layer is evaluated.

5.6 Conclusion

The support of QoS requirements from the different types of application services which are expected to coexist in the same WiMAX network infrastructure, demands the implementation of corresponding packet schedulers. The scheduler must be able to combine metrics from different layers of the protocol stack, in a cross-layer based framework. Although the IEEE 802.16e standard defines the different functionalities needed for the provision of QoS, and it strictly specifies that all service flows must be connection-based (even for packets from services of type best effort, to which no QoS guarantees are established), the scheduler is not defined in the standard and it is left for the manufactures the task of defining appropriate packet schedulers, as a kind of differentiation among system solutions from different suppliers.

The packet scheduler is the core component in the DRA architecture design for wireless networks, namely for wireless networks expecting to be key drivers in the specification of wireless cellular networks of fourth generation (4G). This chapter presents in detail a DRA architecture framework for Mobile WiMAX networks. Its implementation is based on the cross-layer design paradigm, since the inputs arriving from different layers in the protocol stack are combined by means of the control channels implemented in the MAC frame, and used in the exchange of information. The different modules of the DRA are detailed and the parameters of the standard used in the DRA for conducting system-level simulations are presented. Some schedulers, commonly referred in the literature and already implemented in real networks, such as Round Robin, Proportional Fairness and Maximum C/I, are used in the validation of the DRA.

The chapter describes in detail all the steps followed in each frame period by the communication protocol, between the base and the mobile stations, for the transmission of the information associated to the scheduled users. An important step in the implementation of the DRA architecture is the definition of the resource allocation map, which comprises the radio resources assigned for data transmission over the air interface. These resources are defined both in time and frequency domains and can be extended for the space domain also, if smart antennas are implemented for beamforming (this is the subject of chapter 8). The steps followed in resource allocation inside the DRA are presented.

The following chapters elaborate on packet schedulers based on the notion of utility functions. As it will be demonstrated, these schedulers plug in very easily to the proposed DRA framework, without significant modifications to the base architecture.

Chapter 6

System Validation

6.1 Introduction

The system level simulator must be validated before it is used for conducting simulations. The validation is performed by testing the models implemented (for traffic, channel, signal to noise plus interference ratio (SINR) computation, interference generation, etc.), which must be corroborated against theoretical values. Also, for each new packet scheduling algorithm implemented in the tool, its performance must be compared against benchmark figures available in the literature. Bearing this in mind, this chapter conducts a comprehensive overview of the testing methodology followed in the validation of the system level simulator architecture, encompassing the Dynamic Resource Allocation (DRA) scheme, proposed for implementation of packet schedulers in Mobile WiMAX networks. Ideally, the results should be compared against benchmark figures available in the literature. However, the results obtained depend heavily on the type of architecture and on the configuration chosen for the DRA, namely on the definition of the map of resources in the MAC layer. This is because the IEEE 802.16e standard for Mobile WiMAX contains a plethora of possible configuration profiles. Also, differently from proposals available under 3GGP for HSDPA standard, there is no straight-forward solution

available in the literature that can be used to compare against the architecture proposed in this work, and consequently fully support the validation process.

In this chapter, the implemented models are validated by comparing the results obtained from simulations to the theoretical values associated to the different types of models used in the simulator. This is an important step in the process of validation, because it infers the level of accuracy achieved in the implementation of these models in the system level simulator. Also, as an effort to address the trade-off between simulation time and accuracy, the validation methodology has been addressed based on the central cell approach and assuming full load conditions.

This chapter is organized as follows: section 2 describes the procedures followed in the validation of the simulator architecture, namely the theoretical models implemented for modeling signal propagation: path-loss, shadowing and fast fading, under the framework of the Single Input Single Output (SISO) and Multiple Input Multiple Output (MIMO) channel models. Section 3 presents the results for the tests conducted on the validation of the DRA implemented in the system level simulator. The DRA is validated for a basic maximum C/I scheduler with the World Wide Web (WWW) traffic model from 3GPP. Section 4 presents in detail the different scheduling algorithms considered in system level simulations and the DRA under the scope of this thesis. Section 5 presents the results from the performance evaluation of the proposed DRA for different schedulers commonly referred in the literature and for different types of channel and traffic models. Section 6 concludes this chapter.

6.2 Validation of the Basic System Level Simulation Platform

This section presents the results obtained from the validation tests that have been performed for the basic system level simulation platform. The scenario is the central cell approach with combined-snapshot dynamic configuration. No mobility and/or handover and connection admission control algorithms are considered. Users are assumed as active and backlogged (always with information ready in the buffer) since the beginning of each new run.

6.2.1 Fast Fading Channel Model Implementation

Fast fading propagation is simulated for each channel between a given mobile station and each one of its neighboring cells. For the serving cell the channel is simulated by a multi-path channel model (frequency selective model), whilst the channel between the same mobile station and each one of its neighboring cells is simulated by a single path (flat frequency) channel model. The fast fading is generated with a periodicity equal to the frame length, i.e. 5 ms and considered constant along this period.

6.2.1.1 Multipath Power Profile

According to the multipath channel model implemented in the system level simulator, each path power should have a specified value which depends on the type of filter implemented: Vehicular A (VehA), Pedestrian A (PedA) and Pedestrian B (PedB). The following test was performed:

- One simulation is run for a large number of frame periods. For each period the power of the path for the channel between one given mobile station and its serving base station is collected.
- The average power of each path in the multi-path channel is computed to form the path power profile of each type of multi-path radio channel model.
- These power delay profiles are compared to the theoretical value (tap coefficients of the channel model).

Tables 1, 2 and 3 provide a comparison between the theoretical and simulated values for the multi-path channel models implemented in the simulator. It is expected that the average power is near 0 dB and that the standard deviation is low.

ITU Pedestrian B Path Power Delay Profile						
Path 1	Path 2	Path 3	Path 4	Path 5	Path 6	
-4.1123dB	-4.8025dB	-8.8317 dB	-11.8884 dB	-11.6074 dB	-27.7948 dB	Simulation
-3.92dB	-4.82 dB	-8.82 dB	-11.92 dB	-11.72 dB	-27.82 dB	Theory

TABLE 1: ITU PEDESTRIAN B PATH POWER PROFILE

ITU Pedestrian A Path Power Delay Profile				
Path 1	Path 2	Path 3	Path 4	
-0.5240dB	-8.8317dB	-19.3862dB	-23.3068dB	Simulation
-0.51dB	-10.21dB	-19.71dB	--23.31dB	Theory

TABLE 2: ITU PEDESTRIAN A PATH POWER PROFILE

ITU Vehicular A Path Power Delay Profile						
Path 1	Path 2	Path 3	Path 4	Path 5	Path 6	
-3.3368dB	-4.1270dB	-12.1562dB	-13.1129dB	-18.0319dB	-23.1193dB	Simulation
-3.14dB	-4. 4.14dB	-12.14dB	-13.14dB	-18.14dB	-23.14dB	Theory

TABLE 3: ITU VEHICULAR A PATH POWER PROFILE

As can be seen from tables 1, 2 and 3, there is a good agreement between simulation and theory.

6.2.1.2 Rayleigh Distribution of the Path Amplitude

According to the channel model implemented for simulation of fast fading in system level simulations, the Cumulative Distribution Function (CDF) of the amplitude of each path should follow a Rayleigh distribution as given by equation (1):

$$F(x) = 1 - e^{-\frac{x^2}{2\sigma^2}}, \quad x \geq 0 \qquad (1)$$

Where: x is the instantaneous amplitude of the path and σ is the standard deviation.

The path amplitude is generated, collected and stored for one given path in the channel model and for a large number of frame periods. Then the CDF of the amplitude is obtained and compared to $F(x)$. This test can be repeated for different paths, different speeds and different

channels. Figure 1 presents the plot for path 1 of a given test mobile station. As can be seen there is a good agreement between simulation and theory.

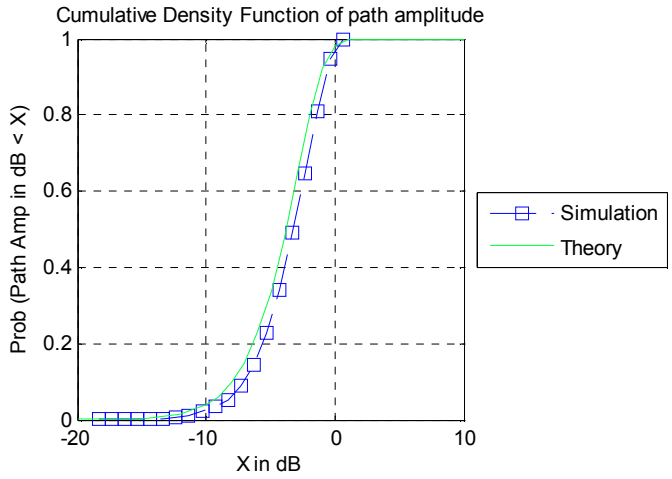


Figure 1 - CDF of Rayleigh channel model for fast fading simulation

6.2.1.3 Time Correlation

According to the channel model implemented for simulation of fast fading in system level simulations, the time-correlation of the amplitude of each path shall follow a Bessel distribution as given by equation (2):

$$R(t) = B(2\pi f_d t) \qquad (2)$$

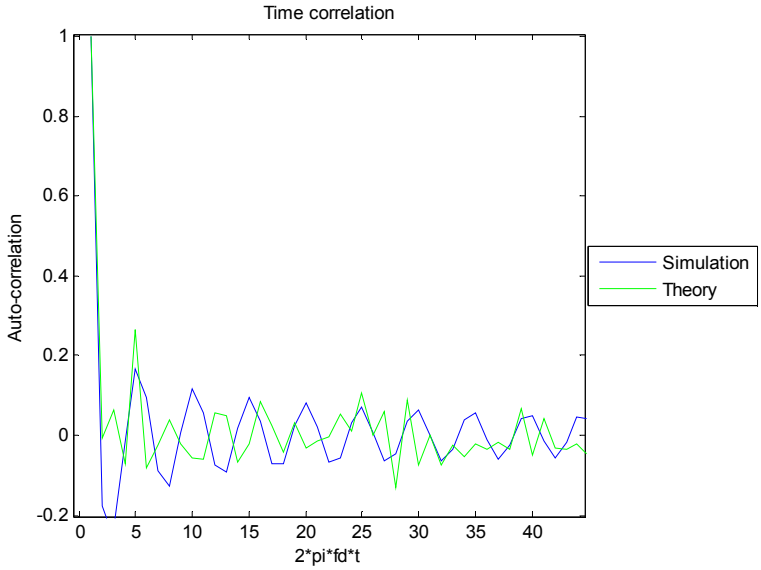


Figure 2 - Time correlation of Rayleigh path amplitude

Where:

- B is the Bessel Function.
- t is the time in seconds

- $f_d = vf_c / c$ is the Doppler Frequency in Hz (v is the mobile speed in m/s, f_c is the carrier frequency in Hz and c is the speed of light (300 000 Km/s).

The path amplitude is generated, collected and stored for one given path in the channel model and for a large number of frame periods. Then a circular time-correlation product is computed and compared to the theoretical formula. This step can be repeated for different paths and different speeds. Figure 2 presents the plot for the theoretical Bessel function and the circular time-correlation of the path amplitude. As can be seen from figure 2 there is a good agreement between simulation and theory.

6.2.1.4 Fast Fading Channel Correlation over Each Frame Period

According to the channel model implemented for simulation of fast fading in system level simulations, the channel shall remain constant over one single frame period for the mobile speed used in simulations (high coherence time). A plot of the Bessel function and of the correlation of the channel over adjacent frames was produced and is plotted. Figure 3 illustrates, for one single path, the value of the correlation between the path amplitude at the beginning of the first frame period and the path amplitude at the end of the first frame period, and also between the beginning of the first frame period and the end of the second frame period, for different speeds.

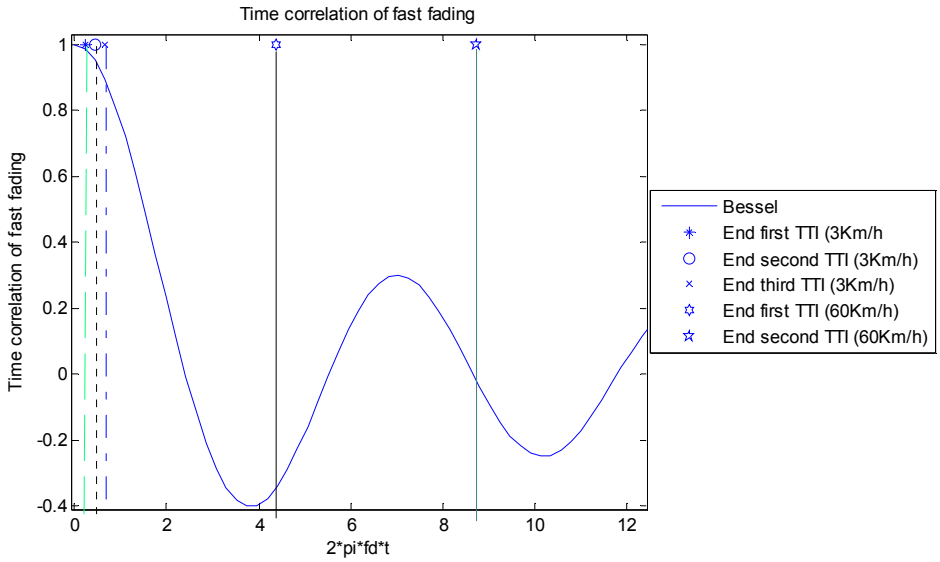


Figure 3 - Channel correlation over the data block time length

It can be noticed that when the speed is equal to 3 Km/h the channel is fully correlated over 2 consecutive frame periods. At 60 Km/h, at the end of the first frame period, the channel is completely uncorrelated. Therefore, the assumption that the channel is constant over frames of period equal to 5 ms is true for the pedestrian speed of 3 Km/h.

6.2.2 Shadowing

The modeling of the shadowing component of signal propagation is implemented according to the methodology presented in chapter 4.

6.2.2.1 Log-Normal Law

According to the channel model implemented for simulation of log-normal shadowing in system level simulations, shadowing follows a log-normal law. If s is the amplitude of the shadowing between one given mobile station and a given base station, $u = 10\log_{10}(s)$ follows a Normal law.

In the simulator, shadowing samples were generated by means of a large number of independent runs (100 runs). Then, based on these samples, the CDF of the shadowing was computed and compared to the theoretical Normal CDF. Figure 4 illustrates the plot of the CDF of shadowing for the test mobile station. As can be seen, there is a good agreement between simulation and theory. This same test can be performed using different values for the standard deviation.

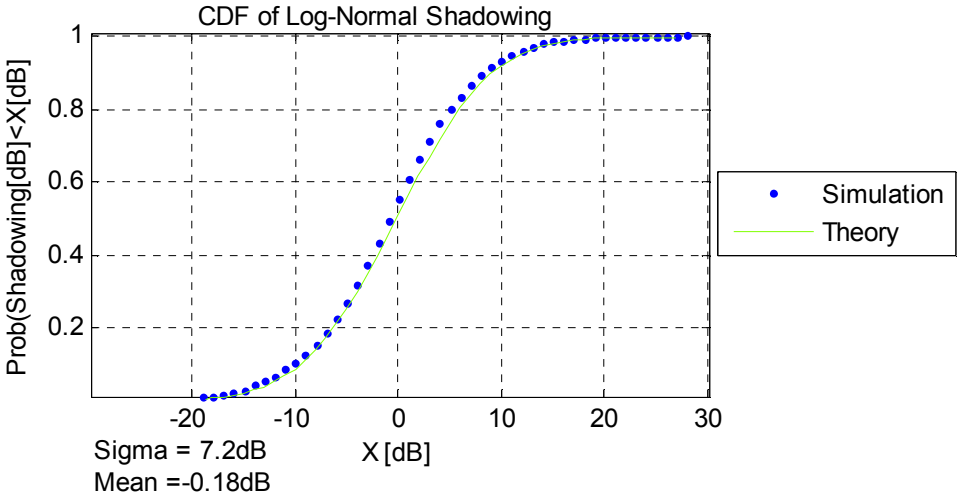


Figure 4 - CDF of amplitude of log-normal shadowing

6.2.2.2 Space Correlation

According to the channel model implemented for simulation of log-normal shadowing in system level simulations, shadowing is spatially correlated and the correlation between the shadowing perceived by one mobile station, MS1, from a given base station and by another mobile station, MS2, and the same base station is given by equation 3:

$$R(dx) = e^{-\ln(2)\frac{dx}{D}} \quad (3)$$

Where: dx is the distance between MS1 and MS2 and D is the shadowing de-correlation length.

The following test was performed on the simulator for verification of this property:

- 10 mobiles were created in the network at positions along a straight line and spaced 10 m apart.
- A large number of simulation runs was performed. For each run *n* and for each mobile *i* the value of the shadowing $s(i,n)$ between the mobile and the serving base station was measured and stored.
- For each step $dx = k * 10\,\mathrm{m}\,(k \text{ in } [0..9])$, the value of the correlation function was computed.

Figure 5 plots the value of the autocorrelation function for samples obtained from all 10 mobiles from simulation and according to theory.

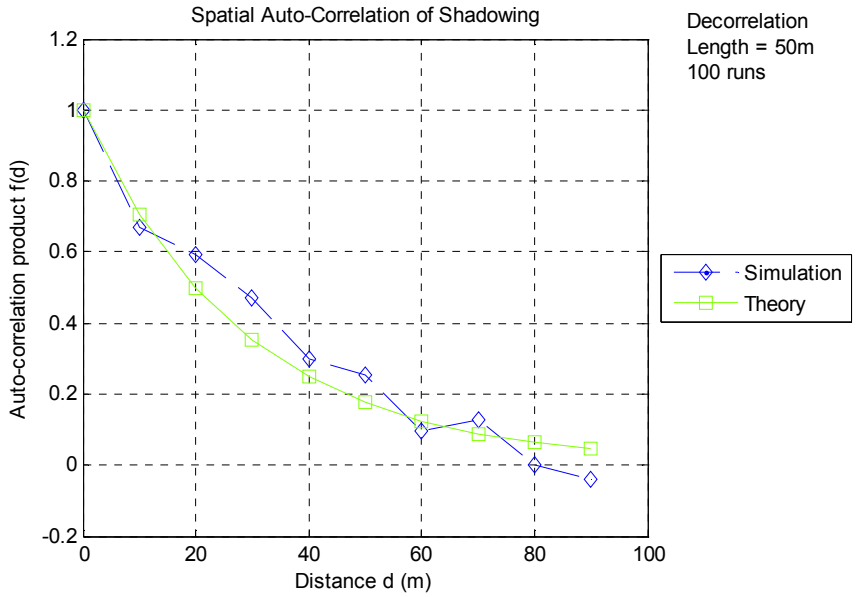


Figure 5 - Spatial correlation of bi-dimensional shadowing

As shown in figure 5 there is a good agreement between simulations and theory.

6.2.2.3 Inter-Site Correlation

According to the channel model implemented for simulation of log-normal shadowing in system level simulations, the shadowing between one mobile station and one base station and the shadowing between the same mobile station and another base station shall be correlated and the inter-site correlation shall be equal to 0.5. 200 independent runs were produced to validate this property. For each run the following actions were performed:

- One mobile station was created.
- The shadowing $s(j,n)$ between this mobile station and each neighboring base station *j* is generated and stored.
- The correlation product between the shadowing terms was computed.

A histogram of the correlations for each one of the neighboring and serving base stations was generated and is plotted in figure 6. The correlation was performed between vectors with shadowing values from the different base stations. As can be seen the results are in complete agreement with theoretical assumptions as the inter-site correlation is equal to 0.5, on average, and there is full correlation inside each cell.

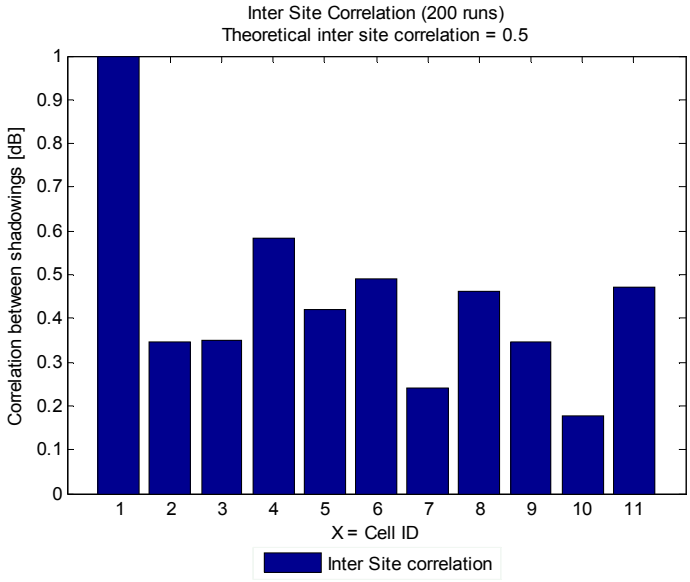


Figure 6 - Inter-site correlation

6.2.3 User Distribution over the Network

In the beginning of each run users are randomly drawn over the central base station (one or three sectors, depending on the configuration). In order to obtain a uniform-like spatial distribution, each user is uniformly dropped in a circle of length *2R* for the tri-sectored case and *R* for the omni-cell case. *R* is the radius of the hexagonal cell (or sector).

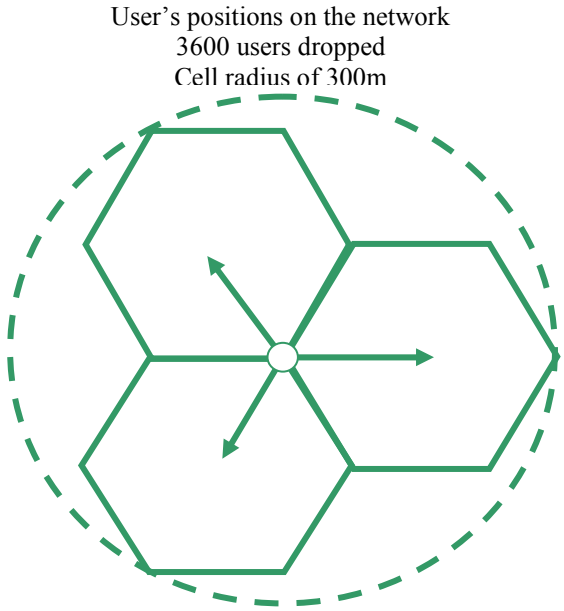


Figure 7 - Users uniform spatial distribution with best cell selection

Then the average coupling gain between the user and each one of its neighboring cells is computed. The average coupling gain includes the path-loss, the shadowing and the antenna gains. If the best cell (having the highest average coupling gain) of the user is one of the central cells, the cell is selected as the serving cell of the user. Otherwise, the position of the mobile is drawn again until the user is served by one of the central cells.

The following test was performed for 100 independent runs: in each run 36 mobiles were dropped over the 3 central cells (sectors) of the network of 19 tri-sectored base stations using the methodology presented before. It is to be expected that the resulting positions of users attached to the central base station are not uniformly distributed in the circle because of the best selection procedure. It seems that there are more users near the base station than at the cell borders. This is due to the fact that, in each run, a fixed number of users is dropped over the network and attached to the central base station. Indeed, if users were really uniformly distributed over the network, although there would be an average number of users attached per base station, there would be a variable number of users attached to each base station because of the particular geometric factor associated to each user.

6.2.4 Validation of the MIMO Channel Model Used in the System Level Simulations

In the simulation platform the MIMO channel is implemented according to the Spatial Channel Model (SCM) from 3GPP. The validation of the channel model that is implemented in the tool follows the guidelines from [85].

6.2.4.1 Root Mean Square (rms) Delay Spread

According to the SCM MIMO channel model the delay spread for each one of the 6 paths of the multi-path channel is a log-normal random variable given by equation (4).

$$\sigma_{DS} = 10\text{^}\left(\varepsilon_{DS}\alpha_n + \mu_{DS}\right), \quad n = 1,...,6 \tag{4}$$

Where:

- α_n is a correlated Gaussian random variable computed according to the correlation between the delay spread, angle spread and shadow fading, as explained in [85].
- $\mu_{DS} = E\left(\log_{10}(\sigma_{DS})\right)$ is the logarithm mean of the distribution of the Delay Spread (DS).
- $\varepsilon_{DS} = \sqrt{E\left[\log_{10}(\sigma_{DS})\right] - \mu_{DS}^2}$ is the logarithm standard deviation of the distribution of the DS.

Figure 8 plots the CDF of the root mean square of the delay spread. It is identical to the plot presented in [85] which corroborates the correctness in the generation of this parameter of the model.

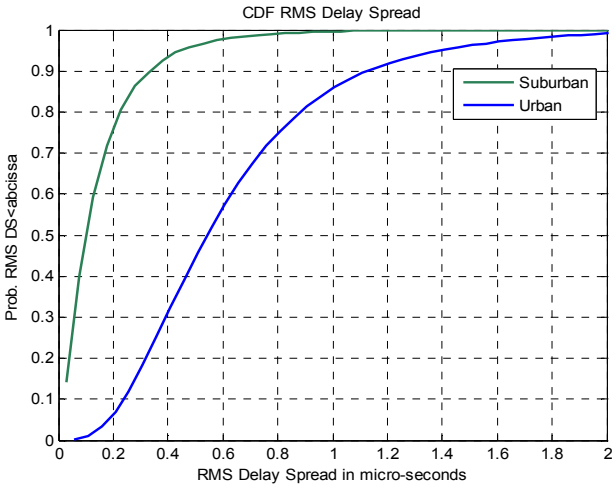


Figure 8 - RMS delay spread at the base station

6.2.4.2 Root Mean Square (rms) of the Angle Spread at Base Station

According to the SCM MIMO channel model the angle spread at the base station for each one of the 6 paths of the multi-path channel is a log-normal random variable given by equation (5)

$$\sigma_{AS,n} = 10\text{^}\left(\varepsilon_{AS}\beta_n + \mu_{AS}\right), \quad n = 1,...,6 \tag{5}$$

Where:

- β_n is a correlated Gaussian random variable computed according to the correlation between the delay spread, angle spread and shadow fading, as explained in [85].
- $\mu_{AS} = E\left(\log_{10}(\sigma_{AS})\right)$ is the logarithm mean of the distribution of the Angle Spread (AS).
- $\varepsilon_{AS} = \sqrt{E\left[\log_{10}(\sigma_{AS})\right] - \mu_{AS}^2}$ is the logarithm standard deviation of the distribution of the AS.

Figure 9 plots the graph for the CDF of the root mean square of the angle spread at the base station. It is identical to the plot presented in [85] which corroborates the correctness in the generation of this parameter of the model.

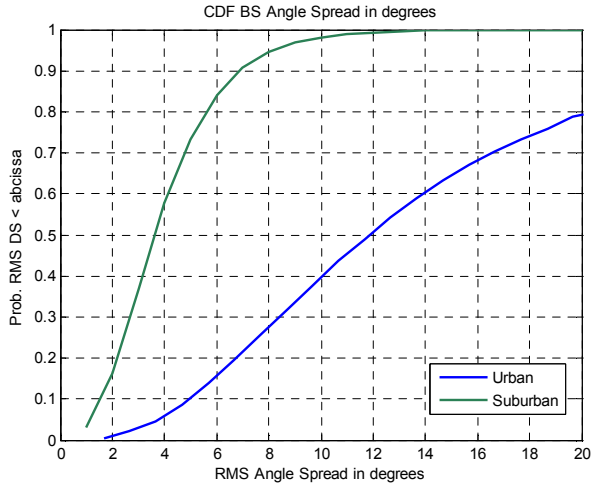


Figure 9 - RMS angle spread at the base station

6.2.4.3 Root Mean Square (rms) of the Angle Spread at Mobile Station

According to the SCM MIMO channel model the per-path angle spread at the mobile station is fixed at 35º for both suburban macro as well as for urban macro. The mean angle spread at the mobile station $E(\sigma_{AS}, MS)$ is equal to 68º for both scenarios.

For the mobile station the AS is computed according to the circular angle spread definition. A number of simulations for the AS at the mobile station are conducted and the circular angle spread is computed according to equation (6):

$$\sigma_{AS} = \sqrt{\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\left(\theta_{n,m,\mu}\right)^2 . P_{n,m}}{\sum_{n=1}^{N}\sum_{m=1}^{M} P_{n,m}}} \qquad (6)$$

Where:

- $P_{n,m} = P_n / 20$ is the power for each sub-pat
- $\theta_{n,m,\mu} = \begin{cases} 2\pi + \left(\theta_{n,m} - \mu_\theta\right) & \text{if } \left(\theta_{n,m} - \mu_\theta\right) < -\pi \\ \left(\theta_{n,m} - \mu_\theta\right) & \text{if } \left|\theta_{n,m} - \mu_\theta\right| \le \pi \\ 2\pi + \left(\theta_{n,m} - \mu_\theta\right) & \text{if } \left(\theta_{n,m} - \mu_\theta\right) > \pi \end{cases}$ and $\mu_\theta = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M}\theta_{n,m} . P_{n,m}}{\sum_{n=1}^{N}\sum_{m=1}^{M} P_{n,m}}$
- $\theta_{n,m,}$ is the AoA or AoD of the m^{th} sub-path.

Figure 10 plots the graph for the CDF of the mobile station circular angle spread for suburban macro and urban macro at the mobile station. It is identical to the plot presented in [85] which corroborates the correctness in the generation of this parameter of the model.

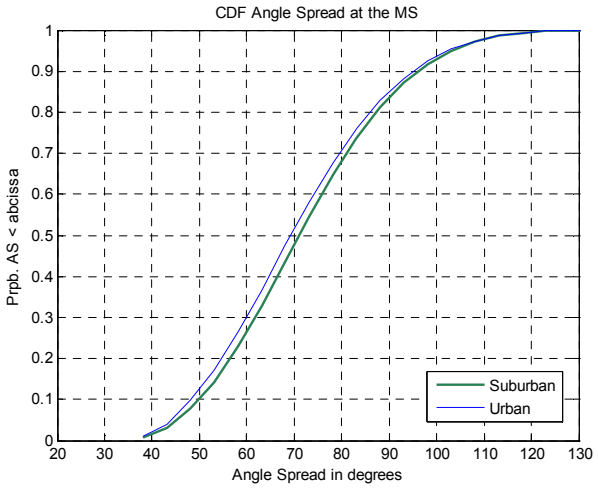


Figure 10 - CDF of MS angle spread (circular AS calculation)

6.2.4.4 CDF Distribution of the Powers from all Paths of the Multi-Path Channel

Figure 11 plots the graph for the CDF of all path powers for urban and suburban macro scenarios. Path power is computed according to step 5 in the SCM MIMO channel model. It is

identical to the plot presented in [85] which corroborates the correctness in the generation of this parameter of the model.

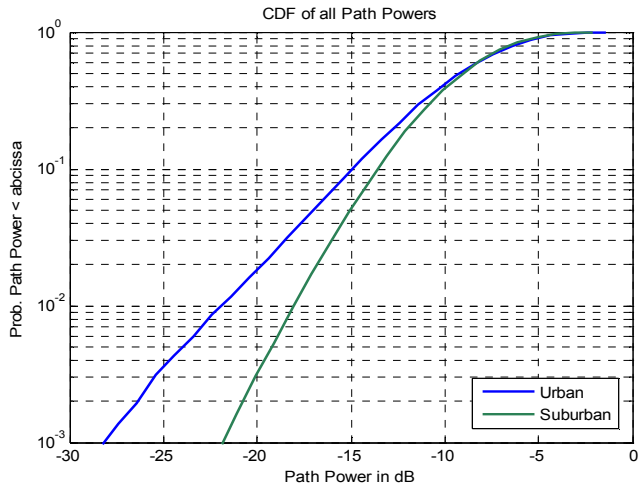


Figure 11 - CDF of all path powers

6.2.4.5 Dynamic Range of Variation for the Powers of all Paths in the Multi-Path Channel Model

Figure 12 plots the CDF of the dynamic range of all path powers for urban and suburban macro scenarios. It is identical to the plot presented in [85] which corroborates the correctness in the generation of this parameter of the model.

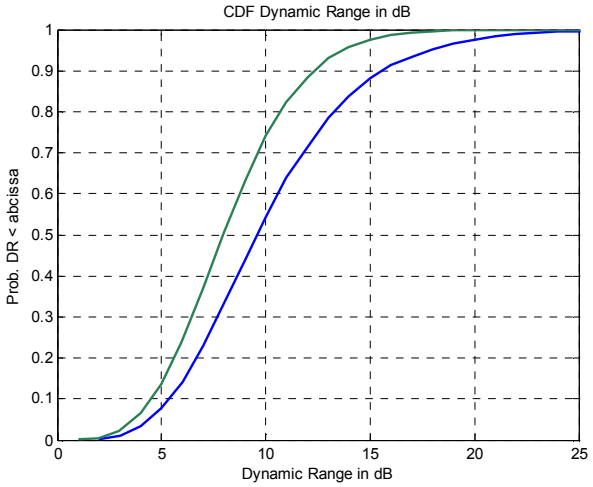


Figure 12 - CDF of dynamic range (dB)

6.3 Validation of the Dynamic Resource Allocation Module

The following sub-sections describe the tests and present the results obtained from the validation of the proposed DRA architecture used in the system level simulations. Each simulation run lasts for 75000 frame periods with 50 runs in total. The cell layout is made up of three tiers of base stations. Within each run 36 users are uniformly drawn along the three sectors of the central base station and transmission is according to the WWW traffic model. Neighboring cells are assumed as transmitting with full power at full load, i.e., they contribute

to the inter-cell interference only. A SISO transmission scheme with a simplified Maximum Ratio Combiner (MRC) is used in the receiver at each mobile station. Simulations are conducted for an urban deployment model with a mobile speed of 3 km/h and with a flat frequency channel model. Each slot in the TDD frame is assigned the same power for transmission, i.e., the maximum transmit power for data is uniformly distributed along the whole set of slots of the frame. It is also assumed that channel quality is perfectly estimated and that there is no error in the feedback of the channel quality indication (CQI) reports by each mobile station. The CQI is reported in every uplink sub-frame.

Cellular Layout	**19 cell, hexagonal grid layout 3 tiers (6 and 12 BS surrounding the central BS)**
BS-to-BS Distance	3000m
Minimum MS-to-BS Distance	35m
BS Antenna Horizontal Pattern	70º (-3dB) with 20 dB front-to-back ratio
BS Maximum Power	46dBm
BS antenna gain	15dBi
MS antenna gain	-1dBi
Antenna Configuration (NtxNr)	1:2
Propagation Model	L=128.1 + 37.6log10(R); R in kilometers
MS Losses	10dB
Specific fast fading model	Jakes spectrum with 3Km/m mobile station speed
Channel Model	Flat Frequency (1 path)
Standard Deviation Shadow Fading	8dB
Correlation Between Sectors	1.0
Correlation Between BS	0.5
Correlation Distance Shadow Fading	50m
Noise Figure (BS)	4dB
Noise Figure (MS)	7dB
Carrier frequency	2.5GHz
Channel bandwidth	10MHz
Thermal Noise Density	-174dBm/Hz
Maximum # of Retransmissions	8
Scheduler	Max C/I with prioritization
MCS Feedback Delay	4 Frame Periods
Fast HARQ scheme	Asynchronous Chase combining
Frequency re-use	1
Simulation Mode	Combined Snapshot-Dynamic
#of MS	12 MS per sector uniformly distributed in the central BS
Number of Runs	50
Number of TTIs	75000
Environment	Urban Macro cell
Traffic Model	3GPP WWW model

TABLE 4: SIMULATION PARAMETERS

The scheduler used is the maximum C/I. A prioritization process was implemented to identify those users whose retransmissions are delayed. In this scheme whenever a block of data is received with error it is classified as of low priority until a certain timer threshold has elapsed. If the block of data is not retransmitted after this time interval that block will be classified as of high priority. High priority users always have precedence over low priority ones. All new data blocks (un-transmitted data) are classified as of low priority.

The computation of the modulation and coding scheme (MCS) used in each transmission attempt is a process lasting for 2 frame periods. This corresponds to a time delay of 1 frame between two consecutive CQI measurements and a time delay of 1 frame between CQI measurement and use of the CQI by the base station. Table 4 presents the simulation setup.

6.3.1 User Geometry

The user geometry depends only on the relative position between the mobile station, its serving cell and each one of its neighboring cells. The higher the geometry factor the better the propagation conditions for the mobile station are and, consequently, the better the improvement in performance (decrease in the ratio of packets received in error to total amount of packets transmitted).

Along simulations the geometry factor was collected for all mobiles in all runs and for all cells. The CDF of the geometry factor samples is depicted in figure 13.

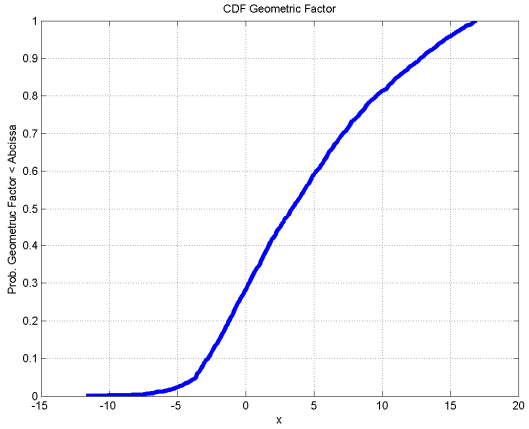


Figure 13 - User geometry factor distribution

It is in close agreement with the results obtained in [118] which are a good indication of the correctness of the system level modeling.

For all users, all runs and all cells the SINR of all packets received with success has been collected and the average for each user, along each run, was computed. Figure 14 plots the average SINR for each mobile along each run, versus the geometry factor, for different values of the Frame Erasure Rate (FER lower than 0.5% and greater than or equal to 0.5%).

As can be seen from the figure the average received actual SINR (taking into account actual fast fading values) is often very close to the user geometry value (that does not take into account fast fading). It can also be observed the linear increase in the average SINR with the geometry factor, which was expected because the average of the SINR values averages out the effect of fast fading. This is a consequence of the fact that the geometry factor behaves much like the received signal with fast fading averaged out.

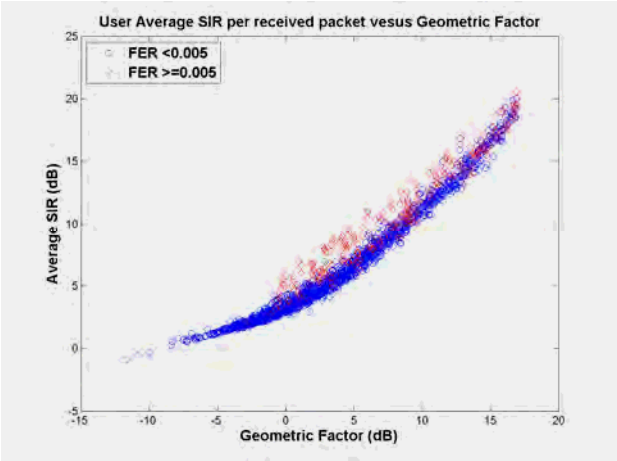


Figure 14 - Users geometry versus users average SINR of received packets

The points corresponding to the FER greater than or equal to 0.5% are associated to average SINR values greater than the ones obtained with the FER lower than 0.5% for the same geometry factor. This is due to the gain from the Chase Combining. By increasing the number of allowable transmission attempts, a significant reduction in the amount of samples with such bad channel states is expected.

6.3.2 User Residual Frame Erasure Rate (FER)

The residual FER was collected for all mobiles in all cells and for all runs. Figure 15 illustrates the complementary CDF (CCDF) of the residual FER.

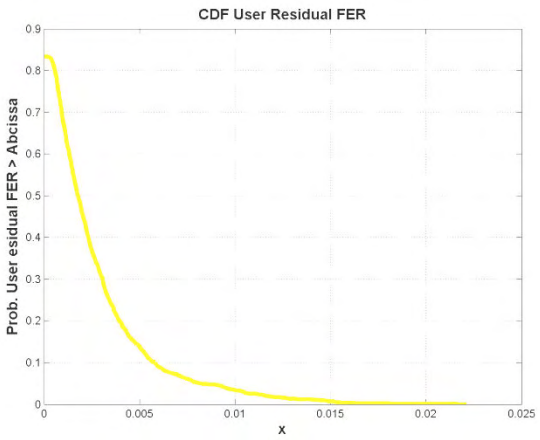


Figure 15 - Residual Frame Erasure Rate (FER)

As can be seen approximately 82% of users have a residual FER equal to 0. This plot corroborates the validation of the Hybrid Automated Repeat Request (HARQ) model with Chase Combining which is implemented in the system-level simulator for error recovery and transmission efficiency improvement. The gain from the Chase Combining results in a significant decrease in the achieved FER for all users. The ones closer to the cell boundary are more subject to bad channel quality due to inter-cell interference and therefore have higher FER values than the users closer to the base station.

6.3.3 Average Number of Transmission Attempts per Packet

The average number of transmission attempts per each packet sent over the air interface was collected for all mobile stations, in all cells, and for all runs. Figure 16 illustrates the plot of the average number of transmissions per packet for each user versus the geometry factor.

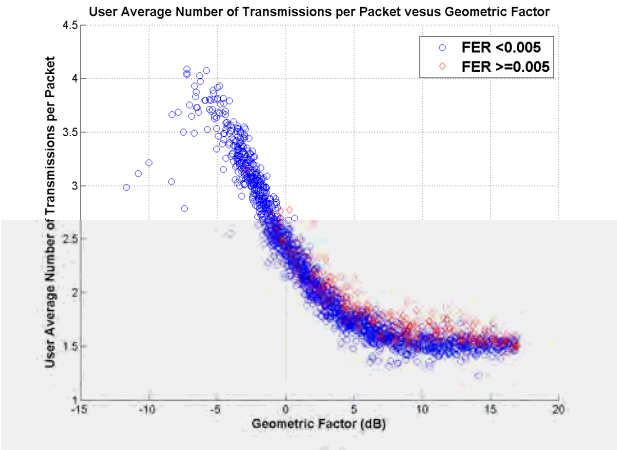


Figure 16 - User average number of transmissions per frame versus user geometry

As expected, the average number of transmission attempts per packet decreases with the increase in the geometry factor. This is because the channel quality becomes higher. With the improvement in the channel quality, the probability of decoding error for each packet received becomes smaller. Mobile stations with good channel conditions manage to receive all packets rapidly and with a small number of transmission attempts, because they manage to use link adaptation to use efficient MCS schemes in the transmission.

Due to the burstiness of WWW traffic mobile stations with good channel conditions do not require resources all the time because they empty their respective buffers rapidly. This results in transmission opportunities for mobiles with bad channel conditions, which need more transmission opportunities to receive their packets with success. The prioritization given to mobiles with packets waiting for a retransmission, after being waiting in queue for a time greater than or equal to the priority timer, result in more transmission opportunities for users with bad channel quality. These users are more sensitive to the inherent geometry factor and a higher number of transmission attempts occur even when they use the most robust MCS scheme due to bad channel quality.

Figure 17 plots the CDF of the average number of transmission attempts per user. It can be seen that most users (90%) have an average number of transmission attempts which is lower than 3. This plot confirms the inherent benefit of using Chase Combining: as the maximum number of transmission attempts is set to 8, only a small fraction of this number is needed in order to receive, with success, most of the packet transmissions over the air interface.

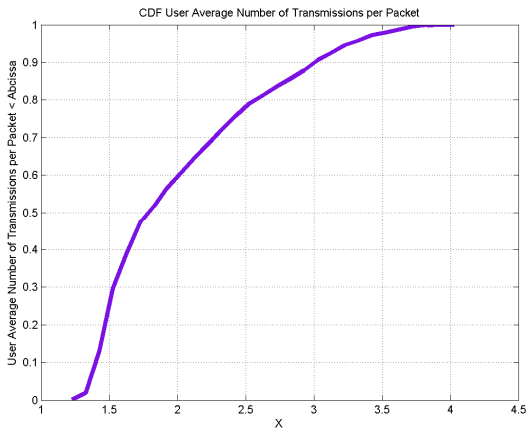


Figure 17 - CDF of user average number of transmissions per frame

6.3.4 Number of Times a User has been Scheduled

Figure 18 is a plot of the number of times a user has been scheduled versus the geometry factor, for users with residual FER lower than 0.5%.

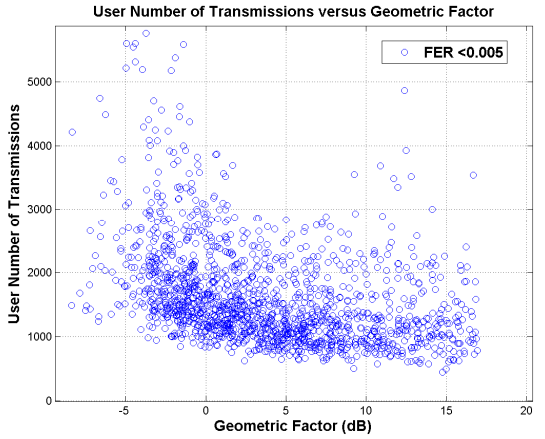


Figure 18 - Average number of transmissions per packet versus user geometry

As could be expected, users corresponding to low values of the geometry factor have more transmission attempts than users with higher values of the geometry factor. This is due to the high percentage of retransmission attempts, due to bad channel quality, for users closer to the cell boundary and more subject to inter-cell interference.

Because of link adaptation users with good channel quality need fewer transmission attempts to empty their buffers for the same amount of information as for users with bad channel quality, where most of the time the most robust scheme must be used, resulting in more transmission attempts to transmit the same amount of information.

Figure 19 illustrates the plot of the CDF of the number of transmission attempts per user. As can be seen the CDF is characterized by a short tail and a fast transition zone. This means that 10% to 90% of the users perform a number of transmission attempts in the range between 1000 and,

roughly, 2800. Fewer users over extrapolate this range. This result corroborates the performance of the DRA, namely the effect resulting from HARQ in recovering packets received with error.

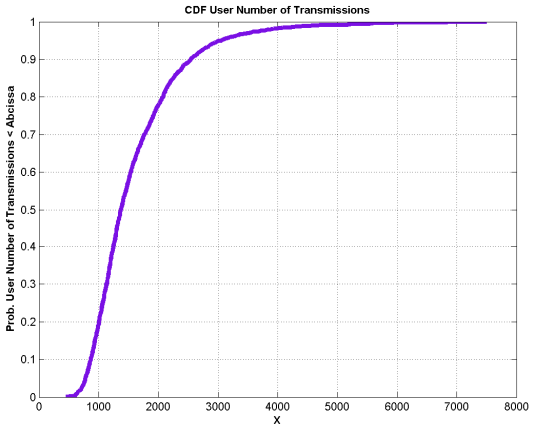


Figure 19 - CDF of the number of transmission attempts per user

6.3.5 Average Received SINR per Packet

For each user and each run the average of the SINR for all packets received with success has been collected and computed. The CDF of the average SINR per packet is plotted in figure 20.

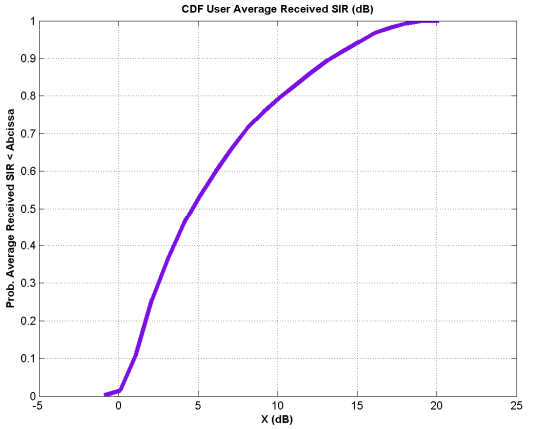


Figure 20 - CDF of average received packet SINR per user

This plot shows the dynamic range of variations for the average SINR and is in accordance with the set of MCS schemes used for the link level interface. The fact that this curve is not steep, albeit the scheduler used being the maximum C/I, is related to the type of traffic model used: WWW with burst packet transmission. After users with good channel finish their transmissions, users with worse channel conditions are given opportunities for transmission attempts. It can also be seen that users with CQI not good enough to keep the estimated Block Error Rate (BLER) higher or equal to 10%, for the most robust MCS scheme, also have opportunities to transmit.

6.3.6 User Service Throughput

The user service throughput has been collected for each user and each run and is plotted as a function of the geometry factor in figure 21 for users with residual FER lower than 0.5%. It can be noticed that all users achieve essentially the same service throughput no matter their channel quality. This is due to the burst nature of the traffic model and also to the low offered load to the system.

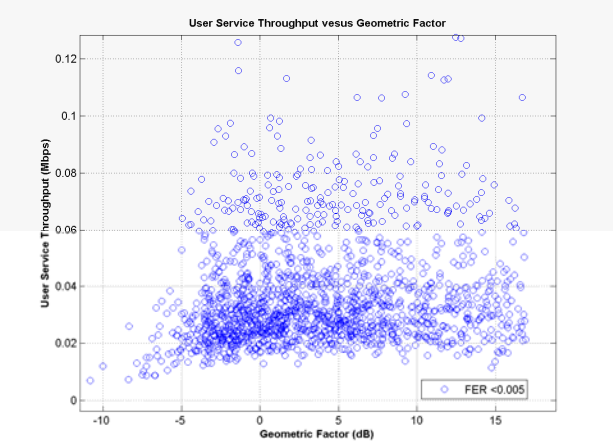


Figure 21 - User service throughput versus user geometry (users with residual FER <0.5%)

The CDF of the service throughput is plotted in figure 22. As can be seen 80% of the users have a service throughput greater than 0.5 Mbps and this can be considered as an estimation of the offered traffic load due to the low residual FER and burst nature of the WWW traffic model used in the simulations.

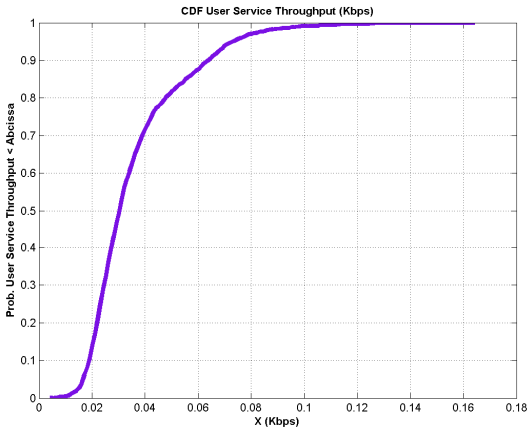


Figure 22 - CDF of user service throughput

6.3.7 Average Packet Delay

The average packet delay has been collected for each user and each run and is plotted as a function of the geometry factor in figure 23 for users with residual FER lower than 0.5%. As could be expected the average packet delay is lower for users with good channel quality. This is

a consequence of the link adaptation mechanism which empties transmission buffers in a much faster pace than users with bad channel quality.

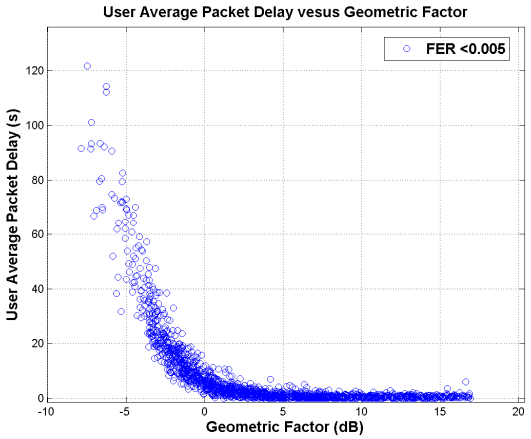


Figure 23 - User average packet delay versus user geometry factor

Figure 24 plots the CDF of the average packet delay for users with FER lower than 0.5%. As can be seen the resulting average packet delay is quite high for WWW application service.

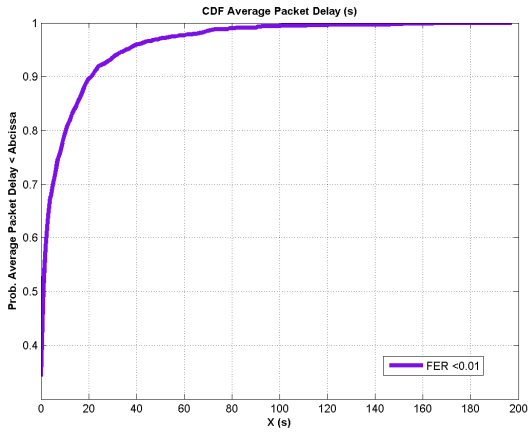


Figure 24 - CDF of user average packet delay

The main reasons for this are:

- The higher values for the packet delay are strongly influenced by those users transmitting with bad channel quality even for the most robust MCS scheme. Due to the burst nature of the traffic model, quite often they perform transmission attempts. As they are not blocked in transmission and there is no packet drop due to time-out, a number of transmission attempts are often performed before they achieve successful decoding of the radio block.
- The average offered load of the WWW traffic model used is 2 Mbps, i.e., with lengthy packets which must be fragmented for transmission. Invariably, each packet occupy all resources available in the MAC frame, which means that it is normal to expect that only a single user is scheduled for transmission in each transmission time interval.

- The scheduler is the Max C/I which results in starvation for remaining users with bad channel quality.
- There is no mechanism for the control of packet delay. There is a prioritization mechanism in the access to system resources, but only for those users which have already attempted their first transmission. This means that packets are kept in buffer waiting for transmission. When users with better channel have their buffers flushed, after performing the transmission, users with worse channel quality have their opportunity to transmit, and send their packets which were waiting for a long time in buffer.
- The simulations conducted for system level validation assumed transmission time intervals of 10 ms, equal to two frame intervals. As a matter of fact each transmission time interval encompasses 4 steps: (i) computation of the resource allocation map (RAM); (ii) broadcasting of the RAM; (iii) transmission of the information in the MAC frame; and (iv) report of the ACK/NACK message. The average length of the packets for the WWW traffic model used, together with the 10 ms TTI results in reduced values for the over-the-air and service throughput and increases the average packet delay.
- The higher values for the packet delay are strongly influenced by those users transmitting with bad channel quality even for the most robust MCS scheme. Due to the burst nature of the traffic model, quite often they perform transmission attempts. As they are not blocked in transmission and there is no packet drop due to time-out, a number of transmission attempts are often performed before they achieve successful decoding of the radio block.

This is also corroborated from figure 23 where it can be seen that users with geometry greater than zero correspond to packet delays closer to the 10 to 20 seconds delay range. It can also be seen that there are many transmission attempts for users with geometry between 0 and -4 dB. These values cannot guarantee the desired BLER even for the most robust MCS scheme.

It is important to mention that the high percentage of users transmitting with negative geometry is due to: (i) the low offered load, (ii) the relatively small number of active users in the system and (iii) the burst nature of the WWW traffic model used.

6.4 Scheduling Algorithms

In the sections that follow the different scheduling algorithms considered in the system level simulations and conducted under the scope of this thesis are described in detail.

6.4.1 Round Robin (RR)

A simple scheduling algorithm which provides maximal fairness in resource allocation and isolation among service flows is the Round Robin (RR), which has its roots in wired networks domain. This is a classic time division multiplexing-based algorithm, where the delay between successive transmissions to the same user is fixed and equal for all. The scheduler gives priority

to each user in a sequential way without taking into consideration any user QoS requirements. It is simple to implement but is often inadequate to meet objectives such as QoS and throughput maximization.

6.4.2 Maximum C/I (CI)

If, in each scheduling period, radio resources of each single cell in the network are attributed to the user with the best channel state, this attribution scheme will result in the maximization of the network throughput in the longer term. A packet scheduler attributing radio resources to the user with the best channel quality among all active users in the cell is called opportunistic (*opportunistic scheduling* -OS) [26]. This is the principle behind the Maximum C/I (CI) packet scheduler operation.

This scheduler is used in the estimation of the total system capacity with the proposed DRA, as it results in the maximization of the throughput achieved over the air interface, assuming a full queue scenario. The CI scheduler opportunistically assigns resources to the user with the highest channel gain by scheduling, at the beginning of each radio frame, *n*, the mobile with the best channel quality indicator among the set of *K* active mobiles. The user selection rule is given by equation (7):

$$k(n) = \arg \max_{i \in \{1,\ldots,K\}} CQI_i(n), \quad n = 0,1,2,... \tag{7}$$

Where:

- $CQI_i(n)$ is the CQI reported by the i^{th} mobile in the n^{th} frame period.
- N is the amount of active users in the cell.

The multiuser diversity gain resulting from OS is enhanced with the increase in the number of active users in the cell, because this results in a higher degree of variability in the composed channel set and, as a consequence, the probability of finding a user with a link quality in a better state is increased. OS is not adequate for scenarios with small channel variability, such as the ones in line of sight propagation. In this case multiple antennas can be used to induce signal variability [111, 126].

6.4.3 Max C/I over Average C/I (AvgCI)

In spite of its efficiency in resource utilization, OS scheduling results in unfairness, namely for those users on the edge of the cell which are affected by bad channel quality. Also, OS is not appropriate for scheduling packets from service flows with stringent delay bounds. This is because the OS scheduler is greedy, in the sense that users closer to the base station can starve users in the edge of the cell, resulting in many packets dropped due to delay bound violation before the channel eventually gets good enough for resource allocation. Bad resource efficiency negatively affects the assumed objective of long term maximization of the cell throughput.

A variation of the CI scheduler, which attempts to overcome its limitations, is the Maximum C/I over Average C/I (AvgCI) scheduler. This scheduler compensates the influence of the mean SINR on the SINR measured by each user. The mean SINR depends on the antenna gains, path-loss and shadowing, and is highly influenced by the distance from the mobile to base station. Mobiles on the cell edge have a much smaller mean SINR due to the distance from the cell center and due to the inter-cell interference, which results in starvation on the transmission opportunities access. This scheduler results in a more symmetrical SINR value for each user, mainly due to fast fading, by compensating the differences on the mean SINR. The mobiles are ranked according to the scheduling rule given by equation (8).

$$k(n) = \arg \max_{i \in \{1,\ldots,K\}} \frac{CQI_i(n)}{\overline{CQI_i}(n)}, \quad n = 0,1,2,... \tag{8}$$

Where $\overline{CQI_i}(n)$ is the SINR from user *i*, averaged over a period of T_{av}, and is computed by a smoothing filer as given by equation (9).

$$\overline{CQI_i}(n) = \lambda x \overline{CQI_i}(n-1) + (1-\lambda) CQI_i(n), \quad n = 0,1,2,... \tag{9}$$

Where λ is the filter averaging coefficient, which is equal to $\lambda = 1 - TTI / T_{av}$. In the simulations $T_{av} = 1.5\ s$ and TTI is the frame period (equal to 5 ms).

6.4.4 Proportional Fairness (PF)

The Proportional Fairness scheduler was proposed in [111, 127-128] as a means to combat the inefficiencies of the CI scheduler. This scheduling algorithm offers a good compromise between many conflicting objectives, including throughput maximization, fairness to all service flows and ease of implementation.

At the beginning of each radio frame *n* the scheduler selects for transmission the mobile with the highest ratio of estimated maximum data rate $R_i(n)$ to current average throughput $T_i(n)$, according to the rule defined in equation (10).

$$k^{CI}(n) = \arg \max_{i \in \{1,\ldots,K\}} \frac{R_i(n)}{T_i(n)}, \quad n = 0,1,2,... \tag{10}$$

The maximum data rate $R_i(n)$ for user *i* is computed from the reported CQI value and from the look-up tables, and is given by equation (11).

$$R_i(n) = DRC_i(1 - BLER) \tag{11}$$

Where $DRC_i(n)$ designates the reported data rate, from the look-up table, for user *i* in frame period *n* and BLER is the pre-defined maximum Block Error Rate (10%).

The average user $T_i(n)$ throughput in frame period *n* is computed according to equation (12).

$$T_i(n) = \lambda T_i(n-1) + (1-\lambda) DRC_i(n) \delta_i(n), \quad n = 0,1,2,... \tag{12}$$

Where $\delta_i(n) = 1$ if user *i* is backlogged.

If the channels are symmetrical and independent and identically distributed the PF scheduler is fair in terms of resource allocation and effective in terms of capacity maximization.

6.4.5 Modified Largest Weighted Delay First (M-LWDF)

The M-LWDF [129] attempts to statistically guarantee a certain percentage of packets dropped due to maximum delay bound violation. This means that this probability must be lower than a given threshold, which depends on the type of service. It is given by equation (13).

$$P\{W_i > T_i\} \leq \delta_i \tag{13}$$

In frame period *n* the scheduler schedules for transmission the user satisfying equation (14).

$$k(n) = \arg \max_{i \in \{1,\ldots,K\}} \left(\gamma_i W_i(n) r_i(n)\right) \tag{14}$$

Where:

- $W_i(n)$ is the delay of the head of line packet on the queue of user *i*.
- $r_i(n)$ is the channel capacity from user *i*.
- γ_i is an arbitrary positive constant.

The delay $W_i(n)$ can be replaced by the queue length $Q_i(n)$ (in bits), which is the amount of data in the queue of each user. The algorithm has the same performance with this slight modification.

In this algorithm the scheduling decision depends on both current channel conditions (by means of $r_i(n)$) and the states of the queues (by means of $W_i(n)$ or $Q_i(n)$). The choice of the constant γ_i enables the control of packet delay distributions for different users: increasing the parameter γ_i for user *i* while keeping the γ's of other users unchanged, reduces packet delays for this flow at the expense of a delay increase for the others.

In order for the packet delay requirement to be satisfied the parameter γ_i must be appropriately defined. A good rule of thumb is given by equation (15).

$$\gamma_i = a_i / \bar{r}_i \tag{15}$$

Where:

- $a_i = -\dfrac{\log(\delta_i)}{T_i}$.
- $\bar{r}_i$ is the average channel rate with respect to user *i*.

This parameter embodies QoS requirements and provides QoS differentiation among flows.

The greater the user current packet delay, channel quality relative to its average level and the higher the QoS requirement, the greater the chance for this user being scheduled. This rule

balances different users' probabilities of delay violation relatively to their maximum allowed values, δ_i

6.4.6 Exponential (EXP)

The exponential (EXP) scheduler [130] is a variation of the PF scheduler that also tries to explicitly equalize the latencies of all users when their differences become large. At frame period *n* the algorithm selects for transmission the user satisfying the condition given in equation (16).

$$j = \arg \max_{i \in \{1,\ldots,K\}} \left(a_i \frac{r_i(n)}{\overline{r_i(n)}} \exp\left(\frac{l_i(n) - \overline{l(n)}}{1 + \sqrt{\overline{l(n)}}} \right) \right) \quad (16)$$

Where:

- $\overline{l(n)} = \frac{1}{K} \sum_{i=1}^{K} l_i(n)$ is the average of the head of line packet delays observed by all *K* users in the system at frame period *n*.
- a_i is the so-called service specific user weight. It is used in the support of multi-service provisioning, by assigning different weights to users based on QoS requirements.

The EXP rule tries to equalize the weighted delays $a_i w_i(n)$ of all queues when their differences are large:

- If one of the queues would have a larger weighted delay than the others by more than order $\sqrt{\overline{aW}}$, then the exponential term becomes very large and overrides channel considerations (as long as its channel can support a non-zero rate), hence leading to that queue getting priority.
- For small weighed delay differences (less than order $\sqrt{\overline{aW}}$), the exponential term is close to 1 and the policy becomes the proportionally fair rule.

The term $\overline{aW}$ in the exponent can be dropped without changing the rule as it is common for all the queues and the factor 1 in the denominator of the exponential is present simply to prevent the exponent from blowing up when the weighted delays are small.

6.5 Performance Evaluation of the Dynamic Resource Allocation Module

This section elaborates on the steps followed in the performance evaluation of the proposed DRA which is implemented in the basic system level simulation platform. Its performance is evaluated for three different types of schedulers. Two of them are commonly referred in the literature: Round Robin (RR) and Maximum C/I (CI), and the third one is a variation of the Maximum C/I, named Maximum C/I over Average C/I (AvgCI). System level simulations were

conducted for the SISO channel model, with the ITU PedB and PedA channels for a mobile speed of 3 Km/h and the ITU VehA for a mobile speed of 30 Km/h; and also for the MIMO channel. The simulation setup is the same one used in the validation tests presented in the previous section, according to table 4.

The following traffic models were considered in the simulations:

- Full Queue.
- 3GPP's Near Real Time Video (3GPP NRTV).
- 3GPP's World Wide Web (3GPP WWW).

Whenever used with the CI scheduler, full queue traffic model is particularly suitable for the estimation of system capacity and the results obtained are compared against the theoretical benchmark values provided by the WiMAX Forum. Users transmitting with full queue traffic are always backlogged, i.e., they always have information in their buffer waiting for a transmission opportunity. As the buffer is never emptied, even if this mobile station is scheduled for transmission, all active mobile stations are always competing for the same resources. This is illustrated in figure 25.

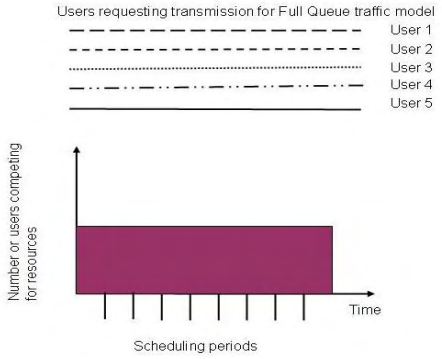


Figure 25 -Number of competing users in Full Queue Traffic Model

Contrary to full queue, web browsing traffic is inherently burst: it is characterized by short periods of packet generation interlaced with long periods of inactivity.

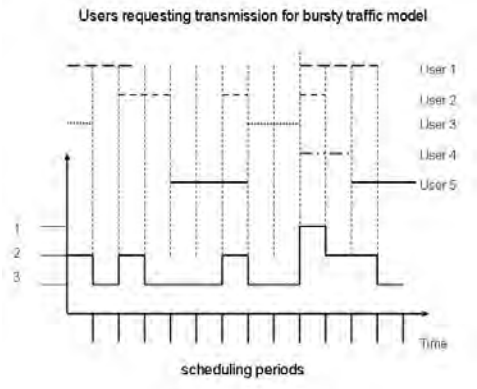


Figure 26 - Number of competing mobiles in Bursty Traffic Model

Even if the mobile is not allowed to transmit for some period of time, as web browsing traffic is insensitive to delay, packets can accumulate in the queue, waiting for a transmission

opportunity. For this reason web browsing is also commonly referred to as “elastic”. Mobile station’s buffers are often empty because the traffic is burst and also because the queue can be emptied after the mobile station has been scheduled. Therefore, at each frame period there is often a different number of mobile stations competing for radio resources. This is illustrated in figure 26.

The following performance metrics were used in the evaluation of the system level performance:

Average Service Throughput per-Cell

The average service throughput per cell is the average throughput effectively received by each active user in the system, assuming the whole simulation duration. By effective one means packets received with success (no error in decoding).

Average Over-The-Air (OTA) Cell Throughput (kbps/cell) (3GPP Definition)

The average over-the-air throughput per cell is a metric very similar to the user service throughput. The only difference is that the average is computed by considering only the total amount of time spent in the transmission of information to each user.

Average Over-The-Air (OTA) Cell Throughput (kbps/cell) (Peak Bit Rate Definition)

This metric is very similar to the OTA throughput. But here all bits (correct and erroneous) are considered in the computation.

Offered Cell Load (kbps/cell) (3GPP Definition)

This metric is used in the evaluation of the data load (in kbps) withdrawn from the base station’s buffers for transmission.

Offered Cell Load (kbps/cell) (Network Definition)

This metric measures the offered load from the core network to the base station for all mobile stations being simulated in the system over the whole simulation run.

User Average Peak Bit Rate at a Given Distance (kbps)

This metric gives the average peak bit rate of a given user at a given distance *d*, in steps of 10m, from the base station.

Per User Service Data Throughput

The user’s service data throughput is defined as the ratio of the number of information bits successfully received by the user and the total simulation run time.

Per-User Average Service Throughput

The average per-user service throughput is defined as the sum of the user service throughput of each user divided by the total number of users in the system.

Cell Edge User Throughput

The cell edge user throughput is defined as the 5^{th} percentile point of the CDF of user’s average packet call throughput.

Packet Delay

For an individual packet the delay is defined as the time elapsed between the instant when the packet enters the queue at transmitter and the time when the packet is received successfully in the receiver. If a packet is not successfully delivered by the end of a run its ending time is the end of the run.

User Average Packet Delay

The average packet delay is defined as the average interval between packets originated at the source station (mobile or base station) and received at the destination station (base or mobile station) in a system for a given packet call duration.

Residual Frame Erasure Rate (FER)

This metric is computed for each user and for each packet service session. A session contains one or several packet calls depending on the application. It starts when the first packet of the first packet call of a given service begins and ends when the last packet of the last packet call of the same service has been transmitted. One packet call contains one or several packets. The Residual FER measures the percentage of the total amount of dropped packets in the packet service session.

Packet Loss Ratio

The packet loss ratio is computed for each user and for each packet service session. It measures the percentage of the total amount of packets discarded due to time-out (delay bound violation and maximum number of transmission attempts achieved).

6.5.1 Performance Evaluation for the Full Queue Traffic Model

Figure 27 is the plot of the CDF of the user service throughput for all three schedulers.

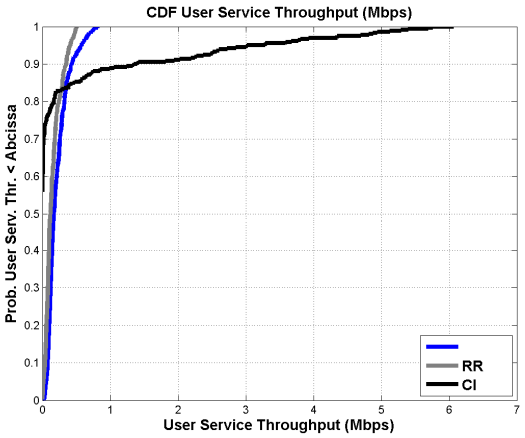


Figure 27 - CDF of user service throughput

As can be seen the AvgCI scheduler is the one with best performance amongst the three proposed schedulers. In terms of fairness, its performance is in between the performance of the RR and CI schedulers. As can be seen from the plot, roughly 75% of the users have no access to radio resources (they have null service throughputs).

The normalization of each user's instantaneous CQI by its average over time compensates the differences among the mean values of the CQI for all users in the cell. As a matter of fact, the channel perceived by the AvgCI scheduler for each active user is symmetric among all users. This is the potential gain of this scheduler as it gives transmission opportunities not only to users close to the base station, and therefore with good channel quality, but also to users in the edge of the cell, and therefore with bad channel quality due to inter-cell interference.

As can be seen from figure 28, when users are symmetric the CI scheduler tends to select users on the peaks of their channels, enabling to achieve the multi-user diversity gain. When users are non symmetric (with different average channel quality) the CI scheduler always selects the same user, the one with the best channel quality, not considering the amount of information in its buffer, and therefore no multi-user diversity gain is achieved. Users in the cell edge use more robust and therefore less efficient MCS schemes in the transmission and this affects the resulted user service throughput. This is the reason for the CDF of the service throughput of the AvgCI being close to the CDF of the service throughput of the RR.

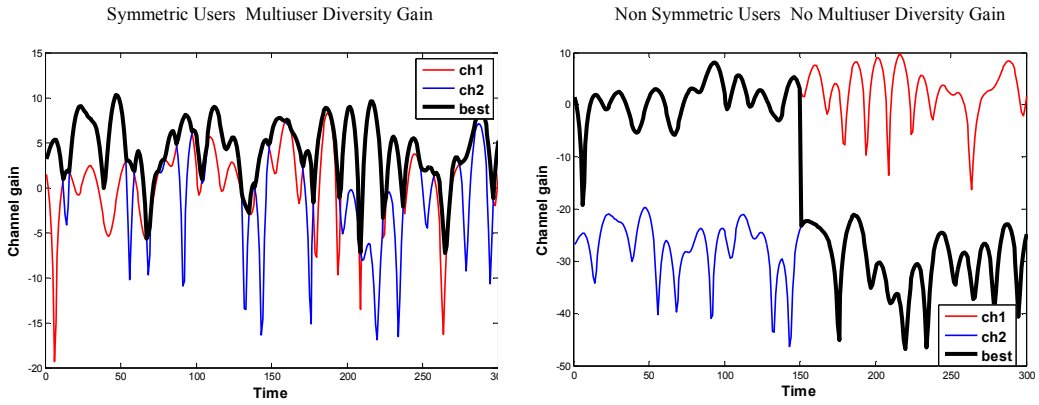


Figure 28 - Multi-user diversity concept

As can be seen from the plot of the MCS distribution in figure 29 the CI scheduler always select the user with the best channel quality which results in transmissions with the most robust MCS scheme.

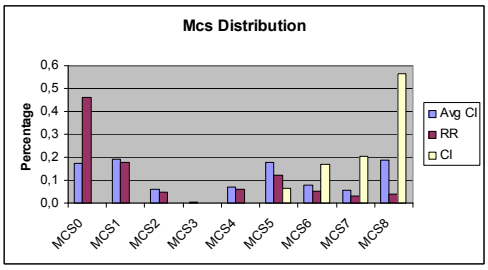


Figure 29 - MCS distribution for full queue traffic model

Differently from the CI scheduler, the AvgCI strives to transmit with more efficient MCS schemes than the RR because, for symmetric channels, it tends to select users on the peaks of

their channels, enabling to achieve a multi-user diversity gain. As transmission opportunities are given for users in the cell edge, more robust MCS schemes are selected for these users. As the RR scheduler does not consider the channel quality in user's selection no multi-user diversity gain is achieved.

Figure 30 is the plot of the peak bit rate averaged over all runs and mobiles versus the distance from the base station, in steps of 10 meters. Most transmissions with the CI scheduler are for users closer to the base station and those few ones, for users closer to the cell edge, are with lower user service throughputs as expected.

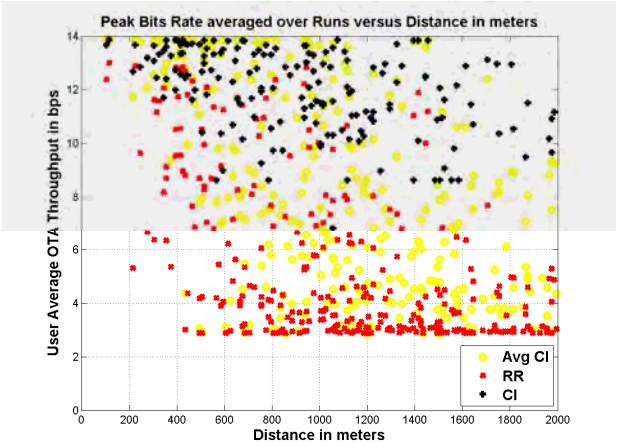


Figure 30 - Peak bit rate vs. distance in meters

The RR scheduler services users no matter their location on the cell, but with a lower user throughput than the one achieved with the AvgCI. This is because this scheduler has no multiuser diversity gain. The AvgCI results in a good compromise between fairness and service rate.

Figure 31 is the plot of average throughput per cell over all runs and mobiles for all three schedulers.

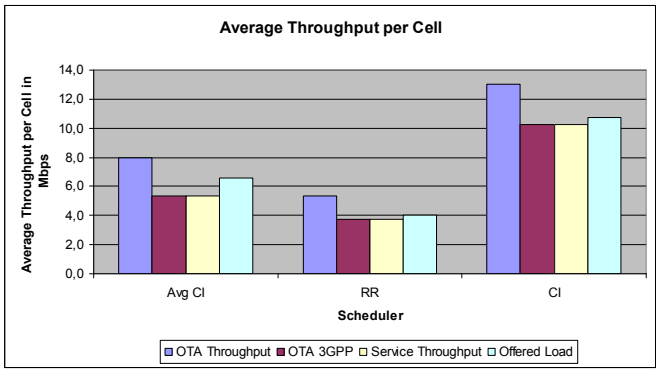


Figure 31 - System throughput for all three schedeulers

As PUSC sub-channelization mode randomly assigns sub-carriers along the symbol spectrum, the fact that some sub-carriers will be more attenuated than others with the PedB channel model, result in smaller values of the compressed Exponential Effective SINR value obtained from the

CQI in the preamble of the frame. As all sub-carriers are roughly attenuated by the same value with the PedA channel model the EESM SINR value will be higher for this channel model. This is illustrated in figure 32.

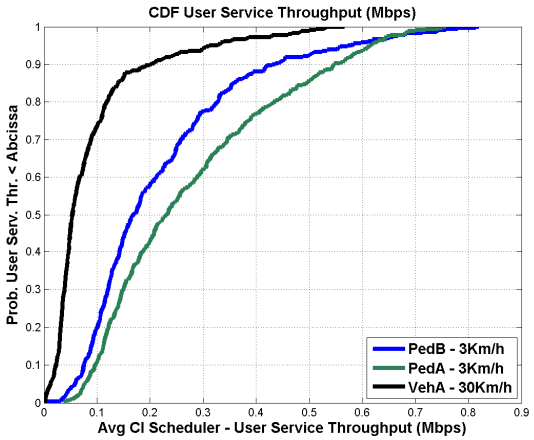


Figure 32 - CDF of user service throughput for the AvgCI scheduler

As can be seen, the PedA channel model results in the best performance in terms of system throughput among all three channel models. In order to overcome the bad performance achieved with the VehA channel mode with 30km/h a different link to system level interface, based on average channel values, should be used as look-up tables, or more robust MCS schemes could be employed.

Also, the PedB channel has a larger delay spread and hence a smaller coherence bandwidth than the PedA channel. This means that with the PedB channel different sub-carriers will be affected differently along the spectrum of the OFDM symbol, while with the PedA channel model all sub-carriers will be roughly affected the same way, i.e. the channel is roughly flat for the PedA channel.

Figure 33 plots the average system throughputs per cell over all runs for the AvgCI scheduler under the three types of channel models used in the simulations.

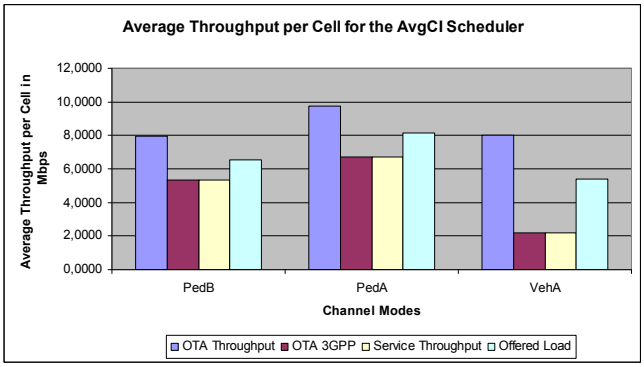


Figure 33 - System throughput for the AvgCI scheduler for all channel modes

Figure 34 is the plot of the throughput fairness, according to the data rate criteria as defined in annex C. The throughput fairness, $F(t)$, is a short-term fairness indicator, updated periodically

every τ ms (τ is suggested to be 20 or 40 ms). The minimum of $F(t)$ can serve as an indication of how much fairness is maintained all the time: the closest $F(t)$ is to one the more fair the scheduler is.

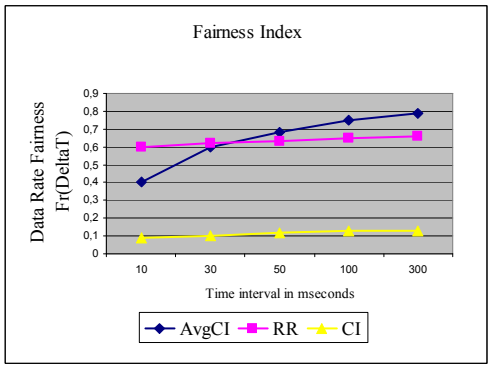


Figure 34 - Short-term data rate fairness vs. time interval

It can be confirmed from this plot that the AvgCI is the fairest scheduler among the three regarding data rate allocation.

6.5.2 Performance Evaluation for the Near Real Time Video (NRTV) Traffic Model

Figure 35 is the plot of the CDF of the user service throughput for all three schedulers and for the 3GPP NRTV traffic model.

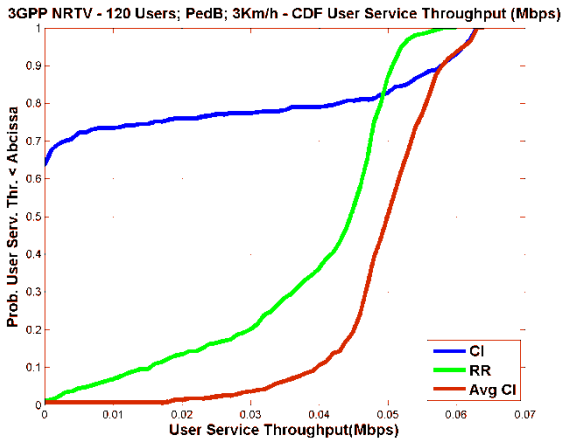


Figure 35 - CDF of User Service Throughput for NRTV users with PedB and 3km/h

The channel model used in the simulations is the ITU PedB with 3 km/h mobile speed. As can be seen from the plot, among the three schedulers the AvgCI presents the best performance, mainly for users in the edge of the cell:

- It can be noticed that about 65% of the users are not serviced with the CI scheduler. The long smooth tail of the CDF distribution is characteristic and indicative of the starvation suffered by most of the users, which do not have access to frame resources for transmission.

- It can also be noticed that the AvgCI scheduler performs much better than the other two schedulers regarding the 10% tile of the CDF of the user service throughput. This is the point corresponding to those users closer to the edge of the cell, which are therefore most affected by interference from neighboring cells. Their average service throughput is 40 kbps.
- Regarding the RR scheduler one can say that the 10% tile of the CDF of the user service throughput is equal to 15 kbps, resulting therefore in a significant degradation of the user service throughput, when compared to the performance of the AvgCI scheduler.
- The CDF of the user service throughput of both the AvgCI and RR schedulers is steep around the medium point of the CDF (50% tile). This is indicative of the fact the both schedulers provide fair access to system, although the AvgCI scheduler results in better performance.
- From the 90% of the CDF distribution one notices that both AvgCI and CI schedulers produce the same service throughput: roughly 60 kbps. This point corresponds to the users which are closer to the center of the cell (closer to the base station).

Figure 36 is the plot of the CDF for the average packet delay under the same conditions. It can be confirmed that the AvgCI scheduler presents the best performance in terms of packet delays as 90% of the users have an average packet delay lower than 15 seconds while the RR and CI result in average packet delays of roughly 40 and 70 seconds respectively.

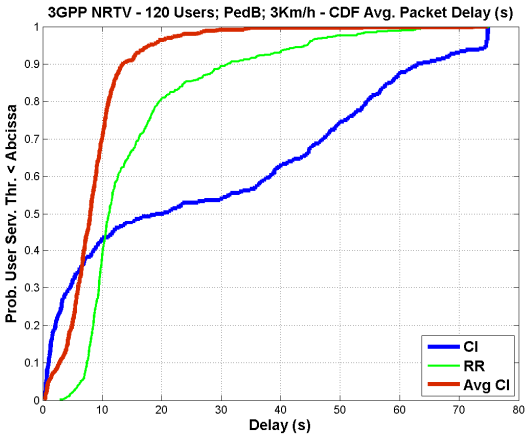


Figure 36 - CDF of average Packet Delay for NRTV users with PedB and 3km/h

Figures 37and 38 plot the MCS distribution and the system throughputs for the three schedulers, respectively. It can be seen that the AvgCI results in the best performance in terms of service throughput.

The gap between the OTA Throughput and the Service Throughput for the AvgCI scheduler is higher than for the CI one. This is because, as the AvgCI results in more transmission opportunities for users in the edge of the cell, and hence with worse channel quality, there is a

higher percentage of retransmission attempts for this scheduler due to bad channel quality resulting from inter-cell interference.

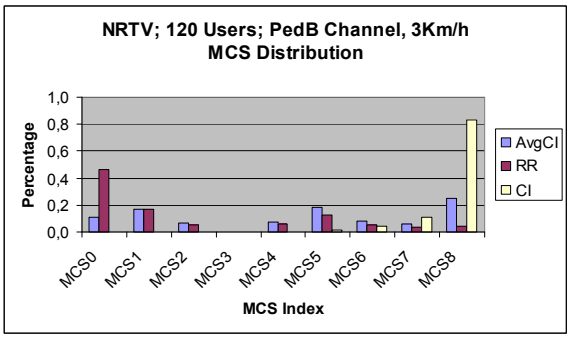


Figure 37 - MCS distribution

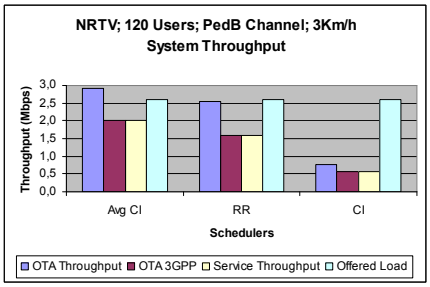


Figure 38 - System Throughput

All results presented up to now are for system level simulations performed with the PedB channel model. As the channel model used influences the performance of each scheduler differently, more system level simulations were conducted with two other types of channel models: PedA and Veh A. These are commonly models proposed by ITU for multipath propagation simulation of SISO channel. Therefore, the impact of multipath propagation regarding the Medium Access Layer of a WiMAX network was evaluated.

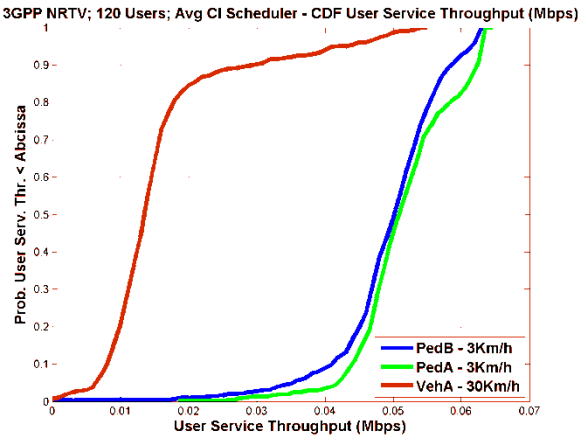


Figure 39 - CDF of user Service Throughput for AvgCI scheduler

Figure 39 is the plot of the CDF of the user service throughput for the AvgCI scheduler resulting from the use of these three channel models. For the reasons pointed out in the Full

Queue Traffic model, as expected, both PedB and PedA channel result in significant better user service throughputs when compared to the VehA channel model.

Figure 40 plots the average system throughput over all three simulated models for the same AvgCI scheduler where it can be noticed the difference in service throughput achieved with PedA or PedB and VehA channel models. This is because of the lack of accuracy in the CQI reported by each mobile station due to coherence time. The OTA Throughput is slightly higher for the VehA–30km/h channel model because this channel induces more variability in channel amplitudes, which is more effective with an opportunistic scheduler such as the AvgCI.

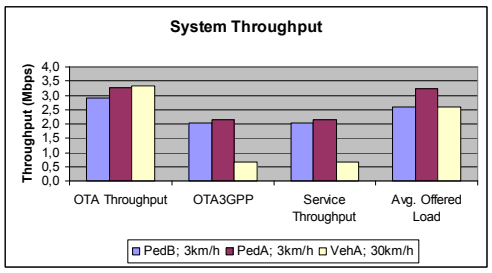


Figure 40 - MCS distribution for AvgCI scheduler

6.5.3 Performance Evaluation for the World Wide Web (WWW) Traffic Model

Due to its burst nature, if channels are not symmetric, the CI scheduler will result in a bad utilization of radio resources. This is because, for a given period of time, it will tend to empty the buffer of the user with highest channel quality and then transmit fewer information bits together with a large amount of padding bits to fill the burst in the frame. Only when the buffer of this user is empty the scheduler will serve the user with a lower channel quality, which results in a bad service throughput for the scheduled user. As the AvgCI scheduler assumes the user's channels as symmetric it will give transmission opportunities more often to different active users in the cell, avoiding the emptiness of the user's buffer in a single row, which results in lower amounts of information bits transmitted along a given period of time, and in an increased service throughput. This is illustrated in figure 41.

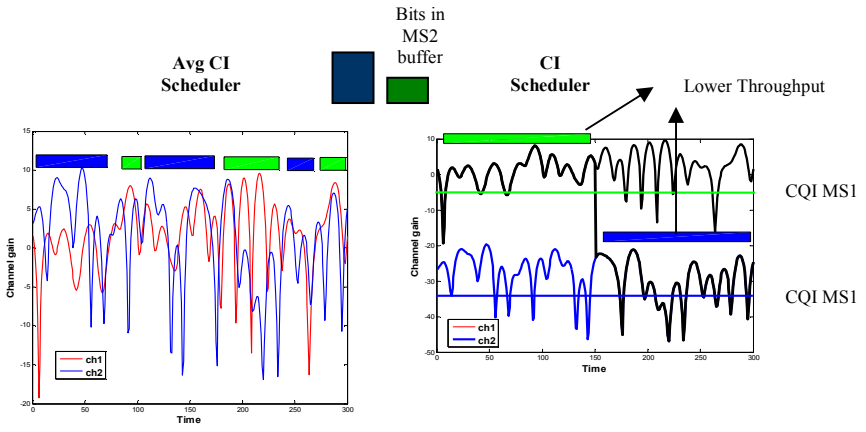


Figure 41 - Multi-user diversity concept for WWW traffic model

Figure 42 plots the CDF function of the user service throughput. As can be seen the CI scheduler is the worst in performance among the three schedulers, with roughly 70% of the users having no access to radio resources in the frame and, as a consequence, with null service throughput. Under the WWW traffic model the RR performance is closer to the one from AvgCI than was verified with the NRTV traffic model. This is due to the burst nature of the WWW traffic: during periods of inactivity there is enough time for the RR to empty the buffer of the user. Nevertheless, the AvgCI scheduler still results in the best performance among the three schedulers, for the reasons presented in previous points.

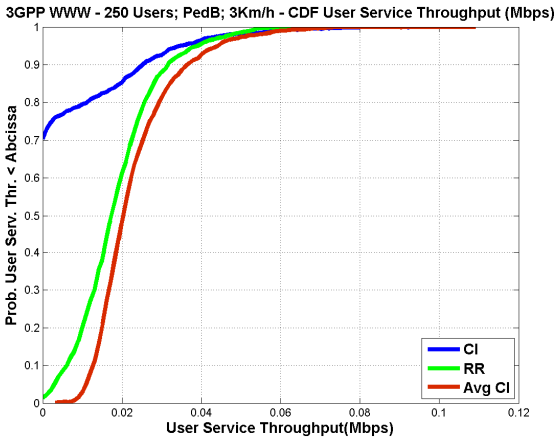


Figure 42 - CDF of user Service Throughput for AvgCI schedule

Figure 43 is the plot of the CDF of the average packet delay per user. As the CI is more effective in empting the users' buffers it presents the best performance until the CDF achieves the 40% tile point, which corresponds to the users with better channel, i.e., closer to the base station. From this point on the CDF has a heavy tail resulting from the unfairness nature of this scheduler. The 90% of the packets are received with a delay lower than 30 seconds and 40 second with AvgCI and RR schedulers respectively.

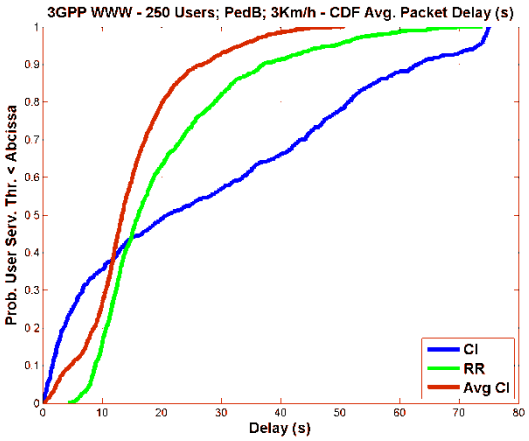


Figure 43 - CDF of user average Packet Delay for AvgCI scheduler

We can conclude that the AvgCI scheduler is more effective in performance with NRTV traffic model and that the performance of the RR approaches the one for the AvgCI with the WWW traffic model.

As can be seen in the plot of figure 44 the gap between the OTA throughput and the service throughput is significantly higher for the AvgCI scheduler than for the CI one. As the OTA throughput includes in its computation also the packets received with error, one can assume that this difference is due to the fact that the AvgCI attempts to serve users in the edge of the cell, and, therefore, with bad channel quality due to interference from neighboring cells. This results in a higher percentage of retransmission attempts resulting from transmission errors than with the CI scheduler.

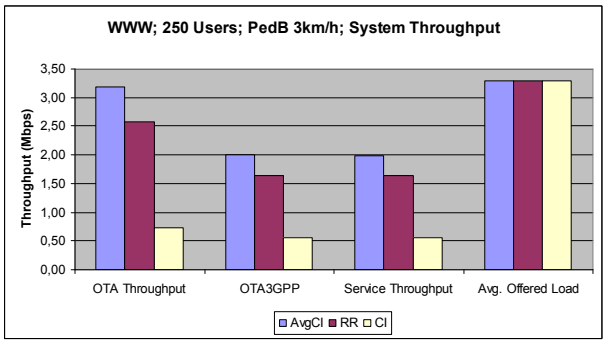


Figure 44 - MCS distribution for AvgCI scheduler

It was mentioned that the AvgCI scheduler results in a normalization (by the mean value) of the channel CQI reported by each mobile. That is: the AvgCI compensates the mean value of the channel, sensed by users as they approach the edge of the cell, and therefore are more affected by the increase in the path loss and interference. This compensation of the channel mean value is more effective in serving the offered load because a higher percentage of users are provided with transmission opportunities to flush their buffers. This can be seen from the small gap between the OTA throughput and the offered load. One can say that the AvgCI is truly an opportunistic scheduler because it results in users with symmetric channels. It can be noticed that a significant percentage of the offered load is not serviced with the CI scheduler due to the starvation nature of this scheduler and for the reasons pointed out above.

6.5.4 Performance Evaluation for Full Queue Traffic Model with MIMO Channel

System level simulations were also performed for the 2x2 MIMO with Alamouti Space Time Block Code (STBC) and under a full queue traffic scenario. The capacity achieved with the MIMO channel was compared against the SISO one. The spatial diversity achieved with 2x2 MIMO Alamouti STBC results in an improvement in the quality of the signal at the reception. Therefore, it is to be expected an increase in the capacity achieved by each cell under this transmission scheme.

System level simulations were performed for a scenario composed by 36 active users (12 users on average per cell), for all three types of schedulers used so far. An *admission threshold* parameter equal to -5dB was assumed in the simulations. This means that users whose CQI is not enough to guarantee at least the most robust modulation and coding scheme, but is greater than -5 dB when this robust modulation and coding scheme is assumed, are considered in the scheduling algorithm. Those users with CQI less than -5 dB, even with the most robust modulation and coding scheme are not considered in the scheduling algorithm.

Figure 45 plots the CDF of the user average service throughput for SISO and MIMO channels, for all three schedulers. In general it can be seen that the use of the MIMO channel results in an increase in the service throughput, although this difference is only marginal for the RR scheduler because this scheduler does not consider the channel quality in the scheduling process. The influence of the MIMO channel is thus only marginal for the RR scheduler.

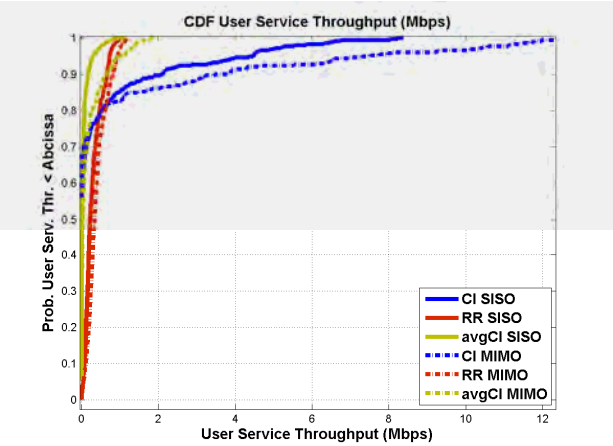


Figure 45 - CDF of the average user service throughput

As expected, MIMO increases the service throughput and, in particular, the CI scheduler achieves the highest service throughput among the schedulers considered. The big tail in its CDF distribution is characteristic of the unfairness behavior of this scheduler, as roughly 70% of the users are not serviced for both types of channels. With the MIMO channel it achieves a service throughput in the order of 12 Mbps and with the SISO channel the achieved service throughput is reduced to roughly 8.5 Mbps. It can also be seen that the AvgCI is the fairest scheduler. This is because the normalization of the CQI by the mean value results in a symmetric channel which gives more transmission opportunities to users in the edge of the cell, and thus with bad channel quality. This has a detrimental effect in the achieved service throughput. Nevertheless, as expected, the service throughput improves the MIMO channel quality.

Figure 46 is the plot of the MCS distribution for all three schedulers for the SISO channel. It can also be seen the detrimental effect that users with bad channel quality have on the AvgCI scheduler as most transmissions are performed with the most robust and coding scheme

enabled. As can be seen, the distribution is biased towards more efficient MCS schemes which corroborate the fact that for the full queue traffic model and for the CI scheduler most transmissions are from user's with good channel quality, i.e., those users closer to the base station.

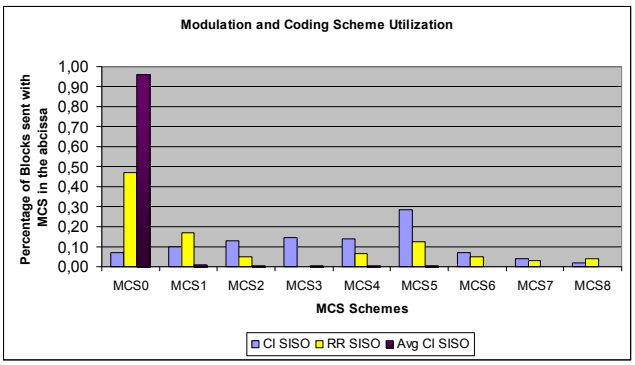


Figure 46 - CDF of the average user service throughput

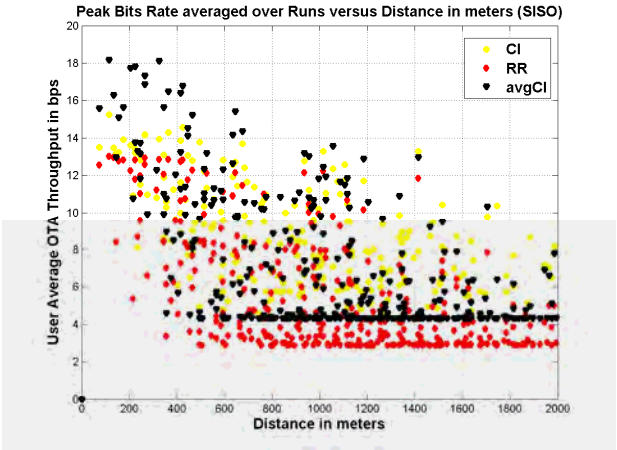


Figure 47 - Peak bit rate average over the distance to the base station in meters – SISO channel

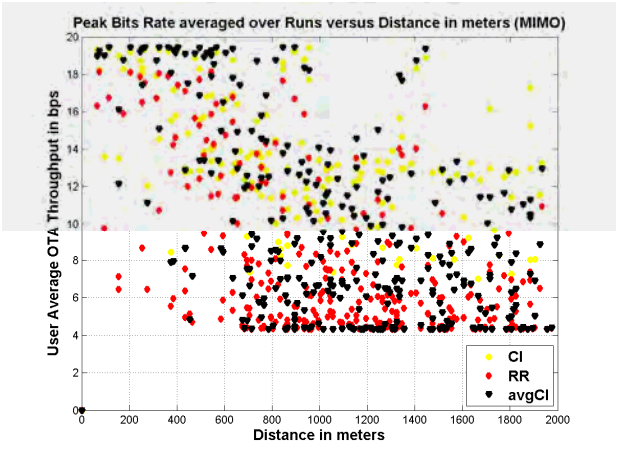


Figure 48 - Peak bit rate average over the distance to the base station in meters – MIMO channel

Figures 47 and 48 are the plot of the peak data rate with the distance from the base station, for SISO and MIMO channels respectively. There are many points in the plot corresponding to the AvgCI scheduling algorithm which are closer to the minimum throughput and from half the

distance from the base station to the cell border. This is because the algorithm opportunistically gives transmission opportunities to the users with the most robust MCS scheme.

Figure 49 is the plot of the system throughput for the different schedulers and SISO and MIMO channels.

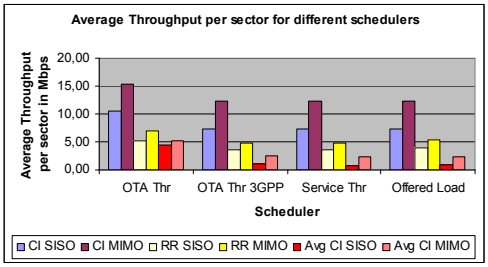


Figure 49 - Figure System distribution

It can be confirmed from this plot the negative effect the admission threshold has on the system performance for the AvgCI scheduler, as users in the edge of the cell and with bad channel quality are also given transmission opportunities, which result in a lower service throughput due to many transmission errors and retransmission attempts.

6.6 Conclusion

This chapter presents results from the tests conducted in the validation of the system level platform implemented for the execution of system level simulations for the IEEE 802.16e standard, which specifies the functionalities of Mobile WiMAX networks. The basic system level platform includes models for channel propagation mechanisms such as path-loss, slow shadow fading and fast fading, multipath propagation and for SISO and MIMO channels. Also, different models for traffic modeling, available in the literature were implemented in the platform.

One first step in the validation process was to verify that the models implemented in the tool produced results which are in accordance with theory. Of particular attention were the channel models used for the implementation of multi-path, fast fading and slow fading. Another set of validation tests was performed for the validation of the model implemented in the tool for MIMO channel simulations. It was also concluded that this model performs according to the SCM model implementation from 3GPP.

In a second step the performance of the dynamic resource allocation (DRA) module, into which packet schedulers and resource allocation algorithms are implemented, was validated. A simple network scenario using the WWW traffic model from 3GPP and the maximum C/I scheduler was considered. The performance of the DRA was mainly evaluated in terms of the user's prioritization, the allocation of resources following the packet scheduler and the protocol implemented inside the system level platform for the exchange of signaling messages. It was

concluded that the DRA performs according to what is expected from available results in the literature.

The last set of validation tests are related with the performance of the DRA for a different set of packet schedulers available in the literature, such as: the maximum C/I, round robin and for a variation of the maximum C/I, namely the maximum C/I over average C/I; for a different set of traffic models, including full queue. This performance evaluation, once again, revealed that the DRA performs according to what could be expected whenever opportunistic and fair schedulers are implemented for the types of traffic models used. The network capacity was also evaluated with the implementation of MIMO and SISO channels and it is evident the gains achieved with the MIMO channel, for both maximum C/I and maximum C/I over average C/I schedulers, which are based on channel state, whereas no significant gain is achieved with MIMO over SISO channel for the round robin scheduler, which does not account channel variation in scheduling decisions.

Having performed the validation, the system level tool is now ready for the implementation of the packet schedulers proposed in this work and more advanced DRA architectures, namely with another degree of freedom for resource allocation: the space domain. These are the subject of the following chapters.

Chapter 7

Utility-Based Packet Scheduling under Mobile WiMAX Network Scenario

7.1 Introduction

The mobile radio channel is a very aggressive medium for the transport of information, especially in mobile environments due to multipath propagation, user's mobility, interference from neighbouring and serving cells and power limitation in mobile devices. Therefore, it poses several constraints to the satisfaction of application's QoS requirements, namely for those types of multimedia applications envisioned for Beyond Third Generation (B3G) network scenarios, with stringent requirements in terms of bandwidth allocation and satisfaction of maximum delay bounds.

Also, as the mobile radio channel is a scarce resource for data transportation, it must be efficiently shared among user's applications, in order to maximize the amount of data conveyed

over the air interface of each cell in the network. Maximization of total cell throughput is achieved by assigning resources to users with the highest channel quality, no matter the QoS requirements. As a matter of fact these are two conflicting objectives:

- In one hand, each user's application poses some QoS constraints to the network operator and expects these to be fulfilled, no matter the state of its connection to the cell. Lack of compliance to these requests result in users' dissatisfaction from the service provided which, in one extreme, can result in user's churn from the network provider.
- On the other hand, maximization of the total amount of data conveyed over the air interface of each cell is of vital importance to the network provider, in order to maximize the service revenue.

These two conflicting objectives must be properly addressed in order to avoid the unfairness arising in resources attribution without observation of QoS requests. Therefore, enabling QoS requirements from the set of high bandwidth demanding, IP multimedia applications, such as those envisioned in Broadband Wireless Networks (BWA) is challenging. In particular, they are driving calls for substantial changes in the network infrastructure of current networks, as they are quite demanding in terms of bandwidth allocation and QoS metrics such as: throughput, loss rate, delay and delay jitter. Also, the architecture of BWA networks must be capable of supporting different levels of service at the same time.

All these constraints and requirements pose high demanding requests in the design of Dynamic Resource Allocation (DRA) modules, from which the packet scheduler is a key element. In particular, packet schedulers must operate inside the DRA, across different service flows, in order to ensure that requested throughputs and bounds on delays and loss rates are met.

In this chapter a new packet scheduler framework for BWA networks is proposed, designed, implemented and validated within the system level simulator implemented for Mobile WiMAX system level simulations. The scheduler framework is based on the cross-layer design paradigm and is fully interoperable with the MAC layer architecture designed for the Mobile WiMAX standard, in line with the principles presented in previous chapters.

This chapter is organized as follows. Section 2 introduces the principles governing the design of a packet scheduler under the cross-layer design paradigm for BWA networks. Section 3 presents in detail the principles behind the definition and implementation of a packet scheduling framework based on the notion of *utility functions* from economics [121]. Section 4 describes the proposed utility-based scheduler implemented in this work and intended for application scenarios comprising users assigned to one of two types of traffic: voice over IP (VoIP) and World Wide Web (WWW). The principles behind the design of the proposed utility functions are presented and the scheduling mechanism is detailed. Simulation results corroborate the efficiency of the proposed scheduler in terms of the ratio of satisfied users and the amount of packets dropped due to delay time-out or maximum number of retransmissions achievement.

This section also describes the packet bundling mechanism regarding the arrival and transmission of VoIP packets. In packet bundling a group of VoIP packets are concatenated in the same resource during transmission, in order to increase the amount of useful data information transmitted over the air interface. The goal behind such strategy is to avoid scheduling users with low amount of data buffered in the base station. This might result in bad efficiency regarding resource utilization. Section 5 elaborates on a variation of the utility-based scheduler which incorporates the token bucket mechanism, in order to provide minimum required average throughput for non-real time (NRT) applications. This new scheduler framework is able to satisfy both types of QoS requirements: (i) maximum packet delay for real-time (RT) users and (ii) minimum sustained average throughput for NRT users. The main motivation behind such proposal is the fact that, unlike the previous version, which cannot guarantee a minimum average throughput, independently of the type of service flow used, the modifications proposed in this new variation of the scheduler comply with the required throughput and can also be considered as an input to the scheduler. Four types of traffic models are considered in the simulations and in the validation of the proposed scheduler: VoIP and Near Real Time Video with 64 kbps (NRTV64Kbps), which are of type RT, WWW with 32 kbps, which is of type NRT, and File Transfer Protocol with 64 kbps (FTP), which is of type Best Effort (BE). Finally, section 6 concludes the chapter.

7.2 Packet Scheduling in Broadband Wireless Access (BWA) Networks

The packet scheduler must efficiently allocate available radio resources in response to burst data traffic, time-varying channel conditions and service's QoS requirements. The scheduler functioning is based on the availability and access to contention-free radio resources from the Medium Access Control (MAC) layer. It is located in the base station to enable rapid response to traffic requirements and channel conditions. The scheduler processes data packets associated to service flows with well defined QoS parameters in the MAC layer, in order to correctly determine the packet transmission ordering over the air interface. Adaptive modulation and coding (AMC), combined with Hybrid Automatic Repeat Request (HARQ), provides robust transmission over the time varying channel.

In Mobile WiMAX radio resources are available in time, frequency and space. For frequency-diverse sub-channels, such as PUSC permutation, sub-channels are of similar quality. Therefore, resources available on time and frequency domains do not benefit from frequency selectivity in slot allocation. With band AMC permutation sub-channels may experience different attenuation and the frequency-selective scheduling can allocate mobile users to their corresponding strongest sub-channels. The frequency-selective scheduling can enhance system capacity with a moderate increase in Channel Quality Information Channel (CQICH) channel overhead in the

uplink sub-frame. The resource allocation is delivered in Mobile Application Part (MAP) messages at the beginning of each radio frame and this allows the modification of the resource allocation pattern for each active mobile scheduled in each frame period.

In its most generic form a MAC scheduler may calculate a metric $M_i(n)$ per service flow *i* that is a function of many attributes specific to the flow and serve the flows in descending order of the metric values, according to equation (1).

$$M_i = f(QoS_i, CINR_i, Delay_i, Throughput_i, Other\ Parameters_i) \quad (1)$$

The scheduler behavior is strongly influenced by the type of traffic model being serviced:

- Real-Time (RT) services pose constraints on the maximum allowable delay for a data packet to be serviced and transmitted through the air interface. If a RT data packet violates the maximum delay bound it is dropped from the queue, because its transmission would result in a squander of the set of resources allocated for its transmission, as it is outdated for the receiver. Data packets of RT service flows are normally generated with a constant bit rate and are of fixed or variable size. The typical approach is to increase the priority of the service flow as its Head of Line (HOL) packet approaches the service deadline.
- Non-Real-Time (NRT) services are more relaxed regarding the satisfaction of a strict maximum delay bound for each packet sent through the air interface. Packets from applications of this type are not generated with a constant bit rate, that is, they are generated in bursts. The best approach to service a NRT user is to keep packets in the user's queue for it to fill up before start transmitting them over the air-interface. Packets can remain for some time in the queue waiting for the best possible user's channel state for transmission. These applications are particularly targeted for opportunistic schedulers, of which the Maximum C/I is a typical example, and benefit from the so called "multi-user diversity gain", achieved when the scheduler selects for transmission the user with the best channel in each frame period [26, 36, 111, 122-125].

Therefore, the resource allocation problem for supporting both RT and NRT traffic in a wireless system is a very challenging task when diverse QoS requirements have to be considered. A logical approach in the design of the packet scheduler is to delay the transmission of packets of RT service type, as long as the delay constraint is not violated, and then make the opportunistic transmission of packets of NRT service type. This strategy will result in a more efficient utilization of the available resources.

The performance of the utility-based packet schedulers presented in this chapter is compared against the performance of standard packet schedulers available in the literature. In particular these are: Maximum C/I (CI), Proportional Fairness (PF), Round Robin (RR), Exponential (EXP) and Modified Largest Weighted Delay First (M-LWDF), which are presented with some

level of detail in section 6.4 of chapter 6, and are used as benchmark figures for performance evaluation.

7.3 Utility-Based Packet Scheduling

7.3.1 Introduction

Wireless networks are characterized by randomness in network topology, and this result in the variation, in time, position and space, of the strength and quality of the signal which propagates along the transmission path between both communication ends. This randomness is normally used in the provision of diversity in all three domains on a point to point link level basis. Diversity mechanisms improve the quality of each individual radio link and enhance the service data rate. At the network or system level this randomness results in multiuser diversity, because the signal impinging on each mobile station antenna array is independent from the others due to the nature of multipath propagation. Examples of opportunistic schedulers, which rely on the multiuser diversity principle to select the user with the best channel quality, at each scheduling instant are: CI and PF. However, provision of diversity mechanisms over the radio link is not enough to satisfy QoS requirements. Actually, CI and PF schedulers normally result in user starvation whenever used in scenarios of mixed real time and non real time application services.

As a matter of fact, provision of QoS requirements demands for a cross-layer based design approach, whose principle relies in the combination of the information coming from: the Physical (PHY), Medium Access Control (MAC) and Network layers (including application layer), in order to optimize some cost or revenue function. A packet scheduler whose architecture is implemented according to this cross-layer design paradigm is illustrated in figure 1.

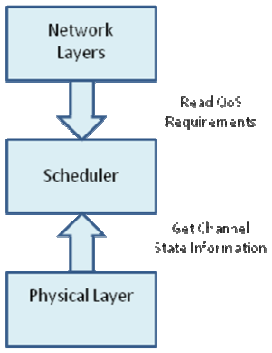


Figure 1 - Model for cross-layer design approach

Examples of schedulers which consider both the state of the radio channel and delay requirements, whenever conducting scheduling decisions, in terms of the delay of the head of line packet in the user's buffer, are: M-LWDF and EXP.

7.3.2 Utility-Based Scheduling Principle and Cross-Layer Design

M-LWDF and EXP schedulers present important shortcomings in their performance and in complying with user's QoS requests. These limitations result from the fact that the definition of the adequate set of values for the calibration of the parameters used in the definition of the respective scheduling algorithms is normally a difficult task, which is not efficient anyway, namely when users can be assigned to one traffic flow from a set of two or more different types of service applications coexisting in the network [129-131].

One approach which could be followed to circumvent these limitations is to map the level of importance associated to each packet into a scalar value, according to a properly defined *utility function*, in agreement to a concept derived from economics. In particular this is the scheduling principle proposed and illustrated in detail in [131]. According to the scheduling principle of a utility function, if one wants to comply to delay constraints associated to each type of service flow in the system, a possible utility function, $U_i(\tau_l^{(i)})$ of the l^{th} packet in the buffer of user i will be a monotonic decreasing function of packet delay, $\tau_l^{(i)}$. In this case the importance of each packet depends on its delay incurred at the given transmission time interval.

Two examples of computation of a user's utility using simple utility functions of packet delay are given here:

Utility function for delay constraints satisfaction

The utility associated to a given active user results from the summation of the utilities of all packets in the buffer. It is computed according to equation (2)

$$U_i(\mathbf{Q}_i) = \sum_{j=1}^{L} U_i(\tau_i^{(j)}) \tag{2}$$

Where:

- $\mathbf{Q}_i = \{\tau_i^{(1)}, \tau_i^{(2)}, ..., \tau_i^{(L)}\}$ represents queue state for user i.
- L is the number of packets in the queue.
- $\tau_i^{(j)}$ is the delay of the j^{th} packet in the queue of user i. Packets are sorted according to the decreasing value of the delay: $\tau_i^{(1)} \geq \tau_i^{(2)} \geq \cdots \geq \tau_i^{(L)}$.

Utility function for jitter constraints satisfaction

For some applications like streaming service flows, satisfaction of packet delay jitter is the most important requirement. As the packet delay jitter refers to the variation of the time elapsed in the transmission of two consecutive packets, the delays from remaining packets, other than the last one transmitted and the next one in the head of line, are not that important to contemplate for jitter compensation. A proposal for the utility function which could quantify the degree of satisfaction regarding jitter constraints is given by equation (3).

$$U_i(\mathbf{Q}_i) = U_i(\tau_i^{(1)} + \Delta_i) + (L-2)\max_{x\in\{L\}}(U_i(x)) \quad (3)$$

Where:

- $\tau_i^{(1)}$ is the delay of the head of line packet from user's *i* buffer.
- Δ_i is the time elapsed since the last packet of user's *i* buffer has been transmitted successfully.
- The term $(L-2)\max_{x\in\{L\}}(U_i(x))$ accounts for the estimation of the delay from remaining packets in buffer, other than the head of line (next one to transmit) and the last one transmitted.

According to (3) the utility of each user decreases with the increase in delay jitter variation and the delay suffered by the head of line packet. Utility of remaining *L-2* packets is assumed as the maximum in the set.

The shape of the utility function depends on the type of application service envisioned (Real Time – RT or Non-Real Time – NRT applications) and on the access priorities defined by the operator [131]. Figure 2 illustrates two types of utility functions: one envisioned for delay-constrained traffic (RT applications) and the other one for the moderate delay-sensitive traffic (NRT applications). As can be seen the utility function decreases with the increase in the delay of the head of line packet in buffer.

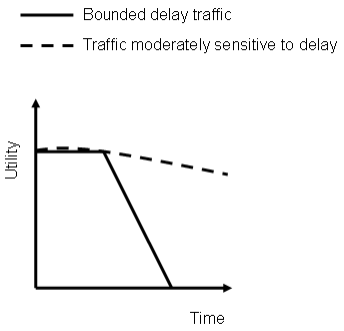


Figure 2 - Utility Functions for bounded and non-bounded traffic models

According to the utility-based scheduling principle proposed in [131], in frame period *n* the user selected for transmission is given by equation (4).

$$k(n) = \arg\max_{i=1,\ldots,N} U_i(\tau_i(n))R_i(n) \quad (4)$$

Where:

- $\tau_i(n)$ is the delay of the head of line packet of user *i* in transmission time interval *n* and $U_i(\tau_i(n))$ is the respective utility.
- $R_i(n)$ is the estimated instantaneous data rate for user i in transmission time interval n.

Depending on the type of utility function used, this approach can degenerate into a greedy algorithm such as the CI. As can be seen from figure 3, for a step utility function with constant gain until the maximum delay bound is achieved, the algorithm is equivalent to an opportunistic CI scheduler.

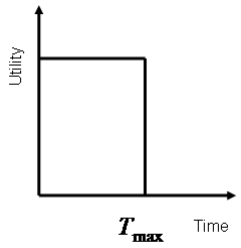


Figure 3 - Utility Function equivalent to max C/I scheduler

However, M-LWDF and EXP, as well as the utility-based scheduler proposed in [131], do present some important limitations in performance, as the cost inherent in deferring transmissions from other users is not estimated or inserted into the computation of the prioritization metric. Whenever the scheduler decides for the transmission of packets from a given user, some others will have their transmission requests postponed. However, deferring transmission attempts reduces the number of opportunities to transmit the packets and this increases the probability that they will be lost by time-out overflow. In order to solve this drawback the scheduling algorithm should estimate the cost resulting from postponement of transmission attempts for all users in each transmission time interval.

The implementation of a function which could quantify this cost would improve resource utilization and satisfy QoS requirements at the same time, namely in a system with different types of QoS demanding applications. Also, the cost or revenue function should be a function of the level of importance each packet has to the network, and this depends on the type of application associated to the service flow and also on the operator's preferences. But the definition of the appropriate cost or revenue function is normally a challenging task.

According to what was said, the principle behind the implementation of a utility-based scheduling algorithm should be resumed in the following way: the utility function definition should account for the transmission requests of active users and include the costs associated with the postponement of transmission requests due to lack of capacity in the channel.

Other examples of utility-based packet schedulers for mixed traffic applications scenarios are provided in [200-201].

7.3.3 Packet Utility and Utility Function

The utility scheduling algorithm is based upon the concept of potential energy, a notion from Physics. The main idea is to assign a *utility* to each packet waiting in buffer for a transmission opportunity. This utility translates to a benefit for the service provider, because the amount of

packets transmitted over the air interface is intrinsically related to the turnover and profit resulting from the application service.

For each packet the utility is quantified by the, so called, *utility function*. Therefore, assuming that in transmission time interval *n* there are *K* active users in the cell, and that for a given user *k* $(k = 1,..., K)$ there are L_k packets in the buffer $(i = 1,..., L_k)$, the utility of the i^{th} packet of the k^{th} active user in the cell is computed as: $u_k(x_k^i(n)), \quad k = 1,...,K;\ i = 1,..., L_k$.

- $u_k(.)$ is the type of utility function assigned to user *k*. The shape and parameters used in the definition of each possible kind of utility function depend on the type of services envisioned in the network. The choice of the utility function for user *k* depends on the type of service provided to the user.
- The argument $x_k^i(n)$ is a scalar value assigned to each packet in the user's buffer. It is used in the quantification of the utility the packet has to the service provider, according to a defined utility function, $u_k(.)$. Its definition depends on the type of application being provided (e.g. Real Time or Non-Real Time) and on the quality of service requirements.

The kind of utility functions considered in this work has two basic properties:

- $0 \le u_k(x_k^i(n)) \le A, \ \forall k, \forall i, \forall n$. A is the maximum possible utility gain.
- $u_k(x_k^i(0)) \le u_k(x_k^i(1)) \le u_k(x_k^i(2)) \le \cdots \le u_k(x_k^i(n)) \le u_k(x_k^i(n+1)) \le \cdots, \quad \forall k, \forall i$. For any packet the utility function decreases monotonically with time.

Although packets have a potential utility, as they represent an amount of revenue for the service provider, utility transfer is only accomplished if packets are truly received with success in the destination node and comply with demands of quality for the type of service provided. The destination node can be either the serving mobile station (downlink connection) or the base station (uplink connection), and quality of service requirements would be for example, commitment to packets deadline or satisfaction of a minimum service throughput.

Here, the argument of the utility function is the packet delay, $x_k^i(n) = \tau_k^i(n)$. This means that the level of importance attributed to each packet is dictated by the delay, τ, it has suffered up to a given transmission time interval. $\tau_k^{(i)}(n)$ represents the delay of the i^{th} packet from the k^{th} user.

Finally, it is worth mentioning that it could be granted the utility function the eventuality of negative values. This would depend on the strategy pursued by the service provider. But this approach was not followed in this work.

7.3.4 Scheduling Algorithm

At the beginning of each transmission time interval the total amount of packets in the cell represents a certain amount of *potential utility* for the service provider. For transmission time interval *n* the potential utility in the cell is denoted as $U_p(n)$. This is the total amount of utility that would eventually be achieved if all packets were successfully transmitted in a given transmission time interval.

Basically, in each transmission time interval, the utility based scheduling algorithm attempts to maximize the fraction of this initial potential utility effectively transferred to the network service provider. However, the amount of utility *actually* transferred is highly affected by the degree of success achieved in the transmission of each individual packet through the air interface:

- The utility of a given packet must somehow be affected by the quality of the mobile radio channel, sensed and reported by each user. This is because packets amounting to higher values of utility, but belonging to a user with a bad channel (users in edge of the cell and, therefore, highly interfered for example), will result in a high probability of bad reception (error decoding) in the destination node. That is: the transfer of the estimated potential utility is, after all, not accomplished.
- The more reliable the mobile radio channel is the fewer amounts of transmissions are attempted and the smaller is the period of time during which radio resources remain assigned to the user. This is because packets which are received with error must attempt another transmission, and therefore, do not relieve the assigned radio resources. Besides, packets attempting another transmission might lose their utility with time, as they must wait in buffer for another transmission opportunity.

For a given user *k*, the influence of the mobile radio channel is considered by means of the channel capacity, $R_k(n)$, in bits. This is because the amount of packets which can effectively be transferred depends on the quality of the mobile radio channel. The capacity of the mobile radio channel for user *k* is a function of the reported Channel Quality Indicator (CQI): $R_k(n) = f(CQI_k)$

For the computation of packet utility some notation must be introduced. In particular:

- Vector $\mathbf{Q}_k(n) = \{\tau_k^{(1)}(n), \tau_k^{(2)}(n), \tau_k^{(3)}(n), \cdots, \tau_k^{(L_k)}(n)\}$ represents the state of user *k*'s buffer in transmission time interval *n*.
- $\tau_k^{(i)}(n)$ is the delay of the i^{th} packet of user *k* up to transmission time interval *n*.
- $\mathbf{Q}_k(n)$ is a First In First Out (FIFO) buffer: index *1* represents the first packet to be transmitted (HOL – Head of Line) and index L_k represents the last packet in the tail of the buffer (EOL – End of Line).

· There are a total of L_k packets in the buffer of user *k*.

The following steps are followed by the algorithm in each cell, at the beginning of transmission time interval *n*:

1. Computation of Potential Utility, $U_P(n)$

The scheduler estimates the cell's potential utility, $U_P(n)$. This is the utility that would be achieved in the cell if all packets were successfully transmitted during this transmission time interval. The Potential utility is given by equation (5):

$$U_P(n) = \sum_{k=1}^{K} U_k(n) = \sum_{k=1}^{K} \sum_{1}^{L_k} u_k(\tau_k^{(i)}(n)) \quad (5)$$

2. Computation of the Transferred Utility for user *k*, $U_k^T\left(R_k(n), \mathbf{Q}_k(n)\right)$

The scheduler estimates the transferred utility for each active user in the cell. For backlogged user *k* (a backlogged user is a user with packets to transmit), the estimated transferred utility is the amount of the fraction of the cell's potential utility, which will be transferred to the service provider if this user is scheduled for transmission. For the k^{th} user $(k = 1,..., K)$ the estimated transferred utility is given by equation (6):

$$U_k^T(R_k(n), \mathbf{Q}_k(n)) = \sum_{l=1}^{M_k} u_k(\tau_k^{(i)}(n)) \quad (6)$$

M_k is the amount of packets used in the computation of the total transferred utility for user *k*. It is determined by the capacity of the channel for this user, which depends itself on the reported channel quality (CQI) and on the number, L_k, of packets in the buffer of user *k*. It is given by equation (7).

$$M_k = f(\mathbf{Q}_k(n), CQI_k), \quad M_k \le L_k \quad (7)$$

The transferred utility is thus a function of the state of the user's queue, $\mathbf{Q_i}(n)$, in transmission time interval *n*, and also on the capacity of the channel, $R_i(n)$, measured by the channel quality feedback channel (CQICH) in the uplink sub-frame.

3. Update remaining Potential Utility if user *k* is scheduled for transmission, $U_P(n+1 \mid k)$

Whenever a decision is taken to transmit a given subset of packets, the ones that are not scheduled in the current transmission time interval will have their delays increased in one unit (one transmission time interval). This may or may not translate to a decrease in their potential utility, depending on the type of utility function used and on the new value for the packet delay. Therefore, assuming a given user $(k = 1,..., K)$ was scheduled in transmission time interval *n*, such postponement will result in a new value for the remaining utility in the cell, in the next transmission time interval, *n*+1. The amount of potential utility in the cell for the next

transmission time interval *n*+1, assuming user *k* was scheduled in transmission time interval *n* is denoted as: $U_p(n+1|k)$.

4. Computation of the Decrease in Potential Utility if user *k* is scheduled, $D_k(n)$

For an appropriate definition of utility functions, the sum of the transferred utility for user *k*, $U_k^T\left(R_k(n),\mathbf{Q}_k(n)\right)$, and the remaining potential utility in the cell, $U_P(n+1|k)$, will not exceed the original potential utility, as given by equation (8).

$$U_k^T\left(R_k(n),\mathbf{Q}_k(n)\right)+U_P(n+1|k)\le U_P(n) \quad (8)$$

In the particular scenario in which all packets are transmitted the whole original potential utility in the cell, $U_P(n)$, will be transferred to the service provider. This means that the remaining utility will vanish, as there are no more packets waiting for a transmission opportunity.

However, in most of the cases, there won't be enough capacity and some packets will be postponed in transmission. Therefore, there will be some remaining utility in the cell for the next transmission time interval. Whenever one or more users are scheduled for transmission, while others remain waiting with their packets in buffer, two possibilities may occur:

- Remaining packets have their delays increased by one transmission time interval, but their utilities are not decreased: *there is no utility loss*. In this case the decrease in the original potential utility, $U_P(n)$, is due to the fraction which is transferred to the service provider, through each user scheduled in the current transmission time interval.
- Remaining packets have their delays increased by one transmission time interval and some of them have their utilities decreased: *there is utility loss besides utility transfer*. In this case the decrease in the original potential utility, $U_P(n)$, is due not only to the percentage transferred to the service provider but also to the decrease in the utility of those packets whose transmission is postponed. Actually this postponement translates to a loss in the original potential utility, as it is not being transferred to any active user in the cell.

The reduction in the original potential utility in the cell if user *k* is scheduled for transmission is given by equation (9).

$$D_k(n)=U_P(n)-U_P(n+1|k)-U_k^T\left(R_k(n),\mathbf{Q}_k(n)\right) \quad (9)$$

For the two scenarios mentioned above, assuming user *k* is scheduled for transmission:

- $D_k(n)=0$, mean that there is no utility loss arising from packets from remaining users, whose transmission is postponed. Therefore, there is no loss in the original potential utility, $U_P(n)$, of the cell. The decrease in its value is equal to the utility transferred to users and we have the equality: $U_P(n)=U_P(n+1|k)+U_k^T\left(R_k(n),\mathbf{Q}_k(n)\right)$.

- $D_k(n) \neq 0$, mean that there is some utility loss arising from packets from remaining users, whose transmission is postponed and we have the inequality: $U_P(n) \leq U_P(n+1|k) + U_k^T\left(R_k(n), \mathbf{Q}_k(n)\right)$.

5. Computation of the Scheduling Metric $M_k(n)$

As mentioned in previous point, the decrease in the original utility, $U_P(n)$, is due to the fraction of the original potential utility which is transferred to the scheduled user, $U_k^T\left(R_k(n), \mathbf{Q}_k(n)\right)$, and to the loss in the utility, $U_P(n+1|k)$, from those packets whose transmission is postponed for the transmission of user *k*.

The selection of a particular user *k* will be a good scheduling decision if it results in the maximization of the transferred utility for this user, $U_k^T\left(R_k(n), \mathbf{Q}_k(n)\right)$, and in the minimization of the decrease, $D_k(n)$, in the original potential utility in the cell, for remaining active users whose transmission is postponed. This is expressed in equation (10).

$$M_k(n) = U_k^T\left(R_k(n), \mathbf{Q}_k(n)\right) - D_k(n) \quad (10)$$

Therefore, the scheduler selects for transmission the user with maximum value of the scheduling metric, according to equation (11)

$$k(n) = \arg \max_{i,(i=1,\ldots,N)} \left(M_i(n)\right) \quad (11)$$

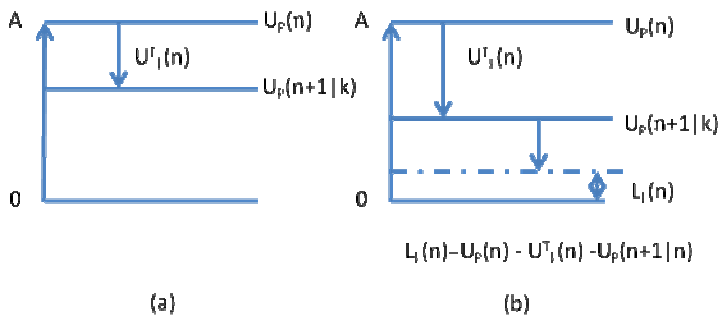


Figure 4 - Illustration of the utility computation in the proposed scheduling algorithm

Figure 4 illustrates the principle behind the proposed utility-based packet scheduling algorithm. Two different scenarios are illustrated:

- (i) There is no loss (packets are not sensitive to delay and, therefore, the decrease in the original potential utility is due only to the transfer of utility to the service provider, through the selection of user *k* (the system is conservative)
- (ii) Another one where there is a loss incurred to the scheduling decision arising from the decrease in the utility for remaining packets, which are not transmitted in the current transmission time interval.

The utility-based scheduling framework is general enough. The key point is the definition of an adequate utility function which is meaningful to the network operator and that do not lead to

excessive computations inside the scheduler. The transfer of potential utility for the service provider is accounted for all cells in the network as the scheduler works for each individual cell.

7.3.5 Guidelines for Utility Function Definition

As mentioned in previous sections, the utility function is constant or decreasing with the increase in packet delay. Accordingly, a possible utility function must satisfy the following properties:

1. $U(\tau):[0,\tau_{\max}[\rightarrow[0,A]$. If the delay achieved is higher than the maximum tolerated its utility is equal to zero in order to express the total user dissatisfaction. The initial gain of the function, A, is the value of the utility for a packet with no delay (when the packet arrives to the buffer).
2. $\lim_{\tau\rightarrow 0^{+}} U(\tau)=A$. This property reflects a maximum user satisfaction (100%) when the packet delay resulting from the transmission of the packet tends to zero.
3. $if\ \tau_i>\tau_j \Rightarrow U(\tau_i)\leq U(\tau_j))$. The utility function is monotonically decreasing with delay.
4. $U^{'}(\tau)\leq 0$ and $U^{''}(\tau)\geq 0$. The first order derivative is negative to reflect the decrease in the level of satisfaction of the user if the packet's transmission is postponed. The second order derivative may increase as the delay approaches the deadline.

One example of a utility function complying with these properties is the sigmoid function.

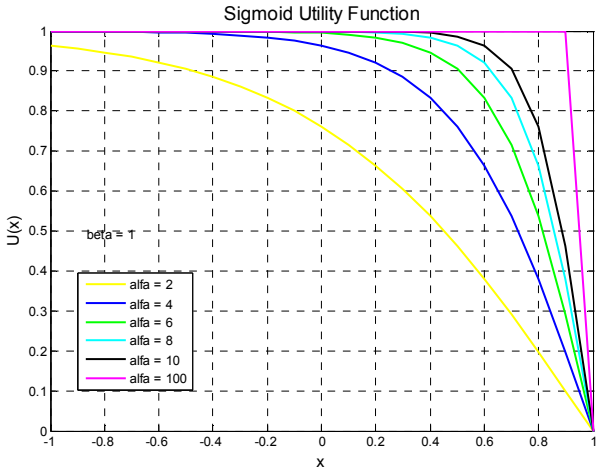


Figure 5 - Different types of Sigmoid functions definition

Sigmoid functions have been used in the literature to approximate the user's satisfaction with respect to the quality of service provided by operators. One example of Sigmoid function is provided by equation (12).

$$U(\tau)=\begin{cases}1-\dfrac{2}{1+\exp\left(-\alpha(\tau-\beta)\right)}, \text{if } \tau\in[0,\tau_{\max}] \\ 0 \qquad\qquad\qquad\qquad\quad , \text{ if } \tau\in[\tau_{\max},+\infty[\end{cases} \tag{12}$$

Parameters α and β determine, respectively, the steepness and the centre of the curve. They can be tuned to customize the function for different types of service. This is illustrated in figure 5.

The important aspects which must be considered in the definition of the type of utility function to be used by a given application service are the following:

- **The delay beyond which the utility becomes zero**. This value depends on the type of service and the maximum allowable packet deadline. After the packet has violated the delay bound required by its application service there is no point in transmitting the packet anymore, as it has no utility for the service provider at all.
- **The maximum utility value**. This is the utility assigned to all packets as they arrive to the buffer. It can be used in the differentiation of the different types of applications coexisting in the cell. Packets labelled with more importance to the service provider can be assigned higher maximal gains when they arrive.
- **The shape of the utility function**. A real time service will have a more steep utility function compared to a non real time service. This is because packets from a real time flow are more demanding for radio resources, as they are more stringent to delay bound satisfaction than non real time ones.

7.4 Multi-Class Utility-Based Packet Scheduling

This section presents a new packet scheduling algorithm based in the notion of utility functions and in the cross-layer architecture paradigm. It was designed and implemented in the system level simulator for Mobile WiMAX, introduced in chapter 4. According to the utility algorithm principle described in the previous section, not only the channel quality and queue information are used in the estimation of the efficiency achieved in resource allocation, but also the cost incurred in the remaining users which have their transmission requests postponed. The proposed scheduler is integrated into the Dynamic Resource Allocation (DRA) module designed for the Mobile WiMAX system, presented in chapter 5.

7.4.1 Proposed Algorithm

The principle behind the utility-based scheduling algorithm was presented in sections 7.3.3 and 7.3.4. It is introduced here with the necessary adaptations for the type of scenario and applications considered in the system level simulations. The following steps are followed by the algorithm in each transmission time interval:

1. At the beginning of transmission time interval *n* compute the total amount of potential utility $U_P(n)$ in the system, as given by equation (13)

$$U_P(n) = \sum_{k=1}^{K} U_k(n) = \sum_{k=1}^{K} \sum_{l=1}^{L_k} u_k(\tau_k^{(l)}(n)) \quad (13)$$

Where $u_k(\tau_k^{(l)}(n))$ is the utility of the l^{th} packet with delay $\tau_k^{(l)}$ in the buffer of the k^{th} user, assuming there are L_k packets in the buffer of user *k* and a total of *K* users in the system.

2. For a given user $k \in \{1,...,K\}$ estimate the amount of utility that will be transferred to the service provider if the user is scheduled for transmission. The amount of utility that can be transferred depends on the user's channel capacity and on the amount of packets in the user's buffer. Assuming user *k* has enough capacity to transmit the first M_k ($M_k \le L_k$) packets, in a total of L_k packets in the buffer, its estimated transferred utility is given by equation (14).

$$U_k^T(R_k(n), \mathbf{Q}_k(n)) = \sum_{l=1}^{M_k} u_k(\tau_k^{(i)}(n)) \quad (14)$$

Where:

- $R_k(n)$ is the capacity, in bits, of the channel from user *k* during transmission time interval *n*. It represents the number of packets that can be transferred if this user is scheduled for transmission in this transmission time interval. It is given by: $R_k(n) = N_k(n) * (1 - PER_k)$, where $N_k(n)$ is the amount of bits in the buffer of user *k* and PER_k is the packet error rate associated with the channel from user *k* during transmission time interval *n*.
- $\mathbf{Q}_k(n) = \{\tau_k^{(1)}, \tau_k^{(2)}, \cdots, \tau_k^{(L_k)}\}$ is a vector representing the delay of each packet from the buffer of user *k*. $\tau_j^{(1)}$ is the delay of the Head of Line (HOL) packet in the user's *k* buffer.

3. The original proposal for the utility based algorithm has no memory associated with the event of packets transmission. That is, there is no information from the state of each user concerning provision of transmission opportunities on behalf of the scheduler.

After conducting some system simulations it was realized that there would be some benefit if some information, namely, previous transfer of utilities from each active user in the cell, up to current transmission time interval, would be used in the computation of the scheduling metric. Therefore, in the computation of the transferred utility, information regarding previous assignments is used. The algorithm computes the average of the utility already transferred to user $k \in \{1,...,K\}$ according to equation (15).

$$\overline{U_k^T(n)} = \lambda . \overline{U_k^T(n-1)} + (1-\lambda), U_k^T\left(R_k(n), \mathbf{Q}_k(n)\right) \quad (15)$$

Where:

- $\overline{U_k^T(n)}$ is the updated value of the average utility transferred to user *k* in transmission time interval *n*. It stores the state of the previous transmissions to this user.
- λ is the forgetting factor.

The principle behind the use of the normalized utility in the computation of the priority metric is the introduction of memory in the scheduling process, which is something similar to the principle behind the Proportional Fairness scheduling algorithm, i.e. to favour those users with a smaller amount of transferred utility, mainly due to poorer channel conditions or lower amount of packets in buffer, in detriment of the ones with more packets or better channel conditions.

4. Assuming user *k* is scheduled for transmission compute the new value for the original potential utility in the cell from remaining users. Here, there is another small adaptation regarding the original formulation of the utility-based algorithm: packets from the user whose transferred utility is being estimated (user *k*), and which cannot be transmitted due to lack of capacity, are used in the computation of the remaining potential utility. The transmission of these packets will be postponed for transmission time interval $n+1$.

Therefore, assuming user *k* has L_k packets and that only M_k packets can be withdrawn from its buffer, the algorithm estimates the remaining potential utility as given by equation (16).

$$U_P(n+1 \mid k) = \sum_{j=1, j\neq k}^{K} \sum_{l=1}^{L_j} u_j(\tau_j^{(l)}) + \sum_{l=M_k+1}^{L_k} u_k(\tau_k^{(l)}) \tag{16}$$

5. The algorithm selects user *k* which results in the maximization of the difference between its transferred utility, $U_j^T\left(R_j(n), \mathbf{Q}_j(n)\right)$, and the decrease, $D_j(n)$, in the utility given that user *k* was chosen, according to equation (17).

$$k(n) = \arg\max_j \left(\frac{U_j^T\left(R_j(n), \mathbf{Q}_j(n)\right)}{\overline{U_j^T}} - D_j(n) \right) \tag{17}$$

And the decrease, $D_j(n)$ in utility is given by equation (18).

$$D_j(n) = U_P(n) - U_P(n+1 \mid j) - U_j^T\left(R_j(n), \mathbf{Q}_j(n)\right) \tag{18}$$

7.4.2 Proposed Utility Functions

The QoS objective is represented by a utility function that quantifies the degree of satisfaction from each user. It depends on the delay incurred in the transmission of each packet and on the amount of packets dropped due to bad channel quality or time-out overflow. Packet delay and loss are incorporated into the utility function by means of:

- Delay bound, D_{th}, beyond which packets lose their utility and are dropped.
- Decreasing rate of the utility with packet delay.
- Shape of the utility function, which reflects the service priority of each traffic class.

In the proposed scheduler since VoIP is of type RT it must be given higher priority over WWW traffic class, which is of type Non Real Time (NRT). However, web packets approaching the service's delay bound must be given priority over voice packets with some remaining time. For both types of traffic packets must be delivered in the MAC layer of the receiver within the maximum allowable delay for the service.

7.4.2.1 Proposal of a Utility Function for Voice over IP (VoIP) Traffic Users

For VoIP traffic model a hybrid utility function is considered. The utility is kept constant until packet delay becomes higher than some priority timer. Whenever the delay becomes higher than this threshold the utility decreases with a significant rate. As VoIP packets are associated to traffic flows with some constant packet generation pattern, their transmission may be postponed until packet delay approaches the maximum delay bound allowed for the service. In this time interval the algorithm essentially behaves like the traditional maximum C/I algorithm, achieving efficiency in resource utilization and throughput maximization. The used utility function for VoIP traffic users is defined according to equation (19).

$$U(\tau_j^l) = \begin{cases} 1 & \text{if } t_l < T_{req} - T_{pri} \\ (T_{req} - t_l)/T_{pri} & \text{if } t_l \geq T_{req} - T_{pri} \\ 0 & \text{if } t_l > T_{req} \end{cases} \quad (19)$$

Where:

- t_l is the waiting interval from the instant of arrival for the l^{th} packet at the j^{th} user's buffer.
- T_{req} denotes the maximum allowable delay for the service. Whenever t_l exceeds T_{req} the packet is dropped due to time-out violation.
- T_{pri} is the priority timer. It denotes the time interval between the instant when the priority of a packet is increased and the time when the packet is deleted.

7.4.2.2 Utility Function for World Wide Web (WWW) Traffic Users

For WWW traffic model, a slowly decaying function of packet delay, with a smaller initial utility gain *A* is proposed. The idea behind this setting is to give more priority to voice packets in detriment of web ones, namely when voice packets approach the deadline.

The proposed utility function for WWW packets is defined according to equation (20).

$$U(\tau_j^l) = \begin{cases} a - b(t_l / T_{req})^{1.3} & \text{if } t_l \leq T_{req} \\ 0 & \text{if } t_l > T_{req} \end{cases} \quad (20)$$

Where:

- *a* is the initial gain of the utility function.

· b is the step from this value at the maximum delay T_{req}.

Figure 6 plots both utility functions and table 1 list the values for the parameters considered in the definition of both utility functions.

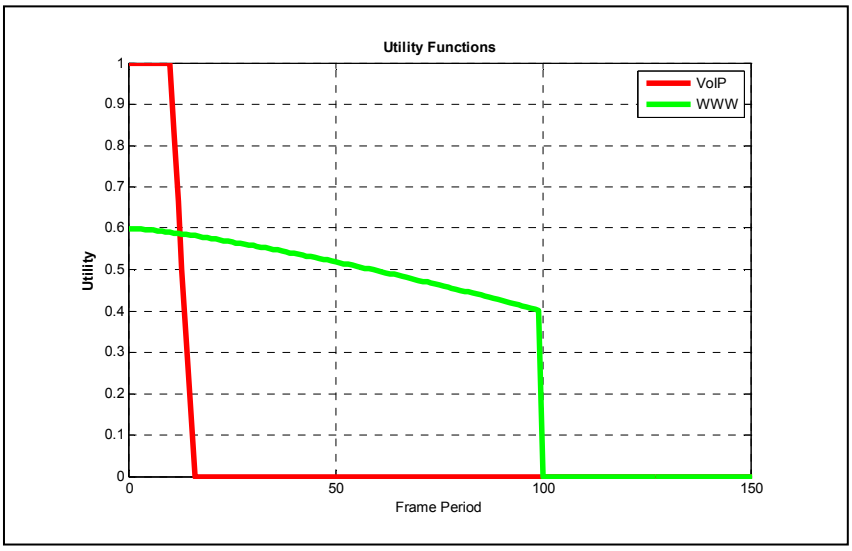


Figure 6 - Proposed utility functions for VoIP and WWW users

	VoIP	WWW
Maximum Utility, A	1	0.6
Maximum Delay Bound T_{req}	80 ms	1 s
Priority Timer T_{pri}	50 ms	-
Utility step from maximum value	0	0.2

TABLE 1 – CONFIGURATION OF THE PROPOSED UTILITY FUNCTIONS

7.4.3 Packer Bundling for VoIP Scheduling Efficiency

Transmission of VoIP packets over 3G and B3G networks such as HSDPA and Mobile WiMAX is a challenging task due to the much tighter delay constraints and the lower source data rate when compared to other types of applications such as web traffic for example. The low source data rate imposes some additional requirements to the DRA such as the need to multiplex as many users as possible over the available slots in each frame period. This translates into an increase in the signalling associated to each burst sent in the frame. Furthermore, even with slot multiplexing there might be not enough data buffered in the base station for a scheduler to fully and efficiently exploit its available air interface capacity, which results in a small efficiency due to the amount of padding bits needed to fill in the resource before transmission. This is particularly true for Mobile WiMAX because there is an inherent trade-off between the size of each slot (which influences the amount of users transmitting in the same frame) and the amount of signalling overhead incurred in the transmission of the information.

For example, in the proposed DRA each resource in DL PUSC mode is composed of 30 slots, which results in a total amount of 15 resources for allocation in downlink sub-frame at each frame period. With the highest MCS scheme this corresponds to an available capacity of 6480

bits per resource and with the lowest MCS scheme this corresponds to an available capacity of 1440 bits per resource. As each VoIP packet consists of 304 bits (38 bytes) there is a lack of efficiency, due the use of padding bits, ranging from 4.7% for the highest MCS scheme to 21% for the lowest one.

Packet bundling is a scheme proposed for HSDPA in [122-124, 132-135], where a group of VoIP packets are concatenated in the same resource during transmission in order to increase the amount of useful data information transmitted over the air interface. Packet bundling is based on the periodic generation of VoIP packets in the respective codecs: VoIP packets arrive at the base station every 20 ms during the ON period (talk spurt), which is equivalent to an average data rate of 15.2 kbps during these periods of conversation.

Packet bundling is implemented in the simulator according to the following algorithm: in every frame period the scheduler defines a set of M_{users} users as forming a *Scheduling Candidate Set* (SCS). The SCS includes active users under one of the following conditions:

1. Users that have at least M_{pkts} VoIP packets buffered in the respective buffer at the base station.
1. Users whose HOL packet delay is equal to or larger than ($M_{pkts} - 1$)*20 ms.
2. Users with pending retransmissions in the HARQ manager, i.e., users with packets belonging to an active HARQ waiting for a retransmission opportunity.

The goal behind such strategy is to avoid scheduling users with low amount of data buffered in the base station, which might cause a loss of system capacity since the supportable radio resource size is much larger than one single VoIP packet.

In the simulations a total of 3 packets were considered in the bundle (M_{pkts} = 3). This increases the efficiency in the transport of data information which now ranges from 14.1%, to the most efficient MCS scheme, to 63%, for the most robust MCS one.

The bundle of packets has an obviously side effect which is the increase in the amount of delay incurred by each packet before transmission: a single VoIP packet in the HOL of the queue will suffer an additional delay of 60 ms before the first transmission. Assuming the same packet has 3 more transmission attempts (each transmission cycle lasts for 10 ms) this corresponds to a total amount of 60+10*2 = 80 ms. Note that the last transmission is not considered in the total amount of delay because the delay must be considered before the scheduling.

According to [125, 136] the estimated available delay budget for the base station processing and the mobile reception ranges from 80 ms to 150 ms, depending on whether the VoIP call is between two mobiles or between a land-line and a mobile. In the simulations the delay bound of 80 ms is assumed for VoIP traffic model.

7.4.4 Algorithm Implementation

The scenario considered in the validation of the proposed utility-based scheduling algorithm considers two service classes, representative of RT and NRT services types: Voice over IP (VoIP) and World Wide Web (WWW), respectively. The utility-based packet scheduler is plugged into the DRA module, whose architecture and implementation into the system level simulation platform is described in chapter 5.

7.4.4.1 Scheduler

Every frame period the scheduler computes the priority metric for each user in the scheduling candidate set (SCS), according to the implemented scheduling algorithm. The conditions a given user must have in order to be included in the SCS are the following:

- The user has at least one HARQ process waiting for a retransmission opportunity.
- The user has new packets in its buffer for transmission, has an inactive HARQ process which can be allocated for new packets and its channel quality (reported in the CQICH in the uplink sub-frame) is higher than a given *admission threshold*.

The purpose behind the definition of the *admission threshold* is to avoid selecting users whose channel quality is in such a bad state that the probability of error in the reception is almost one. In the simulations a threshold of -5 dB was considered, i.e., users with their CQI lower than -5 dB are considered as inactive. Those users whose CQI is not high enough to guarantee at least the most robust MCS scheme but whose CQI value is higher than the admission threshold are assigned the most robust MCS scheme.

Each user in the SCS may have a group of HARQ processes active and waiting for another transmission attempt opportunity. For active HARQ processes the scheduler reads the delay of the packets composing the MPDU stored in the HARQ buffer, in order to compute its utility. The whole MPDU packet in the HARQ buffer must be transmitted in the same frame period if the HARQ process is selected for transmission. The same amount of resources and the same MCS scheme used in the original transmission attempt, must be used in the re-transmission.

7.4.4.2 Resource Allocation

The total amount of resources assigned to each mobile depends on the size of the individual resource, defined by the link adaptation module, and on the amount of bits corresponding to the packets selected for transmission. In the simulations two different cases of packet transmission are considered:

Mobile is attempting the first transmission

Packets are removed from the user's buffer as long as there is capacity available. The capacity available for transmission depends on the MCS scheme, defined by the link adaptation module,

and on the number of radio resources available for data allocation in the frame, as given by equation (21).

$$C_{available} = N_{res\_availale} . B_{resource}(MCS_i) \quad (21)$$

Where:

- $C_{available}$ is the remaining capacity of the frame, in bits.
- $N_{res\_availale}$ is the total amount of resources available in the frame.
- $B_{resource}(MCS_i)$ is the size of each individual resource, which depends on the MCS scheme used.

Parameter	Value
Cell Layout	Hexagonal Grid, 19 cell sites, 3 sector Base Stations, 1 cell reuse in a cloverleaf layout with wraparound to simulate interference to edge cells
Cell radius	900m
Minimum Mobile to Base Station distance	35m
BS Antenna Model (Horizontal)	$A(\theta) = -\min\left[12\left(\frac{\theta}{\theta_{3dB}}\right)^2, A_m\right]$, $\theta_{dB} = 70º$, $A_m = 20dB$, Antenna Gain :15dBi
MS Antenna Gain	Omni-directional with 0dBi
BS Maximum Transmission Power	43dBm
Propagation model	$L = 128.1 + 37.6\log_{10}(R)$, R in Km
Penetration Loss	10dB
Log-Normal Shadowing	Standard Deviation = 8dB
Shadowing Correlation	0.5 for sectors of different BSs and 1 for sectors of the same BS
Channel Model	3GPP SCM MIMO
Number of BS/MS Antennas	4/2
Traffic Models	On-Off Voice and Web Browsing
Duplex Mode	TDD
Operating Frequency	2.5GHz
Bandwidth and FFT size	10MHz; 1024 sub-carriers
Frame Duration	5ms
Number of OFDM Symbols in DL	35
Preamble, FCH, DL/UL MAP overhead	5 symbols
Sub-channelization	DL-PUSC
Burst Size	10 sub-channels x 6 symbols (15 resources)
MS speed	3Km/h
(moving average filter length)	1.5s
T_{CQI} (CQI feedback delay)	2 frame periods
Discard Timer and Priority lengths	15 and 3 frame periods respectively
ACK/NACK feedback	1 frame period in CQICH UL sub-frame
N_{ret} Maximum Number of Retransmissions	4
Number of HARQ processes per MS	4
BLER Threshold for Link Adaptation	10%
Link to System Interface	Actual Value Look-Up Tables
CQI compression method	EESM

TABLE 2 – SIMULATION SETUP CONFIGURATION

The number of resources, $N_{resources}$, used in the transmission of the MPDU is given by equation (22).

$$N_{resources} = \left\lceil \frac{B_{PDU}}{B_{resource}} \right\rceil \quad (22)$$

Where B_{PDU} is the size of the MPDU packet which results from the concatenation of all packets withdrawn from the buffer.

The power available for data transmission is uniformly distributed among the set of radio resources in the frame, no matter the location of the user in the cell. But an adaptation scheme may be followed, in which the power assigned per resource for each user in the edge of the cell may be increased, whereas the power per resource for each user closer to the cell centre is decreased.

It is important to mention that fragmentation is implemented in order to efficiently use the available resources. The MPDU block is concatenated with padding bits, if necessary, processed and mapped onto the resources assigned to the user.

Mobile is attempting a re-transmission

The same MCS scheme used in previous transmissions and the same amount of resources are assumed in the new transmission attempt. Information resulting from multiple transmissions of the same MPDU is softly combined by the Chase Combiner. The combining gain increases the probability of receiving the MPDU with success in the receiver.

7.4.5 Simulation Scenario

The performance of the proposed utility-based scheduling algorithm is compared against the following scheduling algorithms:

- **Maximum C/I (CI)**.
- **Proportional Fairness (PF)** – the moving average filter used has a length of $T_{mean} = 1.5s$.
- **Modified-Largest Delay First (M-LWDF)** – with a violation threshold of $\delta = 0.01$ and maximum delay bound of $W_{\max} = 80ms$ for VoIP and $W_{\max} = 1s$ for WWW traffic models.

The evaluation is performed according to the following performance metrics. For a detailed definition of these metrics please refer to Annex C.

- **Satisfaction Ratio** – this is the ratio between the total amount of satisfied users and the total amount of active users in the cell. A user is deemed as satisfied if the packet drop ratio is less than a given threshold value. In this work, users for both types of services are considered as satisfied if less than 2% of the total amount of packets generated is dropped due to maximum delay bound violation or due to bad channel quality (residual errors).
- **Cell Capacity** – this corresponds to the number of satisfied users in each cell.

- **Average Packet Delay** – this corresponds to the mean delay among all packets received with success in the system.
- **Average Packet Drop Ratio** – this corresponds to the ratio between the number of packets dropped and the total amount of packets transmitted.
- **Service Throughput** – this corresponds to the fraction of packets received with success from the set of packets generated in the system, per user.

Table 2 lists the setup for the scenario used in the system level simulations

7.4.6 Results

System level simulations were conducted for a number of traffic loads, from light to heavier ones, corresponding to different amount of active users in the system.

Satisfaction Ratio

Figure 7 illustrates the evolution of the percentage of satisfied users versus the number of active users in the system for the VoIP and WWW services respectively.

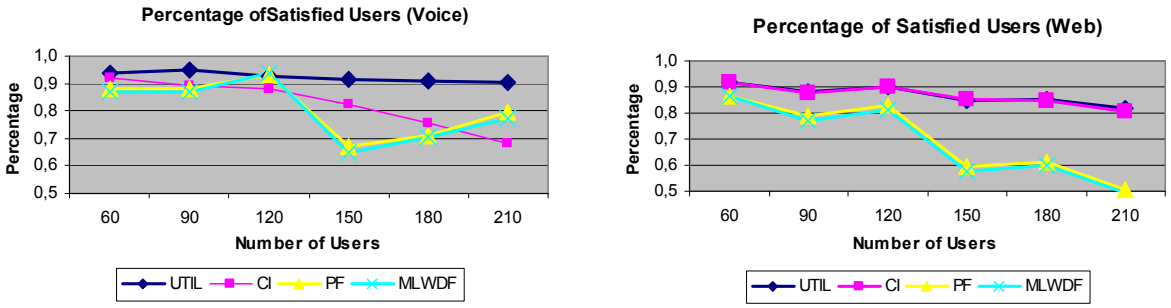


Figure 7 - Percentage of satisfied users versus the number of active users for VoIP and WWW users

As can be seen from both figures, the proposed utility-based scheduling algorithm is somehow insensitive to the amount of active VoIP users in the system. This is because the non-degradation in the satisfaction ratio of VoIP users is compensated by the degradation in the satisfaction ratio of WWW users. This is in accordance with the two types of utility functions defined in the scheduler, which give higher priority for VoIP packets after the priority timer is achieved in the utility function assigned to VoIP users. On the opposite way, packets from WWW service are given less priority and their utility changes very little with the increase in packet delay, behaving very much like the CI scheduler until the maximum delay bound is achieved.

According to the utility-based scheduling principle, WWW packets transmission result in a higher cost if the transmission of VoIP packets is postponed after they achieve the priority timer. For VoIP packet delays lower than the priority timer the scheduling algorithm behaves very much like a CI scheduling algorithm too, but with a higher utility gain than WWW packets, in order to give higher priority to these packets due to delay constraints.

The utility scheduler performs well for all scenarios and its performance is better than the performance of the M-LWDF service. This is because this scheduler tries to equalize the percentage of users violating the maximum delay bound, even for users with poor channel quality, which is detrimental for users with better channel because those users will need more time for the transmission of the same amount of information. The utility-based algorithm avoids this cumbersome because it tries to minimize the cost of postponing the transmission of each user. As could be expected, the CI presents worse performance because the delay of voice packets is not taken into account.

Both PF and M-LWDF algorithms present almost similar performance. This is because the same violation probability was considered in the M-LWDF scheduling algorithm for both types of traffic, which results in the degeneration of this algorithm into the PF one.

Average Packet Delay

Figure 8 illustrates the average packet delay for all four schedulers and for VoIP and WWW traffic models, respectively. As can be seen from both figures, the average packet delay for the utility-based scheduler is insensitive to the increase in the number of active users in the system, at the cost of an increase in the average packet delay for WWW packets. This is due to the intrinsic prioritization scheme resulting from the type of utility functions implemented for both traffic models in the scheduler.

Because the M-LWDF scheduler tries to equalize the delays, no matter the type of traffic model associated, and because the delay bound for VoIP service is much more stringent than the delay bound for WWW service flows, this scheduler presents a worse average packet delay than the CI scheduler.

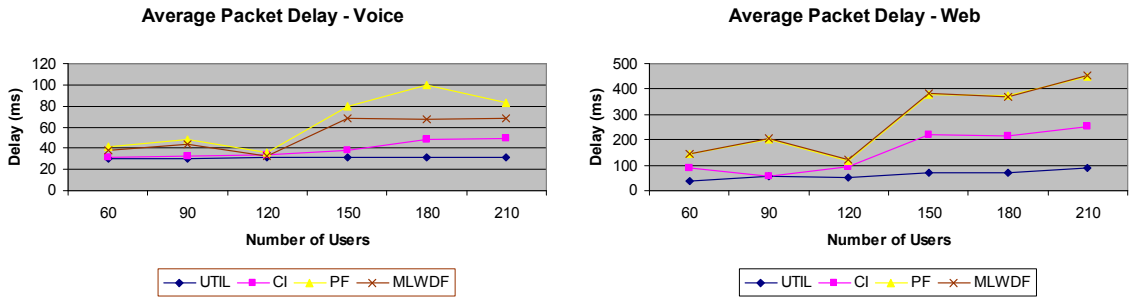


Figure 8 - Average packet delay of VoIP users versus the number of active users

The CI scheduler presents better figures in terms of average packet delay at the cost of an increase in the number of dropped packets for VoIP traffic flows, due to maximum delay overflow, as can be seen in figure 9. As WWW packets have a much less stringent delay bound, the average packet drop rate for this traffic model has better figures than PF and M-LWDF at the cost of an increase in the average packet drop rate for VoIP packets.

Average Packet Drop Rate

Figures 9 illustrate the average packet drop rate for all four schedulers for VoIP and WWW traffic models respectively.

The packet drop rate includes packets dropped due to delay bound violation and due to the violation of the allowable maximum number of transmission attempts. As can be seen from both figures, the average packet drop for the utility-based scheduler performs better than the other remaining three schedulers, for both two types of services. This performance results from the intrinsic prioritization accrued with the utility functions defined for both two types of service.

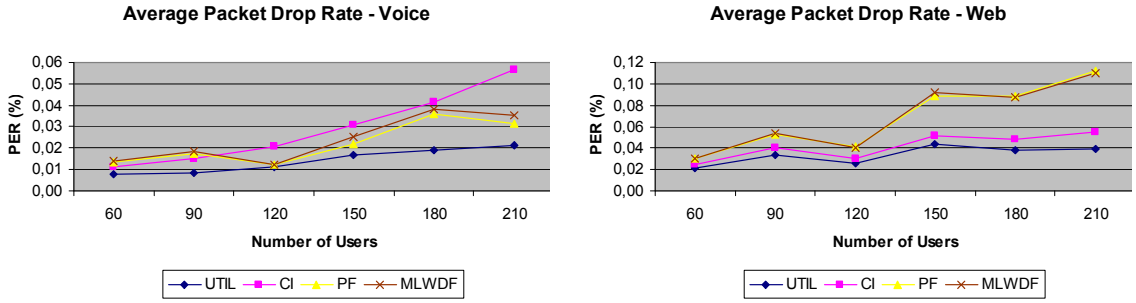


Figure 9 - Average packet drop rate versus the number of active users for VoIP and WWW users

The CI scheduler performs worse than the PF and MLWDF schedulers for VoIP traffic type. This is because it does not consider the packet delay bound in the scheduling metric and also because the delay for VoIP packets is much more stringent than for WWW. As priority is given for users with better channel, and WWW packets have a much lower delay bound constraint, VoIP packets are dropped at a higher pace.

As MLWDF considers the delay of the HOL packet in the computation of the scheduling metric for each user, it presents better performance than PF for both types of service.

Average Service Throughput

Figure 10 illustrates the average service throughput for all four schedulers for VoIP and WWW traffic models respectively.

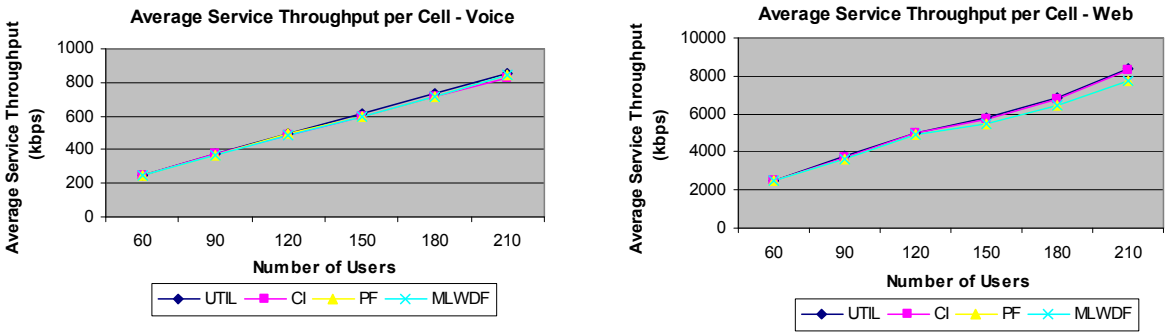
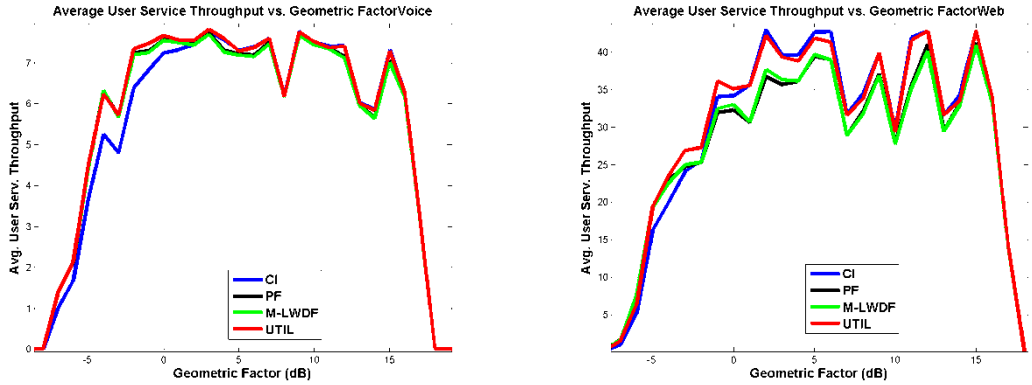


Figure 10 - Average service throughput versus the number of active users for VoIP service

As expected, the service throughput increases with the amount of active users in the system. Also, for higher loads the service throughput of CI scheduler is higher for both types of traffic models among all types of schedulers.

Average Service Throughput Variation with the Geometric Factor

Figure 11 plots the graph of the average service throughput versus the geometric factor for VoIP and WWW traffic models for all four schedulers and for the highest load considered in the system (210 users).



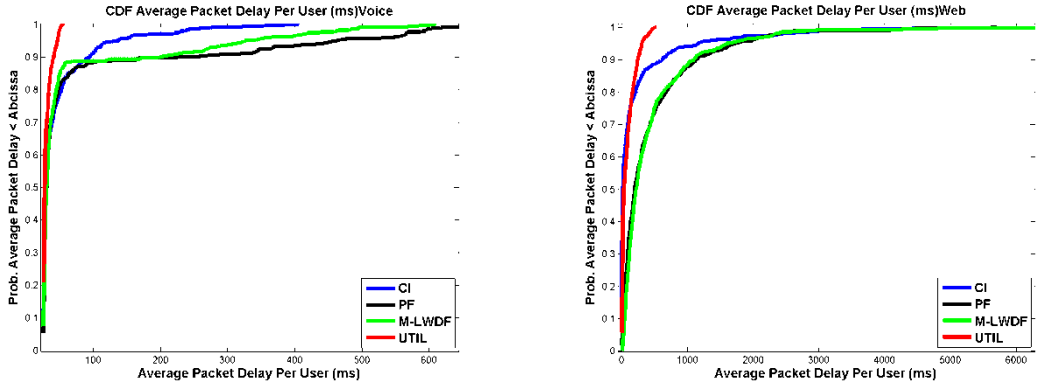
a **b**

Figure 11 - Average service throughput versus the geometric factor for VoIP and WWW users

As expected the CI scheduler has a bad performance for both types of service for low geometric factors. The other three schedulers have a much better performance. This is because the CI is an opportunistic scheduler which gives higher priority to users near the cell's centre.

Cumulative Distribution Functions for Average Packet Delay and Average Packet Drop Rate

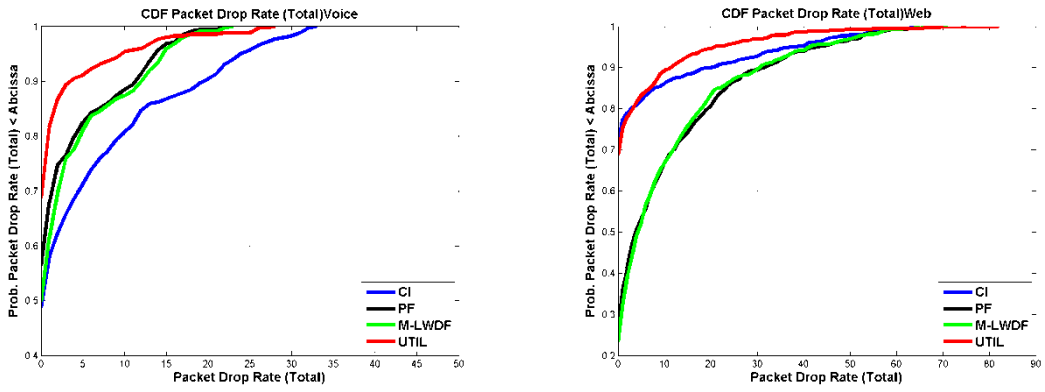
Figures 12 and 13 plot the CDFs of the average packet delay and average packet drop rate for both types of traffic models for all four schedulers and for the highest load considered in the system (210 users).



a **b**

Figure 12 - CDF of the average packet delay for VoIP and WWW users

The utility-based scheduler presents the best performance for both types of traffic models. The CI scheduler has better performance than PF and MLWDF ones because many packets are dropped due to delay overflow, as can be seen from figure 13. It can also be seen that PF and MLWDF have almost similar performance in terms of packet drop rate for both types of traffic models. This is because the delay violation probability parameter in the MLWDF algorithm was set as equal both for VoIP and WWW traffic models, which results in the degeneration of the MLWDF scheduler into the PF one.

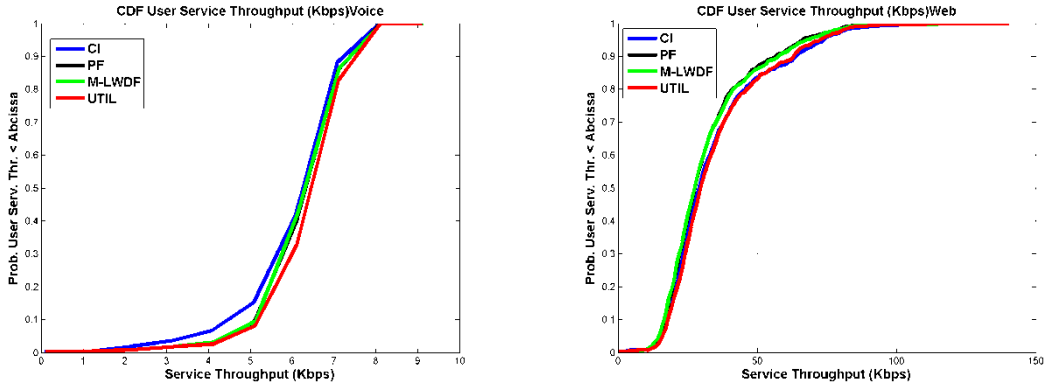


a **B**

Figure 13 - CDF of the average packet drop rate for VoIP and WWW users

Average Service Throughput Cumulative Distribution

From the plot of the cumulative distribution function of the average service throughput in figure 14 it can be noticed that the utility-based scheduler has the best performance among all four schedulers.



a **B**

Figure 14 - CDF of the average service throughput for VoIP and WWW users

7.5 Joint Utility-Token Bucket Based Packet Scheduler

QoS requirements can be defined in a number of ways, depending on the type of traffic model used. Real Time (RT) services typically require that packets must be received within a defined time interval. If packets violate the maximum delay bound allowed for the service flow they are

useless at the receiver and therefore dropped. Examples of RT services are live audio, video streams and voice over IP. A RT user is deemed as satisfied if the percentage of the total amount of packets dropped in the receiver, due to delay bound violation or to bad channel quality, is lower than a given threshold. This can be expressed statistically according to equation (23):

$$\Pr\{W_i > T_i\} \leq \delta_i \quad (23)$$

Where:

- W_i is the delay of the head of line (HOL) packet in the buffer of the i^{th} user.
- T_i is the maximum allowable delay for the service carried by the i^{th} user.
- δ_i is the maximum value for the probability of violation of the delay bound for the i^{th} user.

Non-Real-Time (NRT) services are more relaxed regarding the satisfaction of a strict maximum delay bound, as packets from NRT service flows are not generated with a constant bit rate, i.e., they are generated in bursts. Therefore, the best approach to service NRT flows is to keep packets in buffer for some time, waiting for the best channel condition to begin their transmission over the air-interface. QoS requirements are different for NRT users. Typically NRT service flows are assumed as satisfied if the service throughput provided by the network is above a minimum threshold. This is described by equation (24).

$$R_i > r_i \quad (24)$$

Where:

- R_i is the average service throughput achieved by user *i* over a given time interval.
- r_i is the minimum average throughput required by user *i*.

The joint support of RT and NRT service flows, with such a mix of QoS requirements, is a challenging task for the packet scheduler and resource allocator in BWA networks. As described in the utility-based algorithm, a logical approach in the design of the scheduler is to delay the transmission of packets from RT service flows, as long as the delay constraint is not violated and then make the transmission of packets of NRT service flows. This strategy will result in a more efficient utilization of the available resources with satisfaction of packet delay bounds for both RT and NRT service flows. However, there is no guarantee that the minimum required service throughput is supported by the network for service flows of type NRT.

As the original version of the utility-based scheduling algorithm is not able to guarantee a minimum average throughput for a given user, independently of the type of service flow used, in this section the design and implementation of a new version of the utility-based packet scheduler, which results from some modifications in the original algorithm, is proposed. The

new scheduler is able to satisfy both types of QoS requirements: (i) maximum packet delay for RT users and (ii) minimum sustained average service throughput for NRT ones.

7.5.1 Scheduling Principle

7.5.1.1 Non Real Time (NRT) Users

The problem inherent in the provision of a minimum average service throughput guarantee for NRT users can also be solved by the utility-based algorithm, if it is used in conjunction with a virtual token bucket associated to each NRT service flow [129-130, 137]. In figure 15, a queuing model of two users and two virtual token buckets is used in the demonstration of this principle. In this figure each circle corresponds to the size of each packet in the real queue. The amount of tokens in the queue of user i at the beginning of frame period t is designated by $Q_i(t)$. The figure shows that tokens arrive to each bucket at a constant rate denoted as $r_{i,req}$ (in bits), which corresponds to the minimum throughput required by user $i, i = 1,...,K$. The scheduler can satisfy the minimum required throughput to each user if, in each transmission time interval, each user is prioritized according to the delay of the longest token waiting in token queue. Since tokens arrive at a constant rate, $r_{i,req}$, the delay of the longest token in the token queue is given by equation (25).

$$W_i = \frac{(\text{Number of tokens in the bucket of user i})(\text{in bits})}{r_{i,req}\ (\text{in bits/s})} (\text{in seconds}) \qquad (25)$$

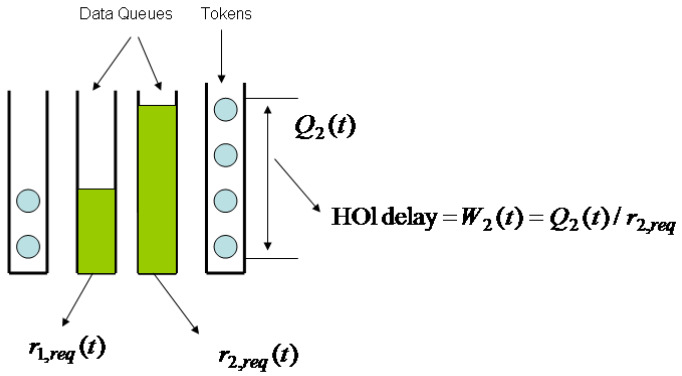


Figure 15 - Virtual Token Bucket Model for two Users

After serving the user the virtual token queue is reduced by an amount of tokens equal to the amount of bits removed from the user's real queue. Thus, the only information the scheduler requires in each frame period is the amount of tokens in each virtual token queue. This mechanism assures that if the token queues are stable (i.e., they do not grow exponentially, which means that they can be bounded), the actual throughput of each user's flow is at least equal to the required minimum throughput, $r_{i,req}$. The token bucket mechanism guarantees flow isolation among users because service is bounded to the minimum throughput required,

according to the constant token rate inputted into each virtual token queue. A large burst of data for one user will not affect the minimum throughput requirements provided to the remaining active users in the system.

As the utility-based scheduler computes the utility of each packet in the queue, which depends on the packet delay, the original token bucket algorithm is changed in order to compute the utility for each packet according to the amount of tokens in the queue. Assuming that user i has N packets in its queue, ordered according to the decreasing value of delay, $d_0 > d_1 > \cdots d_{j-1} > d_j > d_{j+1} > \cdots d_{N-1}$, at the beginning of frame period *n* the delay for each packet is given by equation (26).

$$W_i^{(j)}(n) = \frac{\max\left\{0, V_i(n) - \sum_{k=1}^{j-1} d_k\right\}}{r_{i,req}} \quad i \in NRT \tag{26}$$

Where:

- $W_i^{(j)}(t)$ is the delay of the j^{th} packet in the queue of user *i* at the beginning of frame period *n*.
- $V_i(n)$ is the total amount of tokens (in bits) in the virtual token queue of user *i* at the beginning of frame period *n*.
- $r_{i,req}$ is the minimum required throughput of user *i*.
- d_k is the size, in bits, of the k^{th} packet in the queue $(j = 0,..., N-1)$.

In the computation of the delay for the k^{th} packet the total amount of bits corresponding to previous *k-1* packets must be subtracted from the token counter, $V_i(n)$. If the amount of tokens in the virtual queue is not high enough to guarantee the transportation of all packets, the delay $W_i^{(j)}(n)$ is negative and packets which follow are not scheduled. By these means it is assured the user is not provided more throughput than its minimum requirements, as defined by the QoS metric, $r_{i,req}$.

A variation of the algorithm for the utility-based scheduler, needed for the support of NRT service flows, is the following:

1. First the queue of each user is processed in order to drop all packets which achieve the maximum allowable delay required by the service.
2. For each packet in the real queue the virtual delay $W_i^{(j)}(n)$ is computed.
3. If the virtual delay is negative then a temporary delay value is assigned to the packet. This delay is higher than the maximum allowable delay for the service, which means that the packet has no utility for the system (the utility function returns zero for delays higher than the delay bound) and therefore will not be considered in the scheduler.

4. After transmission of each packet the amount of tokens corresponding to the packet size is subtracted from the token queue.
5. If there is an error in the transmission and the maximum number of transmission attempts has not been reached yet the packet will have another transmission opportunity. Then the amount of tokens is added again to the token queue of the user. The reason for this approach is that the minimum assured throughput must correspond to data that is effectively received with success in the receiver.

It is worth mentioning that in point 3 the delay is artificially increased in order to force the packet not to be considered in the scheduler and therefore not to influence the computation of the priorities of remaining users. However, this does not mean that the packet is dropped. It is kept in the queue until the user has enough credits, measured by the amount of tokens, in order to be transmitted or the maximum delay bound is achieved, in which case the packet is dropped by time-out.

7.5.1.2 Best Effort (BE) Users

The only QoS requirement for services of type Best Effort (BE) is the maximum sustained rate. But there is no guarantee, either in terms of packet delay, or minimum sustained traffic rate. Typically, flows of traffic type BE have the lowest priority among all traffic flows in the cell, which means that they are provided service only if there are still resources available after flows of higher priority traffic classes are serviced.

In the utility-based packet scheduler, packets belonging to a flow of type BE are prioritized according to the shape of the associated utility function and to the priority metric computed to each packet in the flow, according to equation (27)

$$W_i^{(j)}(n) = \left\lceil \frac{B_i(n) - \sum_{k=1}^{j-1} d_k}{T_i^{avg}(n)} \right\rceil \quad i \in BE \tag{27}$$

Where:

- $W_i^{(j)}(n)$ is the delay of the j^{th} packet in the queue of user i at the beginning of frame period n.
- $B_i(n)$ is the total amount of bits in the user's queues, including both the queue with new packets and the queues belonging to HARQ processes which are active and waiting for a new transmission opportunity.
- $T_i^{avg}(n)$ is the average throughput achieved by user i until the beginning of frame period n.

As can be seen from equation (27), the delay $W_i^{(j)}(n)$ decreases with the increase in the average throughput, $T_i^{avg}(n)$. This means that, in order to plug this algorithm into the utility-based packet scheduler, the utility function assigned to flows of type BE must increase with the packet delay. This scheme results into a higher priority for users with lower provided average throughput until frame period *n*. Packets with zero delay have zero utility. From equation (27) it can also be noticed that for the same throughput the first packet (HOL) in the queue will have the highest delay and therefore the highest utility.

As an example of the strategy followed, assume two users with the same amount of packets and the same amount of bits in their queues, but an average throughput for the first user which is half the average throughput achieved by the second one. This means that packets from the first user service flow will have higher priority than packets from the second user service flow.

7.5.2 Utility Functions Definition

Figure 16 plots the utility functions proposed for the four types of traffic models considered in the simulated scenario: VoIP, NRTV, WWW and FTP. The shapes of these four utility functions have to do with the prioritization mechanism followed in the allocation of radio resources for service flows. The general principles governing this choice are the following ones:

- VoIP and NRTV packets have higher priority over packets belonging to service flows of type WWW and FTP.
- When the packet delay from service flows of type VoIP overcomes the priority timer T_{pri}, the rate with which these packets become loosing utility increases significantly, which means that they have the highest priority of service among all types of traffic in the system.
- NRTV packets have an initial utility which is the double of the initial utility of VoIP packets. As delay starts to occur they begin losing utility in an exponential way. The idea behind such an approach is that NRTV packets should be serviced as long as they arrive to the user's queue and until packets of VoIP traffic do not reach the priority timer, T_{pri}.With this approach the scheduler gives priority to NRTV packets initially.
- The utility function for WWW traffic reflects the fact that this is a more relaxed service regarding packet delay constraints. As the transmission of packets is insensitive to delay, they lose utility very softly in order to avoid the system to prioritize their transmission over packets from VoIP or NRTV, whenever packets from these two types of service approach their delay bound. Basically only the channel state is considered in the computation of the priority of this type of service and the utility-based scheduler behaves either as a CI scheduler, if memory is not incorporated in the computation of the user's transferred utility, or as the PF otherwise. WWW service flows are given priority in the

allocation of resources over service flows of traffic type FTP. This is accomplished by attributing a significant higher utility value for each packet of WWW service type over FTP ones.

· Packets from FTP traffic model are also generated in burst mode, like WWW ones, and they are also insensitive to packet delay. But, differently from WWW, they have no assurance in terms of transmission opportunities because they have no minimum service throughput constraints to be satisfied. The delay bound for service flows of FTP traffic model is much higher than for the remaining other service types, which results in more transmission opportunities before dropping a packet due to time-out overflow.

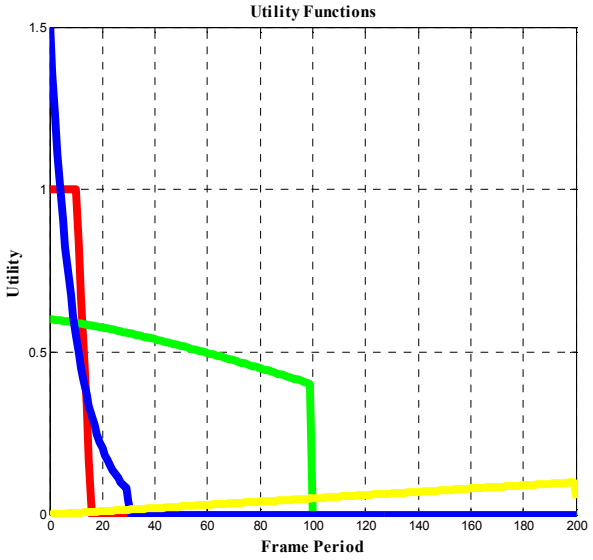


Figure 16 - Utility functions definition

7.5.2.1 Real Time Service Flow – VoIP

Packets from service flows of type VoIP are RT. The utility function is constant until the delay becomes equal to a priority timer designated as T_{pri}, which is the same to say that they do not loose utility in the time interval $0 \le t \le T_{pri}$, and the scheduler behaves basically like an opportunistic CI scheduler. Packets with delays higher than T_{pri} start losing utility with a linear rate equal to $1/T_{pri}$. In the interval $T_{pri} \le t \le T_{req}$ they are given higher priority in the access to system resources. Equation (28) is the mathematical formulation for the utility function of service flows of type VoIP.

$$U(t) = \begin{cases} 1 & \text{if } t \le T_{req} - T_{pri} \\ \dfrac{1}{T_{req}} * t + \dfrac{T_{req}}{T_{pri}} & \text{if } t \le T_{req} \\ 0 & \text{if } t > T_{req} \end{cases} \qquad (28)$$

7.5.2.2 Real Time Service Flow – NRTV

Packets from service flows of type 3GPP NRTV are also RT. The utility function is a negative exponential one with decreasing rate equal to β and initial gain (for zero delay) equal to α. NRTV is a streaming service in which packets are generated with a constant rate. The shape of the exponential function forces NRTV packets to be transmitted as near as possible to the instant of packet generation. This behaviour is more adequate for streaming type of applications such as NRTV and defines the proper set of prioritization between both types of RT traffic models (VoIP and NRTV). Due to the occurrence of inactive periods VoIP packets are kept in buffer until their delay reaches the priority timer, T_{pri}. Equation (29) is the mathematical formulation for the utility function of service flows of type NRTV.

$$U(t) = \begin{cases} \alpha.\exp(-t/\beta) & \text{if } t \le T_{req} \\ 0 & \text{if } t > T_{\text{req}} \end{cases} \tag{29}$$

7.5.2.3 Non Real Time Service Flow – WWW

Packets from service flows of type WWW are NRT. The proposed utility function is softly decaying with packet delay. This reflects the fact that packets are insensitive to delay. The step of decrease in the utility since the initial gain (for zero delay) is defined according to parameters α and β. The utility functions are shaped in such a way that avoids WWW service flows being given higher priority than RT service flows due to a better channel quality. Equation (30) is the mathematical formulation for the utility function of service flows of type WWW.

$$U(t) = \begin{cases} \alpha - \beta.(t/T_{req})^{\gamma} & \text{if } t \le T_{req} \\ 0 & \text{if } t > T_{\text{req}} \end{cases} \tag{30}$$

7.5.2.4 Best Effort Service Flow – File Transfer Protocol (FTP)

Packets from service flows of type FTP are BE, which means that they are assigned resources only when there are no requests from service flows of the other three types of traffic. This is the reason behind the proposed utility function: a linear increasing function with a very slow increasing rate (α / T_{req}), which results in a much lower utility for FTP packets compared to VoIP, NRTV and WWW. The reason for the utility function increase with delay has to do with the computation of the utility of each FTP packet in the scheduling algorithm. Service flows of the same type are prioritized according to the delay of each packet and to the achieved average throughput, as can be seen in the computation of the priority metric given by equation (27). Equation (31) is the mathematical formulation for the utility function of service flows of type FTP.

$$U(t) = \begin{cases} (\alpha / T_{req}).t & \text{if } t \le T_{req} \\ 0 & \text{if } t > T_{\text{req}} \end{cases} \tag{31}$$

	VoIP	NRTV	WWW	FTP
Maximum delay bound	80ms	150ms	0.5s	1s
Minimum service throughput	-	-	32kbps	80kbps
Priority Timer	50ms	-	-	-
Average length for PF scheduler	1.5s	1.5s	1.5s	1.5s
Violation threshold, δ	0.03	0.03	0.01	0.01
Maximum delay for M-LWDF algorithm, W_{max}	80ms	150ms	0.5s	1.5s

TABLE 3 – SYSTEM CONFIGURATION

7.5.3 Simulation Scenario

The same simulation scenario, namely the network setup and configuration of the DRA, as well as scheduling algorithms used as benchmark for performance comparison, are assumed here. Table 3 lists the setup scenario used in the system level simulations.

The user satisfaction ratio for users of type WWW and FTP depends on the accomplishment of the minimum required average service throughput for each type of service.

7.5.4 Results

In order to evaluate the performance of the utility-based scheduling algorithm, benchmark scheduling algorithms such as CI, PF and M-LWDF were simulated and their performance compared against the performance of the proposed scheduling algorithm. Various traffic loads were generated, from light to heavier ones. The performance metrics used in the performance evaluation of all four schedulers are:

- User Satisfaction Ratio.
- Average Packet Delay per User.
- Average Packet Drop Rate per User.
- Average Service Throughput per User.
- Average Over-The-Air Throughput per User.

7.5.4.1 Users Satisfaction Ratio

This metric is defined as the ratio between the total amount of users which are satisfied with the provided service, and the total amount of active users transmitting packets in the network, for each type of service.

- For VoIP and NRTV services a user is considered as satisfied if the percentage of the total amount of packets transmitted which are dropped due to time-out violation and/or violation of the maximum number of transmission attempts allowable is lower or equal to 3%.
- For WWW and FTP services a user is considered as satisfied if the average service throughput provided by the network is greater or equal to 32 kbps and 64 kbps, respectively.

· For WWW and FTP services a user is considered as satisfied if the average service throughput provided by the network is greater or equal to 32Kbps and 64Kbps, respectively.

Figure 17 is the plot of the user satisfaction ratio for (a) VoIP, (b) NRTV, (c) WWW (d) and (d) FTP users respectively. It can be seen that the proposed token and utility based scheduling algorithm has the best performance over all four proposed schedulers, except for high loads of the WWW service. With the increase in the system load, the maintenance of the satisfaction ratio from VoIP and NRTV users is achieved at the cost of degradation in the satisfaction ratio from WWW users. This is because the algorithm attempts to equalize the packet delay for RT services (such as VoIP and NRTV), while attempting to satisfy average service throughput for WWW users, and because it prioritizes access for VoIP and NRTV users according to the proposed utility functions for these three services. It can also be seen that the UTIL scheduler manages to satisfy both types of users, VoIP and NRTV, for different loads, although with a small degradation in the satisfaction ratio of VoIP users as the load increase. This is related to the kind of utility functions implemented for both service models: VoIP packets are kept in the buffer until the delay becomes higher than the priority timer and the negative exponential utility function, associated to NRTV users, gives higher priority whenever NRTV packets arrive in the buffer. That is, both utility functions complement each other.

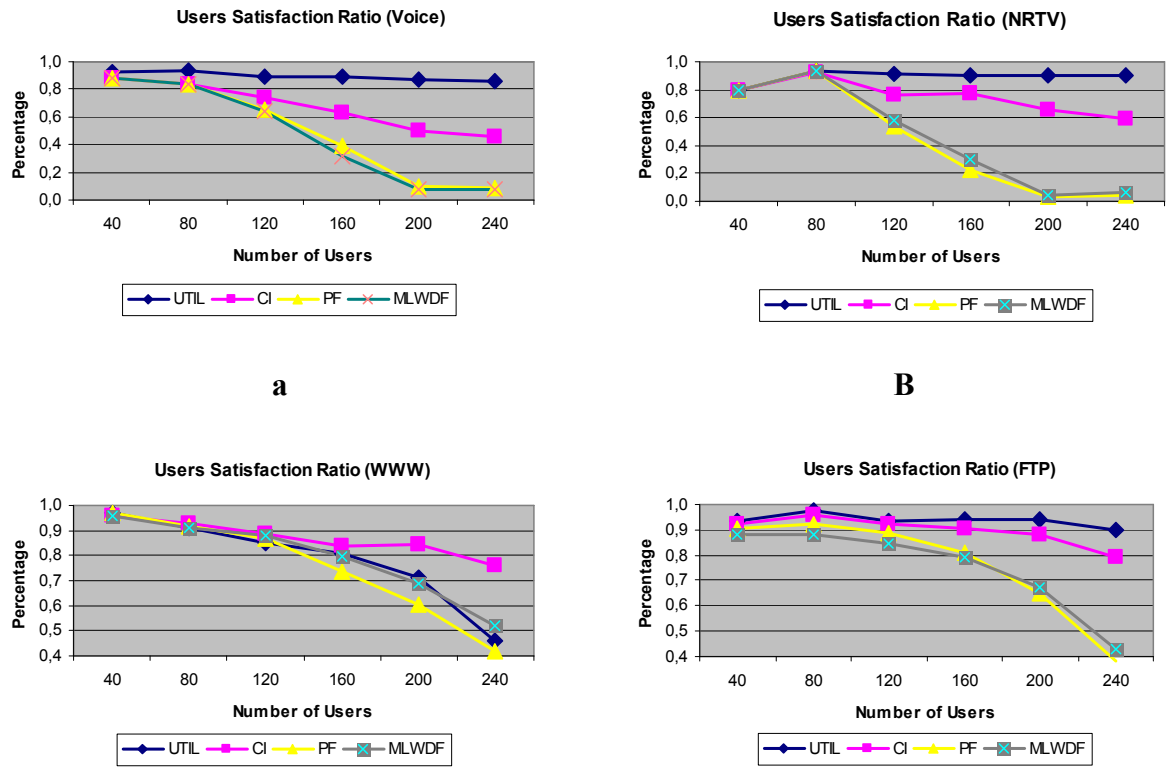


Figure 17 - Satisfaction Ratio for (a) VoIP, (b) NRTV, (c) WWW and (d) FTP users

As the FTP is a burst traffic model with a longer interval between active packet generation the amount of packets in the buffers is enough to be serviced whenever users from other services are not transmitting

It is interesting to observe that the CI scheduler has better performance than the PF and M-LWDF schedulers for all four types of traffic models. This is because both types of schedulers consider the average throughput computation up to the scheduling instant, and this value differs according to the type of service: the average throughput is higher for WWW and FTP and this is reflected into the significant degradation on the satisfaction ratio of users from VoIP and NRTV traffic models, compared to WWW and FTP ones.

Also, the M-LWDF was implemented with a more stringent allowable violation coefficient of the maximum packet delay for WWW and FTP users, which means that packets from these two types of services should be given more priority. It can be concluded that these two schedulers are not efficient in the prioritization of packets from different types of traffic classes. A kind of prioritization metric should be inputted into the priority computation to solve this problem.

7.5.4.2 Average Packet Delay per User

Figure 18 is the plot of the average packet delay per user versus the amount of active users in the network for (a) VoIP, (b) NRTV, (c) WWW (d) and (d) FTP users respectively.

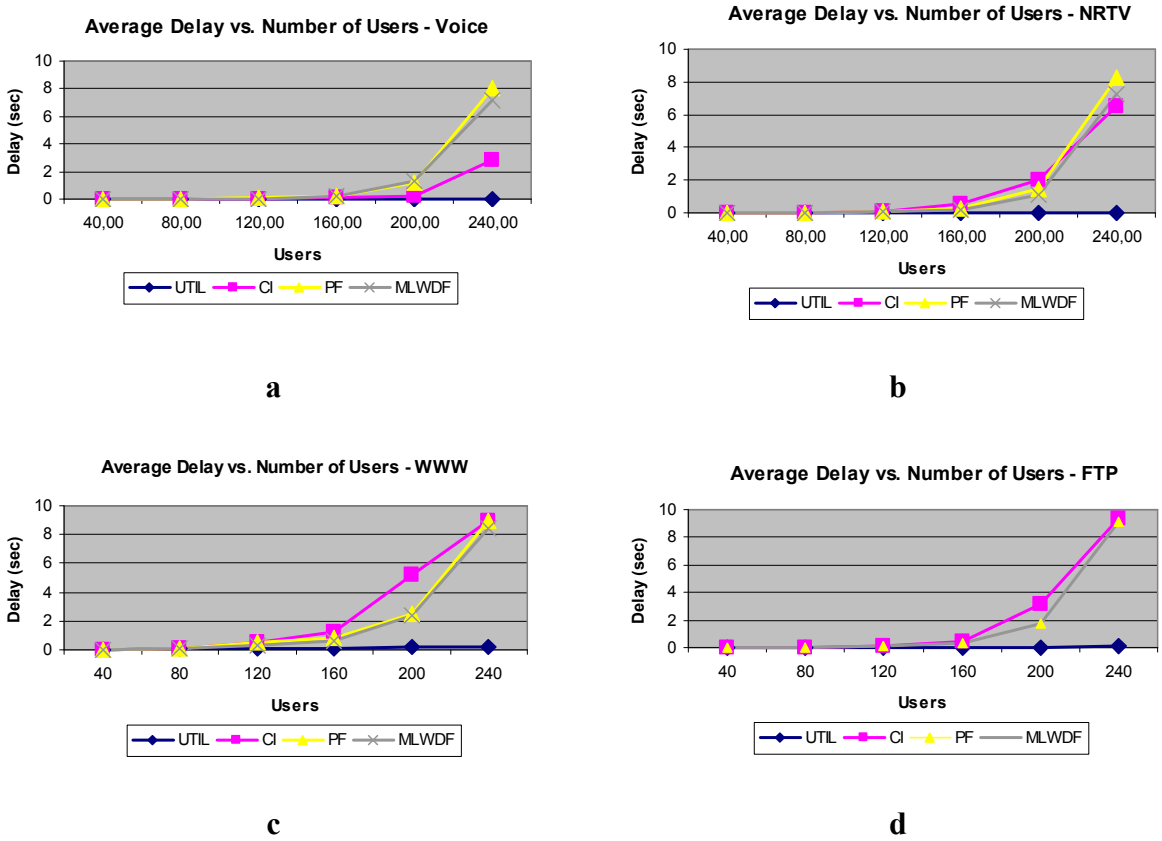


Figure 18 - Average Packet Delay per User for (a) VoIP, (b) NRTV, (c) WWW and (d) FTP users

According to these plots it can be inferred that, with the exception of the UTIL scheduler, the system is congested with loads corresponding to more than 160 users. Because the CI is an opportunistic scheduler which does not consider the packet delay in each scheduling period it presents worse performance than the PF and M-LWDF, especially for WWW and FTP traffic models which are of burst nature and because of the higher delay bound associated to them. This higher delay bound allows packets to remain for a longer period of time in the buffer before they are dropped. This reasoning is corroborated from the analysis of the NRTV traffic in which the CI scheduler has a worse performance than PF and M-LWDF ones. In VoIP traffic model there are periods of inactivity in packet generation, between talk spurts, in which packets accumulated in buffer can be transmitted. Differently from VoIP, NRTV is a streaming traffic model in which packets are generated with a constant rate and where there are no periods of inactivity. Therefore, packets are accumulated in the buffer at a much faster pace than with VoIP and for this reason they must be transmitted as earlier as possible in order not to be dropped, whenever they violate maximum delay bound.

The proposed token and UTIL scheduler presents the best performance for all four types of service classes. This is because packets whose delay becomes equal to or higher than the delay bound are not transmitted (they are dropped) as they lose their utility for the network.

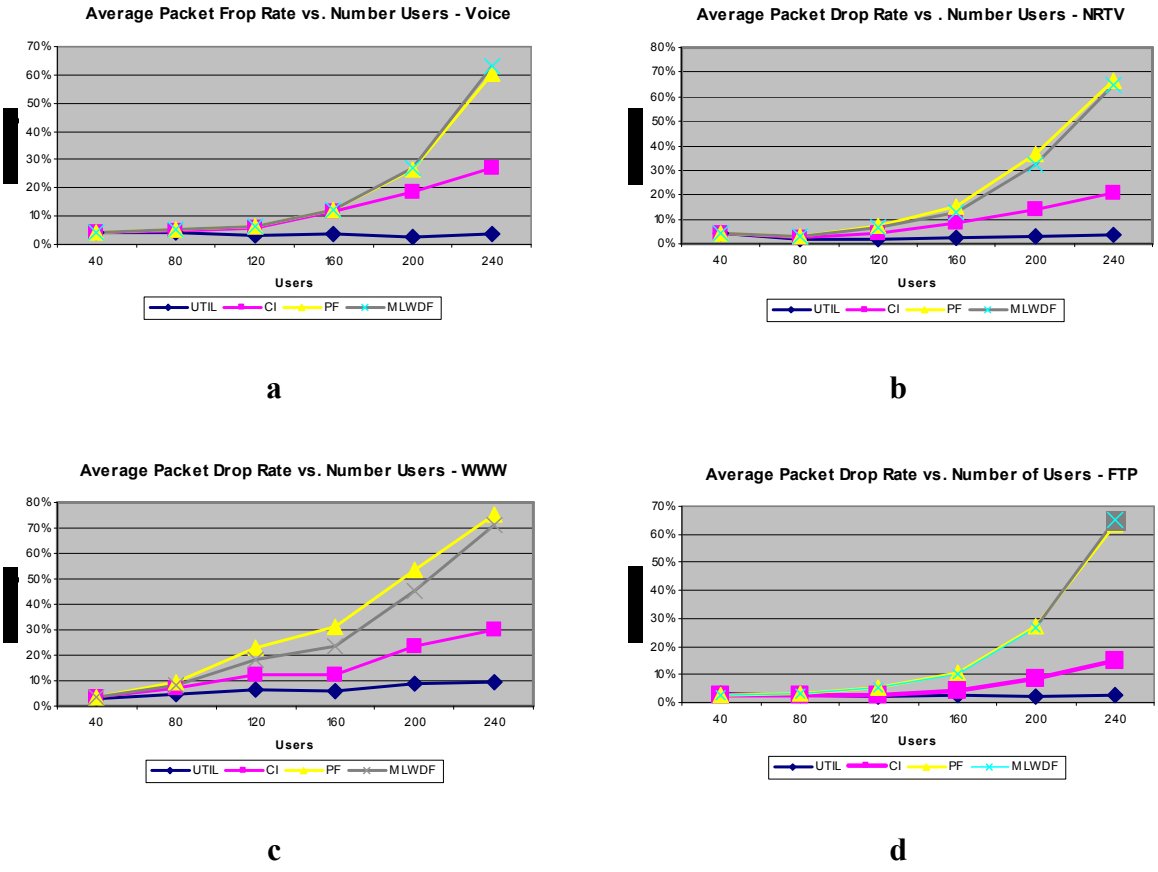


Figure 19 - Average Packet Drop Rate per User for (a) VoIP, (b) NRTV, (c) WWW and (d) FTP users

7.5.4.3 Average Packet Drop Rate per User

Figure 19 is the plot of the average packet drop rate per user versus the amount of active users in the network for (a) VoIP, (b) NRTV, (c) WWW and (d) FTP users respectively. This information regarding the average packet drop rate complements the information regarding the average packet delay per user. For instance: as the CI scheduler results in a higher average packet delay for burst traffic users, such as WWW and FTP, the percentage of dropped packets is lower for both traffic models because packets remain in buffer waiting for transmission for a longer period of time than packets from VoIP and NRTV traffic users.

For WWW users the UTIL scheduler results into a higher percentage of packets dropped due to delay bound violation, compared to the other three types of users. This amount of dropped packets contributes to the decrease in the achieved average service throughput for WWW, as can be seen from the plot in figure 20-c, and in a lower ratio of satisfied users, compared to the CI scheduler. It is worth mentioning here that, in order to decrease computation time, a delay bound of 500 ms was assumed for WWW packets. In the utility functions definition, as priority is given to VoIP and NRTV users, for loads higher than 160 users a higher percentage of users have their packets dropped before achieving the required average service throughput. If a higher delay bound was allowed for WWW packets one could expect a significant increase in the satisfaction ratio of WWW users with the UTIL scheduler.

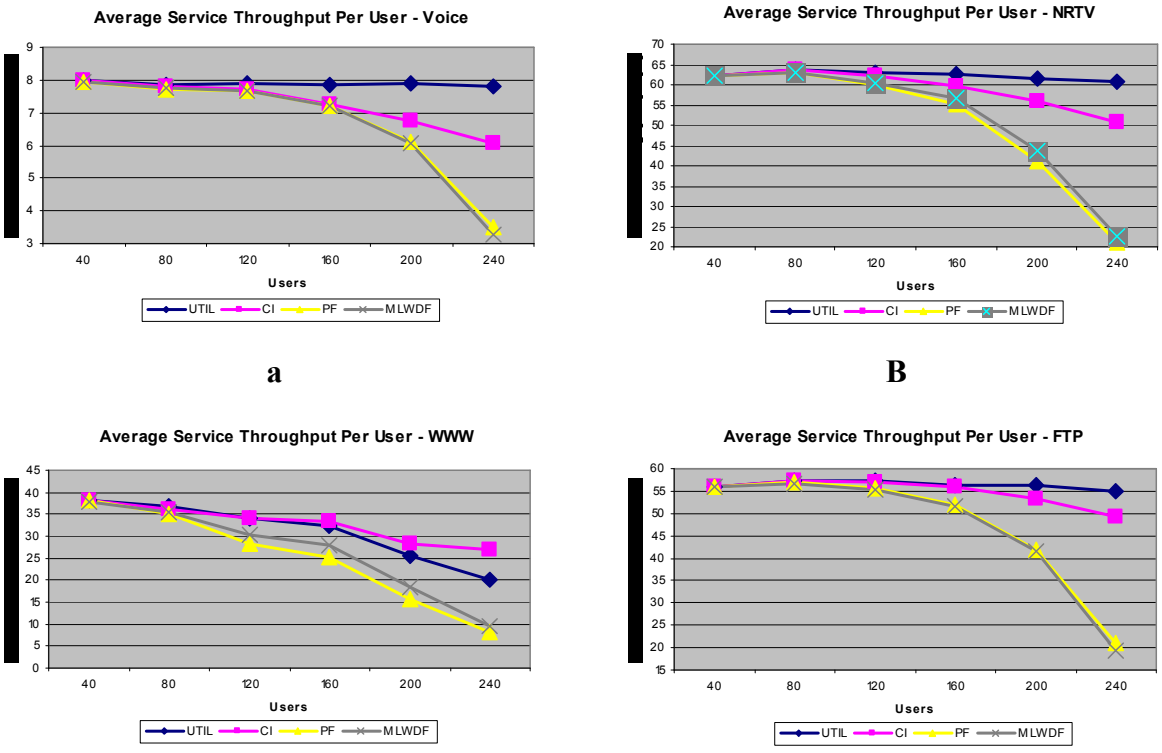


Figure 20 - Average Service Throughput per User for (a) VoIP, (b) NRTV, (c) WWW and (d) FTP users

7.5.4.4 Average Throughput per User

Figure 20 is the plot of the average service throughput for each type of traffic model considered in the simulations. As can be seen, for a load higher than 160 users the system is congested which results in a significant decrease in the service throughput for all types of service, except for the UTIL scheduling algorithm, because this scheduler does not transmit packets with no utility for the network, i.e., packets whose delay is equal to or greater than the maximum allowable delay for the service.

It can also be observed that a simple congestion control mechanism should avoid servicing more than 160 users in the system, because the minimum average service throughput of 32Kbps per user for WWW traffic model is achieved up to this load.

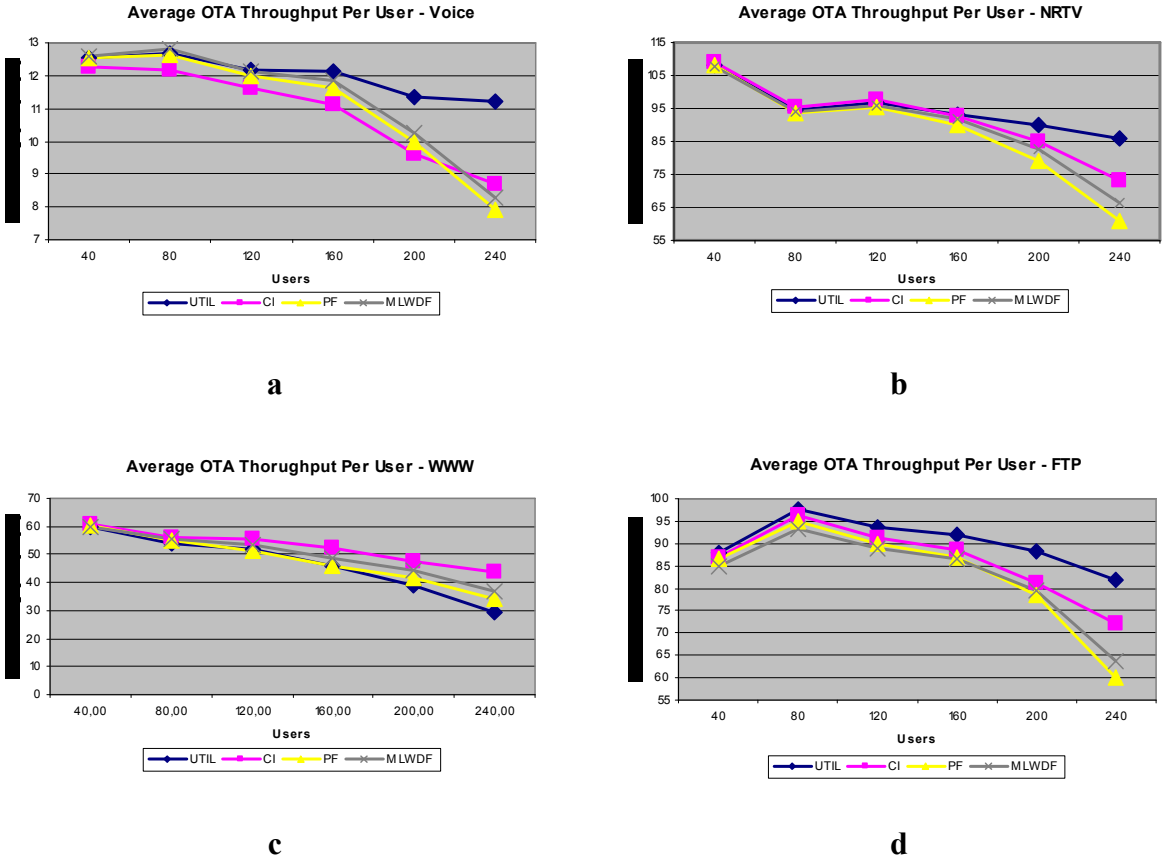


Figure 21 - Average Other-The-Air Throughput per User for (a) VoIP, (b) NRTV, (c) WWW and (d) FTP users

PF and M-LWDF schedulers are much more negatively influenced by the increase in the offered load, to a number higher than 160 users for VoIP and NRTV users, than CI scheduler. This is due to the inherent prioritization mechanism implemented in these schedulers. There is not enough capacity to satisfy packet delay constraints and a higher percentage of packets are dropped in order to satisfy the delay bound, which results in the significant decrease in the achieved service throughput, compared to the CI scheduler. The prioritization scheme implemented in the UTIL scheduler results in the slower decrease in the achieved service

throughput for NRTV users compared to the CI scheduler. It can be noticed that the decrease is less than 3% for the maximum load of 240 users in the system.

Figure 21 is a plot of the over-the-air throughput for each type of traffic model considered in the simulations.

7.5.5 Performance Distributions for Maximum System Load

System performance evaluation is also conducted by analysing the set of plots of CDF functions for each one of the performance metrics considered, for the maximum system load and for all four schedulers. The analysis from these plots is important because the average figures do not show the evolution in the distribution of average packet delay, average packet drop rate or average service throughput. With averaging metrics it is not possible to infer about the behaviour of each type of scheduler regarding the location of each user in the cell. From the CDF plots it is possible to analyse the behaviour of each scheduler for users located in the edge or near the cell's centre.

7.5.5.1 Performance for VoIP Users

Figure 22 (a) is a plot of the average service throughput per user versus the geometric factor. An admission threshold was used to forbid the access to resources from users with bad channel quality. Accordingly, in the simulations the admission threshold was set to -5 dB. Users with CQI lower than this value are not considered in the scheduling process.

As can be seen the UTIL scheduler results in the highest average service throughput per user for users in the edge of the cell, i.e., those users with geometric factor in the range [-5, 0]. It can also be noticed that both PF and M-LWDF schedulers present almost identical average service throughputs per user, in the range [-5, -2], and these figures are better than the ones resulting from the CI scheduler. This means that these two schedulers are more effective in serving users in the edge of the cell and with bad channel conditions than the CI scheduler, something which could not be noticed from the global average performance metrics presented in previous sections.

For higher values of the geometric factor the CI has better performance than these two schedulers and is equivalent to the performance of the UTIL scheduler. From this plot it can be inferred that the degradation in the performance of PF and M-LWDF schedulers, compared to the CI one, is due to the need to serve users in such conditions in the edge of the cell.

Plot 22 (c) shows the CDF of the average packet delay per user. As the M-LWDF and PF schedulers attempt to be fair in the allocation of resources to users from different types of traffic classes, without defining any prioritization other than the head of line packet delay (M-LWDF) or average service throughput (PF), they present a worse performance than the CI scheduler. Figure 22 (b) is the plot of the CDF of the average service throughput per user. This plot

corroborates the previous one, as it can be seen that the 10 percentile is lower for the CI scheduler.

Plot 22 (d) is the plot of the CDF of the packet drop rate. As can be seen, there is a huge difference between the amount of packets dropped by both the M-LWDF and PF schedulers and the amount of packets dropped with the UTIL and CI schedulers. For the 3% figure considered as the QoS requirement, it can be seen that 85% of the users comply with this requirement by means of the UTIL scheduler and 45% by means of the CI scheduler, while almost all users present a packet drop rate higher than 3% for the PF and M-LWDF schedulers.

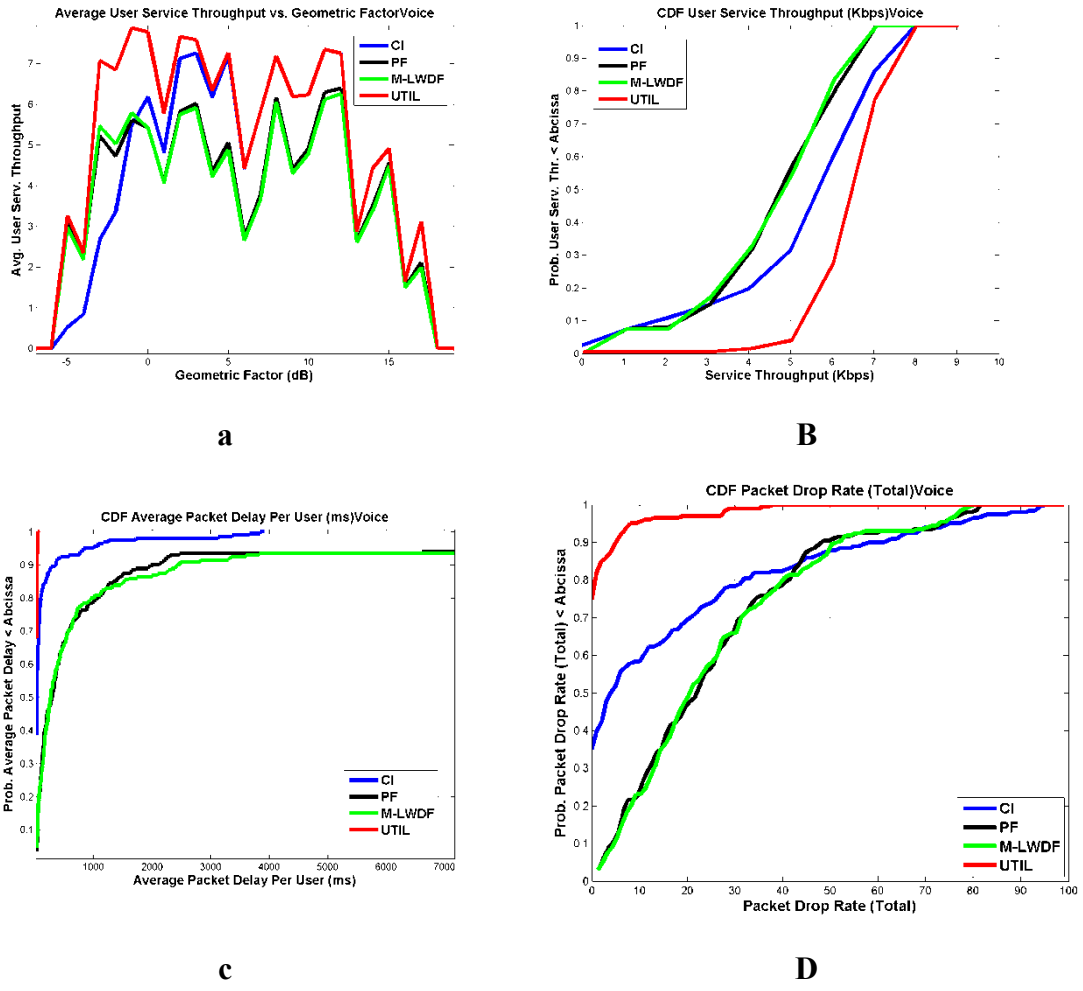


Figure 22 - Performance figures for VoIP service for 200 users of service load

7.5.5.2 Performance for NRTV Users

Figure 23 illustrates the same plots as described in previous point for the NRTV traffic model. As can be seen, the UTIL scheduler is even fairer in sharing resources for the NRTV traffic model, while the other three ones loose fairness as time evolves. This scheduler is also more effective in serving users in the edge of the cell than the CI scheduler. The UTIL scheduler presents better performance than the CI scheduler up to 3 dB of the geometric factor.

The fading deep around the 15 dB point corresponds to a peak in the plot resulting from the FTP traffic model. This behaviour results from the type of utility function implemented in the prioritization among both types of traffic for the UTIL scheduler. For higher values of the geometric factor the UTIL scheduler gives priority to users near to the cell centre. The same reasoning can be applied for the WWW traffic model in the 10 dB point. With the improvement in the channel quality more packets can be considered in the utility which can be potentially transferred to each user from WWW and FTP traffic models. 90% and 65% of the users achieve the required packet drop rate of 3% for UTIL and CI schedulers respectively. Although the M-LWDF scheduler presents better performance than the PF one, both schedulers cannot be considered as potential schedulers for such a high traffic load.

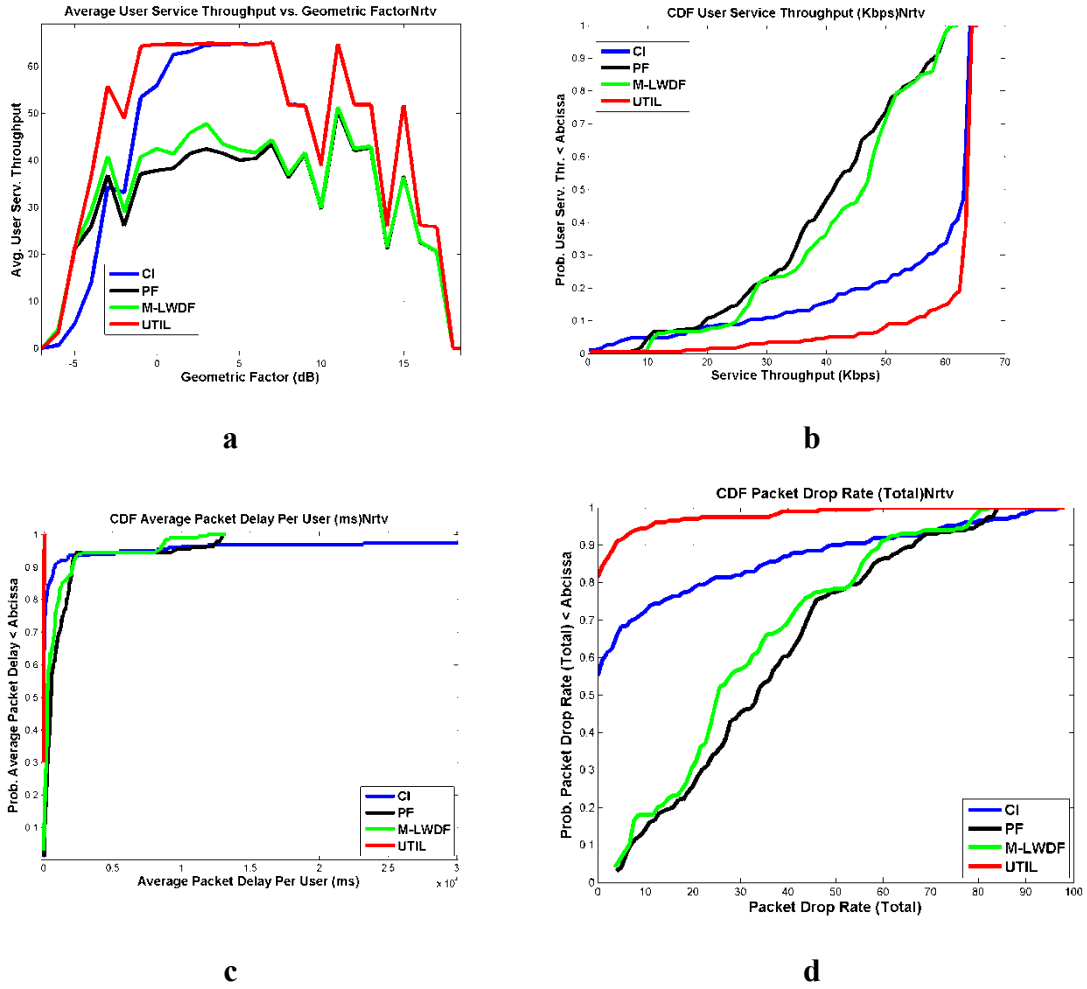


Figure 23 - Performance figures for NRTV service for 200 users of service load

7.5.5.3 Performance for WWW Users

Figure 24 illustrates the same plots for the WWW traffic model. As can be seen from figure 24 (a) both M-LWDF and PF result in the highest fairness among all four schedulers. But this result is achieved at the cost of degraded performance in the other performance metrics. With

WWW traffic model the difference in performance among M-LWDF, PF and UTIL, regarding the average service throughput per user, for users in the edge of the cell, is not that significant. As it was mentioned before, the prioritization resulting from the type of utility functions implemented forces the UTIL scheduler to behave like a CI scheduler as long as packet delays are lower than the maximum delay bound. The type of utility function used for packets of WWW traffic class and the limitation on the provided service throughput per user result in the better performance of the CI scheduler for users near to the cell centre. It can also be seen from the figures that the M-LWDF presents a better performance than the PF scheduler for users of traffic type WWW.

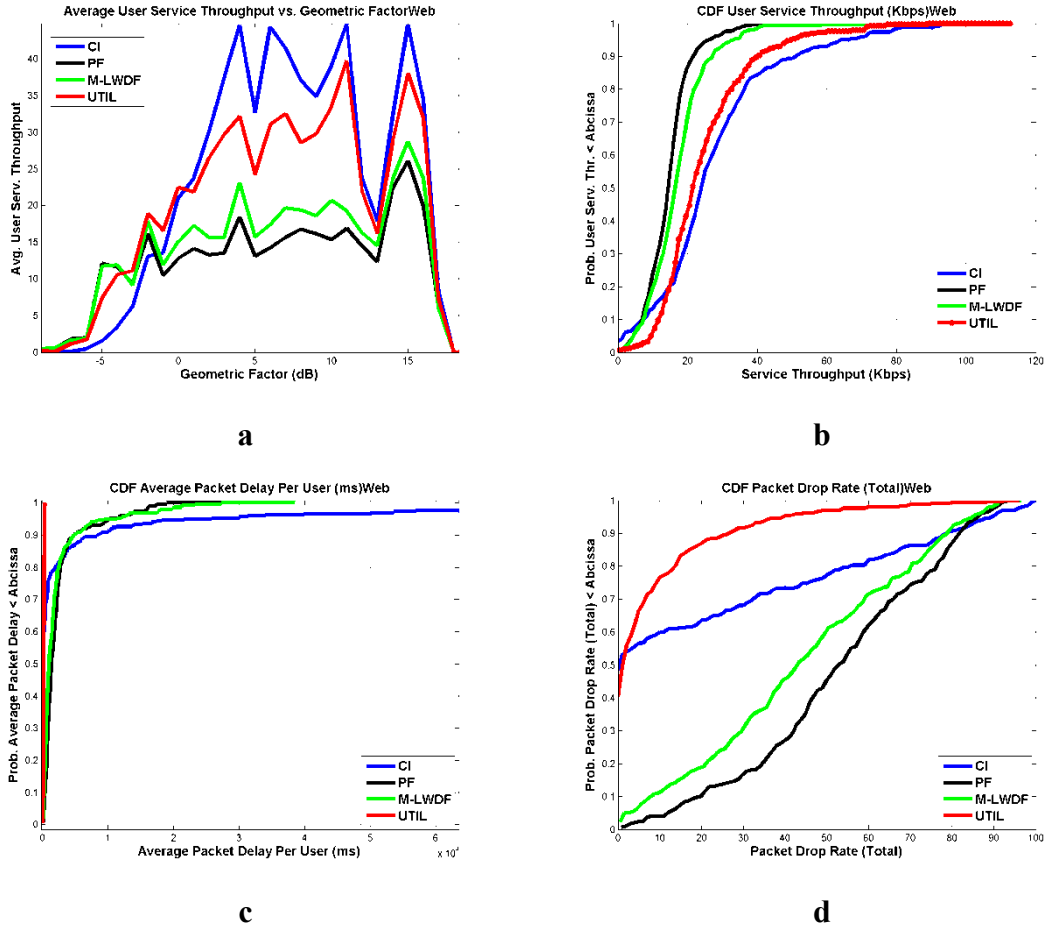


Figure 24 - Performance figures for WWW service for 200 users of service load

7.5.5.4 Performance for FTP Users

Figure 25 illustrates the same plots as described in previous point for the FTP traffic model. As FTP is characterized by long periods of inactivity, without packet generation, the performance of the UTIL and CI schedulers differ only for those users in the edge of the cell, with geometric factor in the range [-5, 0]. Most of the packets dropped are due to bad channel quality for these users. These packets are dropped after the maximum number of transmission attempts is achieved.

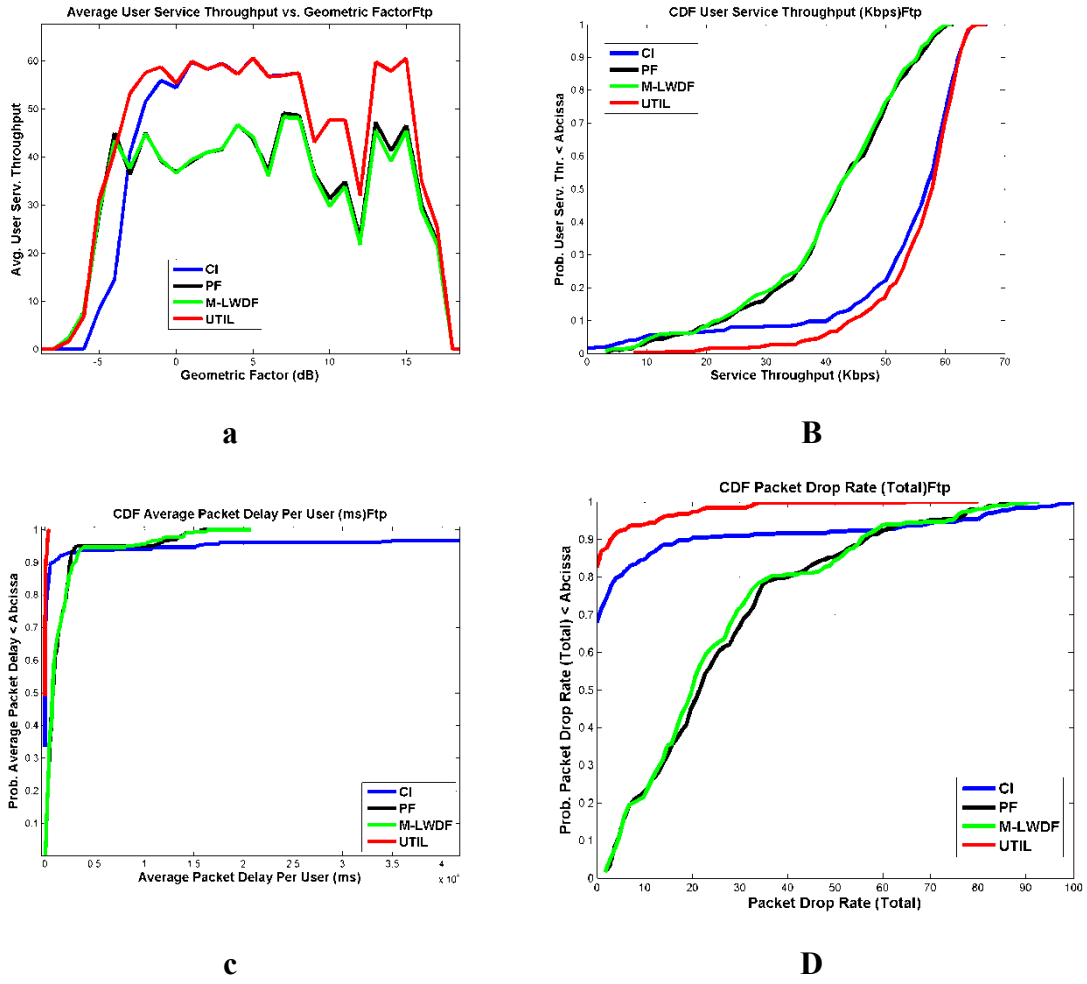


Figure 25 - Performance figures for FTP service for 200 users of service load

7.6 Related Work

There are a number of proposals in the research literature regarding the implementation of packets schedulers for the satisfaction of QoS requirements from different types of systems, such as: HSDPA, LTE, HDR and WiMAX. Many proposals are based on analytical approaches whose performance evaluation is conducted for simplistic scenarios of one single cell (no inter-cell interference) and assuming users as fully backlogged [138-151].

So far, few publications are available in the literature regarding cross-layer based scheduling frameworks which emphasize the level of importance a given application may have to the network and service provider. As it was enforced in this chapter, besides being reactive to changes in the channel state and traffic patterns, packet schedulers for BWA scenarios should take into consideration the level of importance each type of application has to the network operator and the related QoS constraints. The level of importance is fundamental in the operator's revenue, acquired from service provision.

For example in [152] a new QoS framework which uses sigmoid functions to model the degree of user's satisfaction is proposed by defining what is called the "user irritation factors". These factors reflect each user's sensitivity and tolerance to the degradation in the service provided.

Three types of classes of users, which depend on the amount of revenue provided to the network operator, are defined: gold, silver and bronze, and these are quantified according to the shape of the sigmoid functions designed in the scheduler. The attribution of radio resources by the radio resource manager depends on the outputs from these sigmoid functions and on the inputs from the irritation factors measured for each user.

Another example of this paradigm is [153] where the authors are concerned with what is called the "churn rate" i.e. the ratio of users canceling the subscription with their network provider due to the increased and continuous degradation in the perceived service satisfaction. They model the user's decision to join or leave the network and the provider's decision about the satisfaction of user's request, according to strategies from game theory and to the maximization of the revenue they provide. The provision of resources depends on the outcomes of the strategies defined for the game. Users are categorized into multiple classes and offered differentiated services based on the price they are willing to pay and the service degradation they tolerate before leaving the network.

Utility-based scheduling is a hot topic in the research community, regarding the implementation of packet schedulers for 3G and B3G networks, and there are many proposals for schedulers in the literature based on the notion of utility functions of the achievable user's data rate. These utility functions are used in resource allocation (bit, sub-carrier and power) algorithms in order to maximize the spectrum allocation for each user, guarantee fairness in resource allocation and comply to QoS demands, in terms of minimum bit rate requirement with maximum power guarantee (see e.g. [154] and references therein).

In [155] a layered scheduling architecture framework based on utility-based scheduling is proposed. Scheduling is divided into two layers. The first layer is related to the definition of the prioritization among the different types of traffic classes. The second layer deals with the scheduling for all users of the same type of traffic class.

In [156] an optimization scheduling framework is proposed in which the goal is the minimization of the total system's marginal utility of packet delay instead of the conventional objective of maximizing the total system utility. The authors claim that this approach is an easily achievable objective for packet scheduling that runs with significantly low complexity compared to the original utility-based scheduling problem.

In [157-158] a utility-based and channel dependent scheduling algorithm is proposed for the downlink of a cellular packet data system based on HDR standard, using a TDM multiple access scheme. The utility is a function of the user long-term mean throughput and is termed Alpha Rule as it can be parameterized according to a design parameter filling the gap between Max C/I and PF schedulers. They claim that a simple approach is enough to maximize the overall system utility.

In [159-160] an urgency and efficiency based packet scheduling algorithm that is able to schedule RT and NRT service classes is proposed. The mixed service architecture is based on the product of the marginal utility of HOL packet delay and normalized data rate. Properly defined utility functions are used to determine a prioritization in the access to OFDM symbols, along the whole spectrum of each symbol in an OFDMA multiple access scheme. Performance of the proposed framework is compared against standard MLWDF and PF standard algorithms.

In [161] it is proposed a downlink scheduling based on the notion of utility functions for OFDMA multiple access based cellular networks. The authors propose to guarantee both packet drop ratio as well as the play-out ratio of video streaming service flows. The play-out metric measures the fraction of transmitted packets which are received out of the maximum allowable delay jitter and cannot be inserted in the jitter buffer, used to control delay variation among adjacent packets in a streaming service flow. They claim that although standard algorithms such as MLWDF and EXP satisfy packet delay, this compromises delay jitter and they propose a scheduling algorithm based in the utility function principle to circumvent this.

In [162] a two-layered packet scheduling architecture is proposed for a general wireless communication system. The algorithm is defined to provide both QoS guarantees as well as a high throughput performance. Two time periods are considered in packet selection: long term selection selects packets which must be transmitted within the delay bound and short-term is based on channel status. Performance is compared against standard algorithms such as PF, CI and RR.

In [163] the authors propose a set of scheduling structures to support multiple traffic classes over multiple sub-channels, in a scenario based on the IEEE 802.16 OFDMA air interface. Each active user may have multiple connections with multiple traffic classes and must explore inherent diversities both in time and frequency domains. A prioritization is defined among the different classes supported and inside each traffic class urgent packets are transmitted first. The authors claim an increased throughput of up to 50% over standard schedulers such as MLWDF, PF and UEPS [159].

In [164] the authors present a new opportunistic scheduling scheme for an OFDMA-based wireless multimedia network. The scheduling decision is divided into two sub-problems: OFDMA sub-carrier allocation and sub-carrier assignment afterwards. Both steps explore multiuser diversity and are designed to provide fairness in relation to the average service throughput, packet dropping ratio and delay distribution per user. The authors claim that the proposed algorithm outperforms existing ones in the sense of service satisfaction from real-time users, as measured by the degree of satisfaction about QoS constraints: delays, rates and loss ratios.

In [165] a packet-by packet scheduler is proposed by granting priority to each packet instead of a user. Packets are classified into one of three classes: emergency, near emergency and non-

emergency, based on their packet delay, and different scheduling strategies are applied inside each class. According to system simulations the authors claim that their proposed scheme offers higher system throughput and lower packet drop rate than other existing scheduling schemes.

In [166] a cross-layer scheduling framework is proposed for the downlink Mobile WiMAX scenario. The priority metric computed for each connection depends on the QoS requirements as well as channel quality. Both PUSC and FUSC sub-channelization schemes are used in the simulations for performance evaluation.

In [167] an enhanced Proportional Fairness scheduling algorithm is proposed to support different levels of QoS in terms of throughput, delay and frame drop rate. The modified version of the FP scheme includes a parameter for service differentiation.

[168] elaborates on a cross-layer packet scheduler and channel allocation scheme for the IEEE802.16e OFDMA standard and for AMC channelization mode. For each packet, a priority metric is computed which depends on the service priority, channel status on each sub-channel in the frame and QoS requirements. The authors claim their scheme satisfies both maximal delay requirements of rtPS connections and at the same time the minimum reserved rate of nrtPS connections and results in a great enhancement of spectrum efficiency.

In [169] the authors propose a simple allocation scheme for fair allocation of slots in WiMAX frame among users. Both scheduler and slot allocation are based on QoS requirements of each service flow, in terms of data rate and bit error rate. They claim the proposed scheme results in fairness regarding slot allocation among real and non-real time service flows as well as QoS constraints satisfaction and minimizes the number of service flows in outage.

A call admission control and scheduling algorithm framework is proposed in [170] for real-time services for 3G cellular systems such as HSDPA or HDR, using a time slot-based multiple access scheme. The proposed scheduler attempts to minimize a delay-derived cost function. The call admission control module attempts to maximize the number of users supported in the system with satisfied QoS requirements, namely per user expected data rate. In this framework they considered real time scheduling algorithms balancing between system efficiency and user's QoS expectations.

In [171] a variation of the PF scheduler is proposed to overcome the limitations of this algorithm in the support of users in the edge of the cell. The new scheduler improves the cell edge performance.

In [172] an adaptive traffic allocation scheduling algorithm is presented which is able to support multi-traffic. This algorithm defines priority to users according to delay sensitivity from each traffic class. RT and NRT traffic classes are first scheduled according to packet delay. Then packets are allocated in the frame according to the data rate requirements of the service data flow.

7.7 Conclusion

In this chapter a new scheduling framework based in the notion of utility functions from economics is proposed. The scheduler framework is implemented within a fully compatible cross-layer design paradigm by using the signalling control channels available in the WiMAX radio frame. Differently from other schedulers available in the literature, which are based on the notion of utility function, and which attempt to maximize the system utility according to specially designed optimization algorithms, the scheduling principle presented in this work estimates the injury incurred to remaining users in the system, whose transmission is postponed in each scheduling period. The algorithm attempts to maximize the potential utility transferred to the network when deciding which users must be assigned resources for transmission, while at the same time it attempts to minimize the loss in the utility incurred to remaining active users in the cell, if they are postponed for transmission.

The utility-based scheduling algorithm is inserted into the proposed DRA architecture, developed for the conduction of system level simulations for Mobile WiMAX. Radio resources are available in both time and frequency domains, according to the OFDMA frame. Available radio resources are assigned to users according to the list outputted from the scheduling algorithm and according to the amount of information remaining in the user's buffer. That is: scheduling is performed in a user to user basis, and not in a packet to packet basis. Fundamental to the performance of the utility-based packet scheduling algorithm is the definition of the utility function for each type of traffic model considered in the network.

The parameters considered in the design of the utility function are: the initial gain (from which packets start losing utility), the shape of the function (including the rate at which utility is decreased) and the minimum utility value (from which the packet has no utility al all to the network). Prioritization among packets from different users (and different traffic models) is decided according to the assigned utility function.

Two versions of the utility algorithm were proposed. The first version is the basic utility algorithm which considers the packet delay for the computation of the metrics needed in the definition of the prioritization list. The second version is an extension which attempts to map a required minimum service throughput into a properly designed utility function, by means of a token bucket algorithm. Both versions were plugged into the DRA which was implemented in the system level simulator and is specially designed for conducting system level simulations in Mobile WiMAX. Both versions were also compared against commonly proposed schedulers available in the literature, using standard traffic models representing the different types of traffic classes proposed in the Mobile WiMAX standard.

According to the results obtained from system level simulations it was possible to verify the gains achieved in terms of satisfied users ratio and network performance metrics, such as average throughput and average packet delay, compared to the other packet scheduling

algorithms, whose performance metrics were used as benchmark figures. From these simulations and analysis it is possible to conclude that the notion of packet utility can be effectively adapted to be used in packet scheduling algorithms, which turn out to be potential solutions for B3G and 4G networks, regarding the satisfaction of the stringent and demanding QoS requirements, expected from the type of service applications envisioned for these networks in the near future.

Chapter 8

Space Division Multiple Access with Utility-Based Packet Schedulers for Mobile WiMAX

8.1 Introduction

This chapter describes all the steps followed in the implementation of Space Division Multiple Access (SDMA), in the basic Dynamic Resource Allocation (DRA) module presented in previous chapters. In Mobile WiMAX air interface SDMA is a technique resulting from the transmission along the same set of sub-channels and/or OFDM symbols, but using beam patterns spatially separated in the space domain. SDMA provides another degree of freedom in the map of resources of the OFDMA-based air interface, compared to conventional multiple

access schemes such as Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA) or Code Division Multiple Access (CDMA). The underlying idea behind SDMA is to divide the space into a number of orthogonal spatial beams which can be separated in the receiver. SDMA brings out another dimension in resource allocation, from physical (PHY) to medium access control (MAC) layer, in the transmission of information over the air interface.

The new SDMA-based DRA architecture is used in conjunction with the utility-based packet scheduler. Besides time and frequency domains, users can now be assigned spatial beams for information transfer, provided the degradation in signal quality for users already assigned slots in the map of resources, from new users which are assigned the same slots, but with different spatial beams, is not affected to the point of compromising the user satisfaction for the quality of service accomplished by the network. Therefore, in the new DRA architecture, users in the scheduling priority list can now be assigned resources in space, besides time and frequency domains.

The standards of the IEEE 802.16 family adopt Adaptive Antenna System (AAS) as an option to enhance cell capacity and coverage. The benefit incurred with the use of AAS is its ability to reduce interference by steering the beam to a specific user. Adaptive antenna elements can be used to distort the radiation pattern produced by an antenna array at the base station. The idea is to focus (beamform) the transmitted signal energy in the direction of the intended mobile receiver and, at the same time, steer nulls in the directions of co-channel mobiles. This results in an enhancement of the perceived Signal to Interference plus Noise Ratio (SINR) at the desired mobile's receiver. Comparable approaches are currently being standardized by 3GPP for UMTS or by the IEEE for 802.11n. These advanced antenna techniques have a significant impact on the capacity and service quality provided by wireless links and the efficient use of the available spectrum [173].

The beamforming processing is accomplished by applying complex weights to the antenna elements at the base station antenna array (although it theoretically could be done also at the mobile station with the cost of added complexity). The weights are computed by using information regarding the Directions of Arrivals (DoA) of desired and interfering signals at the base station. The amount of main lobes to steer must be lower or equal to the amount of degrees of freedom, corresponding to the amount of antenna elements in the array. Due to the linear nature of the antenna array it is also possible to generate multiple patterns of signals to different mobiles. The resulting pattern will be the linear combination of the different patterns.

In order to assign different spatial beams to mobiles they must be separated in space. This separation depends on the main lobe aperture, which is a function of the amount of antenna elements in the array, and on the spatial signatures from each mobile at the base station. The spatial signature captures the spatial characteristics (direction-of-arrivals or departures, number

of multipath components and attenuation) associated with the mobile's terminal. Of course, in reality, the beams are not completely orthogonal and the amount of intra-beam interference increases with the amount of users being assigned spatial beams for the same radio resource (time and/or frequency for example).

This chapter is organized as follows. Section 2 introduces smart antenna techniques and their use in the SDMA multiple access scheme. Here, these techniques are based on linear processing of the antenna array at the base station, in order to result in non-overlapping antenna beams at each mobile attempting to transmit at the same frame interval. These non-overlapping beams avoid intra-beam interference over the same set of radio resources in the frame and increase cell's capacity. Section 3 details the frame structure for the implementation of AAS in Mobile WiMAX. It introduces different proposals available in the research literature for estimation of user's separability across different spatial beams and describes the steps followed in the SINR estimation, after user's assignment to spatial beams, over the same set of symbols and sub-channels in the TDD frame. Section 4 is the core of this chapter as it explains the principle behind the SDMA-based DRA proposed for Mobile WiMAX MAC. The DRA computes the groups of users to transmit in the same radio resource (same set of slots in time and frequency domains), according to the estimated correlation among the set of channel matrices, over all data sub-carriers composing each resource. Section 5 describes the proposed SDMA-based DRA algorithm which comprises users prioritization and resource allocation, according to SDMA multiple access. The algorithm used in the computation of the groups of users and assignment of spatial beams, over each radio resource, is presented as well as the scenario used in the system level simulations. Section 6 presents results from the performance evaluation of the proposed joint utility-based packet scheduler and SDMA multiple access scheme for Mobile WiMAX. Simulations were conducted separately for three types of traffic models: Full Queue, Voice over IP (VoIP) and 3GPP's World Wide Web (WWW). The scenario used for system level simulations was the one of a 4x2 MIMO channel antenna system configuration with Alamouti Space Time Block Coding (STBC). The gains achieved with this new DRA architecture are compared against standard solutions common referred to in the literature. Section 7 is about the related work available in research literature regarding the implementation of SDMA-based packet schedulers. Section 8 concludes the chapter.

8.2 Spatial Beamforming

Figure 1 depicts a typical antenna array system in which the spatial diversity is exploited for multiple access communications. In the absence of noise the response of an M-element antenna array to a signal $s(t)$ is written as in equation (1).

$$\mathbf{y}(t) = \left(y_1(t), y_2(t), .., y_M(t)\right)^T = \mathbf{a}s(t) \quad (1)$$

The vector $\mathbf{a} = (a(1), a(2), .., a(M))^T$ is the array response that captures the spatial characteristics associated with the terminal and is designated commonly as "spatial signature". Assuming *K* mobiles communicate at the same time with the same base station the total signal output at the antenna arrays is given by equation (2).

$$\mathbf{y}(t) = \sum_{k=1}^{K} \mathbf{a}_k s_k(t) + \mathbf{n}(\mathbf{t}) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(\mathbf{t}) \quad (2)$$

Where $\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_K)$ is the array manifold whose columns represent the spatial signatures (one for each mobile terminal). Each entry in the vector $\mathbf{n}(t)$ represents the Gaussian noise in the respective antenna and each entry in the vector $\mathbf{s}(t) = (s_1(t), ..., s_K(t))^T$ represents the signal transmitted by each one of the *K* terminals.

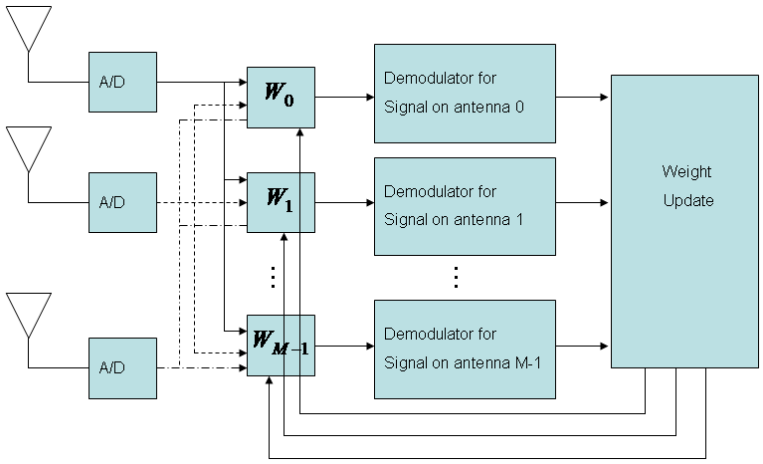


Figure 1 - Spatial beamforming module for a linear antenna array

Although the spatial signatures from the *K* terminals are superimposed at the base station's antenna array, if they are spatially separable (orthogonal) they can be separated through spatial filtering processing. To retrieve the individual signal $s_i(t)$ from mobile *i*, antenna outputs are weighted and summed with a set of complex weight coefficients (computed for each mobile), according to equation (3).

$$\hat{s}_i(t) = \sum_{m=1}^{M} w_i^*(m) y_m(t) \quad (3)$$

The vector of weights $\mathbf{w}_i = (w_i(1), ,..., w_i(M))^T$ is designed to constructively combine the signal of interest from mobile *i* (steer the beam in this direction) and destructively combine the interference-plus-noise from other mobiles. If multiple beams are applied there must be one weight vector per beam.

8.3 SDMA Scheme for IEEE802.16e Mobile WiMAX

This section describes, in great level of details, the model describing the integration of SDMA with the OFDMA multiple access schemes used in Mobile WiMAX.

A cellular network composed of three tiers of tri-sectored base stations is considered. Only the downlink connection is used for beamforming. Each base station is equipped with a linear uniform antenna array with N_T antennas and each mobile terminal is equipped with a linear uniform antenna array with N_R antennas. With N_T antennas in the transmitting array there can be a maximum of $M_n \leq N_T$ beams on the n^{th} sub-carrier of the spectrum of each OFDM symbol. Based on the computed SINR on beam *m* and sub-carrier *n*, the estimated BLER is obtained from the look-up tables which implement the interface to the link layer. The supportable data rate, $r_{n,m}$ on beam *m* and sub-carrier *n* is derived from the estimated BLER.

The amount of feedback reported by each mobile station may be extremely high if there is a need to compute the set of weights for each sub-carrier in the spectrum of each OFDMA symbol of the TDD frame. The periodicity of this report depends also on the type of environment in the mobile radio channel. For mobile terminals moving with high velocity, the coherence time results in the de-correlation among blocks of consecutive OFDM symbols, requesting feedback reports inside the same TDD OFDMA frame. However, for the mobile speeds and channels used in the set of simulations conducted under the scope of this work, the mobile channel is assumed as being constant along each frame period, and the report is performed periodically for each radio frame.

In the frequency domain the degree of correlation among sub-carriers depends on the type of sub-channelization used:

- For DL-PUSC sub-channelization, according to the way data sub-carriers are distributed in frequency it turns out to be not very practical for the implementation of beamforming because reports must be performed for all sub-carriers in each sub-channel and different vector of weights must be computed for each sub-carrier.
- For DL-AMC sub-channelization mode, sub-carriers are allocated continuously to form a bin and there is a significant correlation between adjacent sub-carriers inside the same bin, which means that the vector of weights may be estimated for the pilot sub-carrier and the same vector applied to all data sub-carriers inside the bin. This reduces the feedback rate.

8.3.1 User Spatial Separability

In Mobile WiMAX active users under the area of coverage of each base station are allocated slots, mapped into both time and frequency domains in the map of resources of the TDD OFDMA frame. SDMA expands the available capacity (in number of slots) by allowing multiple mobile stations to transmit in the same slot, through the assignment of different non-overlapping spatial beams to the set of mobile stations. It adds another dimension to the map of resources by expanding each slot into another set of slots in the space domain.

Assuming downlink connection, the size of the space domain is defined by the number of degrees of freedom associated to the number of antenna elements at the base station. If there are N_T antenna elements in the antenna array then up to N_T mobiles can transmit in the orthogonal beams associated to the same slot, in time and frequency. This translates into an N_T -fold increase of system throughout in an ideal scenario (up to N_T orthogonal beams can be created). These spatial beams are steered by the beamforming algorithm in the base station.

However, in most practical situations the number of active users in the area of coverage of a base station is greater than the maximum number of spatial beams which can be created for each slot. Also, in reality, the ability to capture multiple transmissions in different beams of the same slot depends critically on the spatial configuration of the co-slot mobiles, because not all users can be separated in space. Intuitively, by assigning "most-orthogonal" mobile terminals to the same slot, the average performance of the system can be improved.

The performance of the combined SDMA/OFDMA system is highly affected by the strategy followed in the separation of users into non-overlapping groups, in space. Therefore, different schemes were proposed in the literature.

In [177], a spatial grouping algorithm is proposed to group users according to their spatial separability. The result of the grouping is a set of spatial groups of users. For users from distinct groups there is no separability so that different groups have to be allocated in the time domain. For users inside the same group they can be serviced in the same resources in time and/or frequency at the same time.

A different application of user's spatial grouping is presented in [178] for an adaptive resource allocation algorithm implemented on a MIMO/OFDMA multiple access system. The objective is the minimization of the overall transmission power, given user's QoS requirements in terms of bit error rate and data rate. The proposed scheme computes the correlation among users for each sub-carrier in order to control the sub-carrier's sharing according to the user's spatial separability. This makes it possible to decouple the multiuser joint resource-allocation problem into simple single-user optimization ones.

In [179] a heuristic metric is proposed for the computation of the degree of separation for each pair of users. This metric is based on the normalized Frobenius norm of the product of their MIMO channel matrices and is used in the definition of the groups of users who can be separated in space. The metric that estimates the cost of putting two users together in the same group is given by equation (4):

$$\xi_{i,j} = \frac{\left\| \mathbf{H}_i \mathbf{H}_j^H \right\|_F^2}{N_{Ri} N_{Rj}} \tag{4}$$

Where $\mathbf{H}_i$ and $\mathbf{H}_j$ are, respectively, the channel matrices for users *i* and *j*, and N_{Ri}, N_{Rj} are the number of antennas at the receiver of mobiles *i* and *j*, respectively. After computing all $\xi_{i,j}$ for all possible combinations, a Compatibility Optimization Algorithm (COA) is used to find the group that optimizes the space allocation, according to equation (5):

$$\sum_{g=1}^{G}\left(\sum_{i,j\in G_g}\xi_{i,j}\right) \tag{5}$$

Where *G* is the number of groups in the system, corresponding to the number of time or frequency slots available for user multiplexing and G_g is one particular group in this set.

In [180] the same principle is used in the computation of the degree of separation of two users transmitting in the same sub-carrier. Two users are called spatially separable on a sub-carrier if they share the same sub-carrier and the SINR requirements at corresponding receivers are satisfied. The normalized scalar product between the channels of each pair of users is computed according to equation (6):

$$\eta_{i,j}=\frac{\left|\mathbf{H}_{i,n}^{H}\mathbf{H}_{j,n}\right|}{\left\|\mathbf{H}_{i,n}\right\|\left\|\mathbf{H}_{j,n}\right\|} \tag{6}$$

Where: $\mathbf{H}_{i,n}$ and $\mathbf{H}_{j,n}$ are, respectively, the channel matrices for users *i* and *j* on sub-channel *n*.

The larger the metric is the more correlated their channels are, and more power is needed to meet the SINR requirements. For each user to be allocated, the one resulting in the minimum of the set containing the maximum of the scalar products, for all already allocated users is selected, as in equation (7):

$$k^{*}=\arg\min_{i\in K-K_n}\max_{j\in K_n}\eta_{i,j} \tag{7}$$

Where *K* is the set of users in the cell and K_n is the set of users allocated to sub-carrier *n*.

In [181] an SDMA based wireless packet cellular system is proposed. Users are assigned the same time-slot provided their angular separation in space is above a given threshold. The architecture encompasses a packet scheduler and a resource allocator. Users which cannot be spatially separated are assigned to an empty time-slot.

8.3.2 SINR Estimation after Performing Beamforming

After a given mobile station is assigned an empty beam in a given slot in the map of resources, a new SINR value will result for both the new mobile as well as for the mobiles already assigned empty beams in the same slot. The new value for the SINR ratio will be affected from both the intra-cell interference, resulting from all mobiles transmitting on different spatial beams associated to the same slot, as well as from the new value of the desired signal resulting from the beamformed pattern in the direction of the mobile.

As the new antenna pattern is steered into the direction of the mobile, the beamforming process results into an improvement of the SINR measured at the receiver. In order to estimate the correct SINR all relevant signals must be differentiated. Figure 2 illustrates the relevant signals during SDMA transmission.

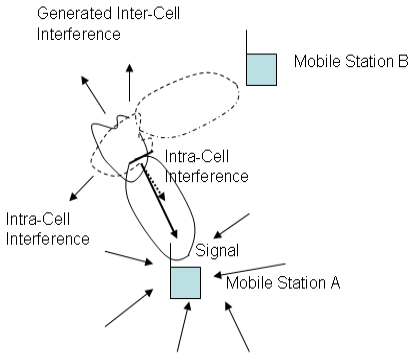


Figure 2 - Estimation of the SINR ratio for antenna beamforming

An optimized antenna pattern is applied at the base station to produce a main lobe (spatial beam) in the direction of mobile station **A** in order to maximize the desired signal at the receiver. The other spatial beams are optimized for concurrent transmission, in order to minimize intra-beam interference in mobile station **A**. Neighboring cells produce inter-cell interference but the attenuated side lobes, resulting from the beamforming operation, result in a small contribution to the interference at mobile station **A**.

The SINR at mobile station is computed according to equation (8):

$$SINR_{SDMA} = \frac{S}{\sum_{allbeams} I_{Intra} + I_{Inter} + N_0} \tag{8}$$

Where S is the desired signal, I_{Inter} is the contribution due to inter-cell interference, I_{Intra} is the contribution to the interference from each other spatial beam in the same cell and N_0 is the power of the thermal noise at the receiver.

8.4 Proposed SDMA-Enabled DRA Architecture

The principle behind the SDMA-based DRA architecture proposed for Mobile WiMAX MAC layer is now described. The SDMA architecture defines the groups of users to transmit in the same radio resource (group of slots in time and frequency domains), based on the correlation among the set of channel matrices over all data sub-carriers composing each resource.

As described in chapter 5, in the DL-PUSC sub-channelization mode the map of radio resources associated to the TDD OFDMA frame is made up of 15 resources. Each resource is comprised of 10 sub-channels and 6 OFDM symbols which result in a total of 30 slots available per resource. The map of resources in the SDMA architecture is illustrated in figure 3.

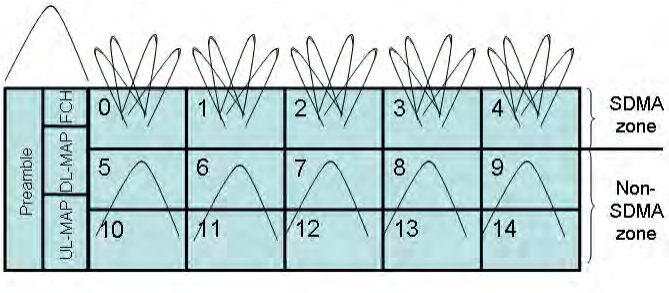


Figure 3 - WiMAX TDD frame for SDMA implementation

For the implementation of SDMA, the original map of resources was divided in two zones:

- **SDMA zone**: each resource can be assigned to different mobiles by means of spatial beams. This zone comprises the resources in the first row of the map (indexes from 0 to 4), and, therefore, they are designated as **SDMA-mode resources**.
- **Non-SDMA zone:** there can be no spatial beam assignment and therefore SDMA is not applied to their resources. Resources in the other two rows of the map (indexes from 5 to 14) constitute the non-SDMA zone and, therefore, they are designated as **non-SDMA mode resources**.

The reason behind this subdivision of radio resources into two zones has to do, mainly, with complexity issues. If all resource units were allowed to transmit in SDMA mode the simulation speed would decrease significantly, because correlations and verifications of SINR compatibility checks would have to be performed for all resources in the frame.

As can be seen from the figure, two types of radiation patterns were used in resource assignment for packet transmission and signalling broadcast:

- One radiation pattern with a larger beamwidth for broadcasting the preamble and control signals in the downlink sub-frame (FCH, DL-MAP, UL-MAP) and the control signals in the uplink sub-frame. This radiation pattern is also used for broadcasting resources in the non-SDMA zone in the downlink sub-frame (see table 1).
- A second radiation pattern with a narrower beamwidth corresponding to the beamformed pattern used in the broadcast of the resources in the SDMA-zone. This beamformed pattern can be steered in the direction of the desired mobile (see table 1).

8.4.1 Computation of Users Correlation

In the uplink sub-frame of the Mobile WiMAX system the base station defines a permutation zone for channel sounding [11]. The correlation is computed from pilot symbols specially programmed for channel sounding in this zone, and for all data sub-carriers in each sub-channel composing each resource in SDMA mode in the map of resources.

Assuming user *i* is to be assigned a resource, the correlation of its channel matrix, $\mathbf{H}_n^i$, with the channel matrix, $\mathbf{H}_n^j$, of user *j* (which is already assigned to one spatial beam in the same

resource), for the n^{th} data sub-carrier, is simply the scalar product of both channel matrices as given by equation (9).

$$\rho_{i,j}^{n} = \frac{\left\| \left(\mathbf{H}_n^i \right)^H \mathbf{H}_n^j \right\|_F}{\left\| \left(\mathbf{H}_n^i \right)^H \right\|_F \left\| \mathbf{H}_n^j \right\|_F} \tag{9}$$

Then, the algorithm performs the average over all data sub-carrier in the vector of correlations, as in equation (10):

$$\overline{\rho_{i,j}} = \frac{\sum_{l=1}^{N} \rho_{i,j}^{l}}{N} \tag{10}$$

Where N is the number of data sub-carriers in the resource.

It is important to mention that a better approach would be to correlate the matrices resulting from stacking the channel matrices for all data sub-carriers as given in equation (11).

$$\overline{\rho_{i,j}} = \frac{\left\| \mathbf{H}_i \mathbf{H}_j^H \right\|_F}{\left\| \mathbf{H}_i \right\|_F \left\| \mathbf{H}_i \right\|_F} \tag{11}$$

Where $\mathbf{H}_i = stack\left[\mathbf{H}_1^i \quad \cdots \quad \mathbf{H}_n^i \quad \cdots \quad \mathbf{H}_N^i \right]$ and $\mathbf{H}_j = stack\left[\mathbf{H}_1^j \quad \cdots \quad \mathbf{H}_n^j \quad \cdots \quad \mathbf{H}_N^j \right]$ are the matrices resulting from stacking the vectors of channel gains for each data sub-carrier.

The problem with such an approach would be the increased degree of complexity in the computations. These would slow down the simulation speed. However, for the DL-PUSC sub-channelization mode there is a kind of frequency diversity in the allocation of data sub-carriers due to their pseudo-random distribution over the spectrum. This seems to be a good reasoning for the proposed metric.

In the Mobile WiMAX standard the base station can reserve a set of pilot sub-carriers and symbols in the TDD OFDMA radio frame for channel sounding of the uplink transmission from the desired mobiles. These pilot sub-carriers are used in the computation of the correlation matrices.

8.4.2 Computation of SINR for Resources in SDMA and in non-SDMA Zones

All active mobile stations in the cell read the preamble of the TDD OFDMA frame to estimate their channel quality (CQI). Also, besides being used for CQI estimation, the preamble is used in time and frequency synchronization between the mobile and its serving cell. It is broadcasted with the most robust MCS scheme for even mobiles in the edge of the cell to be able to decode its modulated pattern. For this reason, the preamble must be transmitted in the non-beamforming configuration, i.e., with the horizontal beam pattern used in previous system-level simulations (a beam with 70 degrees of half power bandwidth and 20 dB of maximum

attenuation). For the computation of the SINR for all mobiles being assigned non-SDMA mode resources this is the antenna pattern that is considered in the serving and interfering base station. For all mobiles being assigned SDMA-mode resources, the SINR will depend on the type of antenna radiation pattern and on the beamforming algorithm. To simplify the simulations, it is assumed that the beamforming algorithm performs optimally in the determination of the line-of-sight direction between each mobile station and its serving cell, and on the computation of the set of weights to point a beam in this direction. The SINR must also be computed for each data sub-carrier in the radio resource. Ignoring time, mobile and used SDMA-mode resource indexes and assuming the desired mobile station is assigned a spatial beam with index l and that the serving base station has index i, the SINR for the n^{th} data sub-carrier is given by equations (12-15).

$$SINR_{SDMA}(n) = \frac{P_{Signal}^{i,l}(n)}{P_{Inter}^{i,l}(n) + P_{Intra}^{i,l}(n) + N_0 W_i F_{MS}} \tag{12}$$

$$P_{Inter}^{i,l}(n) = \sum_{j=1}^{N_{Inter}} \frac{G_{MS} \sum_{k=1}^{N_{beams}} \frac{P_{BS_j}}{N_{beams}} \left| H_{BS_j \to MS}^{k}(n) \right|^2 G_{BS_j}^{k}}{PL_{BS_j \to MS} SH_{BS_j \to MS} L_{loss}} \tag{13}$$

$$P_{Intra}^{i,l}(n) = \frac{G_{MS} \sum_{k=1, k \neq l}^{N_{beams}} \frac{P_{BS_i}}{N_{beams}} \left| H_{BS_i \to MS}^{k}(n) \right|^2 G_{BS_i}^{k}}{PL_{BS_i \to MS} SH_{BS_i \to MS} L_{loss}} \tag{14}$$

$$P_{Signal}^{i,l}(n) = \frac{G_{MS} \frac{P_{BS_i}}{N_{beams}} \left| H_{BS_i \to MS}^{l}(n) \right|^2 G_{BS_i}^{l}}{PL_{BS_i \to MS} SH_{BS_i \to MS} L_{loss}} \tag{15}$$

Where:

- $P_{Signal}^{i,l}(n)$ is the desired signal from the serving cell i for the mobile transmitting in beam l.
- $P_{Inter}^{i,l}(n)$ is the contribution from the neighbouring cells to the inter-cell interference in the mobile station assigned the spatial beam l in cell i.
- $P_{Intra}^{i,l}(n)$ is the contribution from the other spatial beams in the same serving cell i in the same resource to which the mobile station is assigned to.
- $\left| H_{BS_i \to MS}^{k}(n) \right|^2$ is the channel gain from the desired mobile station to base station i in the k^{th} spatial beam.
- $G_{BS_i}^{l}$ is the gain of the antenna array at base station i for the l^{th} spatial beam.

- $PL_{BS_j \to MS}$ and $SH_{BS_j \to MS}$ are, respectively, the path loss and shadowing from base station *i* to the desired mobile station
- L_{loss} combine the losses due to cable attenuation, penetration loss, etc.
- $\frac{P_{BS_i}}{N_{beams}}$ is the power assigned to each spatial beam in each resource for base station *i*. It is important to remember that the power is divided uniformly over the set of resources.

The estimated SINR is computed from the vector of SINR values for each sub-carrier, using the Effective Exponential SINR Mapping (EESM) compression method.

8.5 Joint Scheduling and Resource Allocation Algorithm Using SDMA Multiple Access

The proposed SDMA-based DRA algorithm comprises the following steps:

A. Prioritization (scheduling)

The first step is performed by the scheduler, which outputs a list with mobiles sorted by their decreasing order of priority. The utility-based scheduler is used as a reference.

B. Resource allocation

For the next user returned from the priority list, the number of resources required is computed from the amount of information in its buffer and from the MCS scheme returned from the Link Adaptation module. If there are any SDMA-mode resources available, the resource allocator starts with the resource corresponding to the smallest accumulated spatial correlation to all mobiles in the SDMA-zone. In the sequence the algorithm verifies if the mobile can be assigned to the given resource.

The allocation of resources is conducted in such a way that limits intra-beam interference. Assuming each user is assigned one spatial beam, new users are not allowed to transmit in empty spatial beams of the same resource, to which other users were already assigned into, if they might cause intra-beam interference higher enough to degrade the CQI value corresponding to the selected MCS scheme for these users already assigned. In order words, for users which are too close to be spatially separable, the beam would be assigned only to the user with the highest priority. In order to use resources efficiently the mobile has to operate above the MCS SINR threshold to avoid excessive packet errors, which force the predicted BLER to be lower than the threshold. A new mobile may be allocated the same resource as long as the resulting compressed SINR ratio, after beamforming, is kept above the same threshold value associated to the selected MCS scheme from the CQI of the preamble.

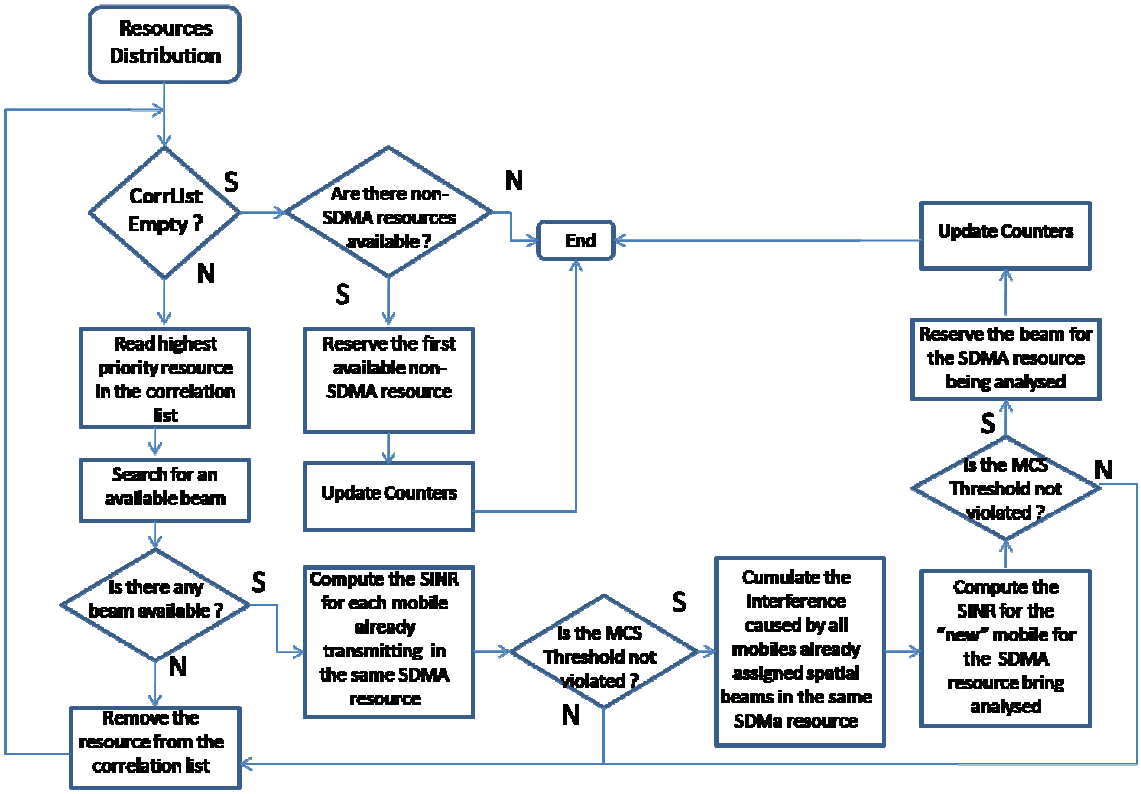


Figure 4 – Fluxogram of the algorithm followed in the atribution of spatial beams in SDMA

The principle behind beam assignment to new users is the following:

1. The resource allocator determines the total amount of resources needed by the next mobile in the priority list.
2. For each resource in the SDMA-zone the resource allocator computes the spatial correlation between its channel matrix and the channel matrix corresponding to each one of the mobiles already allocated in the SDMA-mode resource. The resulted spatial correlations are sorted in increasing order of their value.
3. If there are resources available for allocation in the SDMA zone the resource allocator first tries to assign spatial beams to SDMA-mode resources. Only then it allocates resources in the non-SDMA zone.
4. The process starts with the resource resulting into smaller correlation among channel matrices. For this resource the algorithm computes the new SINR value (with the intra-beam interference) from the new mobile to each one of the already assigned mobiles. If the SINR level of each mobile is degraded to a level lower than the threshold, corresponding to the MCS scheme selected by the mobile already assigned, the process stops and the next resource in the correlation list is checked.
5. If the new mobile can be assigned an empty beam without affecting other already assigned users, the SINR of the new mobile is computed according to the amount of intra-beam interference from the already assigned mobiles into the new one. If the SINR level of the new mobile is degraded to a level lower than the threshold, corresponding to the

MCS scheme selected, the process stops and the next resource in the correlation list is checked. Otherwise, an empty beam is assigned for this mobile in the SDMA-resource. In this way it is assured the quality of the connection for all mobiles transmitting in the same resource: new and already assigned ones.

Cell Layout	Hexagonal Grid, 19 cell sites, 3 sector BSs, 1 cell reuse in a cloverleaf layout with wraparound to simulate interference to edge cells
Number of Users	200 (Traffic Model Used); 72 (Full Queue Used)
Cell radius and Minimum MS to BS distance	900m; 35m minimum distance to the BS
BS Antenna Model for Horizontal Pattern	$A(\theta) = -\min\left[12\left(\frac{\theta}{\theta_{3dB}}\right)^2, A_m\right]$, Antenna Gain: 15dBi
BS Antenna Pattern in non-SDMA zone θ_{dB}, A_m	$\theta_{dB} = 70^{\circ}, A_m = 20dB$
BS Antenna Pattern in SDMA zone θ_{dB}, A_m	$\theta_{dB} = 8{,}75^{\circ}, A_m = 29dB$
BS Antenna bore-sight gain in non-AAS zone	3dBi
BS Antenna bore-sight gain in AAS zone	15dBi
Number of Beams per SDMA resource	4
MS Antenna Gain	Omni-directional with 0dBi
BS Maximum Transmission Power	43dBm
Propagation model	$L = 128.1 + 37.6\log_{10}(R)$, R in Km
Penetration Loss	10dB
Log-Normal and Shadowing Correlation	Standard Deviation = 8dB; 0.5 for sectors of different BSs and 1 for sectors of the same BS
Channel Model	ITU PedB and PedA with 3km/h; ITU VehA with 30km/h
Mobile Station Receiver	MRC
Traffic Models	VoIP and WWW
Duplex Mode	TDD
Operating Frequency	2.5GHz
Bandwidth and FFT size	10MHz; 1024 sub-carriers
Frame Duration	5ms
Number of OFDM Symbols in DL	35 (30 symbols available for data transmission)
Preamble, FCH, DL/UL MAP overhead	5 symbols
Sub-channelization	DL-PUSC
Number of Resources non-SDMA zone	10
Number of Resources SDMA zone	5
Burst Size	10 sub-channels x 6 symbols (15 resources)
Moving average filter length	1.5s
T_{CQI} (CQI feedback delay)	1 frame period in CQICH UL control channel
$T_{ACK,NACK}$ feedback	1 frame period in ACK/NACK control channel
Discard Timer and Priority lengths	15 and 3 frame periods respectively
N_{ret} Maximum Number of Retransmissions	4
Number of HARQ processes per MS	4
BLER Threshold for Link Adaptation	10%
Admission Threshold Parameter	-5dB
Scheduler	Full Queue – Max C/I; VoIP and WWW – Utility

TABLE 1 – SIMULATION SETUP CONFIGURATION

8.6 Results

System level simulations were conducted in the system level simulator for mobile WiMAX. The performance of the SDMA-based DRA was evaluated by incorporating the first proposal for the utility-based packet scheduler, as detailed in chapter 7, with one type of traffic model only.

Simulations were conducted separately for three types of traffic models: Full Queue, Voice over IP (VoIP) and 3GPP's World Wide Web (WWW). The delay bound for VoIP and WWW are respectively 80 ms and 0.5 s. Full queue is used jointly with the simple and opportunistic Maximum C/I algorithm in order to estimate the gains achieved with the SDMA-based DRA, against the system throughput resulting from the non-SDMA-based DRA.

The scenario used for system level simulations was the one of a 4x2 MIMO channel antenna system configuration with Alamouti STBC. Simulations were performed both for the SDMA configuration, with a maximum of 4 beams per SDMA resource, and with the normal mode, without implementation of SDMA. Five resources were configured to be used in the SDMA zone and the remaining 10 resources were configured to be used in the non-SDMA zone.

It was assumed that the direction of arrival from each mobile is perfectly estimated and that the beamforming algorithm is optimal in steering one spatial beam into the direction of the desired mobile. Power is uniformly distributed among the set of beams in each resource and neighbouring cells are assumed as transmitting with maximum power (with full load) in non-SDMA configuration. They are considered only for the generation of inter-cell interference.

Table 1 lists the parameters used in the setup of the scenario used in the system level simulations.

8.6.1 Performance for Full Queue Traffic Users

In order to infer about the system capacity using the SDMA multiple access scheme, system level simulations were conducted for the Full Queue traffic model with the opportunistic Maximum C/I scheduler for a total amount of 72 users. Figure 5 shows the average Over-The-Air (OTA) throughput (peak bit rate), the 3GPP Over-The-Air throughput, the average service throughput and the average offered load, per sector in Mbps. For further details on these performance metrics please refer to Annex C.

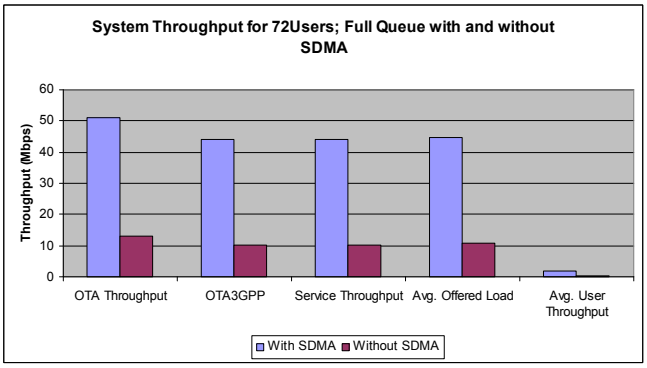


Figure 5 - System throughput: SDMA and non SDMA-based DRA, for full queue and with 72 users load

It is noticed that if SDMA multiple access is implemented into the DRA the OTA throughput for the full queue traffic model can reach up to 50 Mbps, resulting in a throughput gain of roughly 38 Mbps over the simple non SDMA-based DRA scheme. It is worth mentioning that the OTA throughput measures the total amount of bits transmitted over the air interface while the service throughput measures the total amount of successfully transmitted bits over the air interface, both within the same simulation time. The discrepancy between service and OTA throughput, although slightly, is due to users being serviced at the cell edge which will normally experience poor signal quality arising from inter-cell interference from neighbouring cells and also due to intra-cell interference arising from beam assignment in the space domain. It is evident the gain achieved with an SDMA-based DRA architecture in terms of system throughput.

In spite of the significant performance gains achieved these are obtained with the Maximum C/I scheduler, which does not provide QoS requests and for the theoretical full queue traffic model. Therefore, it is to be expected that different figures would result from other types of "more practical" schedulers and traffic models.

8.6.2 Performance for VoIP and WWW Traffic Users

The performance of the SDMA-based DRA was evaluated with the utility-based packet scheduler for both VoIP and WWW traffic models.

8.6.2.1 Voice over IP (VoIP) Traffic Model

Figure 6 plots the user satisfaction ratio for VoIP traffic users using either SDMA or non SDMA based DRA versus the offered load. The user satisfaction ratio is measured from the percentage of packets dropped due to bad channel quality and time-out overflow.

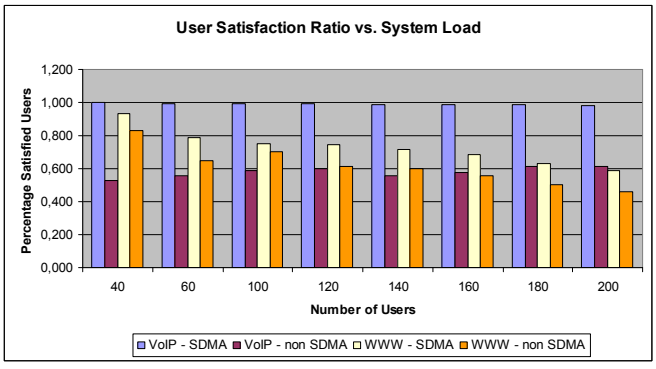


Figure 6 - User satisfaction ratio: SDMA and non-SDMA based DRA for VoIP and WWW traffic models

As can be seen, for the VoIP users the user satisfaction ratio is roughly insensitive to the amount of users in the system as there are virtually no unsatisfied users if the SDMA-based DRA is implemented. This is because the system is relatively under-loaded for VoIP and, as the system load is not enough, additional radio resources cannot improve system performance due to small

packet size and low utilization of radio resources. Nevertheless, one can observe a significant gain in terms of the satisfaction ratio for the non-SDMA based DRA. This is due to the smaller amount of resources available for allocation if the SDMA mode is not implemented.

Figure 7 is the plot of the CDF of the average packet drop rate per user for the highest load used in the simulations (200 users). It can be observed the significant difference in performance between both modes. This has to do with the strategy followed in the allocation of spatial beams in the SDMA-based DRA: users are assigned spatial beams according to the smallest value of the channel matrix correlation and also if the SINR threshold, which corresponds to the selected MCS scheme, is not violated due to the assignment of beams in the same resource.

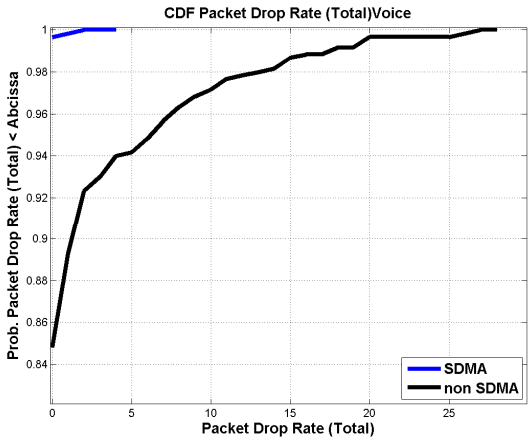


Figure 7 - CDF of average packet drop rate for VoIP users

Figure 8 is the plot of the average packet drop rate per user, averaged over all active users in the network, versus the offered load.

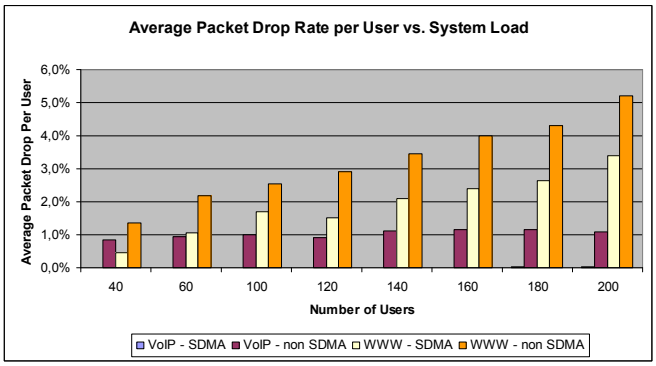


Figure 8 - Average packet drop rate per user for SDMA and non-SDMA based DRA for VoIP and WWW vs number of users

As expected, the ratio of packets dropped due to time-out violation and/or bad channel quality is almost zero for the SDMA-based DRA. The difference in percentage to the non-SDMA mode is also small and remains constant with the increase in the system load, because, as mentioned, the system is under-loaded. The gap between both modes is due to the allocation strategy followed in the SDMA mode resource allocation.

Figure 9 is the plot of the average packet delay per user averaged over all active users in the network versus the offered load. It corroborates this same reasoning as it can be seen that the

average packet delay remains roughly insensitive to the increase in the system load for both SDMA and non-SDMA modes.

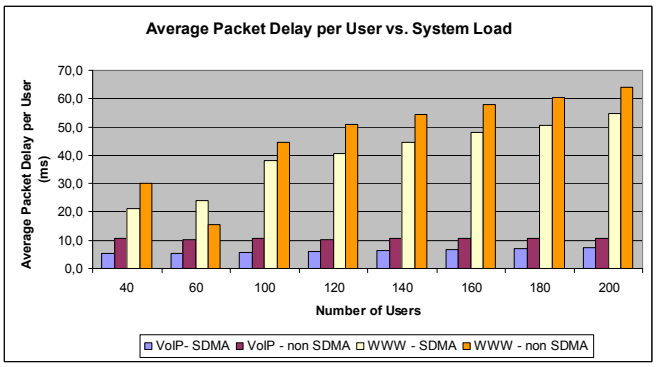


Figure 9 - Average packet delay per user for SDMA and non-SDMA based DRA for VoIP and WWW vs number of users

Figure 10 is the plot of the CDF of the average packet delay per user for the highest load used in the simulations (200 users). As expected, the SDMA-based DRA results in small packet delays per user. But the difference is not that much significant due to: (i) VoIP packets are highly constrained in terms of delay, which means packets cannot wait in buffer for a better transmission opportunity and the utility scheduler prompts them for transmission as the delay approaches the delay bound; (ii) the system is under-loaded.

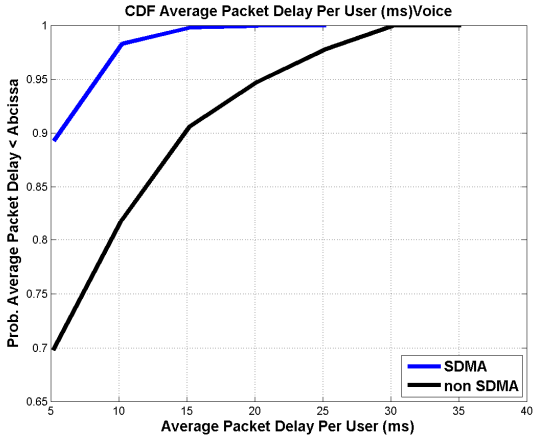


Figure 10 - CDF of the average packet drop rate per VoIP user

Figure 11 is the plot of the average service throughput averaged over all active users in the network versus the offered load.

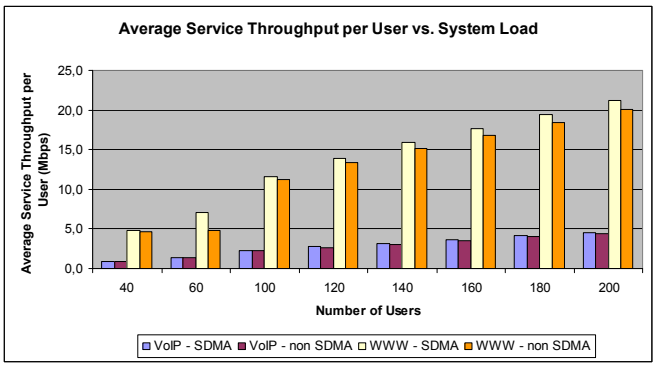


Figure 11 - Average service throughput per user for SDMA and non-SDMA based DRA for VoIP and WWW vs number of users

As expected, the difference in the service throughput is almost zero for both modes and for all traffic loads, although it can be observed a small gap for 180 and 200 users load. This has to do with: (i) under-loaded system; (ii) VoIP is a real time service with stringent delay constraints. Therefore packets cannot remain in buffer waiting for better channel conditions, as they are dropped whenever they achieve their maximum allowable delay. This results in a small gain for the multi-user diversity when using the utility-based scheduler.

Figure 12 shows the average OTA throughput per user averaged over all active users in the network versus the offered load. It can be seen that, for the same load, as the amount of resources increases the resulting multi-user diversity gain is smaller for the SDMA mode. Differently, for the non SDMA based DRA the multi-user gain is higher and this translates into a higher OTA throughput. As expected also, the OTA throughput increases with the amount of active users (system load) in the system.

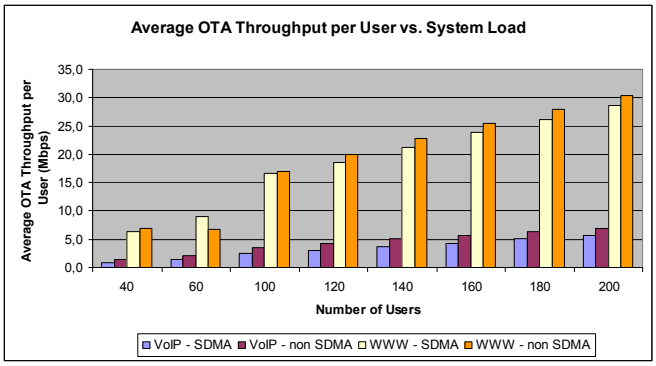


Figure 12 - Average OTA throughput per user for SDMA and non-SDMA based DRA for VoIP and WWW vs number of users

Figure 13 is the plot of the CDF of the average service throughput per user for the highest load used in the simulations (200 users). As can be seen, some users in the edge of the cell have smaller service throughput with the non-based SDMA DRA.

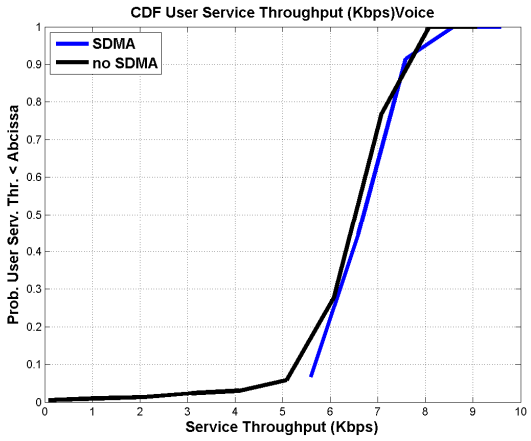


Figure 13 - CDF of the average service throughput per VoIP user

Due to the mechanism followed in the allocation of empty beams in the SDMA-based DRA, users in the edge of the cell have more transmission opportunities and this can be confirmed from the plot in figure 14 of the average service throughput with the geometric factor, for the highest system load used in the simulations (200 users).

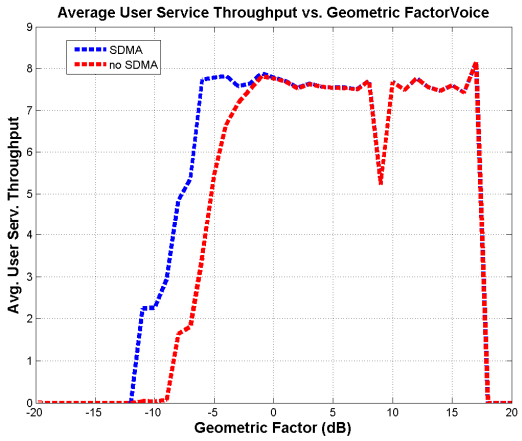


Figure 14 - Average user service throughput versus geometric factor

Figure 15 is the plot of the average SINR per received packet averaged over all active users in the system versus the offered load. As expected, the average SINR is significantly better for the SDMA mode and it is roughly insensitive to the increase in the number of users in the system. This is due to the scheme followed by the utility based scheduler in the elaboration of the priority lists.

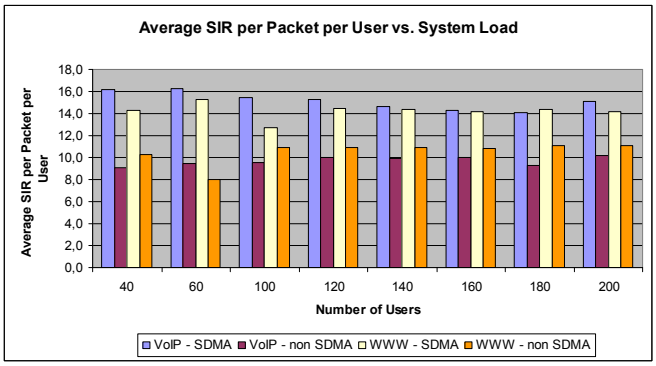


Figure 15 - Average SINR per packet per user for SDMA and non-SDMA based DRA for VoIP and WWW vs number of users

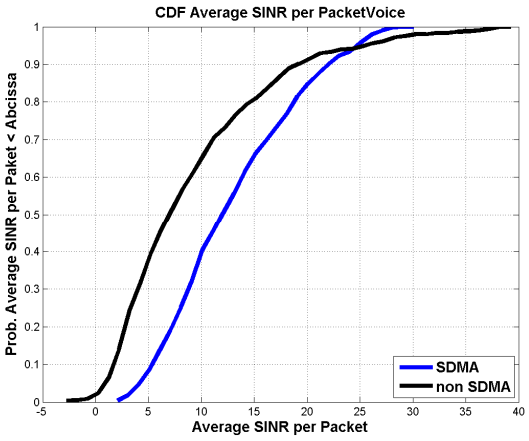


Figure 16 - CDF of user average SINR per packet for VoIP users

The plot in figure 16 is the CDF of the average SINR per received packet averaged over all active users in the system, for the highest load used in the simulations (200 users). The plot resulting from the non-SDMA mode has a heavy tail in the highest SINR zone. This tail translates to a higher maximum SINR for the non-SDMA mode. This is because a DRA with no SDMA mode is inherently more loaded than if SDMA is implemented, which results in a roughly higher gain of multi-user diversity which can pick users with better channel conditions for transmission. Also, the non SDMA mode uses a higher transmission power per resource, which is lower for the SDMA mode.

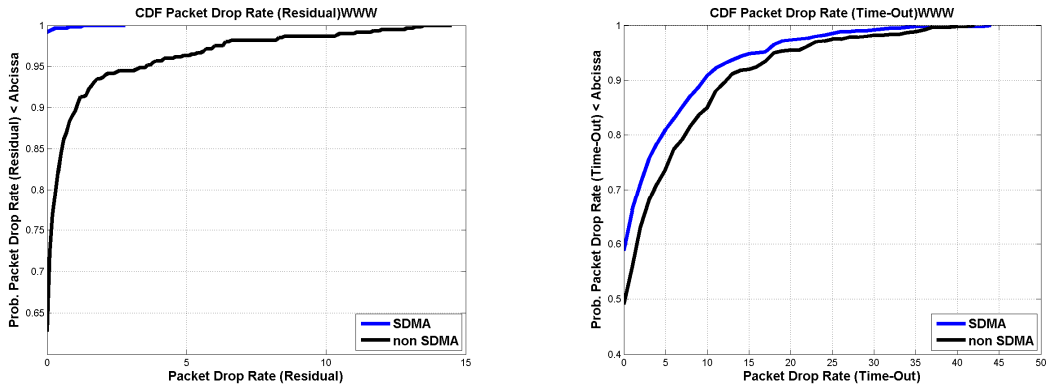
8.6.2.2 World Wide Web (WWW) Traffic Model

Packets generated according to the WWW traffic model are of much larger size than VoIP ones. Therefore, for the same amount of users, the system is more loaded if WWW traffic model is used rather than VoIP. The higher load results in a significant higher multi-user diversity gain. Also WWW traffic model is much more tolerant in terms of packet delay than VoIP. This means that packets can remain for longer periods of time in buffer waiting for better channel conditions and performing more transmission attempts. Therefore, resources are used more efficiently with WWW traffic model. For the WWW traffic model the system can be assumed as having enough load. With the increase in the load there is a corresponding decrease in the amount of satisfied users because more packets are dropped due to bad channel quality and/or delay overflow.

Figure 6 plots the user satisfaction ratio for SDMA and non SDMA-based DRA for WWW traffic model versus the offered load. As expected, the SDMA-based DRA results into a higher percentage of satisfied users, compared to the non-SDMA based DRA. The reasons behind this performance gap between two modes are: (i) SDMA results in more resources available (in spatial domain); (ii) the algorithm followed in the allocation of spatial beams in the SDMA-based scheduler; (iii) WWW packets have to fragmented in order to fill radio resources, which

translates into a small number of users transmitting over spatial beams in the SDMA zone of the map of resources per frame period and, as a consequence, this translates to an improvement in channel quality because of the inherent smaller intra-cell interference.

Figure 17 plots the CDF of the average packet drop rate, residual and time-out respectively, for both SDMA and non SDMA modes, with the WWW traffic model and the highest load assumed in the simulations (200 users). It is evident from these two plots that the main contribution to the difference in the average packet drop rate for the two schedulers is due to the amount of residual packets, i.e., packets which are dropped when the maximum number of transmission attempts is reached, which is significantly higher for the non SDMA-based scheduler. SDMA mode achieves much better quality transmissions according to the reasoning above.



a **B**

Figure 17 - CDF of the average packet drop rate for (a) residual (b) time-out for WWW users

Figure 9 is the plot of the average packet delay per user, averaged over all active users in the network, versus the offered load, for the WWW traffic model. As expected, there is an increase in the average packet delay with the amount of users in the network: as the SDMA-based scheduler provides more resources for allocation, with the SDMA-based DRA the average packet delay is smaller for all values of the load.

Figure 18 plots the CDF of the average delay per packet for all active users in the network and for the highest system load of 200 users used in the simulations.

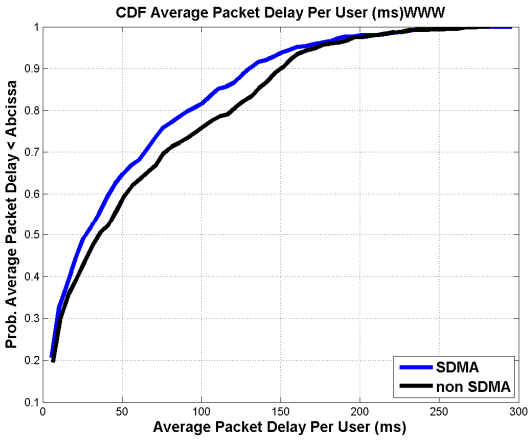


Figure 18 - CDF of the average packet delay for WWW users

As can be seen from the plot in figure 11, representing the average service throughput averaged over all active users in the system, the SDMA-based DRA results into higher service throughput for all system loads assumed in the simulations. The main contribution for this difference arises from the amount of transmission opportunities for users in the edge of the cell, which are allocated empty beams with enough quality for transmission (limiting the intra-beam interference) in the SDMA-based scheduler. This can be seen from the plots in figures 19 and 19 of the CDF of the user average service throughput and of the variation of the average service throughput with the geometric factor, respectively, for the maximum load assumed in the simulations. The difference in performance is not so significant because the size of the transport block with the most robust MCS scheme is very low and users in the edge of the cell (with low geometric factors) are assigned the most robust MCS scheme for transmission.

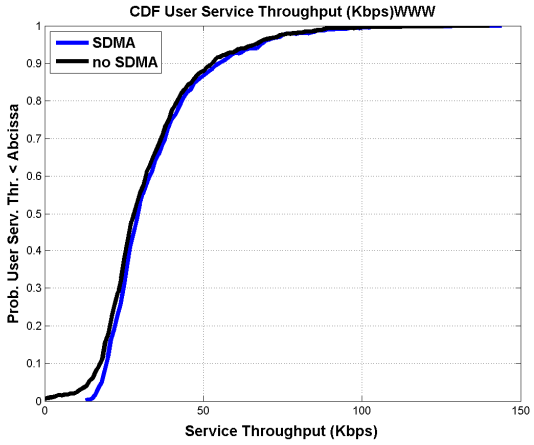


Figure 19 - CDF of the average user service throughput for WWW users

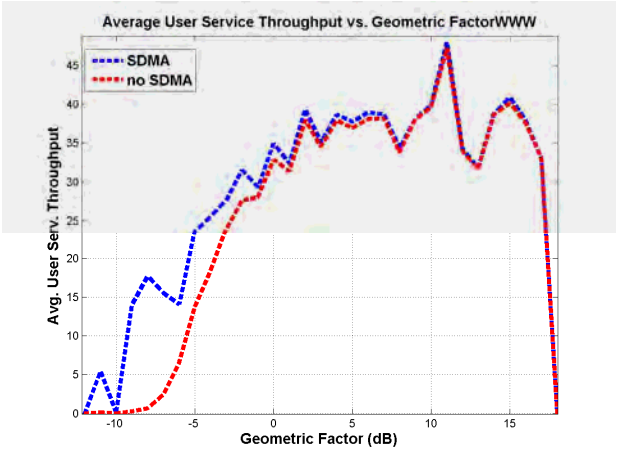


Figure 20 - Average user service throughput versus the geometric factor for WWW users

As can be seen from figure 12 the average OTA throughput increases with the increase in the amount of active users in the cell. This is the result of the multi-user diversity gain arising from the opportunistic channel access. This effect is evident with the WWW traffic model because

WWW packets are more tolerant to delay than VoIP packets. Together with the shape of the utility function used in the utility scheduler, it is a consequence of the fact that under these conditions the scheduler behaves much like the opportunistic maximum C/I scheduler. Another reason for this behavior is that the offered load (in bits) is more significant for WWW packets, which are much larger in size than the packets from VoIP, and this results in a performance enhancement by creation of more radio resources in the SDMA zone.

As can be seen in figure 15, due to the strategy followed in the allocation of empty beams in the SDMA-based DRA, the SINR of packets received correctly in the mobile is much better than without SDMA allocation. This is the same result which was verified with the VoIP traffic model and the same reasoning behind the increase in the service throughput with the amount of users in the system applies.

Figure 21 is the plot of the CDF of the average SINR per received packet for both SDMA modes and for the maximum load assumed in system level simulations.

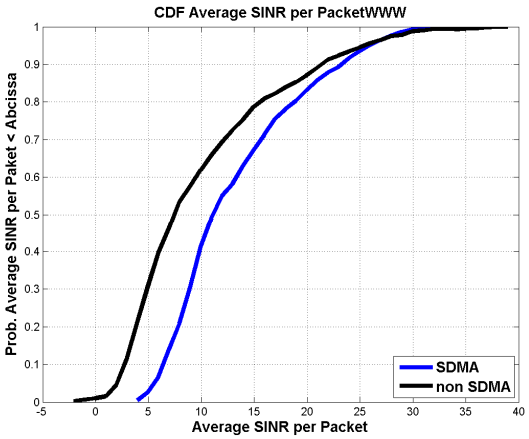


Figure 21 - CDF of the average SINR per packet for WWW users

8.7 Related Work

There are few proposals in the research literature regarding the implementation of SDMA in WiMAX networks, in particular for the Mobile WiMAX standard.

In [174], a multimedia SDMA/TDMA scheduling algorithm is proposed. The algorithm consists of a packet scheduler and a packet allocator. The scheduler prioritizes packets from different users according to the requested QoS, defined in terms of the minimum SINR and time-out values. The allocation is performed for the time-slot which results in the minimum degradation of the SINR, achieved by already allocated users in each time-slot and according to the user's required SINR. The authors claim that their algorithm results in the satisfaction of OoS and in the maximization of the cell's throughput.

In [175-176] an exhaustive analysis, regarding the support of SDMA techniques in the MAC layer of WiMAX standard is presented together with validation results. To the best of the author's knowledge [173] is the only work, up to now, in the literature regarding the proposal of

a practical SDMA algorithm implemented in the Mobile WiMAX system. System level simulations are conducted to reveal the gains resulting from the proposed dynamic resource allocation with SDMA.

8.8 Conclusion

This chapter details all the steps performed in the design and performance analysis of a new DRA architecture based on the addition of space domain as a new degree of freedom for resource allocation in Mobile WiMAX networks.

In the PHY layer radio resources are available in time, frequency and space domains by the joint implementation of OFDMA and SDMA over the same air interface. The MAC layer assigns spatial beams as long as new users do not affect already assigned users in the same set of symbols and frequency channels. SDMA is made possible in Mobile WiMAX with the use of beamforming antennas which is standardized as AAS in the IEEE 802.16e standard for Mobile WiMAX networks. The original utility-based packet scheduler is used in conjunction with the proposed DRA.

A set of system level simulations was performed to infer on the gains achieved over simple scenarios where a non SDMA-based DRA architecture is implemented. The gain in capacity which results from the implementation of a SDMA-based DRA architecture over the non SDMA one is less pronounced whenever traffic models are used in detriment of the theoretical full queue traffic model. Traffic models with a pronounced activity factor and large packet sizes increase the SDMA cell throughput, since users with the highest channel propagation conditions would be continuously served with the highest MCS scheme, avoiding under utilization of resources. This is the reason for the observed differences in the results obtained for WWW and VoIP. Improved results can be attained by using greater density of users to improve the multi-user diversity gain and a priority-based scheduler to control the trade-off between maximization of cell throughput and keeping the number of satisfied users at an acceptable level.

Full queue together with the Maximum C/I algorithm is used in the estimation of the gain in capacity achieved with the SDMA-based DRA. The OTA throughput can reach up to 50 Mbps, a gain of roughly 38 Mbps over the simple non SDMA-based DRA scheme.

For VoIP users the user satisfaction ratio is roughly insensitive to the amount of users in the system. As a matter of fact, the small packet size and low utilization of radio resources from VoIP traffic demands a great amount of users to be simulated, in order to produce figures which could show the potential gain in system capacity from the implementation of an SDMA-based DRA for VoIP. For the same reasoning the average packet delay remains roughly insensitive to the increase in the system load for both SDMA and non-SDMA modes.

In order to have an idea of the potential gain of SDMA in such scenario, simulations were conducted for a load equal to 200 users. Average packet drop rate per user was produced from

these simulations from which it was observed the significant difference in performance between both modes. This is because the implemented algorithm for SDMA attributes spatial beams to already assigned slots in the frame without compromising the channel quality sensed by those already assigned users, in such a way that the selected MCS scheme is not violated due to the assignment of beams in the same resource.

For higher loads SDMA-based DRA results in small packet delays per user. But the difference is not that much significant due to the fact that VoIP packets are highly constrained in terms of delay, which means packets cannot wait in buffer for a better transmission opportunity and the utility scheduler prompts them for transmission as the delay approaches the delay bound. Therefore packets cannot remain in buffer waiting for better channel conditions, as they are dropped whenever they achieve their maximum allowable delay. This results in a small gain for the multi-user diversity when using the utility-based scheduler.

Due to the mechanism followed in the allocation of empty beams in the SDMA-based DRA, users in the edge of the cell have more transmission opportunities. The average SINR is significantly better for the SDMA mode and it is roughly insensitive to the increase in the number of users in the system. This is due to the scheme followed by the utility based scheduler in the elaboration of the priority lists.

Packets generated according to the WWW traffic model are of much larger size than VoIP ones. Therefore, for the same amount of users, the system is more loaded if WWW traffic model is used rather than VoIP. WWW traffic model is much more tolerant in packet delay than VoIP and packets can remain for longer periods of time in buffer waiting for better channel conditions and performing more transmission attempts. Higher load brings and less constringent packet delays brings in higher gains in system capacity due to multi-user diversity gain and, therefore, resources are used more efficiently. With the increase in the load there is a corresponding decrease in the amount of satisfied users because more packets are dropped due to bad channel quality and/or delay overflow. But the decrease in user satisfaction ratio is significantly smaller for the SDMA-based DRA.

It was observed that, for WWW services, which have large packet sizes and delay tolerances, the performance enhancement is contributed by creating more available radio resources in the SDMA zone. However, in the VoIP case, if the system load is not enough, additional radio resources cannot improve system performance due to small packet size and low utilization of radio resources. SDMA was proved to be an efficient means for increasing system capacity whilst still providing QoS requirements from user's applications. The advantages resulting from the implementation of such architecture are higher for higher system loads and/or for traffic models with higher inherent load in packet generation. Last but not least, it is worth mentioning that the architecture implemented with the DRA based on the SDMA access scheme is also innovative and no similar approach was conveyed in the research literature available up to now.

Chapter 9

Joint Time and Frequency Domains Packet Scheduler for Mobile WiMAX

9.1 Introduction

IEEE802.16e standard for Mobile WiMAX system is based on the Orthogonal Frequency Division Multiple Access (OFDMA) scheme and allows sub-channelization in both uplink and downlink connections. For each symbol in the Time Division Duplexing (TDD) frame, different sub-channels may be allocated to different users according to OFDMA, and sub-channels may be constituent using either contiguous sub-carriers or sub-carriers pseudo-randomly distributed across the frequency spectrum corresponding to each OFDM symbol:

· Sub-channels formed using distributed sub-carriers provide more frequency diversity and this is useful for mobile applications. As described in chapter 3 the two representative

channelization modes for sub-carrier pseudo-random distribution are Partial Usage Sub-Channelization (PUSC) and Full Usage Sub-Channelization (FUSC). For these modes of channelization channel quality is estimated from the OFDM symbol conveyed in the preamble of the TDD frame.

· Sub-channels formed using distributed sub-carriers provide more frequency diversity and this is useful for mobile applications. As described in chapter 3, the sub-channelization scheme based on contiguous sub-carriers in Mobile WiMAX is called band adaptive modulation and coding (AMC).

Although frequency diversity is lost with AMC, it allows the exploitation of multi-user diversity over frequency domain, by allocating sub-channels to users based on their frequency response. This multi-user diversity can provide significant gains in overall system capacity if each sub-channel is opportunistically assigned to the user resulting in the highest gain, and therefore using the most efficient modulation and coding scheme. As the frequency diversity is lost with AMC this mode of channelization is more appropriate for fixed and low-mobility applications and/or with the use of Adaptive Antenna Systems (AAS) because these scenarios are associated to better SINR ratios. AMC sub-channelization provides another diversity gain, over frequency, which can improve system capacity.

Until now the proposed versions of the utility-based packet scheduler and the SDMA-based DRA addressed only the downlink PUSC sub-channelization scheme. Although radio resources in different segments can result in different channel states it is not possible to exploit multi-user diversity in the frequency domain with PUSC sub-channelization mode.

Past simulations considered the PUSC scheme in downlink because it is the mandatory scheme of sub-channelization in the system profile available from the WiMAX Forum. This chapter extends the resource definition in the map of resources of the Mobile WiMAX system to include the Adjacent Multi-Carrier (AMC) sub-channelization scheme. Due to the multi-user diversity gain inherent in this scheme it is to be expected an increase in system capacity associated to this type of DRA architecture.

While in PUSC resources are assigned to users no matter the state of each individual sub-channel in the frame, in AMC the state of each sub-channel is considered in the computation of the scheduling metric, as long as users are assigned to their sub-channels with better state. In order to do so, the basic DRA architecture is modified for the inclusion of the map of resources resulting from the implementation of the AMC sub-channelization scheme. Therefore, a different DRA is proposed which is based on the definition of a map of resources according to the AMC sub-channelization mode. AMC sub-channelization mode requires a modification of the utility-based packet scheduler because now there are gains resulting from two distinct domains, time and frequency, which must be considered both in the scheduler as well as in the resource allocator.

This chapter is organized as follows. Section 2 is about the implementation of the map of resources with AMC sub-channelization into the original DRA architecture for Mobile WiMAX system level simulations. With AMC there is another degree of freedom for resource allocation, namely the frequency domain besides time domain. It is important to observe that space domain allocation is not implemented here as it would result in a high complexity system implementation and, above all, simulation execution. Section 3 introduces all the steps followed in the implementation of the new time-frequency domain packet scheduler to be plugged into the proposed DRA architecture for Mobile WiMAX. The scheduler encompasses four different steps in the computation of the list of users selected for transmission: (i) ccomputation of the list of active users eligible for scheduling; (ii) scheduling in time domain by means of the utility-based packet scheduling algorithm; (iii) scheduling in frequency domain according to a scheduling algorithm which takes into account the combined gain of the set of sub-channels, and (iv) combine the outputs from both schedulers in time and frequency. Section 6 presents results from the performance evaluation of the proposed scheduler. Both VoIP and WWW traffic models were used in performance evaluation, for two types of channels: SISO channel with ITU PedB and a 3 Km/h mobile speed, and 2x2 MIMO channel with STBC Alamouti coding also with a 3 Km/h mobile speed. Section 7 is about the related work available in literature. Section 8 concludes the chapter.

9.2 Resource Space for Time-Frequency Domain Scheduler

With band AMC each burst can be allocated to a resource spanning 64 data sub-carriers in frequency domain and 21 OFDM symbols in time domain. In AMC mode pilot and data sub-carriers are organized in bins and each bin comprises 8 data sub-carriers and 1 pilot sub-carrier, which amounts to 12 resources available in the map of resources. Assuming each slot is made-up of 2 bins per 3 OFDM symbols there are in total 7*(8/2) = 28 slots per resource.

Differently from the PUSC mode, in which channel quality estimation is derived from the symbol conveyed in the frame's preamble, in band AMC channel quality must be estimated from the pilot sub-carriers for the bins assigned to each radio resource in the frame. In the simulations the quality of each resource is estimated from the 8 pilot sub-carriers, corresponding to each one of the 8 bins comprising each resource. As the pilots are activated only for those resources which are conveying information, the Channel Quality Indication (CQI) associated to each burst mapped into a resource will depend on the amount of load in the system: a system close to full load will result in more interfered bins and, therefore, in the need to use a more robust MCS scheme; a system under-loaded will result in bursts less interfered and in more efficient MCS schemes. The amount of interference will depend on the position of the mobile in the cell and this negative effect will be somehow compensated by the opportunistic nature of

allocating resources with better SINRs for each user. Figure 1 depicts the map of resources for the band AMC-based DRA.

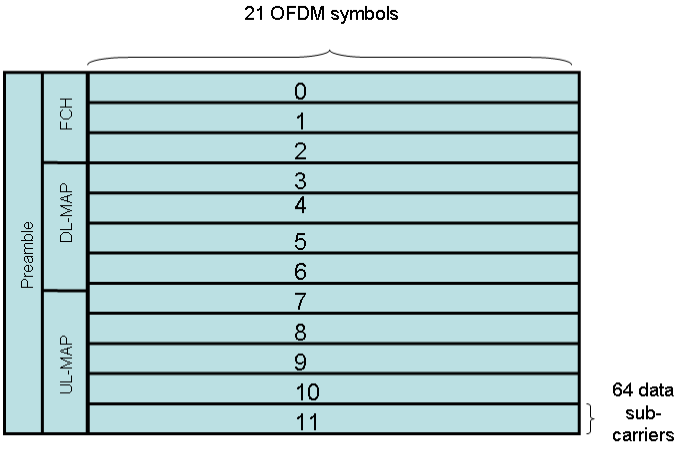


Figure 1 - Average OTA throughput per user for SDMA and non-SDMA based DRA for VoIP and WWW vs number of users

9.3 Proposed Scheduler

The proposed time-frequency packet scheduler encompasses 4 steps:

1. Computation of the list of active users eligible for scheduling. Users are inserted into the active list if they comply with any one of the following constraints: (i) the user has an active HARQ process waiting for transmission; (ii) the user has new packets waiting in the buffer.
2. Perform scheduling in time domain by means of the utility-based packet scheduling algorithm. Only new packets in the user's buffer are considered. The output from step (1) is a list of priorities for the users in the active list. Only QoS requirements, in terms of the maximum allowable packet delay, are considered. Channel quality is not included in the computation of the user's priority, as it would depend on the resource being considered.
3. Perform scheduling in frequency domain. A priority metric is computed for each resource in the map according to a scheduling algorithm which takes into account the combined gain of the set of sub-channels into which each resource is mapped. Different algorithms can be considered such as: the Maximum C/I (CI), Proportional Fairness (PF) or the Maximum C/I over the Average C/I (AvgCI).
4. Combine the outputs from both schedulers, in time and frequency, and compose a list of priorities, one for each resource in the map of resources. The user with the highest priority is selected for transmission.

9.3.1 Time Domain Packet Scheduler

The scheduling algorithm used in time domain is a small modification of the basic utility-based packet scheduler elaborated in chapter 7, as the algorithm does not consider the channel quality

in the estimation of the utility to be transferred. This is because the computation of the user's transferred utility according to the channel state would depend on the position of the resource in the map of resources. And at this stage one does not know in advance in which resource to map the user's packets. One possible solution would be to perform the scheduling for each resource independently. But this would result in an increase in the simulation complexity and execution time. In order to circumvent this all packets in the user's buffer are used in the computation of the transferred utility. If the user does not have any channel with quality good enough to perform data transmission the scheduler in frequency domain will reflect this fact by forbidding channel access to the user, no matter its position in the list of priorities outputted from the time-domain scheduler.

The steps followed in the time-domain scheduler are the following:

1. At the beginning of frame period *n* compute the total amount of potential utility $U_p(n)$ in the cell. This is given by equation (1):

$$U_p(n) = \sum_{i=1}^{N} \sum_{l=0}^{L_i} U_i(\tau_i^{(l)}) \quad (1)$$

Where $U_i(\tau_i^{(l)})$ is the utility of the l^{th} packet with delay $\tau_i^{(l)}$ in the buffer of the i^{th} user, assuming there are L_i packets in the buffer of user *i* and a total of *N* active users in the cell.

2. For a given user, $j \in \{1,...,N\}$, compute the amount of utility that will be transferred to the network if the user is scheduled for transmission. All packets stored in the user's buffer are considered in this computation. This is given by equation (2):

$$U_j^T(n) = U_j(\mathbf{Q}_j(n)) = \sum_{k=1}^{L_j} U(\tau_j^{(k)}) \quad (2)$$

Where $\mathbf{Q}_j(n) = \{\tau_j^{(0)}, \tau_j^{(1)}, \cdots, \tau_j^{(L_j-1)}\}$ is a vector representing the delay of each packet from the buffer of user *j*. $\tau_j^{(0)}$ is the delay of the Head of Line (HOL) packet in the user's *j* buffer.

3. Compute the average of the utility already transferred from user $j \in \{1,...,N\}$, according to equation (3)

$$\overline{U_j^T}(n) = \lambda.\overline{U_j^T}(n-1) + (1-\lambda).U_j^T(n) \quad (3)$$

Where:

- $\overline{U_j^T}(n)$ is the updated value of the average utility transferred from user *j* in frame period *n*. It stores the state of the previous transmissions from this user.
- λ is the forgetting factor. It should be longer than the coherence time in order to average out the influence of fast fading in the radio channel.

4. Estimate the potential utility that will remain for frame period *n*+1 if user *j* is selected for transmission in frame period *n*. This is given by equation (4):

$$U_p(n+1 \mid j) = \sum_{i=1, i\neq j}^{N} \sum_{l=0}^{L_i} U_i(\tau_i^{(l)}) \tag{4}$$

5. Compute the priority metric for each user. The selected user is the one which results in the maximization of the difference between the transferred utility, $U_j^T(n)$, and the loss, $Loss_j(n)$, of utility incurred in this selection, as given by equation (5):

$$M_i(n) = \frac{U_i^T(n)}{\overline{U_i^T}} - Loss_i(n), \quad i \in \{1,...,N\} \tag{5}$$

Where: $Loss_i(n) = U_p(n) - U_p(n+1 \mid i) - U_i^T(\mathbf{Q}_i(n))$.

5. Normalize the priority metric $M_i(n)$ for each user $i \in \{1,...,N\}$, according to equation (6):

$$\begin{aligned} &\text{if } M_i(n) \geq 0 \quad M_i^{norm}(n) = M_i(n) / \max(M_i(n)) \\ &\text{else} \qquad\quad M_i^{norm}(n) = M_i(n) / \min(M_i(n)) \end{aligned} \tag{6}$$

The purpose of this normalization is to force the final priority to be in the range: $[-1,1]$.

9.3.2 Frequency Domain Packet Scheduler

The scheduling algorithm used in frequency domain is the Proportional Fairness (PF), although other scheduling algorithms could be considered as well. The PF metric is computed for each resource in the map of resources, with channel quality strong enough to guarantee the transmission with a high probability of success. To this effect an *admission threshold* parameter is considered:

- If the channel quality is not good enough to support transmission, with at least the most robust MCS scheme, a new value for the CQI is computed, by assuming the most robust MCS scheme is used in the transmission.
- If the new value of the CQI is equal to or greater than the Admission Threshold the PF metric is computed for the given resource and user.
- Otherwise the user is not considered in the list of priorities for this specific resource.

This strategy is followed in order to avoid transmissions with a CQI level which could result in transmission errors with high probability and, therefore, would decrease resource utilization efficiency.

The PF scheduler computes the priority metric for each user and for each resource in the map of resources according to equations (7-9).

$$F_i(n) = \frac{DRC_i^k(n)}{T_i^{avg}(n)}, \quad i \in \{1,...,N\}; k = \{1,...,K\} \tag{7}$$

$$DRC_i^k(n) = \frac{N_i^k}{T_{frame}}(1 - BLER_i^k) \tag{8}$$

$$T_i^{avg}(n+1) = \lambda.T_i^{avg}(n) + (1-\lambda).R_i(n) \quad (9)$$

Where:

- $DRC_i^k(n)$ is the estimated data rate for user *i* on resource *k*.
- N_i^k is the size of the transport block that can be transmitted in resource *k*. It depends on the MCS scheme used.
- $BLER_i^k$ is the estimated Block Error Rate for user *i* on resource *k*.
- T_{frame} is the duration of the radio frame.
- *N* is the number of active users in the active set.
- *K* is the number of resources in the map of resources.
- T_i^{avg} is the average throughput from user i.
- $R_i(n)$ is the amount of information transmitted by user *i* in previous transmission cycle over all resources assigned to it.
- λ is the filter length. It must be higher than the coherence time associated to the Doppler frequency and lower than the period of time corresponding to the de-correlation length. It was defined as 1.5 seconds.

9.3.3 Computation of Final Priority

It is important to mention that resources are first assigned to HARQ processes which are active and waiting for a new transmission opportunity. This is because the same resource must be assigned to the HARQ process for re-transmission, in order to be consistent with the MCS scheme used in the first transmission attempt. Only those resources in the map of resources which remain free for allocation are considered in the computation of the final priority.

In order to decide which user should transmit in each resource a list of priorities must be computed for each one. The user with highest priority is assigned the respective resource for transmission, according to equation (10).

$$u^k(n) = \arg\max_{i\in\{1,\ldots,N\}} \left(M_i.(n)F_i^k(n)\right), \quad k = 1,\ldots,K \quad (10)$$

Where:

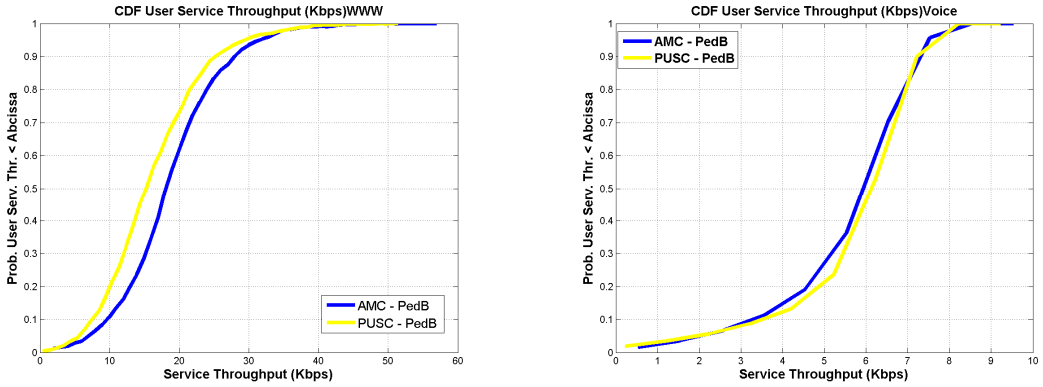
- $M_i(n)$ is the priority returned from the time domain scheduler.
- $F_i^k(n)$ is the priority for resource $k \in \{1,\ldots,K\}$ returned from the frequency domain scheduler.

9.4 Results

This section presents the results obtained from system level simulations conducted for both AMC and PUSC sub-channelization modes with the utility-based packet scheduling algorithm.

All simulations were conducted for a total amount of 200 active users in the network and for two types of traffic models: WWW and VoIP. Simulations were also performed for two types of channels: SISO channel, with ITU PedB 3 Km/h mobile speed and 2x2 MIMO channel, with STBC Alamouti coding also with 3 Km/h mobile speed. The *admission threshold* was set to -5 dB.

Figure 2 is the plot of the CDF of the average service throughput per user for WWW and VoIP traffic models respectively and for both types of sub-channelization modes.



a **B**

Figure 2 - CDF of user service throughput (a) WWW users; (b) VoIP users

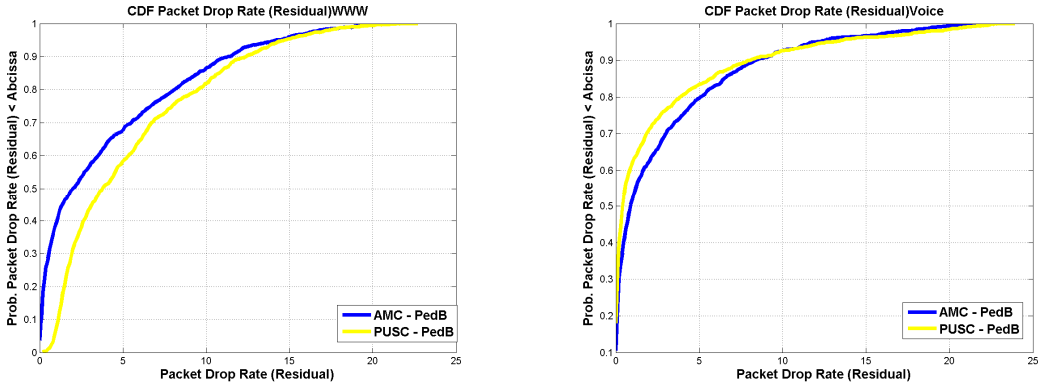
As can be seen from both plots the AMC sub-channelization mode is more efficient for WWW users regarding the achieved service throughput. For VoIP users both sub-channelization modes are very similar in the resulting service throughput per user, although the PUSC sub-channelization mode results in a better performance. This is because the system is overloaded for WWW users and under-loaded for VoIP ones.

As the system is overloaded for WWW, the AMC sub-channelization mode is more efficient in the utilization of the available radio resources. This efficiency results from the multi-user diversity gain over frequency, associated to this sub-channelization mode. We could expect to see an even higher gain in system capacity with the use of the full queue traffic model in conjunction with an opportunistic packet scheduler, such as the maximum C/I, which does not take into account QoS requirements, as the utility algorithm does. As the system is under-loaded for VoIP users there is no resulting multi-user diversity gain over frequency, even with the AMC sub-channelization mode and both sub-channelization modes result into similar behaviour.

It is important to mention that the AMC sub-channelization mode is more sensitive to errors in channel quality reporting and also to inter-cell interference, as there is no randomization in the interference from neighbouring cells over the sub-carriers composing each channel, because sub-channels are not distributed along frequency, in accordance to a random pattern. Therefore, it is to be expected that in an under-loaded scenario where all users are served, even those users

with bad channel quality (as long as the admission threshold parameter condition is satisfied), the PUSC sub-channelization mode results in a better performance than AMC one.

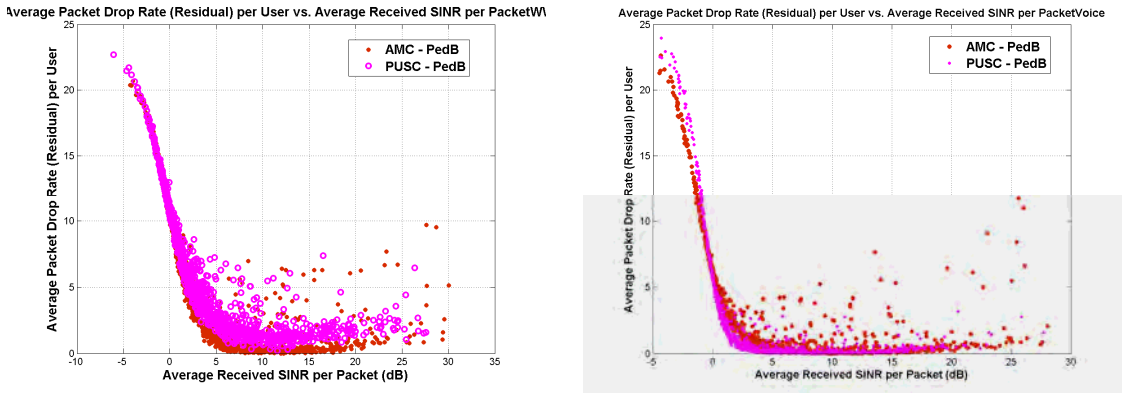
As can be seen in the plot of the CDF of the average packet drop residual rate per user in figure 3, VoIP users present better performance for the PUSC sub-channelization mode over the AMC one. Therefore, the amount of packets dropped due to bad channel quality is smaller for the PUSC sub-channelization mode. This is because the system is under-loaded and PUSC results in better quality over transmission as mentioned above.



a **b**

Figure 3 - CDF of average packet drop rate per user (a) WWW users; (b) VoIP users

Figure 4 is the plot of the average packet drop rate versus the average SINR per packet, for all packets correctly received in the set of mobiles scheduled for transmission. As can be seen, for VoIP and for the set of SINR levels which result in one of the available MCS schemes in the link adaptation process, the PUSC sub-channelization mode presents better performance over AMC. Nevertheless, for SINR levels between 0 dB and -5 dB the AMC results in better performance with a smaller number of packets dropped due to bad channel quality. This is because the AMC mode is more effective for this range of SINR values achieved with the most robust MCS scheme, due to the contiguous allocation of sub-carriers over each sub-channel in each radio resource.



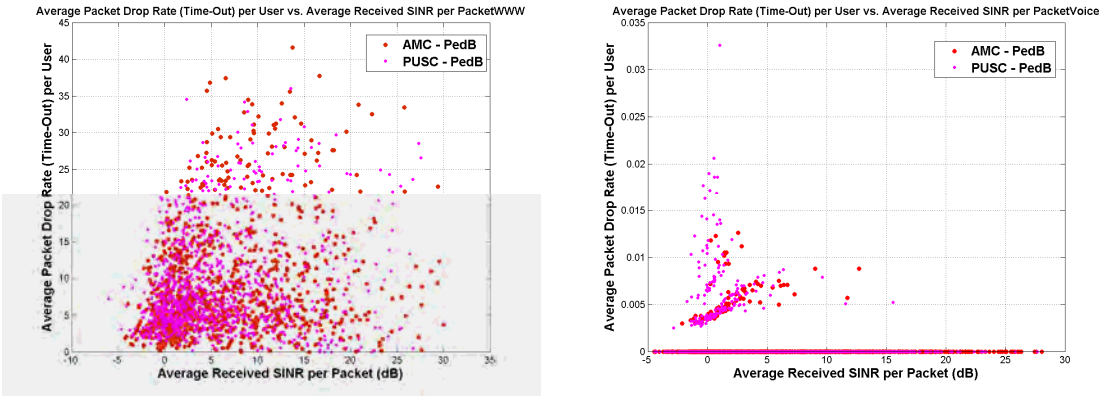
a **b**

Figure 4 - Average packet drop rate per user vs. average received SINR per packet (a) WWW users; (b) VoIP users

Also, a smaller number of users transmit due to the opportunistic nature of the utility-based scheduler for the WWW traffic model.

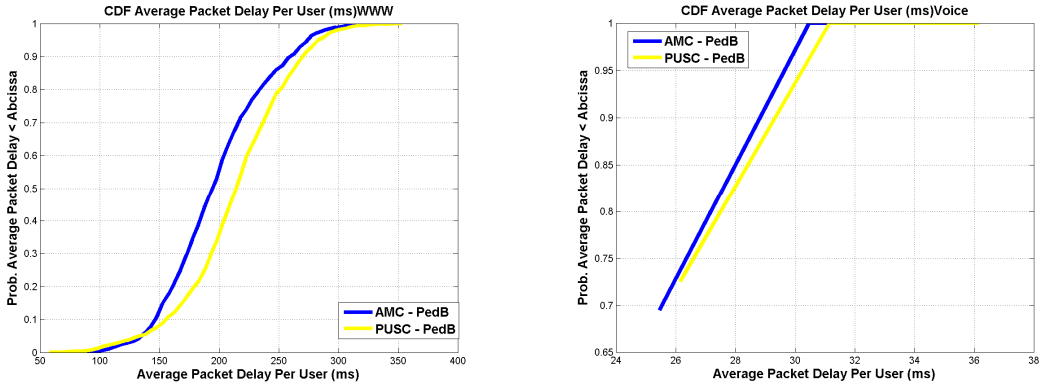
For WWW users, the fact that the system is overloaded and the multi-user diversity gain over frequency from AMC sub-channelization mode, result in a smaller amount of packets dropped due to bad channel quality if AMC sub-channelization mode is used in detriment of the PUSC.

As can be seen from figure 5 there are almost no packets dropped due to time-out violation for VoIP users, in both sub-channelization modes. The few drops which occur are for those users with bad channel quality, in the range of the admission threshold CQI level. This is because the system is under-loaded as already mentioned. It can also be seen from this plot that the overloaded system for WWW result in users with a large fraction of packets dropped due to time-out violation. Therefore it is to be expected a large fraction of users unsatisfied with the service provided under these conditions. Nevertheless the objective pursued with this simulation was to compare the gain achieved in terms of system capacity with both sub-channelization modes, as the utility-based scheduler provides the required QoS in terms of maximum packet delay satisfaction.



a **b**

Figure 5 - Average packet drop rate per user vs. average received SINR per packet (a) WWW users (b) VoIP users



a **b**

Figure 6 - CDF of average packet delay per user (a) WWW users; (b) VoIP users

As can be seen from the plot of the CDF of the average packet delay per user in figure 6 the AMC sub-channelization mode presents the better performance for both types of users: WWW and VoIP.

Figure 7 is the plot of the CDF of the average packet drop rate per time-out violation for WWW users. As can be seen both sub-channelization modes result in the same level of performance. This means that the gain in the performance achieved with AMC sub-channelization mode is due to the multi-use diversity gain over frequency which results into a higher amount of packets being transmitted in the system, empting user's buffer and improving transmission success probability for users with bad channel quality. Also, most of the packets are dropped due to the violation of the maximum number of transmission attempts allowed. This is because the system is overloaded for the WWW users and, therefore, most packets are dropped due to interference from users transmitting in the same resource with other spatial beams. The multi-use gain over frequency results in an increase of the available system capacity for AMC over PUSC sub-channelization schemes.

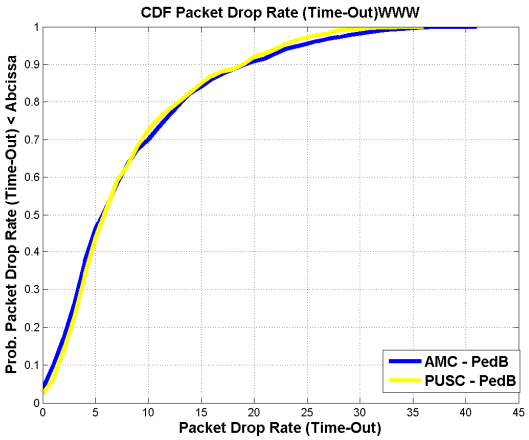
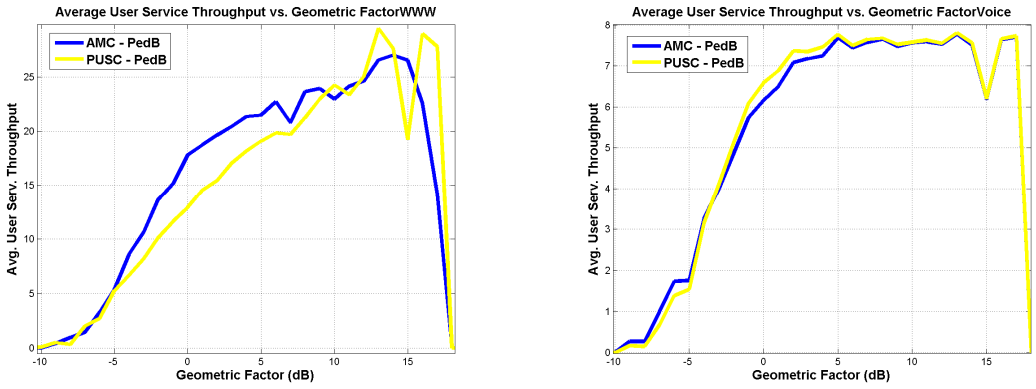


Figure 7 - CDF of average packet drop rate for WWW users

It is important to mention that the utility based scheduler behaves much like a maximum C/I for WWW users until the packet delay starts to approach the packet deadline. Therefore, packets can remain in buffer while transmission opportunities are provided for users with better channel quality. This is particularly relevant for such overloaded system. On the contrary, the fact that the system is under-loaded for VoIP users and the rigid delay constraints force the scheduler to transmit packets in a much faster pace as they cannot remain in buffer for much time (otherwise they are dropped). This contributes to the level of packets dropped due to channel quality and to the insignificant amount of packets dropped due to time-out violation.

Figure 8 plots the average service throughput per user versus the geometric factor in dB. It corroborates previous conclusions as the AMC sub-channelization mode results in better performance than PUSC one for WWW users, thanks to the overloaded system and to the multi-user diversity gain over frequency. AMC results in better service throughput over a large set of

geometry factors. On the contrary, as the system is under-loaded for VoIP users, both sub-channelization modes result in roughly the same level of average user service throughput.



a **b**

Figure 8 - Average user service throughput vs. geometric factor (a) WWW users; (b) VoIP users

Figure 9 is the plot of the average service throughput per user for both sub-channelization modes and for the MIMO channels with 2x2 STBC Alamouti for WWW users. As expected, with the MIMO channel the AMC sub-channelization mode results in an even higher performance over the PUSC sub-channelization mode. This is because the Alamouti STBC encoding with the MIMO channel results in a better channel quality and increases the probability of a packet being received with success.

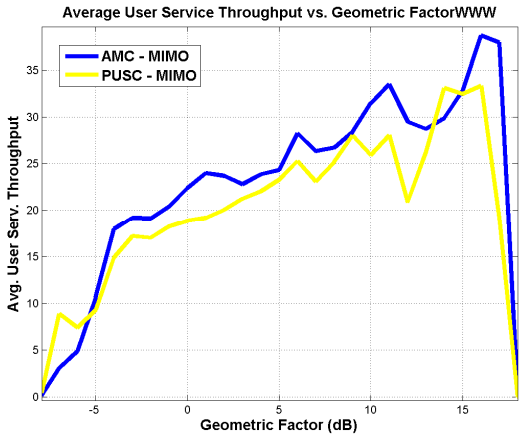
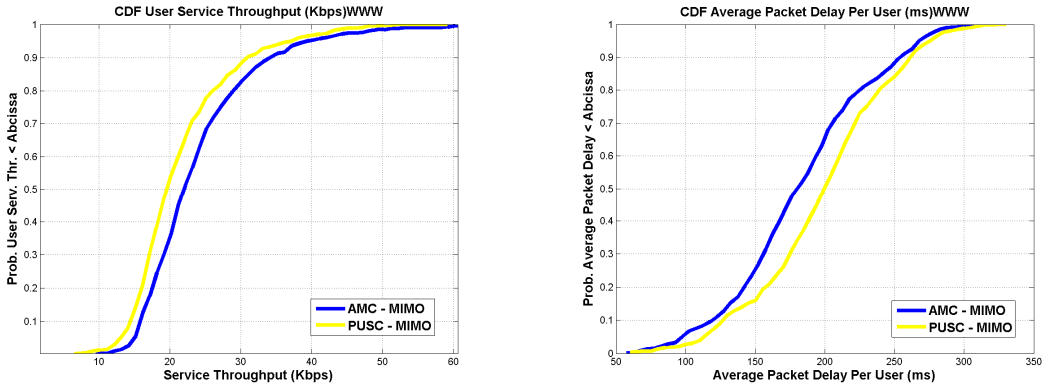


Figure 9 - Average user service throughput vs. geometric factor for WWW users and MIMO channel

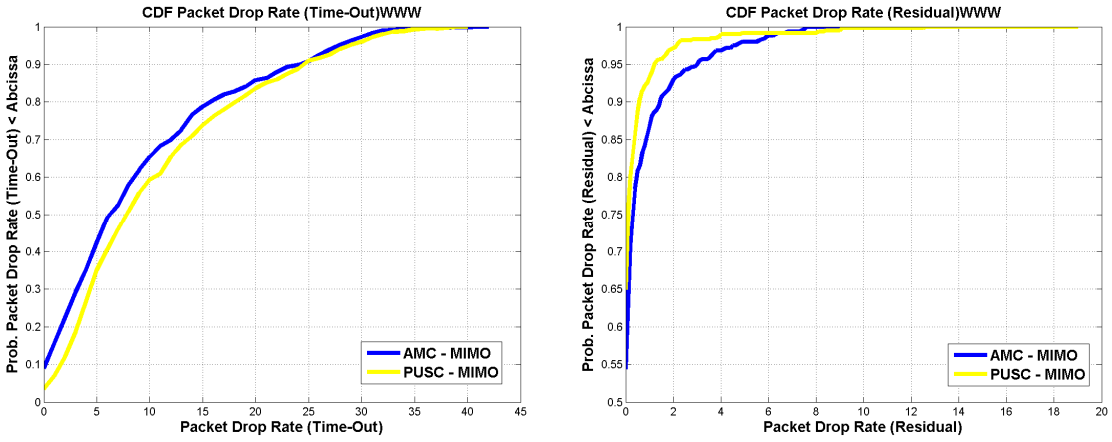
Figure 10 is the plot of the CDF of the average service throughput per user and of the average packet delay per each packet correctly received. These plots corroborate the gains achieved with AMC sub-channelization mode with WWW users over PUSC sub.channelization mode.



a **b**

Figure 10 - CDF of average user service throughput for WWW users; (b) CDF of average packet delay per user for WWW users – MIMO channel

Figure 11 is the CDF of the average packet residual drop rate and of the CDF of the average packet drop rate per time-out violation for WWW users. Although the MIMO channel results in a smaller percentage of packets dropped due to bad channel quality than the SISO channel, the PUSC sub-channelization still results in better performance. Both sub-channelization modes present close performance in terms of the amount of packets dropped due to time-out violation. This is because the utility-based scheduler behaves very much as a maximum C/I scheduler for WWW users, as mentioned before already.



a **b**

Figure 11 - CDF of average packet drop rate per user (time-out) for WWW users; (b) CDF of average drop rate per user (residual) for WWW users – MIMO channel

Figure 12 plots the average service and average over-the air throughput per user for both sub-channelization modes, both SISO and MIMO channels and both WWW and VoIP users.

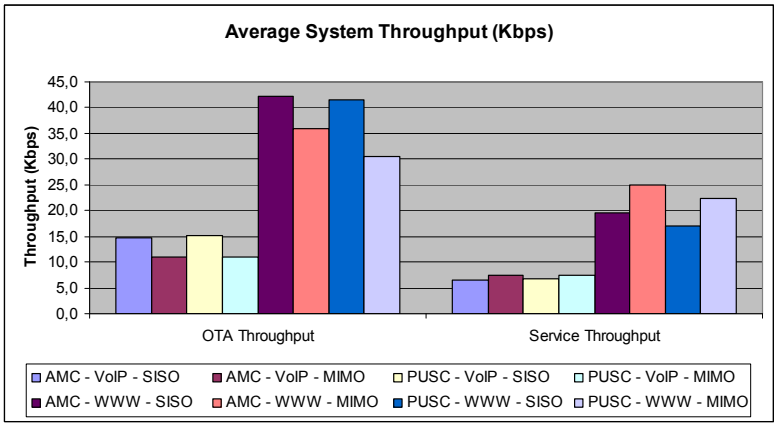


Figure 12 - System throughputs for different simulation scenarios

Some conclusions can be withdrawn from this plot.

- For both sub-channelization schemes service throughput is higher for WWW users if the MIMO channel is used than with the SISO channel.
- For WWW, service throughput is higher if the AMC sub-channelization mode is used, compared to the PUSC one. For AMC with MIMO the average service throughput per user reaches 25 Kbps and for PUSC with MIMO it is roughly equal to 22 Kbps. For AMC with SISO the user service throughput reaches roughly 20 kbps and for PUSC with SISO the user service throughput reaches roughly 16 Kbps.
- OTA throughput is higher if SISO mode is implemented for both types of users than with MIMO. This is because the spatial diversity achieved with the 2x2 Alamouti STBC scheme decreases the degree of variation of the channel amplitude and, therefore, decreases the opportunistic gain for the utility based packet scheduling. As expected, this reduction is more significant for WWW users than for VoIP ones.
- The OTA throughput reduction is more significant for PUSC sub-channelization scheme. This is because the gain achieved with the distribution of sub-carriers according to the pseudo-random sub-carrier distribution in PUSC mode is more affected by the decrease in channel variability resulting from MIMO and Alamouti encoding. As sub-carrier gains are correlated in each sub-channel defined according to the AMC sub-channelization scheme, the decrease in channel variability affects all sub-carriers in each sub-channel by roughly the same way.
- As expected, for all combinations user service, throughput performs roughly the same for VoIP users.

9.5 Related Work

To the best of the author's knowledge there are up to now no publications available in the literature regarding proposals for schedulers in realistic systems, combining both time and frequency domains, associated to the OFDMA multiple access in Mobile WiMAX. The

approach most closely related to this work is the one presented in [61]. However, the authors do not consider realistic traffic models. The proposed DRA assigns each resource in the map to the user with the highest priority, resulting from a modified Proportional Fairness algorithm. Users are assumed as always backlogged, i.e., full queue traffic model is used. This is because the authors compare the gains achieved with AMC sub-channelization scheme over PUSC sub-channelization scheme, in terms of capacity. No QoS requests are considered, also because no traffic model is simulated. System level simulations are performed for both modes and for different types of channel models, under a SISO channel.

Most proposals for packet schedulers which consider the multi-user diversity gain over frequency are based on simplistic scenarios, mainly considering one cell, with no traffic models and/or realistic traffic channels. Invariably these proposals formulate an optimization problem in which sub-carriers are provided to the user which maximizes a given performance metric (for example, a minimum data rate which must be provided to each user) whilst satisfying some simple constraints, such as the maximum power available for data transmission in the cell.

In the last few years there has been some work published in the literature regarding proposals for frequency domain packet schedulers for the 3GPP UTRAN Long Term Evolution (LTE) [182].

In [183] a frequency domain packet scheduler (FDPS) is proposed for 3GPP UTRAN LTE network. The proposed scheduler uses frequency-domain channel quality reports to multiplex users on different portions of the system bandwidth. In LTE the map of resources is organized into blocks of granularity in time and frequency named Resource Block (RB). The FDPS scheduler is based on the proportional fairness principle extended to both time and frequency domains. The authors report a gain of 40-50% in cell throughput compared to a time domain only scheduler (with no multi-user gain in frequency domain).

In [184] another channel dependent scheduler in both time and frequency domains is proposed for 3GPP UTRAN LTE network. The scheduler is divided into a time-domain and a frequency-domain part and both schedulers work independently and different algorithms can be applied in each scheduler part. Schedulers in time domain provide a list of mobiles to the schedulers in the frequency domain. Authors claim that a gain of 35% can be achieved in throughput and coverage over the basic opportunistic time-domain scheduler.

In [185] the same principle is used with the definition of a metric that decouples time and frequency domain schedulers. Two types of traffic models are considered: Best Effort and (BE) and Constant Bit Rate (CBR). Authors claim that this decoupled metric results in a coverage gain of up to 60% for a cell throughput loss of 5% over the time-frequency domain proportional fairness scheduler. The same approach for scheduling in time and frequency independently is followed in [186].

In [187] a frequency domain packet scheduler is considered for the analysis of a downlink OFDMA system using three different forms of CQI reporting schemes. Other proposals for time-domain schedulers are [188-189]. The work in [190] extends the original time-domain scheduler for MIMO channels.

9.6 Conclusion

In this chapter another extension to the original DRA architecture was implemented and its performance analyzed. This new DRA architecture implements the AMC sub-channelization scheme, which results in a multi-user diversity gain over frequency. The original utility-based packet scheduler is used in conjunction with the proposed DRA. A set of system level simulations was performed to infer on the gains achieved over simple scenarios where PUSC only sub-channelization mode is used.

It was verified through simulations that AMC sub-channelization mode is more efficient for WWW traffic model regarding the achieved service throughput in an overloaded scenario, and that in an under-loaded scenario, for VoIP traffic model, both sub-channelization schemes behave essentially in the same way. As the system is overloaded for WWW, the AMC sub-channelization mode is more efficient in the utilization of the available radio resources thanks to the multi-user diversity gain over frequency for AMC. As the system is under-loaded for VoIP users there is no resulting multi-user diversity gain over frequency, even with the AMC sub-channelization mode and both sub-channelization modes result into similar behaviour. AMC could be even more effective in system capacity for a full queue traffic model in conjunction with an opportunistic packet scheduler, such as the maximum C/I.

However, AMC sub-channelization is more sensitive to errors in channel quality reporting and also to inter-cell interference, as there is no randomization in the interference from neighbouring cells over the sub-carriers composing each channel. Therefore, in an under-loaded scenario in which all users are served, even those users with bad channel quality, the PUSC sub-channelization mode results in a better performance than AMC.

For SINR levels between 0 dB and -5 dB, AMC results in better performance with a smaller number of packets dropped due to bad channel quality over PUSC. AMC is more effective for this range of SINR values achieved with the most robust MCS scheme, due to the contiguous allocation of sub-carriers over each sub-channel in each radio resource. For WWW users, the multi-user diversity result into a smaller amount of packets dropped due to bad channel quality if AMC sub-channelization mode is used in detriment of the PUSC.

Concerning the implementation of the AMC sub-channelization scheme it is important to reinforce that the use of an opportunistic packet scheduler such as the Maximum C/I or even the Proportional Fairness schedulers could result in better gains over the PUSC sub-channelization. This is exactly what is available in the literature up to now. The author thinks that the novelty of

the work presented is related to the use of a packet scheduler for QoS provision and with realistic traffic models, something that is not available in the literature.

By means of system level simulations it was demonstrated that the AMC sub-channelization scheme can result in an increase in the system capacity provided the load is high enough, in order to make the best use of the multi-user diversity gain over the frequency domain. However, it is important to enforce that such proposal for a DRA incurs in a high complexity system, namely in the amount of computations, which must be performed for each resource in the frame and in terms of the execution time of the algorithm.

Until now there is no single proposal for such a DRA in the research literature. The work available considers mixes of simple opportunistic schedulers such as variations of the Maximum C/I and Proportional Fairness algorithms. No realistic traffic models are considered as users are assumed as fully backlogged. This is because the gains achieved with the joint time and frequency domains schedulers are more effective under such scenarios, as it was proved in the simulations conducted.

Chapter 10

Conclusions

10.1 Preliminares

This dissertation has investigated the design and implementation of packet schedulers for mobile broadband wireless networks of next generation, which are envisioned to support the quality of service (QoS) requirements of present and future high bandwidth demanding multimedia applications.

In particular the proposed packet schedulers are supported in a new paradigm of mobile communications protocol reference model, named cross-layer design. This new paradigm is of fundamental importance in pursing the satisfaction of these QoS requirements under such hazard means of data transportation such as the mobile broadband wireless channel. The implementation of packet schedulers according to the cross layer design architecture combine information from different layers in the protocol stack, resulting into adaptive algorithms whose functionalities depend on the state of the layers considered. This dynamic behaviour results in an efficient utilization of the scarce radio resources available for data transportation, maximization of QoS satisfaction and fairness in resource allocation to different service flows.

An important research task conducted along this work was the design, implementation and validation of a proper dynamic resource allocation (DRA) module into which the packet schedulers were plugged. The DRA implements the control channels and signalling messages used in the exchange of information regarding the state and functionality of the different layers involved into the cross-layer scheduling algorithms.

This thesis has assessed the IEEE 802.16e standard for Mobile WiMAX networks as a potential radio access technology that could facilitate the evolution of the wireless communication market by serving the demands of the end users and operators. The IEEE 802.16e standard for Mobile WiMAX networks was used as a case study technology for the implementation of the proposed DRA architectures. This is because Mobile WiMAX is currently envisioned as a strong candidate for the mobile broadband wireless networks of next generation. Also, this standard supports the implementation of control channels and signalling messages fully compliant with the cross-layer design paradigm.

Different versions of a basic scheduler based on the notion of Utility Functions, a concept derived from economics, are proposed in this work. The schedulers are inserted into the cross-layer-based DRA architecture for the Mobile WiMAX standard and their performance is validated against commonly packet schedulers available in the literature.

Fundamental was the deployment of a system level platform for the execution of system level simulations. This platform is written in C++ language and implements all major functionalities from the Mobile WiMAX standard, according to the system profile for manufacture's product interoperability and testing issued from the WiMAX Forum.

As system level and link level simulations run in different levels of granularity in time domain, link level performance is encapsulated into properly defined look-up tables, resulting from a set of simulations performed at the link level, thereby emulating the whole transmission chain, and which implements the physical layer of the system, in a point to point configuration. A proper link-to-system level interface is described, which models the link layer transmission chain performance in terms of Signal to Noise Ratio (SNR) versus Block Error Ratio (BLER) pairs of values.

Multiple antenna schemes (MIMO) and beamforming are advanced techniques for data transmission and are included in the Mobile WiMAX set of functionalities. They make it possible the implementation of a spatial division multiple access scheme (SDMA), which adds space domain as another degree of freedom for resource allocation. It increases network capacity whilst satisfying applications QoS requirements. The implementation of joint scheduler and space domain resource allocation schemes is also performed in this work.

In this concluding chapter, a summary of this dissertation is given. Every section of this chapter draws the main conclusions of the investigations of each chapter of the thesis.

10.2 Cross-Layer Design

Chapter 2 concentrated on building understanding on the cross layer design paradigm for the communication protocol model from future generation mobile wireless networks. Cross-layer design is of fundamental importance for the provision of the required set of QoS parameters expected from the kind of multimedia and high consuming bandwidth applications, which are envisioned for these networks. Cross layer design breaks out the stringent modular protocol design proposed with the OSI protocol model and used extensively on the design and implementation of fixed communication networks.

10.3 Mobile WiMAX Networks

Chapter 3 explains in great level of detail the particularities of the IEEE 802.16e standard for the implementation of Mobile WiMAX networks. Mobile WiMAX is considered as a potential strong candidate for the next evolution of mobile wireless networks, as it offers the amount of capacity and bandwidth for the support of the type of applications intended for such scenarios, and provides the stringent QoS requirements which are expected.

Mobile WiMAX offers scalability architecture in both radio access technology and network architecture which provides a great deal of flexibility in network deployment options. Specifically, a detailed description of the physical and medium access control layers functionalities is provided, with a particular emphasis on the mechanisms implemented in the provision of QoS from multimedia applications. The standard specifies a number of control channels and signalling messages which can be used in the exchange of the state regarding each layer in the stack, resulting in the implementation of the cross-layer design framework.

Although the standard does not define the type of scheduler to be implemented, the mechanisms used in the provision of QoS and the different types of traffic classes used in this provision, and by the scheduler implemented in the system are defined in the standard. The chapter introduces the advanced features included in the new version of the standard and which are used to increase capacity.

The IEEE 802.16e standard for Mobile WiMAX networks is used as a case study in this work for the implementation of the kind of packet scheduler algorithms for the support of the kind of multimedia applications in the next generation wireless networks.

10.4 System Level Simulations for Mobile WiMAX

A system level simulator was implemented in this work for conducting system level simulations for a Mobile WiMAX network. The system level simulator was deployed in a C++ environment and the simulation methodology is in accordance to the general methodology followed in the implementation of system level simulators in other existing wireless networks.

In particular, the system level simulator follows the guidelines for performing system level simulations for the Mobile WiMAX standard, from the system profile defined by the WiMAX Forum. The models for the mobile radio channel under such environment, namely the path-loss, shadowing and fast fading for multi-path propagation, according to ITU channel models, for SISO channel and SCM channel model for MIMO, are presented in detail. The models for the emulations of different traffic models are considered.

System level simulations can be performed according to one of two different approaches: a combined snapshot mode, which is basically a Monte Carlo approach for simulation, and a dynamic mode. In this work, as mobility and handover functionalities are not implemented and/or simulated, the combined snapshot methodology is followed in the simulations. In this approach a simulation comprises different runs and each run comprises different transmission time intervals (or frame periods). A number of mobile stations are dropped in the beginning of each run, in the three sectors of the central base station in the network layout. Neighbouring cells are used only for simulating inter-cell interference. Path-loss and shadowing are computed in the beginning of each run and assumed constant throughout all run duration. Fast fading is simulated in each transmission time interval.

10.5 Dynamic Resource Allocation Module Architecture

For the implementation of the proposed packet schedulers in the system level simulator platform and for conducting system level simulations a proper dynamic resource allocation (DRA) module must be deployed. As part of the DRA functionality, the radio resources which are assigned to the set of active users in each cell and used in data transmission were defined. Also, communication protocol architecture for the exchange of signalling messages, used in the transmission of information from the base station to each mobile user was implemented in the DRA. Chapter 5 presents, in great level of detail, all the steps followed in the design of the DRA module.

The DRA is fully compliant to the standard for Mobile WiMAX networks. The packet schedulers used in the support of the different types of applications offered to the network, constitute a single module of the more general DRA architecture framework.

As mentioned in this work, very few of the different approaches for DRAs in the research literature consider practical DRA architectures, which can be implemented in realistic scenarios. The performance of many of the implemented algorithms is based upon analytical models, applied in very simple cellular layouts, whereby a single cell broadcast in a point-to-multi-point configuration to a number of mobile stations. For such simpler application scenarios realistic traffic models are not implemented and mobiles are normally assumed as backlogged all the time.

In this chapter all details regarding the proposed cross-layer based DRA for Mobile WiMAX system-level simulations were presented.

10.6 Validation of the System Level Simulator and Dynamic Resource Allocation Module

Before conducting system level simulations for the architectures implemented for dynamic resource allocators and packet schedulers, the system level simulator must be validated. This validation is performed by testing the models implemented for traffic, channel, Signal to Noise plus Interference Ratio (SINR) computation and interference generation.

Chapter 6 describes the set of simulations which were performed for the validation of the tool and for the validation of the basic DRA architecture described in previous chapter. For each new packet scheduling algorithm implemented in the tool, its performance should be compared against benchmark figures available in the literature. Although this is not possible, as there is no straight-forward solution available in the literature that can be used to compare against the architecture proposed in this work, and consequently fully support the validation process, some figures available in the literature can be used as a basis for comparison.

The implemented models are validated by comparing the results obtained from simulations to the theoretical values associated to the different types of models used in the simulator. This is an important step in the process of validation, because it infers the level of accuracy achieved in the implementation of these models in the system level simulator. Also, as an effort to address the trade-off between simulation time and accuracy, the validation methodology has been addressed based on the central cell approach and assuming full load conditions.

10.7 Utility-Based Packet Scheduling for Mobile WiMAX

This chapter describes the design and implementation of packet schedulers proposed for the support of multimedia, high-consuming bandwidth applications in beyond 3G wireless networks. The scheduler framework is based on the cross-layer design paradigm and is fully interoperable with the MAC layer architecture designed for the Mobile WiMAX standard, in line with the principles presented in previous chapters.

The packet scheduler functioning principle is derived from the notion of utility functions, a concept from economics. In particular, not only the benefit a given user can result for the network provider is considered in the computation of the scheduling metric but also the cost resulting from postponing transmission from other active users in the same transmission time interval. Quality of service requirements are defined mainly in terms of the satisfaction of the maximum packet delay allowed for each type of application service class considered in the system. The performance of the proposed packet scheduling algorithm is compared against other commonly referred schedulers in the literature, namely the Modified Largest Delay First (M-LWDF), the Proportional Fairness (PF) and the Maximum C/I (CI). Performance is

evaluated in terms of the amount of users satisfied by the kind of service provided. A user's satisfaction depends on the amount of packets dropped due to delay bound violation and maximum number of allowable re-transmissions due to bad channel quality. System level simulations were conducted for a number of traffic loads and for a traffic mix composed by VoIP and WWW users.

Satisfaction Ratio

- It was verified from the simulations that according to the type and shape of utility function implemented in the scheduler it is possible keep the user satisfaction rather insensitive to the increase in the load at the detriment of a degradation in the quality sensed by WWW users. This is because the utility function implements a prioritization mechanism in the access to radio resources for VoIP users in detriment of WWW users.
- It was also verified that the proposed scheduler presents better performance than all three schedulers used as benchmark in system performance evaluation. Actually, for WWW users, the utility based scheduler behaves pretty much like the opportunistic Max C/I scheduler and therefore they present equal performance.
- Differently from much scenarios and results available in the research literature, in which, at least to the point of view of the author, most of the proposed schemes do not represent realistic simulation scenarios, M-LWDF, PF and EXP result in worst performance than the CI scheduler, regarding the user satisfaction ratio.

Average Packet Delay/Average Packet Drop Ratio

- The M-LWDF scheduler results in a higher average packet delay for most system loads than the CI scheduler. This is because the M-LWDF tries to equalize the delays, no matter the type of traffic model used: WWW and VoIP, not paying attention to delay requirements from each type of traffic flow. Nevertheless, the smaller average packet delay from CI scheduler for VoIP users is obtained at the cost of an increase in the number of dropped packets. As WWW packets have a much less stringent delay bound, the average packet drop rate for this traffic model has better figures than PF and M-LWDF at the cost of an increase in the average packet drop rate for VoIP packets.

Cumulative Distribution Functions for Average Packet Delay and Average Packet Drop Rate

- Average values are not accurate figures for performance evaluation as they do not reflect performance behaviour for each individual user. Therefore cumulative distributions were produced for the average packet delay and average packet drop rate per user for the maximum load considered in the simulations.
- Regarding the CDF of the average packet delay and average packet drop rate, the utility-based scheduler presents the best performance for both types of traffic models. As the CI scheduler does not consider the delay incurred to each packet until the scheduling instant,

for both types of traffic models, and because VoIP packets are much more demanding than WWW ones concerning packet delay, the CI scheduler results in better delay performance for both traffic models at the detriment of a big amount of packets dropped due to delay bound violation for VoIP packets. The delay violation probability parameter in the MLWDF algorithm was set as equal both for VoIP and WWW traffic models. Therefore it results in the degeneration of the MLWDF scheduler into the PF one and both schedulers result in better performance than CI for VoIP regarding the average packet drop ratio for the maximum system load and for VoIP users.

- Regarding the average packet drop rate per user it was verified that the CI scheduler actually results in a worse performance than all the remaining three scheduler, for VoIP traffic users. Actually, differences of 20% and 10% were achieved, respectively from the utility and PF/M-LWDF schedulers.

The benefits resulting from the implementation of a packet scheduling algorithm according to the notion of utility and utility functions, namely in a scenario of traffic mix, are clear from the resulting system level simulations. A significant gain is achieved in the user satisfaction ratio for the UTIL scheduler compared to the other schedulers used as benchmark.

A variation of the utility-based packet scheduling algorithm was also implemented into the DRA used for conducting system level simulations of Mobile WiMAX. This new packet scheduler is a jointly token-bucket and utility based packet scheduling algorithm and was designed with the objective in mind of providing a minimum service throughput for non-real time applications. The same packet schedulers enumerated above were considered in the system level simulations and four different types of traffic models used: two models representative of services flows of real time: VoIP and NRTV; one representative of NRT: WWW and another one representative of BE (FTP). The definition of the proper shape and configuration parameters of each type of utility function depend on the type of service model envisioned, as the utility function prioritizes user's access, according to the required set of QoS parameters.

In the simulations different amount of system load, in terms of the total amount of active users in the cell, were considered.

Again, the performance of the utility based packet scheduling algorithm was inferred by means of the user satisfaction ratio and system performance metrics, such as: the average packet delay per user, average packet drop rate per user and average service throughput per user.

Users Satisfaction Ratio

Regarding the user satisfaction ratio the following conclusions can be inferred from system level simulations:

- The proposed token and utility based scheduling algorithm has the best performance over all four proposed schedulers, except for high loads of the WWW service.

- The maintenance of the satisfaction ratio from VoIP and NRTV users is achieved at the cost of degradation in the satisfaction ratio from WWW users. This is because the algorithm attempts to equalize the packet delay for RT services while attempting to satisfy average service throughput for WWW users.
- As the FTP is a burst traffic model, with a longer interval between active packet generation, the amount of packets in the buffers is enough to be serviced whenever users from other services are not transmitting.

Average Packet Delay per User

The following conclusions can be inferred:

- With the exception of the UTIL scheduler the system is congested with loads corresponding to more than 160 users.
- The fact that the CI is an opportunistic scheduler results in a better performance than the PF and M-LWDF, for WWW and FTP traffic models, which are of burst nature and do not request the satisfaction of stringent packet delay bounds. As a matter of fact, the associated higher delay bound allows packets to remain for a longer period of time in the buffer before they are dropped. Regarding VoIP traffic model the periods of inactivity in packet generation accumulates packets in buffer before their transmission. This is different from NRTV traffic, which is a streaming traffic flow, in which packets are generated with a constant rate and where there are no periods of inactivity. Therefore, packets are accumulated in the buffer at a much faster pace than with VoIP and for this reason they must be transmitted as earlier as possible in order not to be dropped.
- The proposed token and UTIL scheduler presents the best performance for all four types of service classes. This is because packets whose delay becomes equal to or higher than the delay bound are not transmitted (they are dropped) as they lose their utility for the network.

Average Packet Drop Rate per User

The following conclusions can be inferred:

- The performance of the UTIL scheduler for WWW users is not exactly what could be desired as it results into a higher percentage of packets dropped due to delay bound violation, compared to the other three types of users. This amount of dropped packets contributes to the decrease in the achieved average service throughput for WWW and in a lower ratio of satisfied users, compared to the CI scheduler.
- The main reason for this performance degradation is the choice for the delay bound. A delay bound of 500 ms was assumed for WWW packets to decrease computation time. According to the type of utility functions used, priority is given to VoIP and NRTV users, for loads higher than 160 users, therefore, a higher percentage of WWW users have their packets dropped before achieving the required average service throughput.

- CI scheduler results in a higher average packet delay for burst traffic users, such as WWW and FTP and the percentage of dropped packets is lower for both traffic models because packets remain in buffer waiting for transmission for a longer period of time than packets from VoIP and NRTV traffic users.

Average Throughput per User

- For a load higher than 160 users the system is congested and this results in a significant decrease in the average service throughput for all types of service, except for the UTIL scheduling algorithm.
- PF and M-LWDF schedulers are much more negatively influenced by the increase in the offered load, to a number higher than 160 users for VoIP and NRTV users, than CI scheduler. As there is not enough capacity to satisfy packet delay constraints a higher percentage of packets are dropped in order to satisfy the delay bound. This results in the decrease which was verified in the service throughput, as compared to the CI scheduler. However, the prioritization scheme implemented in the UTIL scheduler results in the slower decrease in the achieved service throughput for NRTV users compared to the CI scheduler. It can be noticed that the decrease is less than 3% for the maximum load of 240 users in the system.

Average figures do not show the evolution in the distribution of average packet delay, average packet drop rate or average service throughput. Therefore, averaging metrics do not reflect the behaviour of each type of scheduler regarding the location of each user in the cell. In particular, the performance for users located in the edge or near the cell's centre is inferred from CDF distributions.

For all four schedulers, an admission threshold of -5 dB was used to avoid the access to resources from users with bad channel quality.

VoIP Users

- Regarding the average service throughput per user the UTIL scheduler results in the highest average service throughput per user for users in the edge of the cell. In the range [-, -2] both PF and M-LWDF schedulers present almost identical average service throughputs per user, and perform better than the CI scheduler. Thus, it can be concluded that these two schedulers are more effective than CI in serving users in the edge of the cell. This was expected due to the opportunistic nature of the CI scheduler, but it could not be noticed from the global average performance metrics. The CI scheduler results in better performance for higher values of the geometric factor and is equivalent to the performance of the UTIL scheduler. The degradation in the performance of PF and M-LWDF schedulers is because they serve users in the edge of the cell. UTIL scheduler achieves gains of roughly 35% and 55% to CI and M-LWDF/PF schedulers respectively. The contribution

for this gap in performance comes from users in the edge of the cell, in the range [-5, 0] of the estimated geometric factor.

- Regarding the average packet delay per user, both M-LWDF and PF schedulers attempt to be fair in resource allocation of resources to users from different types of traffic classes, without defining any prioritization other than the head of line packet delay (M-LWDF) or average service throughput (PF). Therefore they present a worse performance than the CI scheduler.
- Regarding the average packet drop rate per user, there is a huge difference between the amount of packets dropped by both the M-LWDF and PF schedulers and the amount of packets dropped with the UTIL and CI schedulers. For the QoS requirement of 3%, 85% of the users comply with this requirement from the UTIL scheduler and 45% with the CI scheduler. All users present a packet drop rate higher than 3% for the PF and M-LWDF schedulers.

NRTV Users

- The UTIL scheduler is fair in sharing resources for the NRTV traffic model, while the other three ones loose fairness as time evolves.
- It is also more effective in serving users in the edge of the cell than the CI scheduler as it presents better performance than the CI scheduler up to 3 dB of the geometric factor.
- With the improvement in channel quality more packets can be considered in the utility which can be potentially transferred to each user from WWW and FTP traffic models.
- 90% and 65% of the users achieve the required packet drop rate of 3% for UTIL and CI schedulers respectively.

WWW Users

- Regarding the average service throughput, both M-LWDF and PF result in a worse performance than CI and UTIL schedulers. But this result was achieved at the cost of a performance degradation regarding the other metrics. Also, the CI scheduler presents better performance than the UTIL scheduler, due the opportunistic nature of the algorithm. As a matter of fact most of the users are serviced with a service throughput higher than the what is requested in the QoS service profile. On the contrary, and on average, most of the WWW users are provided with the requested service throughput. It was also verified from the simulations that the difference in performance among M-LWDF, PF and UTIL regarding the average service throughput per user, for users in the edge of the cell, is not that significant. The type of utility function used and the limitation on the provided service throughput per user result in the better performance of the CI scheduler for users near to the cell centre.

- The UTIL scheduler is very effective in controlling the amount of packets dropped due to quality. For 10% tile, the gain from UTIL to CI, M-LWDF and PF is respectively: 20%, 70% and 75%.

FTP Users

- FTP traffic is characterized by long periods of inactivity in which there are no packets generated by the source. Therefore, and according to system level simulations it was observed that the performance of the UTIL and CI schedulers differ only for those users in the edge of the cell, with geometric factor in the range [-5, 0]. For this reason there is a gain of roughly 5% regarding the service throughput for the UTIL scheduler from the CI.
- It was also observed that most of the packets dropped are due to bad channel quality. These packets are dropped after the maximum number of transmission attempts is achieved. Actually, 85% of the users have no packets dropped at all for the UTIL scheduler and this figure reduces to 75% for the CI scheduler

Although it is, by itself, a simple concept, the definition and design of the proper type of utility function can be a complex topic as it is intrinsically related to two different and conflicting objectives: (i) profit maximization from the point of view of the operator, which translates resumes basically to the maximization of the traffic carried in each cell, for the minimum cost; and (ii) compliance to user's QoS requirements, namely minimum service throughput, average expected packet delay and average amount of packets dropped, by time out and bad channel quality. Therefore, a trade-off must be followed, as the resulting user's satisfaction for the service provided and, thus, the resulting churn rate depends intrinsically on the definition of the appropriate parameterization to be used in the scheduling algorithm.

It is possible to conclude that the joint utility based and token bucket packet scheduler result in a significant increase in system performance in term of the amount of satisfied users in the system, whilst presenting similar user average service throughputs to the remaining packet schedulers used as benchmark comparisons.

10.8 Space Division Multiple Access with Utility-Based Packet Schedulers for Mobile WiMAX

This chapter presents an extension to the initial DRA architecture proposed in chapter 7. It considers an advanced feature available in the Mobile WiMAX standard, regarding the implementation of Advanced Antenna Systems (AAS), namely: the possibility of implementing multiple antenna arrays for beamforming. Beamforming specifically enables the implementation of space division multiple access schemes (SDMA) which result in another degree of freedom for user allocation, i.e., radio resources can be also defined in space domain.

For this type of multiple access scheme cross-layer design among lower and higher layers in the protocol stack determines how spatial beams are attributed to scheduling users. Differently from

radio resources in time and frequency domains, space beams cannot be assumed as completely orthogonal, and, therefore, some amount of intra-cell interference happens to degrade transmission quality. A properly designed packet scheduler must consider the non-orthogonality of radio resources defined in space domain whenever assigning them for packet transmission.

For the evaluation of system capacity for both types of DRAs, system level simulations were conducted with the full queue traffic model scenario. According to these simulations it was to possible to conclude that DRA performance with SDMA multiple access scheme can result in an OTA throughput of 50 Mbps, resulting in a throughput gain of roughly 38 Mbps over the simple non SDMA-based DRA scheme. The gains achieved with SDMA in resource allocation are of significant importance over the basic DRA scheme.

The implementation of the SDMA-based DRA over the Mobile WiMAX air interface results in an increase in system capacity and into a higher percentage of satisfied users, compared to schemes where DRA is not SDMA based. It is important to mention that this increase in system capacity is not obtained at the cost of degradation in the QoS provided by the network. The basic utility-based scheduling algorithm is considered jointly with the SDMA scheme as it is basically a new form of resource allocation, for users prioritized first according to the specifics of the utility algorithm.

For system performance evaluation system level simulations were also performed with the joint implementation of the SDMA-based DRA and the utility-based packet scheduler, for both VoIP and WWW traffic models. Performance for such scenarios was inferred by the computation of the total amount of satisfied users and other metrics regarding system performance, such as: system throughput, average packet delay per user and average packet drop rate per user. User satisfaction ratio is evaluated by measuring the percentage of packets dropped due to bad channel quality and time-out violation.

Regarding the user satisfaction ratio for VoIP users, according to system level simulations it was verified that the additional radio resources, which are made possible in the SDMA multiple access scheme, do not translate into a verifiable variation of the user's satisfaction ratio with system load. This seems to be due to the small size of VoIP data packets and to the low utilization of radio resources, as the system seems to be un-loaded in such scenario. Therefore, the user satisfaction ratio is roughly insensitive to the amount of users in the system as there are virtually no unsatisfied users if the SDMA-based DRA is implemented. Nevertheless, a verifiable gain of roughly 40% can be observed with the implementation of an SDMA based DRA over non-SDMA, whereby a smaller amount of resources is available for allocation.

Compared to VoIP, in the scenario of WWW traffic, It was verified from system level simulations that SDMA results into a higher percentage of satisfied users compared to the non-SDMA based DRA. The system is more loaded, as the WWW average packet size is much larger than VoIP one. This higher amount of load in the system results in an increase of the

multi-user diversity gain. Also WWW traffic model is much more tolerant in terms of packet delay than VoIP. This means that packets can remain for longer periods of time in buffer waiting for better channel conditions and performing more transmission attempts. Therefore, resources are used more efficiently with WWW traffic model. Also, the fragmentation of WWW packets translates into a small number of users transmitting over spatial beams in the SDMA zone of the map of resources per frame period and, as a consequence, this translates to an improvement in channel quality because of the inherent smaller intra-cell interference. The higher load in the system results in a smaller gap in the amount of satisfied users for both schemes. However, it was verified from the simulations the reduction in this figure with the increase of the load in the system. According to the simulations a gain of roughly 15% is achieved with SDMA over non-SDMA multiple access scheme.

System performance was evaluated for both DRA architectures for a maximum system load of 200 users and for both VoIP and WWW traffic modes:

- Regarding the average packet drop rate per user, there is a significant difference in performance between both modes. In SDMA mode, beam assignment is performed in such a way not to degrade the estimated channel quality from already assigned users to the same radio resource, but in different spatial beams, in order not to violate the selected MCS scheme. Regarding VoIP users roughly 95% of users have a packet drop rate less than 5% without SDMA whereby there are virtually no packets dropped for SDMA. Regarding WWW users it was verified that main contribution to the difference in the average packet drop rate is due to the amount of residual error, which is significantly higher for the non SDMA-based scheduler. WWW traffic model is much less sensitive to delay than VoIP and to the type of slow decreasing utility function utility used by WWW users, whereby the utility scheduler behaves as an opportunistic scheduling algorithm. For example, 90% of the packets have a drop rate less than 4% due to residual errors while almost 100% of the packets are not dropped at all if SDMA mode is implemented.
- For VoIP, SDMA-based DRA results in smaller packet delays per user. But the difference is not that much significant as VoIP packets are highly constrained in terms of delay. In the other way, the fact that WWW is much less stringent regarding delay than VoIP results into a small gap in the performance of both schemes. It was noticed from performance results that for the maximum load of 200 users in the system the maximum difference amounts to 5%.
- Average SINR is significantly better for the SDMA mode and it is roughly insensitive to the increase in the number of users in the system. This is due to the scheme followed by the utility based scheduler in the elaboration of the priority lists. A DRA with no SDMA mode is inherently more loaded than if SDMA is implemented. This results into a roughly higher gain of multi-user diversity which can pick users with better channel conditions for

transmission. Therefore, for WWW a performance gain of roughly 20% can be achieved with WWW users for SDMA.

- Regarding VoIP users, the difference in the service throughput is almost zero for both modes and for all traffic loads, although it can be observed a small gap for 180 and 200 users load. Higher system load results in a higher gap in the performance of both types of multiple access schemes, regarding the service throughput, than what is verified for VoIP. The main contribution for this difference arises from the amount of transmission opportunities for users in the edge of the cell, which are allocated empty beams with enough quality for transmission (limiting the intra-beam interference) in the SDMA-based scheduler. The difference in performance is not so significant because the size of the transport block with the most robust MCS scheme is very low and users in the edge of the cell (with low geometric factors) are assigned the most robust MCS scheme for transmission. Nevertheless, gains of roughly 2% can be verified from system level simulations.
- For non SDMA based DRA the OTA throughput is slightly higher than for SDMA, because there is a higher percentage of packets which are received with error due to bad channel quality. The procedure followed in the assignment of spatial beams results into a smaller number of transmission attempts and, therefore, into an increased service throughput and a smaller OTA throughput, compared to the DRA without SDMA. For the same load, as the amount of resources increases the resulting multi-user diversity gain is smaller for the SDMA mode. Differently, for the non SDMA based DRA the multi-user gain is higher and this translates into a higher OTA throughput. As expected also, the OTA throughput increases with the amount of active users (system load) in the system. This is the result of the multi-user diversity gain arising from the opportunistic channel access. This effect is evident with the WWW traffic model because WWW packets are more tolerant to delay than VoIP packets. Together with the shape of the utility function used in the utility scheduler, it is a consequence of the fact that under these conditions the scheduler behaves much like the opportunistic maximum C/I scheduler.

According to system level simulations it is possible to perceive the gains achieved with the implementation of a SDMA based DRA for both types of traffic models. It could be interesting to analyse the performance of both systems in a mixed scenario with both VoIP and WWW traffic models. This was left for future work.

10.9 Joint Time and Frequency Domains Packet Scheduler for Mobile WiMAX

AMC sub-channelization scheme adds another diversity gain for radio resources organization over the TDD frame of the Mobile WiMAX air interface. This is the frequency diversity gain.

With AMC implementation an increase in system capacity can be expected, provided the best channels are opportunistically assigned to their corresponding users, not paying attention to the QoS demands, such as packet delay bound.

A set of system level simulations was conducted for the initial version of the utility-based packet scheduling algorithm, implemented into the DRA jointly with the AMC sub-channelization scheme. System level simulations were also conducted for PUSC sub-channelization mode, implementing the same utility-based packet scheduling algorithm. Both VoIP and WWW traffic models and both SISO (ITU PedB 3 Km/h) and a 2x2 MIMO channel with STBC Alamouti coding were simulated. A fixed system load of 200 active users was considered in all simulated scenarios. Traffic mix was not considered in any simulated scenario: performance was analysed for VoIP or WWW traffic users only in each set of simulations.

SISO channel

- In the scheduling algorithm an admission threshold was considered. This is a parameter used by the scheduling algorithm in defining the set of users who can attempt transmission and therefore are inputted into the scheduling algorithm. Such level of admission results into the scheduling of users in the edge of the cell, with a bad channel quality and therefore using the most robust MCS scheme. As verified in the simulations a higher percentage of users are given transmission opportunities for transmission while residing in the regions of the cell close to the edge, therefore with bad channel quality. Therefore, regarding WWW users, while performing transmission, the size of each data packet and the small MCS scheme results in a significant load to the system. Therefore, a fraction of packets are dropped due to time-out violation, although the multi-user gain of frequency, from AMC mode, results into a slightly better performance compared to PUSC.
- However, there is a difference in performance for both sub-channelization schemes regarding average packet drop ratio per user of roughly 20% for WWW traffic model. As expected, and for the same reasons pointed above, the difference is not significant for VoIP users.
- For WWW users the utility based scheduler behaves much like a maximum C/I. Therefore packets remain in buffer while transmission opportunities are provided for users with better channel quality. On the contrary, the fact that the system is under-loaded for VoIP users and the rigid delay constraints force the scheduler to transmit packets in a much faster pace as they cannot remain in buffer for much time (otherwise they are dropped). This contributes to the difference in the amount of packets dropped due to channel quality and to the insignificant amount of packets dropped due to time-out violation for both WWW and VoIP traffic models.
- Regarding the average service throughput per user versus the geometric factor, a gain of roughly 5 kbps can be achieved for users closer to the cell edge if AMC sub-channelization

mode is implemented. On the contrary, as the system is under-loaded for VoIP users, both sub-channelization modes result in roughly the same level of average user service throughput. In particular, for the average CQI value of 0 dB, a gain of roughly 10% is achieved for AMC over PUSC mode and for WWW users, whereby there is almost no difference between both schemes for VoIP users.

- Regarding the achieved service throughput, AMC sub-channelization is more efficient for WWW users. AMC can achieve gains of roughly 20% over PUSC. In the opposite way, both sub-channelization modes are very similar regarding the achieved average service throughput per user for VoIP. As a matter of fact, the AMC sub-channelization mode is more efficient in the utilization of the available radio resources. This efficiency results from the multi-user diversity gain over frequency, associated to this sub-channelization mode.
- The AMC sub-channelization mode presents a better performance in terms of the average packet delay per user for both types of users: WWW and VoIP. For WWW and VoIP, a gain of roughly 20% and 5%, respectively, can be achieved with AMC over PUSC.

MIMO channel

- For the MIMO channel and regarding the average service throughput per user the AMC sub-channelization mode results in an even higher performance over the PUSC sub-channelization mode. This is due to the better channel quality resulting from the implementation of the Alamouti STBC encoding jointly with the multi-user gain over frequency domain achieved with the implementation of the AMC sub-channelization scheme. The result is the increase in the probability that the packet is received with success.
- Regarding the achieved service and over-the-air throughput per used, it was verified that, for both sub-channelization schemes, service throughput is higher for WWW in detriment of the SISO mode. Service throughput is also higher if the AMC sub-channelization mode is used compared to PUSC. In particular, the average service throughput per user reaches 25 kbps and it is roughly 22 kbps for PUSC.
- For AMC with SISO the user service throughput reaches roughly 20 kbps and for PUSC with SISO it is roughly 16 kbps. OTA throughput is higher if SISO mode is implemented than with MIMO for both types of users. The reason for such performance difference between MIMO and SISO is that the Alamouti STBC scheme attenuates the variations in the channel amplitude, due to the mechanism of spatial diversity. As a result, a reduction in the opportunistic gain occurs for the utility based packet scheduling. This is more significant for WWW traffic model, since the scheduling of users in the edge of the cell, with the most robust MCS scheme, incurs to a significant higher load from the transmission of the large sized packets from WWW traffic model. The reduction in the OTA throughput

is more significant for PUSC sub-channelization scheme compared to AMC, from MIMO to SISO, due to multi-user diversity gain over frequency for AMC.

It was verified that, for both sub-channelization schemes, service throughput is higher for WWW users if the AMC sub-channelization mode is used, compared to PUSC. As sub-carrier gains are correlated in each sub-channel defined according to the AMC sub-channelization scheme, the decrease in channel variability affects all sub-carriers in each sub-channel by roughly the same way. The difference in performance for both sub-channelization schemes is less significant if MIMO channel is used.

As the system is overloaded for WWW, the AMC sub-channelization mode is more efficient in the utilization of the available radio resources. This efficiency results from the multi-user diversity gain over frequency associated to this sub-channelization mode. We could expect to see an even higher gain in system capacity with the use of the full queue traffic model in conjunction with an opportunistic packet scheduler, such as the maximum C/I, which does not take into account QoS requirements, as the utility algorithm does. As the system is under-loaded for VoIP users there is no resulting multi-user diversity gain over frequency, even with the AMC sub-channelization mode and both sub-channelization modes result into similar behaviour.

10.10Future Research

There exist multiple lines of investigation to continue the research carrier out in this work.

One area of particular interest is the investigation of the performance of the utility-based packet scheduling algorithm for scenarios where mobility and handovers are implemented. In particular handovers incur in an increase in the latency of the service provided which can affect the performance of the utility algorithm. New types of utility functions must be investigated which consider the effect of postponing packet transmission whenever a given user is performing a handover

Also, the uplink connection was not implemented and simulated in the system level simulation platform deployed so far. In particular, Mobile WiMAX implements different schemes for allocating bandwidth for data transmission in the uplink connection. These mechanisms depend on the type of traffic model used and are of fundamental importance in the standard for the provision of QoS. Therefore, it is important to implement the transmission chain for uplink connections in the tool and to modify the implemented utility schedulers, in order to include these QoS mechanisms and also to implement the signalling messages needed in the communication protocol, which implement these mechanisms for QoS control.

Besides the point to multi point configuration, Mobile WiMAX standard defines also the mesh mode where all nodes in the network communicate with each other. Another future work could

be to investigate new versions of the utility-based scheduling algorithm and modifications into the basic DRA architecture for working in a mesh network scenario.

All implementations performed up to now considered only technical and performance measures in taking scheduling decisions. By technical it is meant parameters regarding QoS requests such as the maximum allowable delay bound for packets generated from a specific type of service application or a minimum required service throughput for example. But other constraints could be taken into consideration, such as the amount of revenue resulting for the service provider in accepting to serve a new user, and the cost incurred to already active users, which have their QoS affected due to the reduction on the amount of radio resources available for transmission, or due to the increase in the amount of interference in the system resulting from the new user. These parameters can be included in the set of input parameters for the scheduler's decisions according to a game theoretical approach. New and existing users as well as the service provider can be considered as players in the game which take place in a competitive scenario where players do not cooperate. The idea is to maximize the provider's revenue while satisfying the QoS requests from the new user and limiting the level of degradation in the QoS provided to existing ones.

In all simulations conducted in this work the influence of the signalling overhead was not evaluated to the level of detail it deserves. Future research could focus on the performance achieved by the packet scheduler and resource allocation in a scenario with dynamic overhead, depending on the amount of users and on the type of traffic models used.

Finally, properly designed connection admission control and congestion control algorithms can be also included in the radio resource manager used so far, together with the proposed DRA, in order to improve network efficiency and maximize the number of satisfied users, namely for the scenarios where congestion occurs.

References

[1] Antón-Haro, C.; Svedman, P.; Bengtsson, M.; Alexiou A.; Gameiro, A.; “Cross-layer scheduling for multi-user MIMO systems”, IEEE Comm. Magazine, vo. 44, no. 9,Sept. 2006, pp. 39-45.

[2] Ajib W.; Haccoun, D.; “An Overview of Scheduling Algorithms in MIMO-Based Fourth-Generation Wireless Systems”, IEEE Network Magazine, September/October 2005, pp. 43-47.

[3] Alex S.P.; Jalloul, L.M.; “Performance Evaluation of MIMO in IEEE802.16e/WiMAX”, IEEE Journal on Selected Topics on Signal Processing, vol. 2, no. 2, April 2008.

[4] Catreux, S.; Erceg, V.; Gesbert D.; Heath, R.W.; “Adaptive modulation and MIMO coding for broadband wireless data networks”, IEEE Communications Magazine, vol. 40, no. 6, June 2002, pp. 108-115.

[5] “Vision, framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000". DRAFT NEW RECOMMENDATION ITU-R M. [IMT-VIS] [DOC. 8/110] DRAFTING GROUP PROPOSED MODIFICATIONS; 5th of February 2003.

[6] Chuang J.; Sollenberger N.; “Beyond 3G: wideband wireless data access based on OFDM and dynamic packet assignment”, IEEE Communications Magazine, vol. 38, no. 7, July 2000, pp. 78-87.

[7] 3GPP. Technical Specification Group Radio Access Network. High Speed Downlink Packet Access; Overall UTRAN Description. (3GPP TR 25.855 version 5.0.0).

[8] 3GPP2 (2002), 1xEV-DO evaluation methodology. WG5 Evaluation Ad-Hoc.

[9] 3GPP TSG-RAN1#48 R1-070674, LTE physical layer framework for performance verification, Feb. 2007.

[10] Ghosh, A.; Wolter, D.R.; Andrews, J. G.; Chen, R.; “Broadband wireless access with WiMAX/802.16: current performance benchmarks and future potential”, IEEE Communications Magazine, vol. 43, no. 129, Feb. 2005, pp. 129-36.

[11] Eklund, C.; Marks, R. B.; Stangwood K. L.; Wang, S.; “IEEE Standard 802.16: A Technical Overview of the Wireless MAN Air Interface for Broadband Wireless Access”, IEEE Communications Magazine, June 2002, pp. 98-107.

[12] Yaghoobi, H.; “Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN”, Intel Technology Journal, 8, 201, 2004.

[13] Cicconetti, C.; Lenzini, L.; Mingozzi, E.; “Quality of Service support in IEEE 802.16 networks”, IEEE Network Magazine, vol. 20, March 2006, pp. 50-55.

[14] Yile Guo; Chaskar, H.; “Class-based quality of service over air interfaces in 4G mobile networks”, IEEE Communications Magazine, vol. 40, no. 3, March 2002, pp. 132-137.

[15] Koffman I.; Roman, V.; “Broadband wireless solutions based on OFDM access in IEEE802.16”, IEEE Communications Magazine, vol. 40, no. 4, Apr. 2002, pp. 96-103.

[16] Srivastava, V.; Motani, M.; “Cross-Layer Design: A Survey and the Road Ahead”, IEEE Communications Magazine, Dec. 2005, vol. 43, no. 12, pp. 112-119.

[17] Xi Zhang; Jia Tang; Hsiao-Hwa Chen; Song Ci; Guizani, M.; "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks", IEEE Communications Magazine, vol. 44, no. 1, Jan. 2006, pp. 100-106.

[18] Kawadia, V.; Kumar, P. R. ; "A cautionary perspective on cross-layer design", IEEE Wireless Communications, vol. 12, no. 1, pp. 3-11, Feb. 2005.

[19] Yaxin Cao and Li, V.O.K.; "Scheduling algorithms in broadband wireless networks", Proceedings of the IEEE, vol. 89, no. 1, Jan. 2001, PP. 76-87.

[20] Zhu H.J.; Hafez, R.H.M.; "Scheduling schemes for multimedia service in wireless OFDM systems", IEEE Wireless Communications, vol. 14, no. 5, Oct. 2007, pp. 99-105.

[21] H. Fattah; C. Leung; "An overview of scheduling algorithms in wireless multimedia networks", IEEE Wireless Communications, vol. 9, no. 5, Oct. 2002, pp. 76-83.

[22] Gosh, A; Wolter, D.R.; Andrews, J.G.; Chen, R.; "Broadband Wireless Access with WiMAx/802.16: Current Performance Benchmarks and Future Potential", IEEE Communications Magazine, Feb 2005, vol. 43, no. 2, pp. 129-136.

[23] Bertsekas, D.; Gallager, R.; "Data Networks", 2nd edition, Prentice-Hall, 1992.

[24] Rapparport, T.; "Wireless Communications – Principles and Practice", Prentice-Hall, 2º edition.

[25] Goldsmith, A.J.; Varaiya, P.P.; "Capacity of fading channels with channel side information", IEEE Transactions on Information Theory, Nov. 1997, vol. 43, no. 6, pp. 1986-1992.

[26] Knopp, R.P.; Humblet, P.A.; "Information Capacity and Power Control in Single-Cell Multiuser Communications", Proceedings of the International Conference on Communications, (ICC), June 1995, Seattle, USA.

[27] Tse, D.; "Forward link multiuser diversity through proportional fair scheduling", Aug. 1999, presentation at Bell Labs.

[28] Floyd, S.; "TCP and Explicit Congestion Notification", ACM Computer. Communications Review, vol. 24, Oct. 1994, pp. 10-23.

[29] Balakrishnan, H. *et al*.; "A comparison of mechanisms for improving TCP performance over wireless links", IEEE/ACM Transactions on Networking, Dec 1997.

[30] Shakkottai, S.; Rappaport, T.S.; Karlsson, P.C.; "Cross-layer design for wireless networks", IEEE Communications Magazine, vol. 41, no. 10, pp. 74-80, Oct. 2003.

[31] Girod, B.; Kalman, M.; Liang Y.J. ; Zhang, R.; "Advances in Channel-Adaptive Video Streaming", IEEE International Conference on Image Processing (ICIP' 02), vol. 1, pp. 19-112, Sept. 2002.

[32] Liu H.; Zarki, M.; "Adaptive source rate control for real-time wireless video transmission", Mobile Networks and Applications, vol. 3 no. 1, pp. 49-60, 1998.

[33] Aramvith, S.; Pao, I.M.; Sun, M.T.; "A rate-control scheme for video transport over wireless channels", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 5, pp. 569-580, 2001.

[34] Hsu, C.Y.; Ortega, A.; Khansari, M.; "Rate control for robust video transmission over burst-error wireless channels", IEEE Journal on Selected Areas in Communications, vol. 17, no. 5, pp. 756-773, 1999.

[35] Carneiro, G.; Ruela J.; Ricardo, M.; "Cross-layer design in 4G wireless terminals", IEEE Wireless Communications, vol. 11, no. 2, April 2004, pp. 7-13.

[36] Liu X.; Chong, K.P.; "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks", IEEE Journal on Selected Areas in Communications, vol. 19, pp 2053-2064, Oct. 2001.

[37] Ji Z. *et al*; "Exploiting Medium Access Diversity in Rate Adaptive Wireless LANs", Proceedings of ACM Annual International Symposium on Mobile Computing and Networking, Oct. 2004.

[38] Dimic, G.; Sidiropoulos, N.D.; Zhang, R.; "Medium Access Control – Physical Cross-Layer Design", IEEE Signal Processing Magazine, vol. 21, no. 5, Sept. 2004, pp. 40-50.

[39] Foukalas, F.; Gazis, V.; Alonistioti, N.; "Cross-Layer Design Proposals for Wireless Mobile Networks: A Survey and Taxonomy", IEEE Comm. Surveys&Tutorials, 1st Quarter of 2008.

[40] Kwon, T. *et al* ; "Design and implementation of a simulator based on a cross-layer protocol between MAC and PHY Layers in a WiBro compatible IEEE802.16e OFDMA system", IEEE Communications Magazine, vol. 43, no. 12, Dec. 2005, p. 136-146.

[41] Ferrús, R. *et al.*; "Cross-layer protocol strategy for UMTS downlink enhancement", IEEE Communications Magazine, vol. 43, no. 6, June 205, pp. 24-28.

[42] Balakrishnan, H.; Katz, R.H.; "Explicit Loss Notification and Wireless Web Performance", Proceeding of GLOBECOM Internet Mini-Conference, Sydney, Australia, Nov. 1998.

[43] Liu, Q.; Zhou S.; Giannakis, G.B.; "Cross-layer Combining of Adaptive Modulation and Coding with Truncated ARQ over Wireless Links", IEEE Transactions on Wireless Communications, vol. 3, no. 5, Sept. 2004.

[44] Wu, D.; Ci.; S.; "Cross-Layer Design for Combining Adaptive Modulation and Coding with Hybrid ARQ", Proceedings of International Conference in Communications and Mobile Computing, July 3-6, 2006, pp. 147-52.

[45] Liu, Q.; Zhou S.; Giannakis, G.B.; "Cross-layer scheduling with prescribed QoS guarantees in advanced wireless networks", IEEE Journal on Selected Areas In Communications", vol. 23, no. 5, May. 2005.

[46] Merigeault, S.; Lamy, C.; “Concepts for Exchanging Extra Information between Protocol Layers Transparently for the Standard Protocol Stack”, Proceedings IEEE International Conference on Telecommunication, 2003 (ICT 2003), Feb., March 2003, pp. 981-985.

[47] Khan S.; *et al*., “Application-driven cross-layer optimization for video streaming over wireless networks”, IEEE Comm. Magazine, vol. 44, no. 1 2006, pp. 122-30.

[48] Chan, Y. S.; Modestino, J. W.; “A joint source coding-power control approach for video transmission over CDMA networks”, IEEE Journal on Selected Areas in Communications, vol. 21, no. 10, Dec.2003.

[49] IEEE, Standard 802.16e-2005 Part 16: Air interface for fixed and mobile broadband wireless access systems – Amendment for physical and medium access control layers for combined fixed and mobile operation in licensed band. December 2005.

[50] IEEE, Standard 802.16-2004, Part16: Air interface for fixed broadband wireless access systems, October 2004.

[51] WiMAX Forum, www.wimaxforum.org.

[52] IEEE 802.16, www.ieee802.16.com.

[53] Etemad, K.; “Overview of mobile WiMAX technology and evolution”, IEEE Communications Magazine, vol. 46, no. 10, Oct. 2008, pp. 31-40.

[54] Mobile WiMAX – Part I: A Technical Overview and Performance Evaluation, WiMAX Forum.

[55] Salvekar, A.; Sandhu, S.; Li, Q.; Vuong M.; Qian, X.; “Multiple-Antenna Technology in WiMAX Systems”, Intel Technology Journal, 8, 229, 2004.

[56] Liu, Q.; Wang, X.; Giannakis, G. B.; “A Cross-Layer Scheduling Algorithm with QoS Support in Wireless Networks”, IEEE Transactions on Vehicular Technology, vol. 55, no. 3, May 2006, pp. 839-46.

[57] Nie, C.; Tao, Z.; Mehta, N.B.; Molisch, A.F.; Zhang, J.; Kuze, T.; War, S.; “Antenna Selection for Next Generation IEEE 802.16 Mobile Stations” IEEE International Conference on Communications (ICC 2008), May 2008, pp. 3457-3462.

[58] Muquet, B.; Biglieri, E.; Sari, H.; “MIMO Link Adaptation in Mobile WiMAX Systems”, IEEE Wireless Communications and Networking Conference, 2007, (WCNC 2007), March 2007, pp1810-1813.

[59] Alex S. P.; Jallout, M.; A. L.; “Performance evaluation of MIMO in IEEE 802.16e/WiMAX”, IEEE Journal of Selected Topics in Signal Processing, vol. 2, no. 2, April 2008.

[60] Sung, S.; Hwang I. S.; Yoon, S.; “On the Gain of Data Rate Control in OFDMA Systems”, 1st International Workshop on Broadband Convergence Networks, p. 1, IEEE, 2006.

[61] Balachandran K.; *et al*; “Design and Analysis of an IEEE 802.16e-Based OFDMA Communication System”, Bell Labs Tech. Journal 11(4):53-73.

[62] Wang F.; Ghosh A.; Sankaran C.; Fleming P.; "WiMAX Overview and System Performance", Proceedings IEEE on Vehicular Technology Conference (VTC-2006), pp. 1-5, Fall.

[63] Wang, F.; Ghosh, A.; Sankaran, C.; Fleming, P.; Hsieh, F.; Benes, S.; "Mobile WiMAX systems: performance and evolution", IEEE Comm. Magazine, vol. 46, no. 10, Oct. 2008, pp. 41-49.

[64] Wang F.; Ghosh A.; Sankaran C.; Benes S.; "WiMAX System Performance with Multiple Transmit and Multiple Receive Antennas", Proceedings IEEE on Vehicular Technology Conference (VTC-2007), pp. 2807-2881, Spring.

[65] Hoymann C.; "Analysis and Performance Evaluation of the OFDM-Based Metropolitan Area Network IEEE802.16", Computer Networks, 49(3); pp. 341-363, Oct 2005.

[66] Liu P.; Berry R.; Honig M.L.; "Delay-Sensitive Packet Scheduling in Wireless Networks", Proceedings of the IEEE Wireless Communications and Networking, 2003, (WCNC 2003), vol. 3, March 2003, pp.1627-1632.

[67] Song G.; Li Y.; "Adaptive Resource Allocation Based on Utility Optimization in OFDM", Proceedings of IEEE Global Telecommunications Conference, GLOBECOM '03, Dec. 2003, vol. 2, pp. 586-590.

[68] Song G.; Cimini L.; Zheng H.; "Joint Channel-Aware and Queue-Aware Data Scheduling in Multiple-Shared Wireless Channels", Proceedings IEEE Wireless Communications and Networking Conference, 2004, (WCNC 2004), vol. 3, March 2004, pp 1939-1944.

[69] Song G.; Li, Y.; "Utility-based resource allocation and schedulers in OFDM-based wireless broadband networks", IEEE Communications Magazine, vol. 43, no. 112, pp 127, 2005.

[70] Song G.; Li, Y.; "Cross-Layer Optimization for OFDM Wireless Network – Part I and part II", IEEE Transactions on Wireless Communications, 4(2), 614, 2005.

[71] Huang, J.; Subramanian, V.; Agrawal, R.; Berry, R.; "Downlink Scheduling and Resource Allocation for OFDM Systems", 40th Annual Conference on Information Sciences and Systems, 2006.

[72] Draft IEEE 802.16m Evaluation Methodology, IEEE 802.16 Broadband Wireless Access Working Group, Oct. 2007.

[73] ETSI, "Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of the radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0)", TR 101 112 v3.2.0, April 1998.

[74] Gudmundson, M.; "Correlation Model for Shadow Fading in Mobile Radio Systems", Electronics Letters, vol. 27, pp. 2145-2146, Nov. 1991.

[75] Xiadong Cai; Georgios B.; Giannakis, G.; "A Two-Dimensional Channel Simulation Model For Shadowing Processes", IEEE Transactions on Vehicular Technology, vol. 52.no. 6, Nov. 2003.

[76] W.C. Jakes, Microwave Mobile Communications, Wiley, New York, 1974.

[77] Yunxin Li; Xiaojing Huang; "The generation of independent Rayleigh faders", Proceedings of the IEEE International Conference on Communications (ICC 2000), vol. 1, June 2000, pp. 18-22.

[78] ITU, "Guidelines for evaluation of radio transmission technologies for IMT-2000" Recommendations ITU-R M.1225, 1997.

[79] 3GPP R1-030224, Nortel Networks, "Update of OFDM SI simulation methodology".

[80] Medbo J.; Anderson H.; Schramm, P.; Asplund, H.; "Channel models for HIPERLAN/2 in different indoor scenarios", COST259 TD(98), Bradford, UK, April 23-24 1998.

[81] 3GPP TR 25.892 (2004) "Feasibility study for OFDM for UTRAN enhancement, V1.1.0.

[82] Wang, F.; Ghosh, A.; Love, R.; Stewart, K.; Ratasuk, R.; Bachu, R.; Sun, Y.; Zhao, Q.; "IEEE 802.16e system performance: analysis and simulations", IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2005, (PIMRC 2005), vol. 2. pp. 900-904.

[83] Hamalainen, S.; Slanina P.; Hartman, M.; Lappetelainen, A.; Holma, H.; Salonaho, O.; "A novel interface between link and system level simulations", Proceedings of ACTS Summit 1997, Oct. 1997.

[84] Brueninghaus, K.; Astély, D.; Saltzer, T.; Visuri, S.; Alexiou, A.; Karger, S.; Seraji, G.; "Link performance models for level simulations of broadband radio access systems", Proceedings IEE International Symposium on Personal, Indoor and Mobile Radio Communications, March, 2005.

[85] "Spatial channel model for multiple-input multiple-output simulations (Release 6), 3GPP TR 25.996, 2003-05.

[86] Calcev, G.; Chizhik, D.; Goransson, B.; Howard, S.; Huang, H.; Kogianthis, A.; Molisch, A.F.; Moustakas, A.L.; Reed, D.; Xu Hao; "A Wideband Spatial Channel Model for System-Wide Simulations", IEEE Transactions on Vehicular Technology, vol. 56, no. 2, March 2007.

[87] Alamouti, S.M.; "A simple transmit diversity technique for wireless communications", IEEE Journal on Selected Areas in Communications, vol. 16, no. 8, pp 1451-1458, Oct 1998.

[88] Foschini, G.; "Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multi-Element Antennas", Bell Labs Technical Journal, vol. 1, no. 2, pp. 41-59, 1996.

[89] Huang H.; Venkatesan S.; Kogiantis A.; Sharma N.; "Increasing the peak data rate of 3G downlink packet data systems using multiple antennas", Proceedings of the IEEE Vehicular Technology Conference, 2003, (VTC 2003) Spring, April 2003, pp. 311-315

[90] Floros C.; Thompson J. S.; McLaughlin S.; "Packet Scheduling in Wireless Systems Using MIMO Arrays and VBLAST Architecture", Proceedings of the IEEE Vehicular Technology Conference, 2007, (VTC 2007), Spring, April 2007, pp. 2781-2785.

[91] Hermosilla, C.; Valenzuela, R.; Ahumada, L.; Feick, R.; "Empirical Comparison of MIMO and Beamforming Schemes", IEEE International Conference on Communications, 2008, (ICC 2008), pp. 4226-4229.

[92] Sheng Chiu, C.; "System Simulations for MIMO Enhancements to HSDPA".

[93] Yongquan Qiang; Vivier, G.; Jing Yang; Ning Xu, "Inter-Cell Interference Modeling for OFDMA Systems with Beamforming", IEEE Vehicular Technology Conference, 2008, (VTC 2008), Fall, Sept. 2008, pp. 1-5.

[94] Belghith A.; Nuaymi, L.; "WiMAX Capacity Estimations and Simulation Results", IEEE Vehicular Technology Conference, 2008, (VTC 2008), Spring, May 2008, pp. 1741-1745.

[95] Maltsev, A.A.; Pudeyev, A.V.; "Multi-user frequency domain scheduling for WiMAX OFDMA", 16th IST Mobile and Wireless Communications Summit, 2007, July 2007, pp. 1-4.

[96] Ball, C. F.; Humburg, E.; Ivanov, K.; Treml, F.; "Comparison of IEEE802.16 WiMAX Scenarios with Fixed and Mobile Subscribers in Tight Reuse", available online: http://www.eurasip.org/Proceedings/Ext/IST05/papers/147.pdf.

[97] Fernekess, A.; Klein, A.; Wegmann, B.; Dietrich K.; Humburg, E.; "Performance of IEEE 802.16e OFDMA in Tight Reuse Scenarios", IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007, (PIMRC 2007), Sept. 2007, pp. 1-5.

[98] Ball, C.F.; Humburg, E.; Ivanov K.; Treml, F.; "Performance analysis of IEEE802.16e based cellular MAN with OFDM-256 in mobile scenarios", IEEE Vehicular Technology Conference, 2005, (VTC 2005), Spring, May/June 2005, pp. 2061-2066.

[99] Huy, D.T.P.; Rodriguez, J.; Gameiro, A.; Tafazolli, R.; "Dynamic Resource Allocation for Beyond 3G Cellular Networks", Journal on Wireless Personal Communications (2007) 43:1727-1740.

[100] Cicconetti, C.; Erta, A.; Lenzini L.; Mingozzi, E.; "Performance Evaluation of the IEEE 802.16 MAC for QoS Support", IEEE Transactions on Mobile Computing, vol. 6, no.1, Jan 2007, pp. 26-37.

[101] Braga A.; Rodrigues E.B.; Cavalcanti, F.R.P.; "Packet Scheduling for Voice over IP over HSDPA in Mixed Traffic Scenarios with Different End-to-End Delay Budgets", IEEE VI International Telecommunications Symposium (ITS2006), Sept 2006, Fortaleza, Brazil.pp754 – 759.

[102] Braga A.; Rodrigues E.B.; Cavalcanti, F.R.P.; "Novel Scheduling Algorithms Aiming for QoS Guarantees for VoIP over HSDPA", IEEE VI International Telecommunications Symposium (ITS2006), Sept 2006, Fortaleza, Brazil, pp. 94-99.

[103] Ball, C.F.; Treml, F.; Gaube, X.; Klein, A.; "Performance analysis of temporary removal scheduling applied to mobile WiMax scenarios in tight frequency reuse", IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2005, (PIMRC 2005), vol. 2, Sept. 2005, pp. 888-894.

[104] Srinivasan, R.; Timiri, S.; Davydov A.; Papathanassiou, A.; "Downlink Spectral Efficiency of Mobile WiMAX", IEEE Vehicular Technology Conference, 2005, (VTC 2007), Spring, April 2007, pp. 2786-2790.

[105] Bian Y.Q.; Nix, A.R.; "Mobile WiMAX: Multi-Cell Network Evaluation and Capacity Optimization", IEEE Vehicular Technology Conference, 2008, (VTC 2008), Spring, May 2008, pp. 1276-1280.

[106] Jain, R.; Chakchai, So-In; Al Tamimi, A-k; "System-level modeling of IEEE 802.16e mobile WIMAX networks: key issues", IEEE Wireless Communications Magazine, vol. 15, no. 5, Oct. 2008, pp73-79.

[107] WiMAX Forum, "WiMAX System Evaluation Methodology V2.1", July 2008.

[108] Gao, Y.; Zhang, X.; Jiang, Y.; "Performance Evaluation of Mobile WiMAX with Dynamic Overhead", EEE Vehicular Technology Conference, 2008, (VTC 2008), Fall, Sept. 2008, pp. 1-5.

[109] Chase D.; "Code Combining – A Maximum Likelihood Decoding approach for Combining an Arbitrary Number of Noisy Packets," IEEE Transactions on Communications, vol.33, pp.385 – 393, May 1985.

[110] Bender P.; Black P.; Grob M.; Padovani R.; Sindhushayana N Viterbi S, "CDMA/HDR: a bandwidth efficient high-speed wireless data service for nomadic users", IEEE Communications Magazine, vol. 38, no. 7, Jul. 2000, pp. 70-77.

[111] Viswanath P.; Tse D.N.C.; Laroia, R.; "Opportunistic beamforming using dumb antennas", IEEE Transactions on Information Theory, vol. 48, no. 6, June 2002, pp. 1277-1294.

[112] Wha Sook; Jeon Dong; Geun Jeong; Bonghoed Kim; "Packet scheduler for mobile Internet services using high speed downlink packet access", IEEE Transactions on Wireless Communications, vol. 3, no. 5, September 2004.

[113] Chingvao Huang; Hung-Hui Juan; Meng-Shiang Lin; Chung-Ju Chang; "Radio resource management of heterogeneous services in mobile WiMAX systems", IEEE Communications Magazine, vol. 14, no. 1, Feb. 2007", pp 20-28.

[114] Einhaus M.; Klein O.; Walke B.; Halfmann, R.; "MAC Level Performance Evaluation of Downlink Resource Allocation Strategies for an OFDMA System Based on IEEE 802.16", IEEE Vehicular Technology Conference, 2007, (VTC 2007) Spring, April. 2007, pp. 2796-2800.

[115] Ali, S.H.; Ki-Dong Lee; Leung, V.C.M.; "Dynamic resource allocation in OFDMA wireless metropolitan area networks", IEEE Communications Magazine, vo. 14, no. 1, Feb. 2007, pp. 6-13.

[116] Xhafa, A. E.; Kangude S.; Xiaolin Lu; "MAC Performance of IEEE 802.16e", IEEE Vehicular Technology Conference, 2005, (VTC-2005), Fall, vol. 1, Sept 2005, pp. 685-689.

[117] Lengoumbi, C.; Godlewski, P.; Martins, P.; “Subchannelization Performance for the Downlink of a Multi-cell OFDMA System”, IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, 2007, (WiMOB 2007), Oct. 2007.

[118] Chunchang, T.; Jing J.; Xin, Z.; “Evaluation of Mobile WiMAX System Performance”, IEEE Vehicular Technology Conference, 2008, (VTC 2008), Fall, Sept. 2008, pp. 1-5.

[119] Najah Abu Ali; Pratik Dhrona; Hossam Hassanein; “A performance study of scheduling algorithms in point-to-multipoint WiMAX networks”, .Computer Communications Magazine, vol. 32, no. 3, Feb. 2009, pp. 511-521.

[120] Sook Jeon, W.; Geun Jeong, D.; “Combined Connection Admission Control and Packet Transmission Scheduling for Mobile Internet Services”, IEEE Transactions on Vehicular Technology, vol. 55, no. 5, Sept 2006, pp. 1582-93.

[121] Shenker S.; “Fundamental design issues for the future Internet”, IEEE Journal on Selected Areas in Communications, vol. 13, no. 7, Sep. 1995, pp. 1176-1188.

[122] Wang B.; Pedersen, K.I.; Kolding T.E.; Mogensen, P.E.; “Performance of VoIP on HSDPA”, IEEE Vehicular Technology Conference, 2005, (VTC 2005) Spring, vo. 4., May/June 2005, pp. 2335-2339.

[123] Persson, F.; “Voice over IP Realized for the 3GPP Long Term Evolution”, IEEE Vehicular Technology Conference, 2007, (VTC 2007) Fall, Setp./Oct. 2007, pp. 1436-1440.

[124] Puttonen, J.; Henttonen, T.; Kolehmainen, N.; Aschan, K.; Moision M.; Kela, P.; “Voice-Over-IP Performance in UTRA Long Term Evolution Downlink”, IEEE Vehicular Technology Conference, 2008, (VTC 2008) Spring, May 2008, pp. 2502-2506.

[125] Hernandez-Valencia, E.J.; Chuah, M.C.; “Transport delays for UMTS VoIP”, IEEE Wireless Communications and Networking Conference 2000, (WCNC 2000), vol. 3, Sept. 2000, pp 1552-1556.

[126] Avidor, D.; Ling J.; Papadias, C.; “Jointly opportunistic beamforming and scheduling (JOBS) for downlink packet access”, IEEE International Conference on Communications, 2004, (ICC 2004), June 2004, vol. 5, pp 2959-2964.

[127] Jalali A.; Padovani R.; Pankaj R.; “Data Throughput of CDMA-HDR- A High Efficiency-High Data Rate Personal Communication System”, Proceedings of the IEEE Vehicular Technology Conference (VTC 2000), vol. 3, pp 1854-1857, May 2000.

[128] Wengerter, C.; Ohlhorst, J.; Golitschek A.; Elbwart, A.G.E.; “Fairness and throughput analysis for generalized proportional fair frequency scheduling in OFDMA”, IEEE Vehicular Technology Conference, 2005, (VTC 2005), Spring, May/June 2005, vol. 3, pp. 1903-1907.

[129] Andrews, M.; Kumaran, K.; Ramanan, K.; Stolyar A.; Whiting, P. Vijayakumar, R. “Providing quality of service over a shared wireless link”, IEEE Comm. Magazine, vol. 39, no. 2, Feb. 2001, pp. 150-154.

[130] Sanjav S.; Alexander L.S.; “Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data In HDR”, Proceedings of International Teletraffic Congress (ITC), 2001.

[131] Young-June Choi; Jin-Ghoo Choi; Saewoong Bahk; “Upper-level scheduling supporting multimedia traffic in cellular data networks”, Computer Networks: The International Journal of Computer and Telecommunications Networking, vol. 51, issue 3, Feb. 2007, pp 621-631.

[132] Y-Seok Kim; “VoIP Service on HSDPA in Mixed Traffic Scenarios”, IEEE International Conference on Computer and Information Technology (CIT’06), 2006.

[133] Rittenhouse, G.; Haitao Zheng; “Providing VoIP service in UMTS-HSDPA with frame aggregation”, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, (ICASSP ’05), vol. 2, March 2005, pp. 157-160.

[134] Lunden, P.; Kuusela, M.; “Enhancing Performance of VoIP over HSDPA”, IEEE Vehicular Technology Conference, 2007, (VTC 2007), Spring, April 2007, pp. 825-829

[135] Shuping Chen; Wengbo Wang; Jin. Han; “Providing VoIP Service over TD-SCDMA HSDPA”, IEEE Vehicular Technology Conference, 2008, (VTC 2008), Spring, May 2008, pp. 2066-2070.

[136]ITU Recommendations G.114, “One-way Transmission Time, 2003.

[137] Hua Wang; “Priority-Based Resource Allocation for RT and NRT Traffics in OFDMA Systems”, International Conference on Wireless Communications, Networking and Mobile Computing, 2007, WiCom 2007, Sept. 2007, Sept. 2007. pp. 791-794.

[138] Ofugi, Y.; Abeta, S.; Sawahashi, M.; “Fast packet scheduling algorithm based on instantaneous SIR with constraint condition assuring minimum throughput in forward link”, .IEEE Wireless Communications and Networking, 2003, (WCNC 2003), March 2003, vol. 2, pp. 860-865.

[139] Ofugi, Y.; Abeta S.; Sawahashi, M.; “Unified fast packet scheduling method considering fluctuations in frequency domain in forward link for OFCDM broadband packet wireless access”, IEEE Vehicular Technology Conference, 2004, (VTC 2004), Fall, Sept. 2004, vol. 4, pp. 2724-2729.

[140] Ofugi, Y.; Abeta S.; Sawahashi, M.; “Unified Packet Scheduling Method Considering Delay Requirements in OFCDM Forward Link Broadband Wireless Access”, IEICE Trans. Communications, vol. E88-B, no. 11 Jan. 2005, pp. 170-182.

[141] Ofugi, Y.; Morimoto, A.; Atarashi H.; Sawahashi, M.; “Sector throughput using frequency-and-time domain channel-dependent packet scheduling with channel prediction in OFDMA downlink packet radio access”, IEEE Vehicular Technology Conference, 2005, (VTC 2005), Fall, Sept. 2005, vol. 3, pp. 1589-1593.

[142] Yahiya, T.; Beylot A.; Pujolle, G.; “Cross-Layer Multiservice Scheduling for Mobile WiMAx Systems”, IEEE Wireless Communications and Networking Conference, w008, (WCNC 2008), March/April 2008, pp.1531-1535.

[143] Wang, H.; Iversen, V.; “Hierarchical Downlink Resource Management Framework for OFDMA Based WiMAX Systems”, IEEE Wireless Communications and Networking Conference, 2008, (WCNC 2008), March/April 2008, pp. 1709-1715.

[144] Mehri Mehrjoo; Xuemin Shen; Naik, K.; “A Joint Channel and Queue-Aware Scheduling for IEEE 802.16 Wireless Metropolitan Area Networks”, IEEE Wireless Communications and Networking Conference, 2007, (WCNC 2007), March 2007, pp. 1549-1553.

[145] Hoi Kim, D.; Han Ryu, B.; Gu Kang; “Packet Scheduling Algorithm Considering a Minimum Bit Rate for Non-realtime Traffic in an OFDMA/FDD-Based Mobile Internet Access System”, ETRI Journal, Volume 26, Number 1, Feb. 2004.

[146] Xining Zhu; Jiachhuan Huo; Xiaoxi Xu; Cjunxiu Xu; Wei Ding; “QoS-Guaranteed Scheduling and Resource Allocation Algorithm for IEEE 802.16 OFDMA System”, IEEE International Conference on Communications, 2008, (ICC 2008), May 2008. pp. 3463-3468.

[147] Torres, J.; Morillo-Velarde, V. Soret, B.; Aguayo-Torres, M.C.; Entrambasaguas, J.T.; “Cross-layer user multiplexing algorithms evaluation in MIMO OFDM wireless systems”, 16th Mobile and Wireless Communications Summit, 2007, July 2007, pp. 1-5.

[148] Kontouris, M.; Pandharipande, A.; Hojin Kim, Gesbert, D.; “QoS-based user scheduling for multiuser MIMO systems“, IEEE Vehicular Technology Conference, 2005, (VTC 2005) May/June 2005, vol. 1, pp. 211-215.

[149] Mugen Peng; Wenbo Wang; “Advanced Scheduling Algorithms for Supporting Diverse Quality of Services in IEEE 802.16 Wireless Metropolitan Area Networks”, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007, (PIRMC 2007) Sept. 2007, pp. 1-6.

[150] Shuo Chao; Ma Nan; Wu Tong; Wang Ying; Zhang Ping; “QoS Differentiation Adaptive Retransmission Limits ARQ for IEEE 802.16e BWA System”, IEEE Vehicular Technology Conference, 2007, (VTC 2007), Fall, Sept/Oct. 2007, pp. 1533-1538.

[151] Badia, L.; Baiocchi A.; Todini, A.; Merlin, S.; Pupolin, S.; Zanella, A.; Zorzi, M.; “On the impact of physical layer awareness on scheduling and resource allocation in broadband multicellular IEEE 802.16 systems”, IEEE Wireless Communications, vol. 14, no. 1, Feb. 2007, pp. 36-43.

[152] Sourav Pal; Mainak Chatterjee; Sajal K. Das; “A Two-level Resource Management Scheme on Wireless Networks Based on User-Satisfaction”, Mobile Computing and Communications Review Magazine, vol. 9, no. 4.

[153] Haitao Lin; Mainak Chatterjee; Sajal K. Das; “ARC: An Integrated Admission and Rate Control Framework for Competitive Wireless CDMA Data Networks Using Noncooperative Games”, IEEE Transactions On Mobile Computing, vol. 4, no. 3, May/June 2005.

[154] Jiang *et. al.*; “Max-utility wireless resource management for Best-Effort traffic”, IEEE Transactions on Wireless Communications, vol. 4, no. 1, Jan 2005, pp.100-111

[155] L. Zhe, L.; Z. Je., W. Wu, "A Simplified Layered QoS Scheduling Scheme in OFDM Networks", Proceeding of IEEE Vehicular Technology Conference, 2007, (VTC 2007), Fall, Sept. 2007, pp. 1842-1846.

[156] Choi, Y.J.; "Delay-Sensitive Packet Scheduling for a Wireless Link", IEEE Transactions on Mobile Computing, vol. 5, no. 10, Oct. 2008, pp 1374-1382.

[157] Sang, A.; Wang, X.; Madihian M.; Gitlin, R.; "A Flexible Downlink Scheduling Scheme in Cellular Packet Data Systems", IEEE Transactions On Wireless Communications, vol. 5, no. 3, March 2006.

[158] Sang, A.; Wang, X.; Madihian M.; Gitlin, R.; "Downlink scheduling schemes in cellular packet data systems of multiple-input multiple-output antennas", Global Telecommunications Conference, 2004, GLOBECOM '04, Nov. 2003, vol. 6, pp. 4021-4027.

[159] Ryu, S.; Ryu, B-Han; Seo, H.; Dhin M.; Park, S.; "Wireless Packet Scheduling Algorithm for OFDMA System Based on Time-Utility and Channel State", ETRI Journal, vol. 27, no. 6, Dec. 2005, pp. 777-787.

[160] Ryu, S.; Ryu, B.; Seo H.; Shin, M.; "Urgency and Efficiency based Packet Scheduling Algorithm for OFDMA wireless System", IEEE International Conference on Communications, 2005, (ICC 2005), vol. 4, May 2005, pp. 2779-2785.

[161] Lei, H.; Fan C.; Zhang, X.; Yang, D.; "QoS Aware Packet Scheduling Algorithm for OFDMA Systems", IEEE Vehicular Technology Conference, 2007, (VTC 2007), Fall, Sept. 2007. pp. 1877-1881.

[162] Niida, S.; Inoue T.; Takeuchi, Y.; "Fundamental Analysis of Two-layered Scheduling Algorithm for a Wireless Packet System", IEEE Vehicular Technology Conference, 2006, (VTC 2006), Spring, vol. 1, May 2006, pp. 446-450.

[163] Park, W.H.; Cho S.; Bahk, S.; "Scheduler Design for Multiple Traffic Classes in OFDMA Networks", IEEE International Conference on Communications, 2006, vol. 2, June 2006, pp. 790-795.

[164] Khattab A.K.F.; Elsayed, K.M.F.; "Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks", International Symposium in a World of Wireless, Mobile and Multimedia Networks, 2006, WoWMoM 2006.

[165] Oh, J.; Hwang J.; Han Y.; "A Packet-by-Packet Scheduling Algorithm for Wireless Multimedia Systems", IEEE Vehicular Technology Conference, 2007, (VTC 2007), Fall, Sept./Oct. 2007, pp. 1782-1786.

[166] Ali-Yahiya; Beylot T.; A.-L Pujolle, G.; "Channel Aware Scheduling for Multiple Service Flows in OFDMA Based Mobile WiMAX Systems", IEEE Vehicular Technology Conference, 2008, (VTC 2008), Fall, Sept. 2008, pp. 1-5.

[167] Wand, Y.; Paranchych D. Li, X.; "Enhanced scheduling schemes to integrate QoS support in 1xEVDO", IEEE Vehicular Technology Conference, 2004, (VTC 2004), Fall, pp. 2668-2672.

[168] Lihua Wan; Wenchao Ma; Zihua Guo; "A Cross-layer Packet Scheduling and Subchannel Allocation Scheme in 802.16e OFDMA System", IEEE Wireless Communications and Networking Conference, 2007, (WCNC 2007), March 2007, pp. 1865-1870.

[169] Yariya, T.; Beylot, A.; Pujolle, G.; "Radio Resource Allocation in Mobile WiMAX Networks using Service Flows", IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007, (PIMRC 2007), Sept. 2007, pp. 1-5.

[170] Sang, A.; Wang X.; Madihian, M.; "Real-Time QoS in Enhanced 3G Cellular Packet Systems of a Shared Downlink Channel", IEEE Transactions on Wireless Communications, vol. 6, no. 5, May 2007.

[171] Xu Ning; Guillaume; V. Zhou; Wen Yongquan, Q.; "A Dynamic PF Scheduler to Improve the Cell Edge Performance", IEEE Vehicular Technology Conference, 2008, (VTC 2008), Fall, Sept. 2008, pp. 1-5.

[172] Jin-Yup Hwang; Youngnam Han; "An Adaptive Traffic Allocation Scheduling for Mobile WiMAX", IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2007, (PIMRC 2007), Sept. 2007, pp. 1-5.

[173] Ching Yao Huang; Chieh-Yao Chang; Po-Han Chen; Ming-Hsien Wu; "Performance Evaluation of SDMA based Mobile WiMAX Systems" IEEE International Wireless Communications and Mobile Computing Conference, 2008, (IWCMC 2008), Aug. 2008, pp. 1042-1046.

[174] Hujun Yin; "Performance of Space-Division Multiple-Access (SDMA) With Scheduling", IEEE Transactions on Wireless Communications, vol. 1, no. 4, Oct. 2003.

[175] Hoymann C.; "MAC Layer Concepts to Support Space Division Multiple Access in OFDM based IEEE 802.16", Wireless Personal Communications Magazine, pp. 363-385, Sep. 2006.

[176] Hoymann C.; Meng H.; Ellenbeck J.; "Influence of SDMA-Specific MAC Scheduling on the Performance of IEEE 802.16 Networks", Proceedings of the 12th European Wireless Conference 2006, Athens, Greece, April 2006.

[177] Pabst R.; Ellenbeck J.; Schinnenburg M.; Hoymann C.; "System Level Performance of Cellular WiMAX IEEE802.16 with SDMA enhanced Medium Access", IEEE Wireless Communications and Networking Conference, 2007, (WCNC 2007), March 2007, pp. 1820-1825, March 2007.

[178] Ying Jung Zhang; Khaled Ben Letaief; "An Efficient Resource-Allocation Scheme for Spatial Multiuser Access in MIMO/OFDMA Systems", IEEE Transactions in Communications, col. 53, no. 1, Jan 2005.

[179] Spencer Q. H.; Swindlehurst, A.L.; “Channel allocation in multi-user MIMO wireless communications systems”, IEEE International Conference on Communications, 2004, (ICC 2004), vol. 5, June 2004, pp. 3035-30-39.

[180] Kai Sun; Ying Wang; Tan Wang; Zixiong Chen; Guona Hu; “Joint Channel-Aware and Queue-Aware Scheduling Algorithm for Multi-User MIMO-OFDMA Systems with Downlink Beamforming”, IEEE Vehicular Technology Conference, 2008, (VTC 2008), Fall, Sept. 2008, pp. 1-5.

[181] Shi S.; Amano Y.; Kawamoto K.; Yamagughi A.; Inoue T.; Takeuchi Y.; Kawazawa T.; “Forward link throughput evaluation of a SDMA based wireless packet cellular system”, IEEE Vehicular Technology Conference, 2004, (VTC 2004) Fall, vol. 2, Sept. 2004, pp. 949-953.

[182] 3GPP Technical Report 25.814, “Physical Layer Aspects for Evolved UTRA”.

[183] Pokhariyal, A.; Kolding T.E.; Mogensen, P.E.; “Performance of Downlink Frequency Domain Packet Scheduling for the UTRAN Long Term Evolution”, IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2006, (PIMRC 2006), Sept. 2006, pp. 1-5.

[184] Na Wei; Pokhariyal, A.; Sorensen, T.B.; Kolding T.E.; Mogensen, P.E.; “Performance of MIMO with Frequency Domain Packet Scheduling in UTRAN LTE Downlink”, IEEE Vehicular Technology Conference, 2007, (VTC 2007), Spring, April 2007, pp. 1177-1181.

[185] Pokhariyal, A.; Pedersen, K. I.; Monghal, G.; Kovacs, I.Z.; Rosa, C.; Kolding T.E.; Mogensen, P.E.; “HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution”, IEEE Vehicular Technology Conference, 2007, (VTC 2007), Spring, April 2007, pp. 1177-1181.

[186] Pedersen, K. I.; Monghal, G.; Kovacs, I.Z.; Kolding, T.E.; Pokhariyal, A.; Frederiksen F.; Mogensen, P.E.; “Frequency Domain Scheduling for OFDMA with Limited and Noisy Channel Feedback”, IEEE Vehicular Technology Conference, 2007, (VTC 2007), Fall, Sept./Oct. 2007, pp. 1792-1796.

[187] Monghal, G.; Pedersen, K.I.; Kovacs, I.Z.; Mogensen, P.E.; “QoS Oriented Time and Frequency Domain Packet Schedulers for the UTRAN Long Term Evolution”, IEEE Vehicular Technology Conference, 2008, (VTC 2008), Spring, May 2008, pp. 2532-2536.

[188] Kian Chung Beh; Armour S.; Doufexi, A.; “Joint Time-Frequency Domain Proportional Fair Scheduler with HARQ for 3GPP LTE Systems” IEEE Vehicular Technology Conference, 2008, (VTC 2008), Fall, Sept 2008, pp. 1-5.

[189] Assaad, M.; Mourad, A.; “New Frequency-Time Scheduling Algorithms for 3GPP/LTE-like OFDMA Air Interface in the Downlink”, IEEE Vehicular Technology Conference, 2008, (VTC 2008), Spring, May 2008, pp. 1964-1969.

[190] Nonchev, S.; Venalainen J.; Valkama M.; “New Frequency Domain Packet Scheduling Schemes for UTRAN LTE Downlink”, ICT Mobile Summit 2008.

[191] Andrews, J.G.; Ghosh, A.; Muhamed, R; “Fundamentals of WiMAX: Understanding Broadband Wireless Networkin”, Prentice Hall.

[192] 3GPP TSG-RAN 1, Nortel Networks, “OFDM exponential effective SIR mapping validation, document R1-04-0089, Espoo, Finland, Jan 2004.

[193] 3GPP TSG-RAN 1, Ericsson, “system level evaluation of OFDM – further considerations, Document R1-03-1303, Lisbon, Portugal, Nov 2003.

[194] IEEE802.16 Broadband Wireless Access Working Group, Alvarion, “CINR measurement using the EESM method, document IEEE C802.16e-05/141r3.

[195] Caire G., Tarisco G. and Bigliei E., “Capacity of bit-interleaved channels”, Electronic Letters, vol. 32, no.12, pp 1060-1061, June 1996.

[196] IST WINNER Project WP2, “Link to System Interface Methodology v2.0, September 2004.

[197] Tsai S.; Soong, A.; “Effective SIR mapping for modeling frame error rates in multiple-state channels, 3GPP2 submission C30-20030429010, Apr. 2003.

[198] Lampe M.; Giebel, T.; Rohling, H.; Zirvas W.; “Per-prediction for PHY mode selection in OFDM communication systems”, Global Telecommunications Conference, 2003, GLOBECOM ‘03, Dec. 2003, vol. 1, pp. 25-29.

[199] ETSI SMG2 Universal Mobile Telecommunications System (UMTS); “Selection procedures for the choice of radio transmission technologies of the UMTS”, TR 101 112.

[200] Lei, H.; Zhang, L.; Yang, D.; “A Packet Scheduling Algorithm Using Utility Function for Mixed Services in the Downlink of OFDMA Systems” Proceedings of the IEEE VTC Conference, Fall Sep. 2007, pp 1664-1668.

[201] Huang, V.; Zhuang, W.; “QoS-Oriented Packet Scheduling for Wireless Multimedia CDMA Communications”, IEEE Transactions on Mobile Computing, vol. 3, pp. 73-85, Jan-Feb 2004.

Annex A

Methods for Effective SINR Mapping

A.1 Introduction

In the following sub-sections some proposed methods for SINR mapping proposed in the literature are presented.

A.1.1 One-Dimensional Data Compression and Mapping

A.1.1.1 Mean Instantaneous Capacity Mapping Method (MIC)

In the Mean Instantaneous Capacity Mapping Method (MIC) compression method [191] the mapping is performed by computation of the AWFGN channel capacity for each one of the N SINR values corresponding to the data sub-carrier set into which the data block is mapped into. Then an average over the set of capacity values is performed in order to find the effective SINR. The compressed SINR value is given by equations (1) and (2).

$$C(SINR_{eff}) = \frac{1}{N}\left(\sum_{k=1}^{N} C(SINR_k)\right) \quad (1)$$

$$C(SINR_k) = \log_2(1 + QSINR_k) \quad (2)$$

The parameter Q is a channel-specific correction factor that accounts for the variations in SINR between transmissions.

A.1.1.2 Exponential Effective SINR Mapping (EESM)

Among the different proposals for effective SINR mapping, this is the one which seems to have more acceptance in the research community, standardization bodies and equipment manufacturers. Its derivation and performance analysis is described in detail in [192-194]. The proposed approach for the mapping/compression of the SINR values, one for each sub-carrier, into a single effective (scalar) SINR value is the Exponential Effective SINR Mapping (EESM). The EESM is used together with link level results for the different MCS schemes on AWGN channels to determine the BLER. The EESM is given by the equation (3).

$$SINR_{eff} = -\beta \ln\left(\frac{1}{N}\sum_{k=1}^{N} e^{-\frac{SINR_k}{\beta}}\right) \quad (3)$$

Where β is a correction parameter used to adapt the formula to the different types of scenarios used in the simulations (different MCS schemes and MIMO techniques) and is independent of the channel model used. It must be optimized from link-level simulations for each type of modulation and coding rate combination used. Also, a subset of the data sub-carriers space can be used to evaluate the effective SINR for reasons of computational efficiency.

A.1.1.3 Effective SINR Mapping Based on Mutual Information (MI-ESM)

The Effective SINR mapping based on mutual information was proposed on [195-196]. It is a one-dimensioning effective SINR mapping, very similar to EESM. It employs Bit Interleaved Coded Modulation (BICM) as the compression function. The MI-ESM is computed according to equation (4).

$$SINR_{eff} = I_{m_{ref}}^{-1}\left(\frac{1}{P}\sum_{p=1}^{P} I_{m_p}(SINR_p)\right) \quad (4)$$

The term $I_{m_p}(x)$ is the mutual information function for a given modulation and coding scheme and $I_{m_{ref}}(x)$ is the mutual information function for the modulation and coding scheme used as reference for the whole block. The value of m_{ref} can be set to the average number of transmitted bits per resource element.

The mutual information function for BICM [195] is computed according to equation (5).

$$I_{m_p}(x) = m_p - E_Y\left\{\frac{1}{2_{m_p}}\sum_{i=1}^{m_p}\sum_{b=0}^{1}\sum_{z\in X_b^i}\log\frac{\sum_{\hat{x}\in X}\exp\left(-\left|y-\sqrt{\frac{x}{\beta}}(\hat{x}-z)^2\right|\right)}{\sum_{\tilde{x}\in X_b^i}\exp\left(-\left|y-\sqrt{\frac{x}{\beta}}(\tilde{x}-z)^2\right|\right)}\right\} \quad (5)$$

Where:

- I_{m_p} is the mutual information of the applied modulation alphabet of size 2^{m_p} data symbols at the p^{th} data symbol.
- m_p is the number of bits per symbol transmitted in the elementary resource.
- X is the constellation set of 2^{m_p} data symbols.
- X_b^i is the set of symbols for which bit i equals bit b.
- Y is a zero-mean unit variance Gaussian random variable.
- β is a parameter whose optimal value is found such that it results in the minimization of the gap between the predicted and measured BLER.
- E_Y is the average of the mutual information. The averaging is performed along the set of reserved resource elements.

Due to its complexity, calculation of the mutual information at real time is not efficient. A better approach is to compute it offline by means of Monte-Carlo simulations and storing the results in proper look-up tables.

In [196] the performance of the MI-ESM method was conducted where the authors suggest that the performance of both methods EESM and MI-ESM are similar. However, it is important to mention that a point in favor for the use of this method is that, differently from the EESM mapping method, the MI-ESM includes the possibility of performing adaptive modulation and coding inside the coded block while it is being transmitted.

A.1.2 Two-Dimensional Data Compression and Mapping

Proposals for the SIR mapping also include the two-dimensioning data compression and mapping of the vector of received SINRs. In this case one more degree of freedom is available for the optimization and a better approximation, in terms of BLER, can result. As an example, in [197-198] it is proposed to use the mean, as given by equation (6) and the normalized variance (by the mean value), as given by equation (7) of the vector of SINR values of the frequency selective fading channel in the computation of the compressed effective SINR.

$$SINR_{\text{var},\beta} = \frac{1}{P}\sum_{p=1}^{P} SINR_p \quad (6)$$

$$SINR_{\mathrm{var},\beta} = \frac{1}{P}\sum_{p=1}^{P}\left(\sqrt{\frac{SINR_p}{SINR_{avg}}} - E\left(\sqrt{\frac{SINR_p}{SINR_{avg}}}\right)\right)^{\beta} \quad (7)$$

Once again, the parameter β is a calibration parameter for the minimization of the difference between the predicted and measured values of the BLER. Both $SINR_{avg}$ and $SINR_{\mathrm{var},\beta}$ are used as input parameters of a two-dimensional mapping function to determine the BLER, $SINR_{avg} = f(SINR_p, SINR_{\mathrm{var},\beta})$. See [198] for details regarding the definition of the proper two-dimensional mapping function.

A.2 Calibration of the Link to System Level Interface Model

The calibration and validation of the Link to System Level Interface Model is based on the execution of link level simulations assuming the channel is AWGN only. For each type of modulation and coding scheme used in system level simulations a separate set of link level simulations must be performed in order to generate an independent look-up table. The output of each look-up table is the average BLER, as a function of the Signal to Noise Ratio (SNR) per bit, E_b/N_0. A large number of BLER simulation points have to be generated in order to calibrate the interface for each MCS, so that the averaged BLER is statistically valid (a minimum of 100 block error events is the typical value considered in the simulations). The procedure must be repeated for different noise variances over the range of interest for the SNR per bit.

The optimization of the tuning parameter β considered in commonly used effective compressing and mapping functions is achieved by the training of the obtained BLER values from link level simulations. Assume a total amount of N_{sim} simulated BLER points lying in the range of interest (between 0.03 and 0.3). Assume also that the EESM method, as given by equation (1.15) is the compression function for the effective SINR. Then the effective SINR is computed as a function of the parameter β for each one of the N_{sim} simulated BLER points and the corresponding BLER is looked-up, resulting in a predicted BLER denoted as $BLER_{pred,k}(\beta)$. This predicted BLER is compared with the BLER value obtained from the link level simulations for the same point k $BLER_k$, and the parameter β is adjusted in order to minimize the sum of squares of the difference between both values, (least squares algorithm) as given by equation (8).

$$\sum_{k=1}^{N_{sim}} \left| BLER_{pred,k}(\beta) - BLER_k \right|^2 \quad (8)$$

Annex B

Traffic Models

B.1 Introduction

This annex describes the steps followed in the implementation of the traffic models used in the system level simulations performed in this work.

B.2 Voice over IP (VoIP) Traffic Model

VoIP refers to real-time delivery of voice packets across networks using the Internet Protocols. A VoIP session is assumed to last for the whole simulation period in each run, since the activation of the user.

A typical phone conversation is marked by periods of active talking/talk spurts (ON periods) interleaved by silence/listening (OFF periods) as illustrated in figure 1.

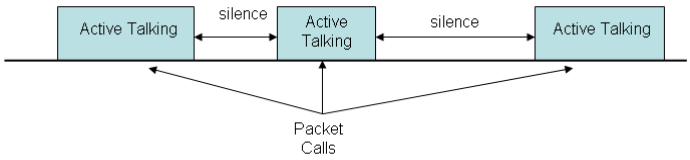


Figure 1 - VoIP traffic model: active talking silence periods

VoIP traffic model is modelled by a simple 2-state Markov chain model as shown in figure 2 [9].

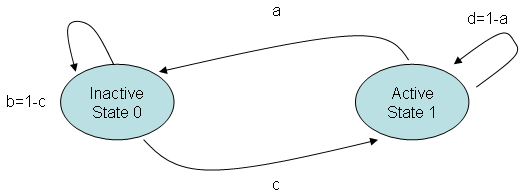


Figure 2 - State transition for VoIP traffic model

Model Statistics

- The conditional probability of transitioning from active speech state (state 1) to the inactive or silent state (state 0) while is state 1 is *a*.
- The conditional probability of transitioning from inactive state 0 to the active state 1 is *c*.
- The model is updated at the speech encoder frame rate *R=1/T*, where *T* is the speech encoder frame duration, assumed as 20 ms.
- The steady-state equilibrium requires that the probabilities of being in state 0 and 1 are given respectively by $P_0 = \frac{a}{a+c}$ and $P_1 = \frac{c}{a+c}$.
- The voice activity factor (VAF) $\lambda = P_1 = \frac{c}{a+c}$.
- A talk-spurt is defined as the time period τ_{TS} between entering and leaving the active state. The probability that a talk-spurt lasts during *n* speech frames is given by $P_{\tau_{Ts}=n} P_{\tau_{TS}}(n) = a(1-a)^{n-1}, \quad n = 1,2,...$
- The probability that a silence period lasts during *n* speech frames is given by $P_{\tau_{SP}=n} P_{\tau_{SP}}(n) = c(1-c)^{n-1}, \quad n = 1,2,...$.
- The mean talk-spurt duration in speech frames is given by $\mu_{TS} = E[\tau_{TS}] = \frac{1}{a}$
- The mean silence period in speech frames is given by $\mu_{SP} = E[\tau_{SP}] = \frac{1}{c}$.
- Since the state transitions from state 1 to state 0 and vice-versa are independent, the mean time between active state entries is given by the sum of the mean time in each state: $\mu_{AE} = \mu_{TS} + \mu_{SP}$ and the mean rate of arrival of transitions into the active state is given by: $R_{AE} = \frac{1}{\mu_{AE}}$.

During the active state, packets of fixed size are generated. The size of the packet and the rate at which the packets are sent depends on the corresponding voice codec and compression schemes. In all simulations conducted in this work an AMR codec with a fixed bit rate of 12.2 kbps was assumed. As the encoder frame duration is equal to 20 ms, the packet payload is equal to 244

bits. Although AMR codecs vary the encoding rate from 4.75 kbps to 12.2 kbps, according to the quality of the speech frames reported, link adaptation is not considered in this work.

The length of the VoIP packet must also include the overhead from the TCP/IP layers. Also, the overhead depends on the use of header compression or not. In this work the following configuration is assumed: 6 bytes of MAC header and 2 bytes of HARQ CRC in the IEEE802.16e reference system, 5 bytes for protocol headers (assuming AMR with header compression) and 33 bytes of packet payload. This configuration results in a packet length of 44 bytes. Although the model assumes the generation of packets with silence description as comfort noise, during each inactive state this is not included in the simulations. During each call session the mean duration of an ON state is equal to 1second and the mean duration of an OFF state is equal to 1.5 seconds. Both states follow exponential distributions. During the ON period packets of fixed length are generated at intervals of 20 ms. Table 1 provides the relevant parameters of the VoIP traffic assumed in the simulations.

Parameters	Values
Codec	RTP AMR 12.2k kbps Source Rate 12.2 kbps
Encoder frame length	20 ms
Voice Activity Factor (VAF)	40%
Payload	33 bytes
Protocol overhead with compressed header	RTP/UDP/IP: 3bytes IEEE802.16: Generic MAC header: 6bytes CRC for HARQ: 2 bytes
Total voice payload on air interface	44 bytes

TABLE 1: DETAILED DESCRIPTION OF THE VOIP TRAFFIC MODEL FOR IPV4

For the parameters considered in the VoIP model used in simulations the model statistics are the following:

$a = 0.0066$

$c = 0.0044$

$\mu_{TS} = \frac{1}{a} = 152 \approx 3s$

$\mu_{SP} = \frac{1}{c} = 227 \approx 4.54s$

$\mu_{AE} = 7.54s$ (1508 frame periods)

$R_{AE} = 0.132$ talk spurts per second.

Provided a single reservation is made per user per talk-spurt each user will request resources for transmission of VoIP packets in average every 7.54 seconds or 1508 frame periods.

B.3 3GPP Near Real Time Video (NRTV) Traffic Model

Figure 3 describes the steady state of video streaming traffic from the network [81]. Latency at call set-up is not considered in this steady-state model.

A video streaming session is defined as the entire video streaming call time, which is equal to the simulation time for each run in combined snapshot-dynamic mode. Each frame of video arrives at a regular interval *T* determined by the number of frames per second and each frame is decomposed into a fixed number of slices, each one transmitted within a single packet. The size of each slice is determined according to a Pareto distribution. The encoding delay *D* introduces delay intervals between the slices of a frame and are modelled by a truncated Pareto distribution. The parameter T_B is the length in seconds of the de-jitter buffer window in the mobile station. The de-jitter buffer is used to guarantee a continuous display of video streaming data. It is not used in the generation of data for this model.

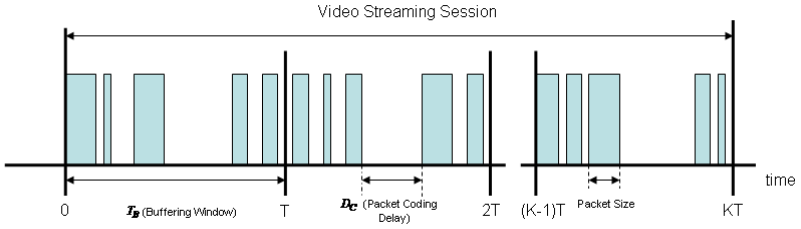


Figure 3 - NRTV packet streaming model

The video traffic model parameters are defined in table 2.

Information types	Inter-arrival time between the beginning of two consecutive frames	Number of slices in a frame	Slice size	Inter-arrival time between two consecutive slices in a frame	Source bit rate (kbps)
Distribution	Deterministic (10 frames per second)	Deterministic (fixed) N	Truncated Pareto (mean: mu, max: ma)	Truncated Pareto (mean: mu, max: ma)	N*mu*8/T
Parameters	100 ms	8	K=40bytes; alpha=1.2 mu=50bytes; ma=250bytes	K=2.5ms; alpha=1.2, mu= 6ms, max=12.5ms	32kbps
	"	"	mu=600bytes; ma=3125bytes	"	2Mbps
	"	"	mu=3125bytes; ma=15625bytes	"	10Mbps

TABLE 2: NRTV PARAMETERS FOR TRAFFIC MODEL

B.4 3GPP World Wide Web (WWW) Browsing Traffic Model

The traffic model for WWW bursty traffic used in the simulations is based on the ETSI WWW browsing model [199], but was tailored to reduce simulation run time by decreasing the number of mobile stations required to achieve peak system loading. The main modification is the reduction of the reading time between packet calls. Figure 4 illustrates the trace of a typical web browsing session. Each packet session consist of multiple packet calls representing web page downloads. Each session is divided into ON and OFF periods. The ON periods correspond to the instants where a web page is downloaded from the server. These are referred as packet calls. The OFF period represent the intermediate reading times required to digest the downloaded web

page. The size of each packet call (in bytes) is modelled by a truncated Pareto distribution producing a mean packet call size of 25 Kbytes. The reading time is modelled by a geometrically distributed random variable with a mean of 5 seconds. The reading time begins when the mobile station has received the entire packet call. Each packet call is segmented into individual packets. The time interval between two consecutive packets is modelled by a geometric distribution with a mean equal to the ration of the maximum packet size divided by the peak link speed. The size of each packet is fixed and equal to 12000 bits. The "slow-start" TCP/IP rate control mechanism for pacing packet traffic is not implemented.

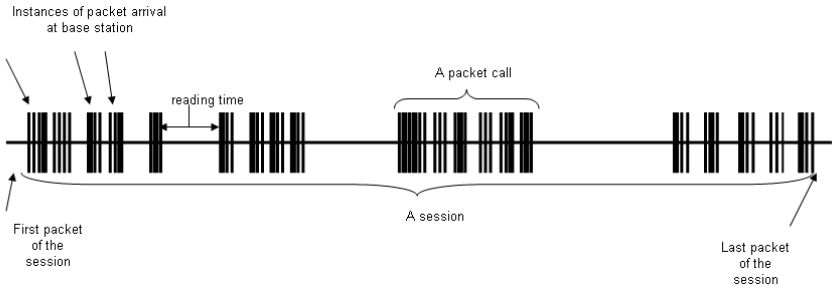


Figure 4 - Web browsing session

Table 3 illustrates the parameters used in this model.

Process	Distribution	Parameters
Packet Call Size	Truncated Pareto	α =1.1 k=4.5Kbytes m=2Mbytes μ =25Kbytes
Time between packet calls	Geometric	μt =5 seconds
Packet Size	Deterministic	12000 bits
Packet per packet call	Deterministic	Based on packet call size and packet size
Packet inter-arrival time	Geometric	μ = packet size/peak link speed Peak Link Speed of 2Mbps

TABLE 3: WEB BROWSING PARAMETERS FOR TRAFFIC MODEL

The other versions of the same traffic model for higher bit rates are obtained for the parameters in table 4

Process	Distribution	Parameters	PBR= μ / μt
Packet Call Size	Truncated Pareto	μ =1280 Kbytes (10,240,000 bits)	2Mbps
Packet Call Size	Truncated Pareto	μ =64000 Kbytes (51,200,000 bits)	10Mbps

TABLE 4: WEB BROWSING PARAMETERS FOR TRAFFIC MODEL

B.5 3GPP File Transfer Protocol (FTP) Traffic Model

In FTP applications a session is a sequence of file transfers, separated by reading times. Figure 5 illustrates a sample packet trace of an FTP session.

The file size in bytes is modelled according to a Log-Normal distribution and the reading time between the end of a download of the previous file and the user request for the next one is

modelled according to an exponential distribution with a mean of 180 seconds. The size of each packet is deterministic and equal to 12000 bits.

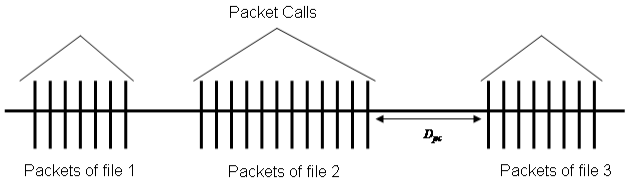


Figure 5 - FTP session

Process	Distribution	Parameters
File Size	Truncated Log-Normal	Mean=2Mbytes Standard Deviation=0.722Mbytes Maximum=5Mbytes (μ =14.45; σ =0.35)
Reading Time	Exponential	Mean=180seconds (λ =0.006)

TABLE 5: FTP PARAMETERS FOR TRAFFIC MODE

Annex C

Performance Metrics

C.1 Introduction

In this annex the performance statistics, generated as an output from the system level simulations, and used in the performance evaluation of the used scenarios and proposed algorithms are described. Metrics are collected along a simulation run.

For a simulation run:

- Simulation time per run: T_{sim}
- Number of simulation runs: D
- Total number of cells being simulated: N_{cells}
- Total number of users in cells of interest (cells being simulated): N_{users}
- Number of packet calls for user u: p_u
- Number of packets in the i^{th} packet call of user u: $q_{i,u}$

C.2 Throughput Performance Metrics

Average Service Throughput per-Cell

The average service throughput per cell is defined as the sum of the total amount of bits successfully received by all active users in the system, divided by the product of the number of cells simulated and the simulation duration.

$$R_{service}^{DL(UL)} = \frac{\sum_{u=1}^{N_k^{usersDL(UL)}} \sum_{i=1}^{p_{u,k}^{DL(UL)}} \sum_{j=1}^{q_{i,u,k}^{DL(UL)}} b_{j,i,u}}{N_{cells} T_{Sim}} \quad (1)$$

Where $N_k^{users,DL(UL)}$ is the number of users transmitting in DL(UL) in the k^{th} cell, $p_{u,k}^{DL(UL)}$ is the number of packet calls for user u in cell k, $q_{i,u,k}^{DL(UL)}$ is the number of packets for the i^{th} packet call for user u in cell k and $b_{j,i,u}$ is the number of bits received with success in the j^{th} packet of packet call i for user u in cell k.

Average Over-The-Air (OTA) Cell Throughput (kbps/cell) (3GPP Definition)

The average OTA throughput per cell is defined as the sum of the total amount of bits being successfully received by all active users in the system divided by the product of the number of cells being simulated in the system and the total amount of time spent in the transmission of these packets.

$$R_{OTA}^{DL(UL)} = \frac{\sum_{u=1}^{N_k^{users,DL(UL)}} \sum_{i=1}^{p_{u,k}^{DL(UL)}} \sum_{j=1}^{q_{i,u,k}^{DL(UL)}} b_{j,i,u}}{N_{cells}.T_{Trans}} \quad (2)$$

Where T_{trans} is the time required to transmit these packets.

Average Over-The-Air (OTA) Cell Throughput (kbps/cell) (Peak Bit Rate Definition)

This metric is very similar to the OTA throughput. But here all bits (correct and erroneous) are considered in its computation.

$$R_{OTA\_PBR}^{DL(UL)} = \frac{\sum_{u=1}^{N_k^{users,DL(UL)}} \sum_{i=1}^{p_{u,k}^{DL(UL)}} \sum_{j=1}^{q_{i,u,k}^{DL(UL)}} b_{j,i,u}^{Trans}}{N_{cells}.T_{Trans}} \quad (3)$$

Where $b_{j,i,u}^{Trans}$ is the number of bits received (with error or success) in the j^{th} packet of packet call i for user u in cell k.

Offered Cell Load (kbps/cell) (3GPP Definition)

This metric is used in the evaluation of the data load (in kbps) withdrawn from the base station's buffers for transmission, i.e., the influence of the channel in the transmission of the data is not being considered.

$$R_{OL3GPP}^{DL(UL)} = \frac{b_{Sent}^{DL(UL)}}{N_{Cells} T_{Sim}} \quad (4)$$

Where $b_{Sent}^{DL(UL)}$ is the total amount of data bits that have been withdrawn from the base station's queues and sent over the air interface for DL(UL) connection, for all mobile stations being simulated over the whole simulation run.

Offered Cell Load (kbps/cell) (Network Definition)

This metric measures the offered load from the core network to the base station for all mobile stations being simulated in the system over the whole simulation run.

$$R_{OLNetwork}^{DL(UL)} = \frac{b_{Network_{net}}^{DL(UL)}}{N_{Cells} T_{Sim}} \quad (5)$$

Where $b_{Network}^{DL(UL)}$ is the total amount of data bits that have arrived to the base station's queues from the core network over the whole simulation run. Figure 1 illustrates the relation among these different metrics and figures.

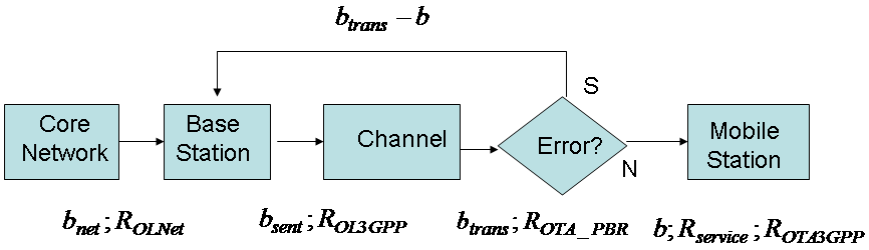


Figure 1 - Traffic metrics interdependence

User Average Peak Bit Rate at a Given Distance (kbps)

This metric gives the average peak bit rate of a given user at a given distance, *d*, in steps of 10 m, from the base station. For one user *i* the average peak bit rate is defined by equation (6).

$$R_{PBR}(i) = \frac{\sum_{k=1}^{N(i)} R_k(i)}{N(i).T} \quad (6)$$

Where $N(i)$ is the total amount of frames received by the mobile station *i* (both transmitted and retransmitted frames are taken into account) and $R_k(i)$ is the total amount of bits in the k^{th} frame received by mobile station *i*.

Per User Service Data Throughput

The user's service data throughput is defined as the ratio of the number of information bits successfully received by the user and the total simulation run time. If user *u* has $p_u^{DL(UL)}$ downlink (uplink) packet calls with $q_{i,u}^{DL(UL)}$ packets for the i^{th} downlink (uplink) packet call and $b_{j,i,u}$ bits in the j^{th} packet the average user throughput for user *u* is given by equation (7).

$$R_u^{DL(UL)} = \frac{\sum_{i=1}^{p_u^{DL(UL)}} \sum_{j=1}^{q_{i,i}^{DL(UL)}} b_{j,i,u}}{T_{\cdot Sim}} \quad (7)$$

Per-User Average Service Throughput

The average per-user service throughput is defined as the sum of the user service throughput of each user divided by the total number of users in the system.

$$\overline{R_u^{DL(UL)}} = \frac{\sum_{u=1}^{N_{users}} R_u^{DL(UL)}}{N_{\cdot users}} \quad (8)$$

Average Packet Call Throughput for a User

If there are N_{users} in the cell of interest and $R_{k,u}^{DL(UL)}$ is the service throughput for the n^{th} user in the cell, the DL or UL service throughput for the cell is given by equation (9)

$$R^{DL(UL)} = \sum_{u=1}^{N_{users}} R_u^{DL(UL)} \quad (9)$$

The packet call throughput is equal to the total amount of bits per packet call received with success divided by the duration of the packet call. If user u has $p_u^{DL(UL)}$ downlink (uplink) packet calls with $q_{i,u}^{DL(UL)}$ packets for the i^{th} downlink (uplink) packet call and j^{th} packet call then the average packet call throughput is given by equation (10)

$$R_u^{pc,DL(UL)} = \frac{1}{p_u^{DL(UL)}} \left(\sum_{i=1}^{p_u^{DL(UL)}} \frac{\sum_{j=1}^{q_{i,u}^{DL(UL)}} b_{j,i,u}}{\left(T_{i,u}^{end,DL(UL)} - T_{i,u}^{start,DL(UL)}\right)} \right) \quad (10)$$

Where $T_{i,u}^{start,DL(UL)}$ is the time instant at which the transmission of the first packet of the i^{th} DL(UL) packet call for user u starts and $T_{i,u}^{end,DL(UL)}$ defines the time instant at which the last packet of the i^{th} DL(UL) packet call for user u is received with success. For uncompleted packet calls this parameter is set to the simulation end time.

Average per-User Packet Call Throughput

The average per-user packet call throughput is defined as the sum of the average packet call throughput of each user divided by the total number of users in the system.

$$\overline{R_u^{pc,DL(UL)}} = \frac{\sum_{u=1}^{N_{users}} R_u^{pc,DL(UL)}}{N_{\cdot users}} \quad (11)$$

Throughput Outage

The throughput outage is defined as the percentage of users with service data rate $R_u^{DL(UL)}$ less than a pre-defined minimum rate R_{min}.

Cell Edge User Throughput

The cell edge user throughput is defined as the 5th percentile point of the CDF of user's average packet call throughput.

C.3 Performance Metrics for Delay Sensitive Applications

Packet Delay

For an individual packet the delay is defined as the time elapsed between the instant when the packet enters the queue at transmitter and the time when the packet is received successfully by the mobile station. If a packet is not successfully delivered by the end of a run its ending time is the end of the run. Assuming the j^{th} packet of the i^{th} packet call destined for user u arrives at the base station (mobile station) at time $T_{j,i,u}^{arr,DL(UL)}$ and is delivered with success to the mobile station (base station) at time $T_{j,i,u}^{dep,DL(UL)}$, the packet delay is defined as in equation (12).

$$Delay_{j,i,u}^{DL(UL)} = T_{j,i,u}^{dep,DL(UL)} T_{j,i,u}^{arr,DL(UL)} \tag{12}$$

Use Average Packet Delay

The average packet delay is defined as the average interval between packets originated at the source station (mobile or base station) and received at the destination station (base or mobile station) in a system for a given packet call duration. The average packet delay for user u is given by equation (13).

$$D_u^{avg,DL(UL)} = \frac{\sum_{i=1}^{p_u} \sum_{j=1}^{q_{i,u}} \left(T_{j,i,u}^{dep,DL(UL)} - T_{j,i,u}^{arr,DL(UL)} \right)}{\sum_{i=1}^{p_u} q_{i,u}} \tag{13}$$

Residual Frame Erasure Rate (FER)

This metric is computed for each user and for each packet service session. A packet service session contains one or several packet calls depending on the application. A packet service session starts when the first packet of the first packet call of a given service begins and ends when the last packet of the last packet call of the same service has been transmitted. One packet call contains one or several packets. The Residual FER is given by equation (14).

$$FER_{residual} = \frac{\eta_{dropped\_packets}}{\eta_{packets}} \tag{14}$$

Where $\eta_{dropped\_packets}$ is the total amount of dropped packets in the packet service session and $\eta_{packets}$ is the total amount of packets in the packet session. A dropped packet is the one in which the maximum number of transmission attempts has been achieved without the packet being successfully decoded.

Packet Loss Ratio

The packet loss ratio is computed for each user and for each packet service session and is defined as in equation (15).

$$PDR = \frac{\eta_{discarted\_packets}}{\eta_{packets}} \quad (15)$$

Where $\eta_{discarted\_packets}$ is the total amount of packets discarded due to time-out (delay bound violation and maximum number of transmission attempts achieved).

Spectral Efficiency (bps/Hz)

This is the ratio of correctly transmitted bits over the radio resources to the total amount of available bandwidth. The average cell spectral efficiency is defined as in equation (16).

$$SE = \frac{R}{BW_{eff}} \quad (16)$$

Where R is the aggregate cell throughput, BW_{eff} is the effective channel bandwidth, defined as $BW_{eff} = BW * TR$, where BW is the used channel bandwidth and TR is the time ratio of the link. For example for TDD with DL:UL=2:1, TR = 2/3 for DL and 1/3 for UL.

System Outage

A user is said to be in outage if more than a given percentage of packets experience a delay greater than a certain time. The system is said to be in outage if any individual users are in outage.

System Capacity

System capacity is defined as the maximum number of users that can be serviced without making the system exceed the maximum allowed outage probability.

C.4 Fairness Criteria

It may be an objective to have uniform service coverage resulting in a fair service offering for best effort traffic. A measure of fairness under the best effort assumption is important in assessing how well the proposed solution performs.

The fairness is evaluated by determining the normalized cumulative distribution function (CDF) of the per user throughput. The CDF is to be tested against a predetermined fairness criterion under several specific traffic conditions.

Let $T_{put}(k)$ be the throughput for user *k*. The normalized throughput with respect to the average user throughput for user *k* is given by equation (17).

$$\widetilde{T}_{put}(k) = \frac{T_{put}(k)}{\underset{i}{avg}\, T_{put}(i)} \tag{17}$$

Moderately Fair Criteria

The CDF of the normalized throughput with respect to the average user throughput for all users is determined. This DCF shall lie to the right of the curve given by the three points in table 1.

Normalized Throughput w.r.t average user throughput	CDF
0.1	0.1
0.3	0.2
0.5	0.5

TABLE 1: MODERATELY FAIR CRITERION CDF

Short Term Fairness Indication

During the simulation, the following short-term fairness indicator should be computed and recorded every τ ms (τ is suggested to be 20 or 40):

$$F(t) = \frac{\left|\sum_{i\in A} \hat{T}_i(t)\right|^2}{|A| \sum_{i\in A} \hat{T}_i^2(t)} \tag{18}$$

Where $\hat{T}_i(t)$ is the amount of service received by the i^{th} user in time interval [t,t+ τ).

A is the set of users with nonzero buffers in [t,t+ τ) and $|A|$ is the cardinality of *A*. The minimum of $F(t)$ during the simulation time, defined as $F_{\min} = \min_{t\in\{0,\tau,2\tau,3\tau,\ldots T_{slot}\}} F(t)$ can serve as an indication of how much fairness is maintained all the time.