UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

CRISTIANO BONATO BOTH

A New Cross-Layer Adaptive Architecture to Guarantee Quality of Service in WiMAX Networks

Thesis presented in partial fulfillment of the requirements for the degree of Doctor of Computer Science

Dr. Lisandro Zambenedetti Granville
Advisor

Dr. Juergen Rochol
Coadvisor

Porto Alegre, December 2011

CIP – CATALOGING-IN-PUBLICATION

Both, Cristiano Bonato

A New Cross-Layer Adaptive Architecture to Guarantee Quality of Service in WiMAX Networks / Cristiano Bonato Both. – Porto Alegre: PPGC da UFRGS, 2011.

153 p.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2011. Advisor: Lisandro Zambenedetti Granville; Coadvisor: Juergen Rochol.

I. Granville, Lisandro Zambenedetti. II. Rochol, Juergen. III. Título.

CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

ATM	Asynchronous Transfer Mode
ARM	Adaptive Multi Rate
BE	Best Effort Service
BER	Bit Error Rate
BPSK	Binary Phase Shift Keying
BR	Bandwidth Request
BS	Base Station
BWA	Broadband Wireless Access
CAC	Connection Admission Control
CC	Convolutional Code
CDMA	Code Division Multiple Access
CID	Connection Identifier
CQI	Channel Quality Indication
CTC	Convolutional Turbo Codes
ertPS	Extended Real Time Polling Service
DCD	Downlink Channel Descriptor
DFPQ	Deficit Fair Priority Queuing
DIUC	Downlink Interval Usage Code
DL	Downlink
DRR	Deficit Round Robin
DSA	Dynamic Service Addition
DSC	Dynamic Service Change
DSD	Dynamic Service Deletion
DSL	Digital Subscriber Line
DWRR	Deficit Weighted Round Robin
EDF	Earliest Deadline First

FCH	Frame Control Header
FDD	Frequency Division Duplexing
FEC	Forward Error-Correction
FFT	Fast Fourier Transform
FIFO	First In First Out
FTP	File Transfer Protocol
GPS	General Processor Sharing
HR	Hand-Off Ranging
HTTP	HyperText Transfer Protocol
IE	Information Element
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IR	Initial Ranging
ISWCS	International Symposium on Wireless Communication Systems
JNCA	Journal of Network and Computer Applications
LBS	Location Based Services
LDPC	Low-Density Parity Check Codes
LoS	Line of Sight
MBS	Multicast and Broadcast Services
MAC	Medium Access Control
Max	Maximal value
MCS	Modulation and Coding Scheme
Min	Minimal value
MiSP	Micro Scheduling Problem
ML	Maximum Latency
MPEG	Moving Pictures Expert Group
MRTR	Minimum Reserved Traffic Rate
MS	Mobile Station
MSID	Mobile Station Identifier
MSTR	Maximum Sustained Traffic Rate
NLoS	Non-Line of Sight
nrtPS	Non Real Time Polling Service
OCSA	One Column Striping with non-increasing Area first mapping
OFDM	Orthogonal Frequency Division Multiplexing

OFDMA	Orthogonal Frequency Division Multiple Access
PCM	Pulse Coded Modulation
PDA	Personal Digital Assistant
PDF	Probability Density Function
PDU	Protocol Data Unit
PF	Proportional Fair
PHY	Physical
PMP	Point-to-MultiPoint
PPP	Point-to-Point
PR	Periodic Ranging
OSI	Open System Interconnection
PUSC	Partially Used Subchannelization
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
REG-REQ	Register Request
REG-RSP	Register Response
RF	Radio Frequency
ROHC	Robust Header Compression
RM	Reference Model
RNG-REQ	Ranging Request
RNG-RSP	Ranging Response
RR	Round Robin
RS	Reed-Solomon
RTG	Rx/Tx Transmission Gap
RTP	Real-time Transport Protocol
rtPS	Real Time Polling Service
RTS	Recursive Tiles and Stripes
SAP	Service Access Point
SBRC	Brazilian Symposium on Computer Networks and Distributed Systems
SC	Single Carrier
SD	Standard Deviation
SDRA	Simple Data Region Allocation
SDU	Service Data Unit

SFID	Service Flow Identifier
SINR	Signal to Interference plus Noise Ratio
SMTP	Simple Mail Transfer Protocol
SS	Subscriber Stations
TDD	Time Division Duplexing
TJ	Tolerated Jitter
TTG	Tx/Rx Transition Gap
UCD	Uplink Channel Descriptor
UDP	User Datagram Protocol
UGS	Unsolicited Grant Service
UL	Uplink
UIUC	Uplink Interval Usage Code
VoIP	Voice over Internet Protocol
WFQ	Weighted Fair Queuing
WF^2Q	Worst-case Fair Weighted Fair Queueing
WiMAX	Worldwide Interoperability for Microwave Access
WMAN	Wireless Metropolitan Area Network
WPF	Weighted Proportional Fair
WRR	Weighted Round Robin

LIST OF SYMBOLS

b_{symb}	Amount of data per symbol
c	Index of a connection
C	Number of CIDs
CP	Cyclic Prefix
$C_{available}$	Available capacity of the RF channel
C_{tot}	Capacity of the RF channel total
C_{used}	Capacity in bits of the network used
$Data$	Amount of bits DL or UL transmitted in the frame
DLY	Delay of a PDU or BR
$\overline{DLY}$	Average delay of a PDU or BR
DL_{CID}	Information of a connection
DL_{head}	Header of the DL-MAP
D	Deadline value
$D^{‘}$	Normalized deadline
DL_{IE}	Amount of DL bursts profile within the OFDMA frame
D_{max}	Maximum deadline
DL_{MAP}	Size of the DL-MAP
f	Index of an OFDMA frame
f_{in}	Incoming OFDMA subframes
f_{out}	Outgoing OFDMA subframes
F	Number of frames in an interval t
F_{bits}	Amount of bits transmitted in each OFDMA subframe
j	Index of the MCS level
J_b	Backoff window
JTR	Jitter of a PDU or BR
$\overline{JTR}$	Average jitter of a PDU or BR

H	Height of a rectangular data area
$\hat{\mathrm{H}}$	Height of a data column
i	Index of the PDU or BR
k	Number of DL connections in a DL_{IE}
K	FEC encoding
l	Index of the BR
L	Size of the subframe
m	Index of a slot to mapping overhead
M	Quantity of bits per modulation symbol
$\overline{MCS}$	Historical average of MCS
n	Number of symbols inside a QAM constellation
N	Number of MCS levels
N_A	CDMA allocation
N_B	CDMA bandwidth request
N_{br}	Number of BRs
N_C	Number of connections in each burst
N_{data}	Amount of N_{pdu} and N_{br} for a given connection
N_D	UL data
N_{fft}	Number of OFDM subcarriers total
N_{used}	Number of OFDM subcarriers used
N_{pdu}	Number of PDUs
N_{R-DL}	Number of data bursts in the DL subframe
N_{R-UL}	Number of data bursts in the UL subframe
N_{sc}	Number of subcarries per subchannel
N_{slots}	Number of slots associated with an amount of data
p	Index of UL resource allocation
q	Sampling factor
r	Index of the PDU
R	Number of UL resource allocated with an UIUC association
s	Index of a slot to PDUs and BRs of a subframe
S	Subset of slots
S_{slot}	Amount of symbols per slots
t	A give instant of time
T_c	Creation or arrival time

T_s	Symbol duration
T_{sys}	System up time
T_f	Duration time of a frame
UL_{IE}	Amount of UL bursts profile within the OFDMA frame
UL_{MAP}	Size of the UL-MAP
UL_{UIUC}	Uplink Interval Usage Code
W	Total available bandwidth
$\hat{\mathrm{W}}$	Width of a data column
α	Constant smoothing factor to the φ_{UGS_ertPS}
μ	Mean of a probability distribution
Λ	Adjustment factor of the threshold criteria
π	A give packing configuration
ρ	Constant smoothing factor to the φ_{rtPS}
σ	Standard deviation of a probability distribution
τ	Current throughput value
$\tau^{`}$	Normalized throughput
τ_{max}	Maximum throughput
$\overline{\tau}$	Backlogged information about the throughput
$\phi(\ )$	Function that creates the maps in the OFDMA frame
Φ	Dynamic map overhead
ε	Efficiency of scheduler and allocator components
φ_{nrtPS_BE}	Scheduling factor of classes of service nrtPS and BE
φ_{rtPS}	Scheduling factor of class of service rtPS
φ_{UGS_ertPS}	Scheduling factor of classes of service UGS and ertPS
ϱ	Channel state
Ψ	Static map overhead

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

ABSTRACT

Wireless networks must provide quality of service to voice, video and data applications. A standard defined to offer quality of service in these networks is the IEEE 802.16 document. In order to improve the quality of transmission, this standard uses two main physical mechanisms: (i) Orthogonal Frequency Division Multiple Access as physical interface and (ii) the possibility of adjusting the transmission robustness to face the physical impairments that may compromise the transmission. Moreover, the standard defines a set of components in the base station, such as allocators, schedulers, and connection admission controllers that must be modeled to provide an architecture that guarantees quality of service. However, the standard does not define either the algorithm running inside each one of the components nor the integration among them.

Investigations aiming to provide quality of service have been proposed in the context of IEEE 802.16 networks. The literature on mobile IEEE 802.16 networks shows that the current research is focused on specific solutions for each component or in solutions with partial integration. The focus of those solutions is to provide the best alternative for individual problems of a particular component. However, to the best of our knowledge, no research addressing the overall quality of service architecture considering both the diversity of applications traffic requirements and the propagation conditions of the radio frequency channel has been proposed so far.

In this context, this thesis proposes a new architecture to guarantee quality of service in the base station that must be modeled using a cross-layer infrastructure able to adapt to the dynamics of traffic requirements as well as to the radio frequency channel conditions. The aim is to integrate the components defined by the standard with the physical mechanisms. Another objective is to evaluate the proposed architecture, through an evaluation methodology that is defined following the specification of the system evaluation of the Worldwide Interoperability for Microwave Access forum. Therefore, the analysis on the new cross-layer adaptive architecture is performed and the results show efficient data allocation as well as a minimal delay and jitter for real-time applications.

Keywords: Adaptive, Architecture, Cross-layer, IEEE 802.16, Quality of Service, WiMAX.

Uma Nova Arquitetura Adaptativa entre Camadas para Garantir Qualidade de Serviço em Redes WiMAX

RESUMO

Redes sem fio devem prover qualidade de serviço para aplicações de voz, video e dados. Um padrão definido para oferecer qualidade de serviço nessas redes é o documento IEEE 802.16. Com o objetivo de melhorar a qualidade de transmissão, este padrão utiliza dois principais mecanismos físicos: (i) *Orthogonal Frequency Division Multiple Access* como interface física e (ii) a possibilidade de ajustar a robustez da transmissão em relação as imperfeições físicas que podem comprometer a transmissão. Além disso, o padrão define um conjunto de componentes na estação base, tal como alocadores, escalonadores e controles de admissões que devem ser modelados para prover uma arquitetura que garanta qualidade de serviço. Entretanto, o padrão não define nem os algoritmos de cada componente, nem a integração entre estes componentes.

Investigações objetivando prover qualidade de serviço tem sido propostas no contexto de redes IEEE 802.16. A literatura sobre redes IEEE 802.16 móveis mostra que as atuais pesquisas estão focadas em soluções específicas para cada componente, ou em soluções com integrações parciais. O foco destas soluções é prover a melhor alternativa para problemas individuais para um componente particular. Entretanto, em todos os estudos realizadas nesta tese, não encontrou-se nenhuma pesquisa endereçando propostas sobre a qualidade de serviço global considerando a diversidade dos requisitos de tráfegos das aplicações e as condições de propagação do canal de rádio frequência.

Neste contexto, essa tese propõe uma nova arquitetura para garantir qualidade de serviço em uma estão base que deve ser modelada usando uma infraestrutura entre camada para adaptar-se aos requisitos dinâmicos do tráfego, bem como as condições do canal de rádio frequência. O objetivo é integrar os componentes definidos pelo padrão com os mecanismos físicos. Outro objetivo é analisar a arquitetura proposta, através de uma metodologia de avaliação que é baseada segundo a especificação do sistema de avaliação do fórum *Worldwide Interoperability for Microwave Access*. Assim, a análise da nova arquitetura adaptativa entre camadas é realizada e os resultados mostram a eficiência na alocação dos dados, bem como o mínimo atraso e *jitter* gerado nas aplicações de tempo real.

Palavras-chave: Arquitetura, Adaptativa, Camadas, IEEE 802.16, Qualidade de Serviço, WiMAX.

1 INTRODUCTION

The advent of mobile devices, such as Personal Digital Assistant (PDA), laptops, and next generation mobile phones has changed the access profile of computer networks, creating a constant need for letting information available at anytime and anywhere. In this context, wireless networks became fundamental entities because they provide fast access deployment and high scalability with low costs of infrastructure, maintenance, and upgrade (PAPAPANAGIOTOU et al., 2009).

A key point in wireless networks is the increasing number of applications requiring communication services able to aggregate data, voice, and multimedia traffics. Usually, such applications demand a lot of bandwidth and high data rates for each user/client of the network. To deal with these issues, networks must guarantee the quality of their services, *i.e.*, they must provide Quality of Service (QoS).

A standard defined to provide QoS-enabled services in wireless networks is IEEE 802.16 (IEEE, 2009), also known as Worldwide Interoperability for Microwave Access (WiMAX), named after the forum of devices manufacturers for this standard. To improve the quality of transmission, WiMAX networks use Orthogonal Frequency Division Multiple Access (OFDMA) as physical interface, which is designed for both nomadic (fixed) and mobile accesses. The use of this multiple access technique is an effective approach to support broadband wireless communications. However, Radio Frequency (RF) channels are susceptible to physical impairments that compromise the transmission, leading to performance problems and low availability of network services.

To minimize the performance problems, WiMAX networks allow adjustments in the transmission robustness to face the physical impairments (KO et al., 2010). The adjustment process is based on adaptive Modulation and Coding Scheme (MCS), offering a trade-off between transmission robustness and data rates. This trade-off affects the quantity of available resources in terms of bandwidth, leading to a variable quantity of data that can be sent within a transmission opportunity and consequently affecting the overall network QoS.

The IEEE 802.16-2009 (IEEE, 2009) standard does not specify how the management of radio resource (*i.e.* bandwidth and transmission power) should be performed. Moreover, as described by Rong *et al.* (RONG; QIAN; CHEN, 2007a), there exists a fundamental trade-off between bandwidth and power resources in wireless networks, due to physical characteristics of the RF channel, such as noise and interference. These characteristics of RF channels must then be considered when proposing a QoS-enabled architecture, especially in wireless networks. The IEEE 802.16-2009 standard defines a set of components in the Base Station (BS), such as allocators, schedulers, and Connection Admission Controllers (CAC) that must be modeled to provide a QoS-enabled architecture. However, the standard does not define either the algorithm running inside each one of the components nor the integration among them. These definitions are, in fact, left opened on purpose to allow future technological advancements and vendors specific solutions (BACIOCCOLA et al., 2010).

Research to provide QoS has been proposed in the context of IEEE 802.16 networks. So-In *et al.* (SO-IN; JAIN; TAMIMI, 2009a) have presented a simple analytical method for capacity evaluation of mobile WiMAX networks. That investigation considered realistic parameters for OFDMA and two RF channel propagation conditions: error-free and imperfect channel. Sekercioglu *et al.* (SEKERCIOGLU; IVANOVICH; YEGIN, 2009) have proposed a partial architecture for uplink transmissions, proving a solution composed of CAC and scheduler components. Each component of the architecture has been designed for fixed WiMAX using Orthogonal Frequency Division Multiplexing (OFDM) disregarding channel conditions. On the other hand, Rong *et al.* (RONG et al., 2008) have conducted an investigation about CAC algorithms designed for downlink transmissions. Other relevant proposals concerning uplink and downlink scheduling have been described by Sched and So-In (SO-IN; JAIN; TAMIMI, 2009b).

The literature on mobile WiMAX networks presents that the current research is focused on specific solutions for each component (COHEN; KATZIR, 2010) (CHOI; JEON; JEONG, 2009) (CICCONETTI et al., 2010) or in solutions with partial integration (RONG; QIAN; CHEN, 2007b) (SHE et al., 2009) (HOU et al., 2009). The focus of those solutions is to provide the best alternative for individual problems of a particular component. However, to the best of our knowledge, no research addressing the overall QoS architecture considering the diversity of applications traffic requirements and the propagation conditions of the RF channel has been proposed so far. In this context, this thesis establishes a line of investigation based on hypothesis and fundamental questions, which are described as follows.

1.1 Hypothesis and Fundamental Questions

This thesis investigates how the architecture of a BS in mobile WiMAX networks should be modeled to guarantee QoS. To help in this investigation, it is defended the hypothesis as follows.

Hypothesis: **"The architecture to guarantee QoS in WiMAX BSs must be modeled using a cross-layer infrastructure able to adapt to the dynamics of traffic requirements as well as to the RF channel conditions."**

To provide an integrated and adaptive solution, this thesis proposes a new cross-layer architecture in BS that considers both the traffic requirements and physical characteristics of WiMAX networks. This investigation aims to integrate QoS components *e.g.* allocators, schedulers, and CAC with the physical mechanisms *e.g.* OFDMA and MCS. Moreover, a cross-layer infrastructure is defined so that it enables to adapt the operation of the new architecture in face of the dynamics of the network traffic requirements and the time variant RF channel conditions. Based on this hypothesis, three fundamental questions are defined in order to guide the investigation:

- How to provide accurate information to the new cross-layer adaptive architecture?

- How to manage the available resources considering the dynamics of network traffic requirements and RF channel conditions?

- What is the impact of integrating adaptive QoS components in WiMAX networks?

1.2 Contributions

The aim of this thesis is to propose a new cross-layer adaptive architecture to guarantee QoS in WiMAX networks, considering application traffic requirements and RF channel conditions. The specific contributions expected as outcome of this investigation are listed as follows.

- Classify the cross-layer architecture proposals published in the literature. The related work about such issue in WiMAX networks presents some proposals based on the communication between Physical (PHY) Layer and Medium Access Control (MAC) Layer. Therefore, this thesis organizes the architecture proposals in three groups according to the usage of the bandwidth, multiplexing technique, and the frame structure;

- Specify the integration among allocator, scheduler, and CAC components. Current proposals already focused on specific solutions for individual problems of each component. However, the design of an integrated set of components is a demand for the correct operation of the cross-layer adaptive architecture;

- Design of an information infrastructure that enables the management of QoS components feedbacks, applications traffic requirements, and time variant propagation conditions of the RF channel. The IEEE 802.16 standard defines all kinds of information that should be exchanged between the Mobile Stations (MSs) and the BS. However, this standard does not define how information should be managed (IEEE, 2009). Despite of the relevance of providing QoS for the management of such kind of information has not been the focus of the related work in the literature;

- Specify a CAC component for controlling the admission of both uplink and the downlink connection requests. The CAC component is responsible for the correct operation of the cross-layer adaptive architecture, by admitting and managing a number of connections that can be served with guaranteed QoS (RONG et al., 2008). The information provided by the infrastructure may be used to adapt the behavior of the CAC component according to the current status of the radio resources, *i.e.*, reception power, and available bandwidth. In this context, the accurate identification of the amount of available resources is fundamental for the correct usage of the CAC component;

- Choose a specific methodology to evaluate the QoS components of the new cross-layer adaptive architecture considering the dynamics of network traffic requirements and the RF channel conditions. The integration among the QoS components should be analyzed to assess the performance of the overall architecture. This methodology should be oriented by the WiMAX Forum recommendations (WiMAX Forum, 2008).

Summarizing the context of this thesis, the design of the new cross-layer adaptive architecture considers a BS in a single-cell. Therefore, in this scenario, handover characteristics, such as the reservation of resources for a MS to change cell are not considered. Moreover, single layer architectures are not analyzed because this approach do not include the diversity both traffics and physical characteristics. Taking into account the scope of this investigation, in next the section the main contributions are presented and discussed.

1.3 Organization

Chapter 2 describes background aspects of the IEEE 802.16 standard. The PHY and MAC layers are addressed with an overview about applications and topologies that can be used in WiMAX networks. In addition, the three main components to guarantee QoS in IEEE 802.16 networks are detailed.

Chapter 3 depicts the state-of-the-art on each QoS component for IEEE 802.16 networks. Moreover, the organization of the cross-layer architecture proposals published in the literature and the challenges to design a cross-layer adaptive architecture are defined.

Chapter 4 presents a new cross-layer adaptive architecture proposal for IEEE 802.16 networks. A RF channel model used to inform the changes on RF channel conditions to the new cross-layer adaptive architecture is described. The design of the new channel-aware architecture is detailed, considering a Joint Information structure used to permit QoS components to be aware of changes on the RF channel conditions and of available resources. In addition, the design of the CAC component with an information infrastructure is proposed to provide a new cross-layer adaptive architecture.

Chapter 5 shows the evaluation of the results to analyze the performance of allocator, scheduler, and CAC components in the new cross-layer adaptive architecture. To evaluate the performance of the proposed architecture, the simulation parameters are also discussed.

Chapter 6 describes the final remarks and conclusions associated to this thesis. The answers for the guide questions are exposed and justified. Finally, directions for future work are identified and detailed.

2 IEEE 802.16 STANDARD

The goal of this chapter is to present aspects on background the IEEE 802.16 standard. The chapter starts with an overview about IEEE 802.16 networks focusing mainly in the applications and the topologies that can be used in these networks. Next, PHY Layer features are described, considering OFDMA multiplexing technique and MCS configuration. In addition, MAC Layer is discussed in two parts: (i) the Convergence Sublayer, where it is described the details of a connection-based network, as well as the definition of service flows for the classes of service specified in the IEEE 802.16 standard and (ii) the MAC Common Sublayer, where the OFDMA frame structure and bandwidth requests mechanisms are discussed. Understanding the OFDMA frame is of fundamental importance for this thesis because the bi-dimensional matrix structure of the frame influences the flexibility and complexity of the IEEE 802.16 wireless networks technology. Finally, the three main components to guarantee QoS in IEEE 802.16 networks are detailed. In the context of this thesis, these components are important because they should be designed considering the PHY and MAC Layers features.

2.1 Applications, Topologies and Overview

The goal of this section is to present an overview about IEEE 802.16 networks, focusing mainly in its applications and topologies. Specific details about this technology and the current release of the standard can be found in the literature (KIM, 2009) (AHMADI, 2009) (LI et al., 2009) (SYDIR; TAORI, 2009) (ETEMAD; WANG, 2009) (VENKATACHALAM et al., 2009) (YANG et al., 2009).

The first version of the IEEE 802.16 standard was published in 2002. The document defined the WirelessMAN air interface for Wireless Metropolitan Area Network (WMAN). The goal of the standard was to define a technology for Broadband Wireless Access (BWA) for fixed users, as an alternative to cabled network access, such as Digital Subscriber Line (DSL). The IEEE 802.16 standard specification originally defined a Point-to-MultiPoint (PMP) topology, where resources are shared among a BS and a set of Subscriber Stations (SSs). Moreover, downlink (DL) and uplink (UL) transmissions can share the same transmission frequency, or use different frequencies. The Frequency Division Duplexing (FDD) technique uses different frequencies for DL and UL transmissions. On the other hand, the Time Division Duplexing (TDD) technique shares the same frequency, through the alternation between DL and UL transmissions.

A version of the standard with support to fixed and mobile operations in licensed bands was released in 2005. This version, called IEEE 802.16e, added several features related to mobile operation, including power-saving, idle mode, handover, etc. Finally, on May 29th 2009, the current release of this standard was published - the IEEE 802.16m

(IEEE, 2009). The main contributions added in this version are: load balancing, Robust Header Compression (ROHC), enhanced mechanism for resource allocation, support for Location Based Services (LBS) and Multicast and Broadcast Services (MBS).

The IEEE 802.16 standard defines, since the first version, a data plane (IEEE, 2009) composed of two layers: PHY and MAC, as can be seen in Figure 2.1. This data plane is related to the two bottom layers of the Open Systems Interconnection (OSI) Reference Model. The PHY Layer supports several multiplexing techniques and a MCS configuration that are described in Section 2.2. On the other hand, the MAC Layer specifies two sublayers: Service Specific Convergence and MAC Common that are presented in Section 2.3. The standard also defines the Security Sublayer within the MAC Common Sublayer.

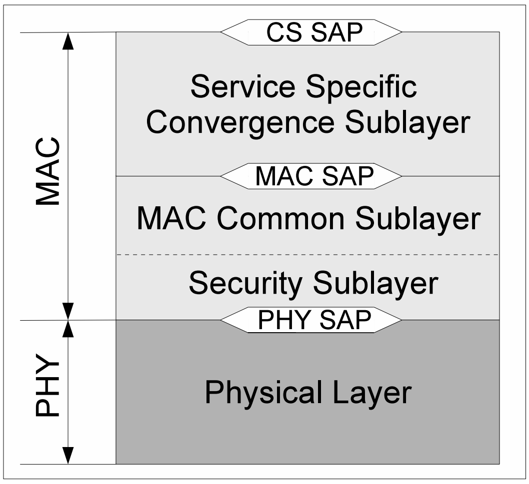


Figure 2.1: Data Plane of the IEEE 802.16 Standard

Figure 2.1 shows three Service Access Points (SAP) between layers and sublayers in the data plane of the IEEE 802.16. The CS SAP receives Service Data Units (SDU) from the higher layers of the RM/OSI and the Convergence Sublayer associates these SDUs to the IEEE 802.16 technology. The standard specifies two types of service: encapsulation of cells and encapsulation of packages. On the other hand, the MAC Common Sublayer receives data from the Convergence Sublayer, via MAC SAP. This data is related to different connection types. The MAC Common Sublayer offers guarantee of QoS by providing procedures as frame construction, bandwidth request mechanisms, allocation, scheduling, and connection admission control of the user's data. In the last access point, called PHY SAP, the data, control information, and statistics are transferred between the MAC Common Sublayer and the PHY Layer.

PHY and MAC layers were defined in IEEE 802.16 standard to provide implementation flexibility (BACIOCCOLA et al., 2010). Therefore, the standard leaves several procedures as optional and allows various implementation ways in BS and MS. In this context, the WiMAX Forum was created to allow conformity and inter-operability of MSs and BSs based on the IEEE 802.16 standard. The WiMAX Forum is composed of documents that specify the network reference model and procedures to evaluate the overall network QoS. The most recent recommendation published by the WiMAX Forum is the System Evaluation Methodology version 2.1 (WiMAX Forum, 2008).

In the next section, physical features of mobile IEEE 802.16 standard are described. Thus, these next subsections are focused in the OFDMA technique that is designed to support mobility.

2.2 Physical Layer

IEEE 802.16 standard specifies multiple physical interfaces using different multiplexing techniques and modulation and coding schemes, that are designed for several scenarios. Therefore, this section is divided in two subsections. A brief evolution of three main multiplexing techniques is presented in Subsection 2.2.1, focusing on the OFDMA since this technique is considered in this thesis due to its the best spectral efficiency. Then, the Subsection 2.2.2 shows the modulation and coding scheme that can be transmitted using OFDMA multiplexing technique.

2.2.1 Channel Multiplexing

Initially, Single Carrier (SC) physical layer was introduced by the IEEE 802.16 standard. In this multiplexing technique, an user transmits data using the whole available bandwidth, as can be seen in Figure 2.2 (a). SC is designed to operate in the range of frequencies between 10 and 60 GHz. Due to the characteristics of this range of frequencies, it is prone to channel impairments such as attenuation, and multipath delay spread, which affect the overall transmission reliability. Therefore, SC technnique requires Line of Sight (LoS) for correct operation.

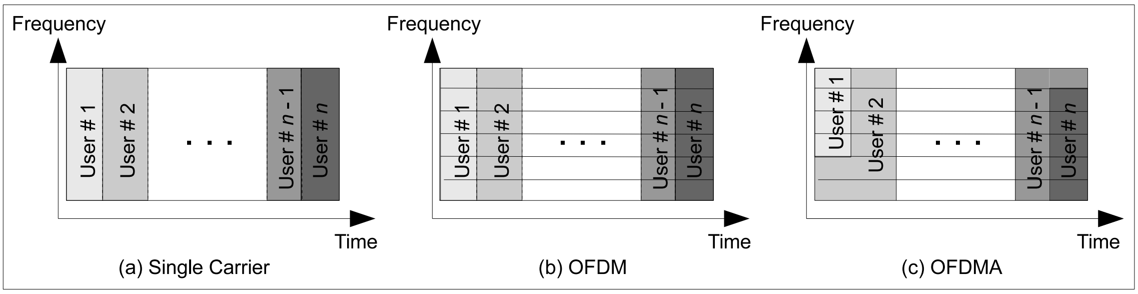


Figure 2.2: RF Channel Multiplexing Techniques

In order to turn possible the operation in Non-Line of Sight (NLoS) scenarios, the IEEE 802.16 standard, in 2004, proposed the implementation of OFDM using frequencies below 11 GHz, wherever the RF channel bandwidth can vary from 1.25 MHz to 28 MHz. OFDM technique was designed for nomadic (fixed) operation. This technique divides the RF channel into multiple subcarriers, *e.g.*, a 10 MHz channel is divided into 1024 subcarriers some of which are used for data transmission while others are reserved for synchronization. During transmission a given user receives complete control of all subcarriers. Therefore, since an user transmits data using all available frequencies, RF channel sharing in OFDM is essentially based on time division (CHIEOCHAN; HOSSAIN, 2008). The OFDM technique is shown in Figure 2.2 (b).

IEEE 802.16e standard was designed with the goal to permit operation in both nomadic and mobile scenarios. Toward this goal, the implementation of OFDMA as physical interface became mandatory in mobile IEEE 802.16 networks and thus it is the physical interface considered in this thesis. OFDMA is a multiplexing technique that combines time division and frequency division multiple access. In this technique, only a subset of subcarriers is allocated to each user, as can be seen in Figure 2.2 (c).

The available subcarriers are grouped into subchannels and the user's data is allocated using one or more subchannels during a specified number of transmission symbols. This technique is called subchannelization method. The standard specifies seven methods which can be used for subchannelization. However, only Partially Used Subchannelization (PUSC) is mandated by the IEEE 802.16 standard. The subchannelization process makes the bandwidth allocation mechanism more flexible, because the allocations are performed in the time (OFDM symbols) and frequency (subchannels) domains, using a data allocation unit, called slot. The exact size of a slot depends on the subchannelization method and on the transmission mode, *i.e.* UL or DL. The structure of DL and UL PUSC can be viewed in Figure 2.3. The DL slot (a) is composed of 2 blocks, each one formed by 14 subcarriers over 2 symbol times. Therefore, the DL slot is organized using 28 subcarriers forming a subchannel over two symbol times (SO-IN; JAIN; TAMIMI, 2009a). Furthermore, the combination between subcarriers and symbols creates 24 elements in each DL block, 20 elements are used for data and 4 are used for synchronization (pilot). On the other hand, the UL slot (b) is defined considering 6 blocks where each block is composed of 4 subcarriers and 3 symbol times. Thus, the UL slot consists of 24 subcarriers, assembling a subchannel over 3 symbol times. Moreover, the combination between subcarriers and symbols creates 12 elements in each UL block, 8 elements for data transmissions and 4 pilot elements.

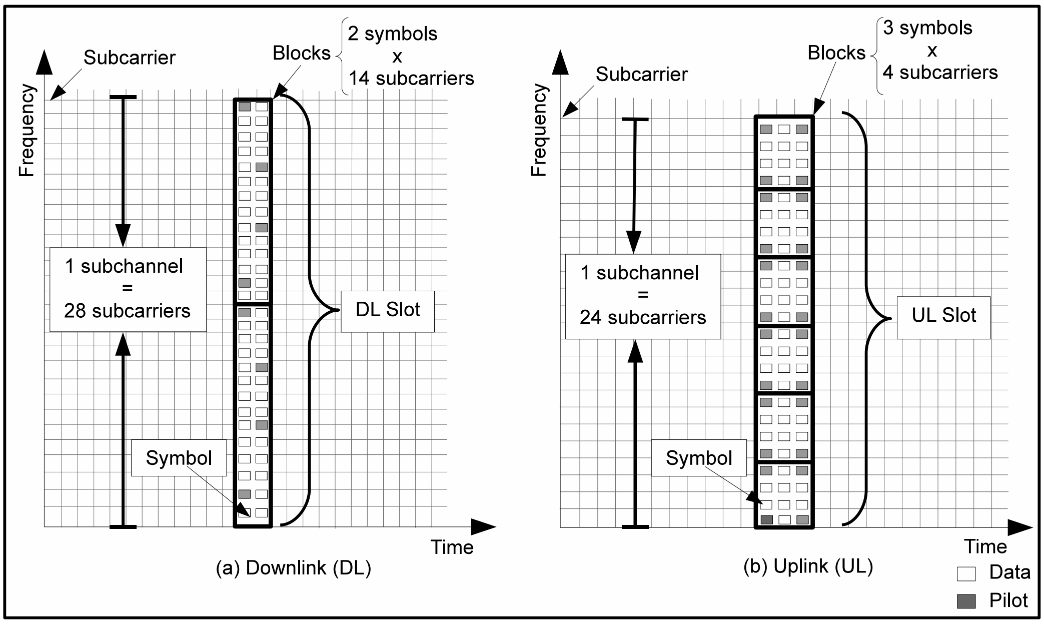


Figure 2.3: Downlink and Uplink PUSC

2.2.2 Modulation and Coding Scheme

The data are encoded using Forward Error-Correction (FEC) techniques before being transmitted. During FEC encoding, redundant information is added to the transmitted data to allow error detection and correction by the receiving device. FEC capabilities (*i.e.*, error detection and error correction) in IEEE 802.16 networks are obtained through the application of Reed-Solomon (RS), Convolutional Codes (CC), Convolutional Turbo Codes (CTC), or Low-Density Parity Check Codes (LDPC) (XU; LIANG; LEON, 2008). In fixed IEEE 802.16 networks, FEC is deployed through the concatenation of an inner RS code with an outer CC of compatible rate (IEEE, 2009). However, the implementation of concatenated codes demands high processing power, which generally is not available in mobile network devices.

Due to the processing and power consumption issues that result from the hardware limitations of mobile network devices, FEC encoding in mobile IEEE 802.16 networks is based only on the implementation of a standalone code. The mobile IEEE 802.16 standard mandates the implementation of CC for devices interoperability, while the deployment of other codes is optional.

After FEC encoding, the data bits to be transmitted must be mapped into a modulation constellation. A modulation constellation is formed of symbols that define the quantity of data that can be transmitted. The mobile IEEE 802.16 standard allows the usage of many Quadrature Amplitude Modulation (QAM) constellations. Therefore, it is possible to vary the quantity of bits transmitted per modulation symbol, which affects transmission capacity. The combination of a modulation constellation and a FEC encoding rate yields a MCS configuration. MCS configurations change dynamically to adapt the transmission reliability to the time variant propagation conditions typically found in the context of RF channels.

The transmission capacity of a RF channel varies according to the number of symbols that form a modulation constellation, as well as to the quantity of bits transmitted within a modulation symbol. The number of symbols that comprise a modulation constellation has a relationship with the amount of data that can be sent within each modulation symbol. This amount of data per symbol (b_{symb}) is typically represented in unit of bits and can be calculated through the equation 2.1, where n is the number of symbols inside a QAM constellation.

$$b_{symb} = log_2(n) \tag{2.1}$$

The modulation constellations defined in the mobile IEEE 802.16 standard are Quadrature Phase Shift Keying (QPSK), 16-QAM, and 64-QAM. The standard mandates the use of QPSK 1/2 MCS configuration to encode control information exchanged among BS and MSs, due to its robustness for ensuring successful transmissions to all MSs within a network. To meet interoperability criteria, the WiMAX Forum mandated that all devices must support both QPSK and 16-QAM, while the implementation of 64-QAM is optional. In total, seven MCS configurations are defined, depending on the modulation constellation and FEC encoding rates used. These MCS configurations, as well as the quantity of bits per symbol for each configuration, are shown in Table 2.1.

Table 2.1: IEEE 802.16 MCS Configurations for OFDMA Multiplexing

Modulation	**Bits per Symbol**	**Convolutional Code Rate**
QPSK	2	1/2
QPSK	2	3/4
16-QAM	4	1/2
16-QAM	4	3/4
64-QAM	6	1/2
64-QAM	6	2/3
64-QAM	6	3/4

MCS causes a trade-off between transmission robustness and data rates. It means that using a more robust configuration, the overall network data rate is reduced. However, in this case, there is a gain on transmission reliability, because there is a higher capability to detect and correct errors. On the other hand, applying a less robust MCS configuration reduces the transmission reliability, while increases data rate, since more data will be transmitted with less redundant information (WANG et al., 2008).

Changes in MCS configuration generally reflect time-variant propagation conditions of the RF channel. The propagation conditions are affected by physical impairments that cause errors in the transmission. The error correction capability depends on the amount of redundant data that is transmitted within a frame (KUNST et al., 2011). In the next sections, the frame structure, as well as service flows of the different classes of service transmitted in the IEEE 802.16 network are presented.

2.3 Medium Access Control Layer

The data plane of the IEEE 802.16 standard defines the MAC Layer into Convergence Sublayer and MAC Common Sublayer. These sublayers should support some components to provide QoS for the users of the network. This section is organized in three subsections to reflect the organization proposed by the standard. The Convergence Sublayer is presented in Subsection 2.3.1 and specifies the service flows and classes of service. The Subsection 2.3.2 describes the MAC Common Sublayer explaining the OFDMA frame structure and the bandwidth request mechanisms. Finally, Subsection 2.3.3 presents three main components used to provide QoS in IEEE 802.16 networks.

2.3.1 Convergence Sublayer

The Convergence Sublayer receives SDUs from the higher layers of the RM/OSI, that are classified and encapsulated into Protocol Data Units (PDUs). After, PDUs are sent to the MAC Common Sublayer. The IEEE 802.16 (IEEE, 2009) specifies two convergence interfaces. The first interface works with Asynchronous Transfer Mode (ATM) services. This interface receives cells from the ATM layer, classifies them into ATM SDUs according to the service flow and encapsulates them into PDUs, that are delivered to the MAC Common Sublayer. The second interface manages packet-based protocols such as Internet Protocol (IP), Point-to-Point Protocol (PPP), and Ethernet frames. The IEEE 802.16 defines a classification procedure where SDUs are mapped to specific connections associated with a service flow (MSADAA; C?MARA; FILALI, 2010). In this context, the definitions about connections and service flows are depicted as follows.

Connections and Service Flows

IEEE 802.16 is a connection-oriented wireless network, *i.e.*, each connection, identified by a unique Connection Identifier (CID), is associated to a service flow which characteristics describes its QoS requirements. Therefore, the CID associates an unidirectional connection between a BS and a MS. A connection only provides network access to one type of service, *e.g.* voice and e-mail cannot be associated with the same MAC connection (NUAYMI, 2007). In this context, each service flow is uniquely identified by a Service Flow Identifier (SFID) that can assume three states: provisioned, (ii) admitted, and (iii) active. The first one is usually used in the initial state of the connection admission. The second one is known as intermediate state of a connection request, where the resources for a connection are admitted and an end-to-end negotiation is performed. The last state occurs when the resources are reserved by the BS.

These three states can be combined to configure two different actions. Figure 2.4 shows the possible transitions among these different states. A BS can decide to activate a service flow directly from the admitted state, or can chose the service flow to pass by the provisioned state, and them by the admitted state before been activated. In this thesis, the service flow is considered to be activated directly by the admitted state, due to scope this work that does not consider handover characteristics. This decision is performed by the CAC component that is discussed in Subsection 2.3.3.

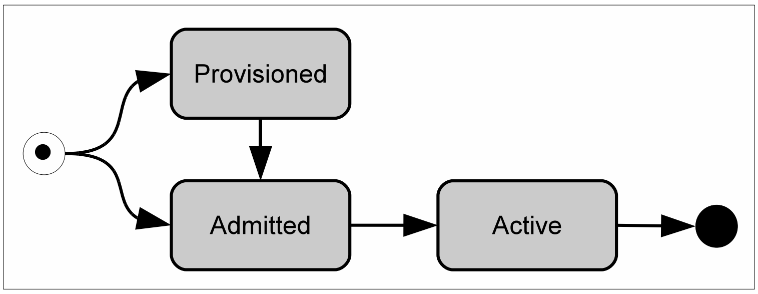


Figure 2.4: Transitions between the service flow states

A service flow is characterized by a set of QoS parameters that include details on how a MS may send bandwidth request messages carrying a traffic descriptor. The four main QoS parameters of the IEEE 802.16 standard (IEEE, 2009) are show as follows and the description of the bandwidth request mechanisms are presented in Subsection 2.3.2.

- **Maximum Sustained Traffic Rate** (MSTR) represents the peak information rate of a service flow. The rate is expressed in bits per second and is defined by SDUs that arrive in the network through encapsulated data from the higher layers. The MSTR parameter does not consider the protocols and network overhead, *e.g.* MAC headers.

- **Minimum Reserved Traffic Rate** (MRTR) specifies the minimum transmission rate reserved to a service flow. The rate is expressed in bits per second and represents the minimum amount of data to be transported within a service flow. However, the BS should reallocate the exceeding bandwidth for other purposes if less bandwidth than MRTR is requested for a connection. The value of this parameters should be determined after excluding the MAC overhead.

- **Maximum Latency** (ML) indicates the maximum time value between the reception of a packet at the Convergence Sublayer of the BS or the MS and the forwarding of the SDU to the PHY Layer.

- **Tolerated Jitter** (TJ) defines the maximum delay variation for a given connection.

ML and TJ represent a service commitment and should be guaranteed after the connection is accepted by the BS. Furthermore, the QoS parameters should be associated with a class of service of a specific connection (GHAZAL; BEN-OTHMAN, 2010). The IEEE 802.16 standard defines five classes of service that are describe as follows.

Classes of Service

IEEE 802.16 networks offer powerful tools to achieve different levels of QoS requirements. The MAC Layer provides QoS differentiation for several types of applications that may operate over these networks. The standard defines five classes of service: Unsolicited Grant Service (UGS), Extended Real Time Polling Service (ertPS), Real Time Polling Service (rtPS), Non Real Time Polling Service (nrtPS), and Best Effort Service (BE) (SO-IN; JAIN; TAMIMI, 2009b). The goal of this section is to present the traffic characteristics of the five classes of services such as fixed/variable packet sizes and types of applications related to each class of service. The characteristics of these classes of service are describe as follows. The bandwidth request mechanisms used by theses classes are presented in Subsection 2.3.2.

- **UGS** is designed to support real-time applications that generate fixed-size data packets at periodic intervals. For example, this would be the case of Pulse Coded Modulation (PCM) for phone signal transmission and Voice over Internet Protocol (VoIP) without silence suppression. The four QoS parameters described in the "*Connections and Service Flows*" topic are important for this class of service, as can be seen in Table 2.2.

- **ertPS** is a service introduced by the IEEE 802.16e-2005 amendment. ertPS is suitable for real-time service flows that generate variable size data packets on a periodic interval, such as VoIP with silence suppression, where no traffic is transmitted during silent periods. The QoS parameters are the same as those in UGS.

- **rtPS** is designed to support real-time data streams that generate variable-size data packets on periodic intervals, such as Moving Pictures Expert Group (MPEG) video transmissions. The QoS parameters are similar to those guaranteed by UGS, but MSTR and MRTR need to be specified separately. For UGS and ertPS services, these two parameters have the same value, when applicable.

- **nrtPS** is defined to carry delay-tolerant data streams consisting of variable-size data packets. Only MRTR is guaranteed. For example, this class of service would support File Transfer Protocol (FTP) transmissions.

- **BE** is specified to support data streams for which no minimum service guarantees, such as throughput and delay, are required. HyperText Transfer Protocol (HTTP) and Simple Mail Transfer Protocol (SMTP) are examples of applications using this class of service.

The classification of traffic flows into these classes of service facilitates bandwidth sharing between different users. According to the QoS parameters, the BS manages the amount of bandwidth required by each application. This mechanism allows an efficient and adapted distribution of the available resources. Each class of service has a mandatory set of QoS parameters that is associated with a service flow. Table 2.2 shows the mandatory QoS parameters for each one of the five classes of service (MSADAA; C?MARA; FILALI, 2010) (LU; MA, 2010).

After this classification, the PDUs are sent to the MAC Common Sublayer. In this context, the main goal of the next subsection is to discuss this sublayer that belongs to the data plane of the IEEE 802.16 standard.

Table 2.2: QoS Parameters for Each Class of Service

Classes of Service/ QoS Parameters	UGS	ertPS	rtPS	nrtPS	BE
MSTR	✓	✓	✓	✓	✓
MRTR	✓	✓	✓	✓	
ML	✓	✓	✓		
TJ	✓	✓			

2.3.2 MAC Common Sublayer

The MAC Common Sublayer receives PDUs from Convergence Sublayer, through the MAC SAP. In this sublayer the QoS requirements are guaranteed considering the service flow of each connection. The IEEE 802.16 standard defines a set of messages that are exchanged between the BS and the MSs to allow data transmission. The transmission of both control messages and data must follow the OFDMA frame structure. Therefore, the OFDMA frame is explained in the "*Frame Structure*" topic. Another important function of the MAC Common Sublayer is to manage the bandwidth request mechanisms among the BS and MSs. These mechanisms are discussed in the "*Bandwidth Request Mechanism*" topic.

Frame Structure

The structure of the OFDMA frame can be represented as a bi-dimensional matrix, in which x-axis represents the time domain of OFDM symbols, while y-axis indicates the frequency domain, *i.e.* the logical subchannels (COHEN; KATZIR, 2008). The structure of the OFDMA frame considering PUSC subchannelization method is shown in Figure 2.5. The frame is divided in DL and UL intervals, which are also called subframes. DL subframe starts with a preamble, which is 1 symbol long, and is used for synchronization purpose. The preamble is followed by Frame Control Header (FCH), allocated in the first 4 subchannels of the second OFDM symbol, carrying information regarding MCS configurations used for transmission of control information within the subframe, commonly known as Burst Profile.

The next part of DL subframe is DL-MAP which defines the burst start time, as well as the first subchannel allocated to the transmission of each burst used to transmit data. Furthermore, DL-MAP informs the number of OFDMA symbols and subchannels allocated for data transmission using a structure called Information Element (IE). A DL IE is composed of the Downlink Interval Usage Code (DIUC), the list of CIDs that will be transmitted with the same burst profile. Following the list of CIDs, the first data burst is transmitted, containing UL-MAP, Downlink Channel Descriptor (DCD) and Uplink Channel Descriptor (UCD). UL-MAP carries information about start time and duration of ranging subchannel and the duration of transmission opportunities granted to be used during the next UL subframe. DCD and UCD broadcast physical configurations of the RF channel for DL and UL subframe, respectively. The remainder of DL subframe is used for transmitting data. However, padding can be added in cases where there is not enough data to complete the subframe duration. Finally, Tx/Rx Transition Gap (TTG) is defined between DL and UL subframes, and Rx/Tx Transition Gap (RTG) is specified between UL and following DL subframe (belonging to the next frame) to allow MS terminals to turn from transmission to reception and from reception to transmission, respectively (BOTH et al., 2008).

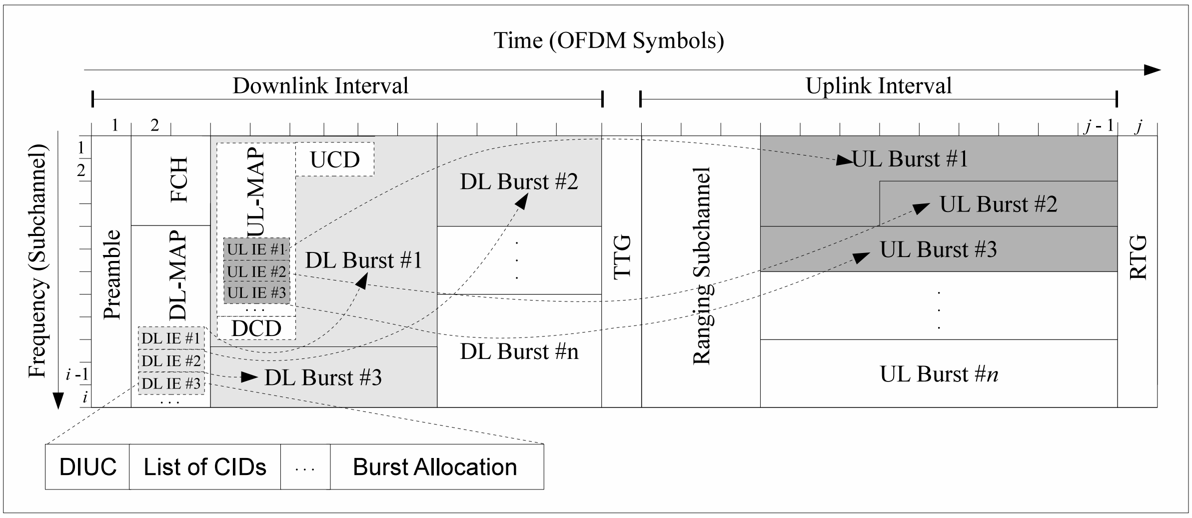


Figure 2.5: OFDMA Frame Structures

UL subframe is divided in a ranging channel, bandwidth request mechanisms, and data bursts, which carry data transmissions previously scheduled by the BS. Ranging subchannel access is shared among all MSs within the coverage area of a given BS (LEE; CHO, 2005). MSs can choose which type of request to use: Initial Ranging (IR), Hand-Off Ranging (HR), Periodic Ranging (PR) or Bandwidth Request (BR) ranging, and inform its interest using Code Division Multiple Access (CDMA). Upon the reception of a code, BS recognizes the type of the request and then takes the proper decisions.

The control information and the maps described previously are fundamentals for the arrangement of the user's data into the OFDMA frame. Due to the bi-dimensional structure of the OFDMA frame, the size occupied by this control information cannot be disregarded in the allocation procedure of the OFDMA frame (SO-IN; JAIN; TAMIMI, 2009a). Table 2.3 shows the size of the maps of the OFDMA frame, as discussed in Figure 2.5. In this context, the control information and maps also are called the overhead.

Table 2.3: Overhead of the OFDMA Frame

Message type	**Fixed (bits)**	**Variable (bits)**	
FCH	96	-	
DCD	24	$24.N_{R-DL}$	
UCD	48	$24.N_{R-UL}$	
DL-MAP	104	$(12.DL_{IE})$ + $(48.N_C)$	
UL-MAP	64	$(20.UL_{IE})$ +	$(32.N_B)$
			$(40\ .N_A)$
			$(10.N_D)$

Except for FCH, which has fixed length of 96 bits, the remaining messages are composed of two types of information: fixed and variable. The channel descriptors (DCD and UCD) are related to the number of data bursts, N_{R-DL} e N_{R-UL}, respectively. In other words, for each new data burst allocated within the OFDMA frame, 24 bits of overhead are inserted. The DL-MAP size is associated with the number of connections (N_C) in each burst and the amount of DL bursts within the OFDMA frame DL_{IE}. Finally, the size of the UL-MAP is related to the information type transmitted, *i.e.* CDMA bandwidth request (N_B), CDMA allocation (N_A), or UL data (N_D) and the amount of UL bursts within the OFDMA frame UL_{IE} sent by the MSs to the BS. This information makes up the bandwidth request mechanisms of the IEEE 802.16 standard that is discussed as follows.

Bandwidth Request Mechanisms

The PMP topology of the IEEE 802.16 is a centralized architecture, where the BS controls the QoS parameters. In the DL transmission, the BS has complete knowledge of the user's traffics, such as arrival and data queue lengths. In this topology, the control information is sent to MSs through broadcast transmissions. On the other hand, in the UL transmission, the MSs need to send BR messages to the BS, which then may grant a transmission opportunity to each MS in the subsequent UL subframe (SO-IN; JAIN; TAMIMI, 2009b). In this context, UL transmissions use one of the four mechanisms presented as follows.

Unsolicited request is a mechanism that periodically allocates resources within the OFDMA frame for a class of service dispensing the need of a BR by the MS. This mechanism is useful for applications that require a fixed rate data stream. Thus, unsolicited request is used only by UGS and ertPS classes of service, as can be seen in Table 2.4. The main advantage of this mechanism is to guarantee a minimum latency to a given MS for real-time service without additional overhead. On the other hand, the main disadvantage is the waste of resource when the bandwidth is granted and the service flow has no data to send (CHUCK; CHEN; CHANG, 2010).

Piggybacking sends a BR inside a data transmission from MS to BS. The piggyback uses the grant management subheader which is transmitted in a generic MAC frame. Thus, the main advantage of this mechanism is to reduce the overhead by avoiding the transmission of a complete BR request message, decreasing the waiting time for a pool of the transmission opportunity as a consequence. This mechanism can be used by all classes of service, except UGS. In addition, this mechanism is not mandated by the IEEE 802.16 standard (CHOU et al., 2010).

Contention-based is the mechanism that uses the ranging subchannel resource, more specifically the BR service, described in Subsection 2.2.1. For this resource, contention-based BR is used by ertPS, nrtPS, and BE classes of service. Initially, the BR message using CDMA code is sent from the MS to the BS in the ranging interval. The MS chooses the transmission time and the CDMA code randomly and then requests the service to the BS (LEE; CHO, 2005). In this case, collisions can occur when more than one MS selects the same transmission time and the same CDMA code. Thus, the *backoff* algorithm should be used to reduce the probability of new collisions. The *backoff* algorithm defined in the IEEE 802.16 standard is based on a truncated binary exponential, where *backoff* window (J_b) is equal to (2^b) and b can assume values between 0 and 15. After this initial process, the BS receives the BR CDMA code from MS and allocates a transmission opportunity. Therefore, the MS receives the transmission opportunity in the next UL subframe. This transmission opportunity is called of N_A or N_B being addressed to only one MS.

Polling-based reserves a BR request opportunity to a given BS in order to avoid collisions. The polling-based BR is known as a contention-free mechanism, which is designed for high priority traffic, where part of data bursts within the UL subframe is used to send BR messages (SAYENKO; ALANEN; HAMALAINEN, 2007). This mechanism can be used by all classes of service, except UGS, as shown in Table 2.4. Furthermore, the polling-based BR can be granted to a group of MSs (multicast polling) or for all MSs belonging to the network (broadcast polling). In this cases, collisions may occur, thus the MSs may use the *backoff* algorithm described previously. However, these cases are not mandatory mechanisms according to the IEEE 802.16 standard.

Table 2.4: Bandwidth Request Mechanisms for Each Class of Service

Classes of Service/ BR Mechanisms	**UGS**	**ertPS**	**rtPS**	**nrtPS**	**BE**
Unsolicited request	✓	✓			
Piggybacking		✓	✓	✓	✓
Contention-based		✓		✓	✓
Polling-based		✓	✓	✓	✓

The association of bandwidth request mechanisms with the QoS parameters of the service flows are crucial to design components that provide QoS in IEEE 802.16 networks, *e.g.* admission of flow services provisioned, admitted, and active (CHUCK; CHEN; CHANG, 2010). These QoS components are discussed in the following subsection.

2.3.3 Quality of Service Components

The IEEE 802.16 standard (IEEE, 2009) defines a set of components that must be implemented to provide guaranteed QoS. The three main QoS components specified for the MAC Layer in the BS are: CAC, scheduler, and allocator. Figure 2.6 shows a global view of these components that use the services of the Convergence Sublayer and MAC Common Sublayer. These components should be integrated with information of the PHY Layer. However, the standard does not specify any implementation detail, which is intentionally left opened to allow future technological advances and vendors specific solutions (AHMADI, 2009). The goal of this subsection is to discuss the main features of CAC, scheduler, and allocator components.

Connection Admission Control

The CAC is the first QoS component employed to guarantee the requirements of the users' applications. This component accepts or denies connections taking the network loading conditions into consideration. To admit a connection, BS must receive information from the MS requesting the connection. The IEEE 802.16 standard defines ranging and register messages to provide such information to the CAC component. As presented in Figure 2.7, ranging and register messages are exchanged between the MS (through the ranging manager and the connection manager) and the BS.

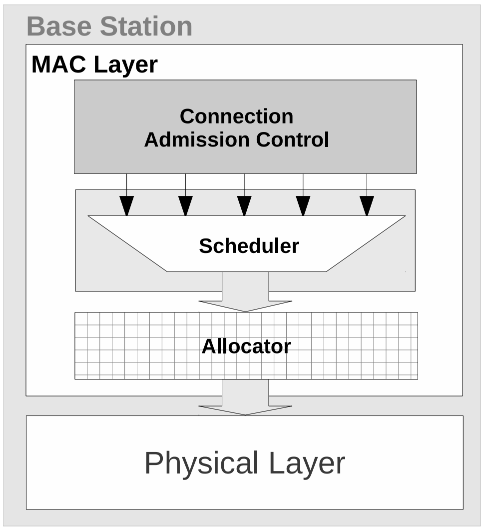


Figure 2.6: QoS Components

Ranging messages are exchanged continuously between a given MS and the BS. This type of message is not associated with a connection request, but to the physical conditions of the RF channel. Ranging mechanism is divided in two phases: initial and periodic ranging. The goal of the initial phase is the negotiation of transmission power and time synchronization between the MS and the BS. In this phase, the MS receives a basic CID. The periodic ranging phase starts after the initial negotiation between MS and BS. In this phase the MS receives a primary CID to exchange two types of messages: Ranging Request (RNG-REQ) and Ranging Response (RNG-RSP). These messages are used to handle time variant conditions of the RF channel between MS and BS. Examples of information inside these messages are signal power, modulation scheme, and coding rate.

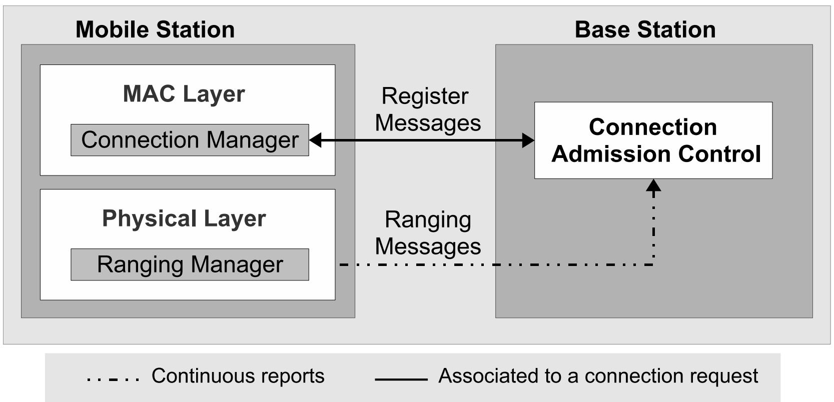


Figure 2.7: IEEE 802.16 Messages Related to CAC Component

Register messages are directly related to requests to admit a new connection that is associated with a service flow provisioned, admitted, or active. To establish a new connection, a given MS must request a registration to the BS. The registration is granted after exchanging two register messages: Register Request (REG-REQ) and Register Response (REG-RSP). The register messages enclose traffic descriptors that contain QoS parameters, associated with the service flow of the new connection. These messages can deploy three types of actions to manage the QoS parameters of the connections: Dynamic Service Addition (DSA), Dynamic Service Change (DSC), and Dynamic Service Deletion (DSD). Based on these actions, the MS can request modifications on its QoS parameters to the BS to handle changes on service conditions.

After the CAC component receives and analyzes the ranging and register messages, the service flow of the accepted connection is forwarded to the scheduler component. The scheduler is the second QoS component which is discussed in the following subsection.

Scheduler

The scheduler is the main component of the MAC Layer that helps to guarantee QoS to several service types. The scheduler component can be organized in queues and scheduling algorithms. The scheduling queues are used to classify the resources within the classes of service defined by the IEEE 802.16 standard. The resources are different according to the transmission types. In DL transmissions, the scheduler works with the users' data queues and in UL transmissions, the scheduler serves BR queues of the MSs, as discussed in Subsection 2.3.2. In this context, the scheduling algorithms can be organized considering DL and UL transmissions, as can be seen in Figure 2.8.

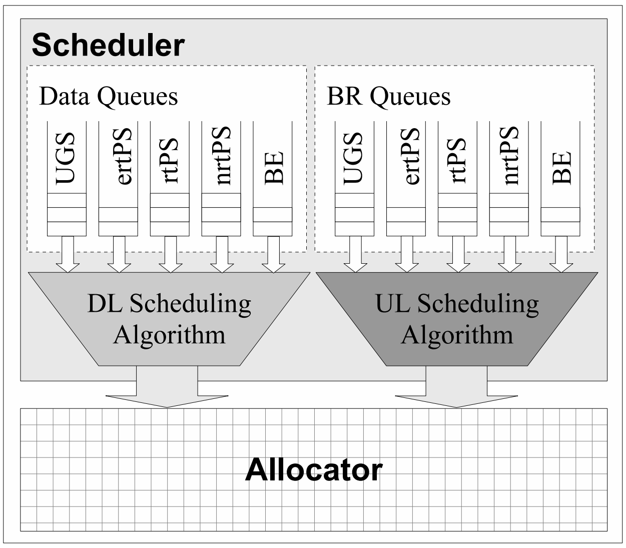


Figure 2.8: Scheduler Component

The scheduling algorithms have the function of serving the scheduling queues, *i.e.* to serve the queues considering mainly priority-based scheme for the classes of service. For example, the priority-base scheme can be: UGS, ertPS, rtPS, nrtPS and BE, respectively, or the resources with the largest delay can be considered at the highest priority. To guarantee QoS to different classes of service many proposals have been developed in the scientific community regarding to queues management and scheduling algorithms that are represented in Subsection 3.1.2.

The last function of the scheduler component is to send the user's data and BRs scheduled in the queues to the allocator component. The allocator aims to organize the DL and UL resources using different methods that are discussed in the following topic.

Allocator

The main function of the allocator component is to arrange the user's data and BRs received from the scheduler component within the OFDMA frame. As described in Subsection 2.2.1, the rectangular data regions formed by burst profiles are defined on a two-dimensional domain, composed of time and frequency. The resources (slots) must be allocated within these regions. However, the allocation mechanism implementation in the DL subframe is different from the implementation in the UL subframe, due to the method used to access the RF channel in each case. During DL period, BS broadcasts information to the MSs, while in UL period the access to the RF channel is shared among MSs. The allocation mechanisms in DL and UL subframes are described as follows.

The DL allocation mechanism should be performed using an allocation algorithm to arrange the burst profiles within the DL subframe. This algorithm uses the technique known as bin packing. Some proposals regarding to allocation algorithms based on OFDMA frame are presented in Subsection 3.1.3. After the bin packing algorithm calculates the rectangular data region for allocation, the slots that form the burst profile must be allocated, as presented in Figure 2.9 (a).

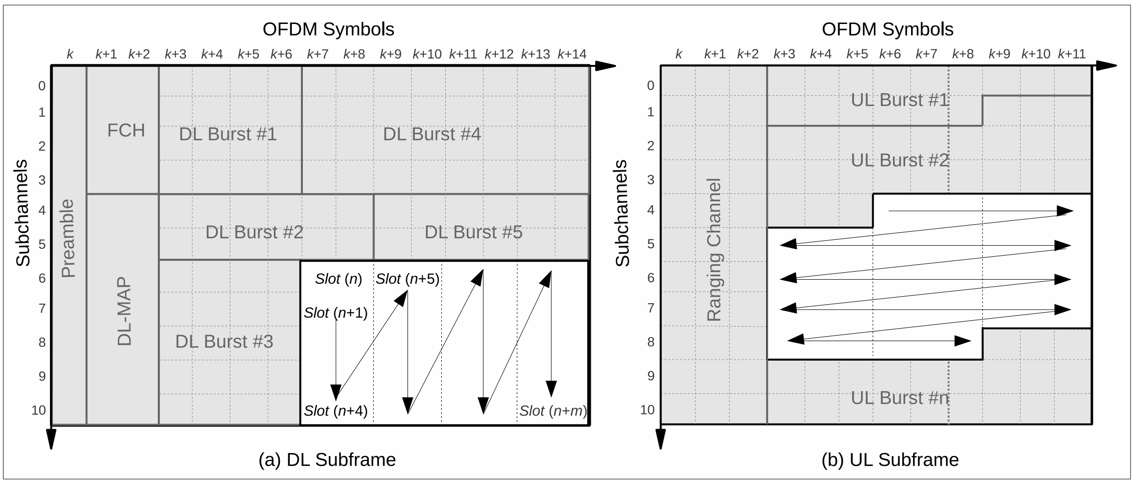


Figure 2.9: Allocation Mechanisms

On the other hand, the UL allocation mechanism arranges the burst profiles horizontally from the left to the right beginning in the top of the UL subframe. The slots of the burst profile are allocated using subchannels (y-axis), when the last selected subchannel is reached, the allocation continues in the next available OFDM symbol, as highlight in Figure 2.9 (b). This method for allocation of burst profiles in UL transmissions is defined by the IEEE 802.16 standard.

2.4 Summary

This chapter presented an overview of the IEEE 802.16 standard, summarizing the evolution since its first release in 2002, up to IEEE 802.16m amendment, published in 2009. Next two sections were organized according to the layers defined in the data plane of the standard. In the PHY Layer, a brief evolution of three multiplexing techniques were presented, highlighting the OFDMA technique that has the best spectral efficiency and is the base of this thesis. Moreover, the MCS schemes were discussed, since they cause a trade-off between transmission robustness and data rates in IEEE 802.16 networks. Still, in the MAC Layer, the features of the Convergence Sublayer and MAC Common Sublayer have been discussed. The former refers to the classification procedure where SDUs are mapped into specific connections associated with its service flow. The latter aims to guarantee the QoS requirements considering the service flow of each connection. Therefore, the OFDMA frame and the bandwidth request mechanisms were described in this sublayer. Finally, the main features of the CAC, scheduler, and allocator components that provide guaranteed QoS specified for the MAC Layer in the BS were discussed in this chapter.

3 STATE-OF-THE-ART

This thesis focuses on the integration of the main QoS components in a new cross-layer adaptive architecture for IEEE 802.16 networks. To understand the state-of-the-art associated with the integration of the QoS components, this chapter is organized as follows. First, the related work on each QoS component is presented in Section 3.1. Second, the proposals published in the literature on cross-layer architectures are discussed in Section 3.2. Finally, the challenges to design a cross-layer adaptive architecture considering the related work are defined in Section 3.3.

3.1 QoS Components

The goal of this section is to present the related work to the three main QoS components of IEEE 802.16 networks, *i.e.* CAC, scheduler, and allocator. The features of the CAC component are discussed in Subsection 3.1.1. The organization of the scheduler component, as well as the scheduling algorithms are presented in Subsection 3.1.2. In addition, the main algorithms of the OFDMA frame allocator component are described in Subsection 3.1.3.

3.1.1 Connection Admission Control

The function of the CAC component is to decide whether to accept or not new connections, while making sure that the available resources are sufficient for both the ongoing and the incoming connections. According to Msadaa, Câmara and Filali (MSADAA; C?MARA; FILALI, 2010), to take such decision, two degradation policies can be adopted when no resources are available for the new connections, as can be observed in Figure 3.1: (i) the policy can be flexible, *i.e.* consists in degrading existing connections to manage the resources for the new one, or (ii) can be conservative, maintaining the QoS provided for ongoing connections and simply rejecting new connection. To discuss the issues about CAC component, this subsection is organized in four main topics. The first topic presents the policy with degradation of the ongoing services. The second topic discusses the policy without degradation of existing services. The third topic addresses specific investigations about some CAC solutions without degradation policy. Finally, a comparative analysis of different aspects of the CAC solutions discussed in this subsection is performed.

Policy with Degradation of Services

The main goal of the policy with degradation of service is to decrease the resources provided to the ongoing connections when necessary and possible. This policy permits the CAC to admit new connections. Msadaa, Câmara and Filali (MSADAA; C?MARA;

FILALI, 2010) classify this policy into three categories: (i) services degradation, (ii) bandwidth borrowing and (iii) bandwidth stealing. These categories can be combined with a threshold-based strategy to avoid starvation, or guard channel strategy that reserves a dedicated amount of bandwidth for several classes of service, *e.g.* UGS and handover services. The three categories of CAC algorithms are described as follows.

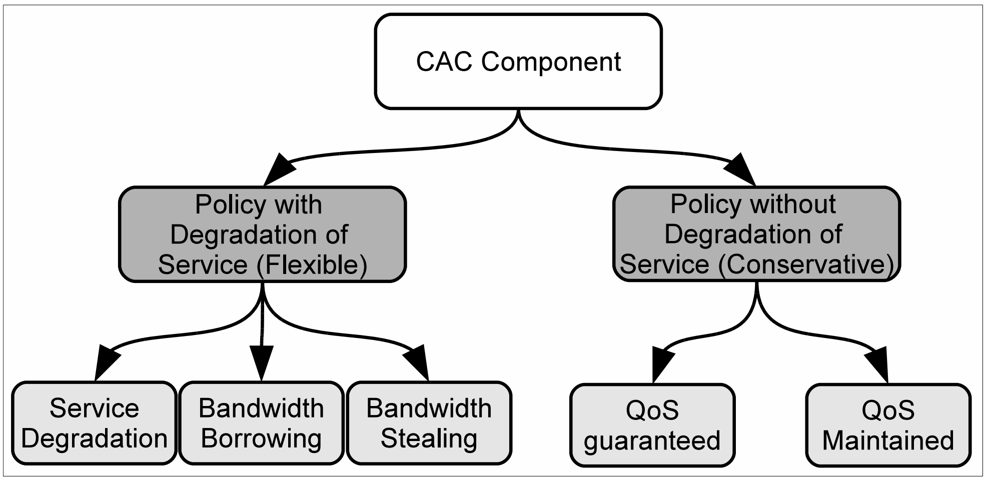


Figure 3.1: Classifications of the CAC Component

Service degradation consists in decreasing the bandwidth assigned to existing service flows which priority is lower than a new connection. Ge and Kuo (GE; KUO, 2006) propose the use of this category of CAC algorithms only in handover services. In other words, if the available bandwidth has not requirements of handover, the service degradation is applied and a new connection is accepted, only if the current available bandwidth guarantees the minimum bandwidth requirement.

Bandwidth borrowing corresponds to the category of the CAC algorithms in which the bandwidth assigned to some rate-adaptive connections is decreased in case of network congestion. In this category, several works were proposed. For example, Niyato and Hossain (NIYATO; HOSSAIN, 2007) defined the use of the bandwidth borrowing considering a non-cooperative game strategy. The players in this game are the rtPS and nrtPS connections which aim to maximize their QoS performance. The goal of this work is to find the equilibrium point between the two types of connections to offer bandwidth for the new connection and guarantee the QoS requirements of both ongoing and new connections. Another approach was proposed by Wang, Li and Agrawal (WANG; LI; AGRAWAL, 2005) integrating bandwidth borrowing and stepwise degradation. In this CAC algorithm a predetermined amount of bandwidth is exclusively reserved for UGS connections. An UGS connection is accepted if there is enough bandwidth to accommodate its requirements, otherwise it is rejected. In addition, Wang *et al.* (WANG et al., 2007) proposed the CAC algorithm combining a guard channel policy and a proportional bandwidth borrowing. Indeed, a guard channel corresponds to a percentage of the channel capacity that is reserved for handover connections. Thus, a new connection is blocked if the available bandwidth is smaller than a threshold, while a handover connection is blocked only if no bandwidth is available. A proportional bandwidth borrowing is applied when the required bandwidth is not available.

Bandwidth stealing is the category based on token-bucket. Jiang and Tsai (JIANG; TSAI, 2006) projected a CAC algorithm where each UL connection is characterized by two parameters: a token rate and a bucket size. Moreover, the rtPS services have an extra parameter corresponding their delay requirements. When a MS requests a new connection to the BS, the CAC algorithm is applied as follows. If the required bandwidth is smaller than the remaining UL capacity, the connection is accepted. Otherwise, a bandwidth stealing policy with a threshold-based strategy is applied. When connections belonging to low priority classes of service are using more bandwidth than their respective thresholds, the new connection can be accepted, if the UL capacity is greater than or equal to its bandwidth requirement. In another way, the capacity occupied by connections belonging to the same class of the new one is checked. If the capacity is greater than the threshold, the new connection request is rejected. Otherwise, a bandwidth stealing is attempted from connections belonging to high priority classes. This last step is possible only if the capacity of these higher classes exceed their thresholds.

Policy without Degradation of Services

The QoS guarantees are defined in the admission of a new connection. In this sense, the degradation of QoS of the ongoing connections should be avoided because this reduces user satisfaction. The disregard in the QoS guarantees may result in eventual revenue loss for the service provider (FERREIRA; CARVALHO; LIMA, 2010). In this context, a new connection is accepted based on a policy without degradation of services considering two requirements: (i) the new connection should receive QoS guarantees, *e.g.* real time services should have guarantees of bandwidth and delay, and (ii) the QoS of existing connections should be maintained.

A related work based on policy without degradation of services was proposed by Chandra and Sahoo (CHANDRA; SAHOO, 2007). The main goal of this work is to ensure QoS guarantees, in terms of bandwidth, delay and jitter. Therefore, an hyper interval is defined to test the admissibility of the requests, because the connections have different QoS requirements. The authors consider the delay and jitter requirements for UGS, rtPS, and even nrtPS connections which may cause the blocking of nrtPS connections due to the jitter requirement. This blocking technique is one of the main features of the policy without degradation of service.

Other CAC Solutions

This topic introduces some specific investigations about CAC solutions without degradation policy. For example, Yang and Lu (YANG; LU, 2006) have proposed a CAC solution specifically dedicated for real-time video applications. To maximize the throughput and to minimize the delay of the ongoing connections, the authors have integrated CAC and scheduling algorithms. The main limitation this work is that it only addresses a specific type of application, *i.e.* real-time video.

Another important investigation in CAC is to take into account the RF channel conditions. However, this investigation is scarce in the literature about IEEE 802.16 networks despite of the importance of this feature to the CAC component. Indeed, a CAC solution needs accurate information regarding to the available resources to admit a new connection. In this context, Kwon *et al.* (KWON et al., 2006) proposed a solution that incorporates MCS into the CAC process. The proposal is based on a Markovian model that considers handover and connections which can change modulation scheme. The model supports only two types of modulations and is built based on the assumption that all the connections have fixed and equal bandwidth requirements which limits its applicability (MSADAA;

C?MARA; FILALI, 2010) . On the other hand, the work proposed by Rong, Qian and Chen (RONG; QIAN; CHEN, 2007a) considers the power of the RF channel in the CAC component. The authors have proposed an adaptive power allocation mechanism to share the limited power resource of BS among different SSs. However, this investigation is based on fixed WiMAX and the bandwidth of each SS is also fixed.

The aim of the previous topics in this subsection was to present the state-of-the-art about CAC components. To complete this investigative study in the next topic a comparative analysis of main CAC solutions discussed in this subsection is performed.

Comparative Analysis of CAC Solutions

Table 3.1 summarizes the different aspects taken into account in the CAC proposals. The comparative study among the works highlights a degradation policy and the metrics used to provide QoS guarantees, *e.g.* data rate, delay, and jitter used by CAC solutions. Moreover, analyzing Table 3.1 it is possible to observe that the RF channel conditions have been considered in two investigations only. However, it is a key feature that should not be ignored in the CAC component.

Table 3.1: Comparasion on CAC proposals

CAC proposals	**Degradation policy**	**Data rate**	**Delay**	**Jitter**	**RF Channel conditions**
Ge and Kuo, 2006	✓	✓			
Niyato and Hossain, 2007	✓	✓	✓		
Wang *et al.*, 2007	✓				
Jiang and Tsai, 2006	✓	✓	✓		
Chandra and Sahoo, 2007		✓	✓	✓	
Kwon *et al.*, 2006					✓
Rong, Qian, and Chen, 2007					✓

The new cross-layer adaptive architecture presented in this thesis supports all metrics used to provide QoS guarantees, as well as the policy without degradation of services. This policy was chosen due to main goal this work, *i.e.* the architecture must guarantee QoS to increase the user satisfaction. In this context, the degradation policies belonging to the CAC component should receive information from other QoS components of the MAC Layer and of the PHY Layer. One of the main components that exchange information with the CAC is the scheduler. This component helps to assure QoS to various classes of service and is still an open research area. Therefore, in the following subsection the main related work about the scheduler component are described.

3.1.2 Scheduler

The aim of this subsection is to present the scheduler proposals for IEEE 802.16 networks. Special attention is given to related work associated with the DL and UL scheduling algorithms that should be employed in the BS. According to So-In, Jain and Tamimi (SO-IN; JAIN; TAMIMI, 2009b) the scheduler component can be classified into two main categories: "*Channel-Unaware Schedulers*" and "*Channel-Aware Schedulers*", as can be seen in Figure 3.2. The former uses no information regarding to the channel state condition to take the scheduling decision. On the other hand, the latter considers the channel state conditions in scheduling decisions.

Channel-Unaware Schedulers

The category of channel-unaware schedulers does not use information about channel conditions such as power level, transmission errors and loss rates. In other words, in this approach, the researchers assume perfect RF propagation conditions, *i.e.* no loss, and unlimited power source. This category can be organized in Intra-class and Inter-class scheduling, as presented in Figure 3.2. This organization also can be classified considering a hierarchical structure of two levels. Indeed, simple structures combining Intra-class and Inter-class are also found in the literature. However, So-In, Jain and Tamimi (SO-IN; JAIN; TAMIMI, 2009b) does not consider the simple structure approach because the proposals usually are restricted to a specific problem and do not analyze all features of the scheduler component for IEEE 802.16 networks. The two classes of channel-unaware scheduling are discussed as follows.

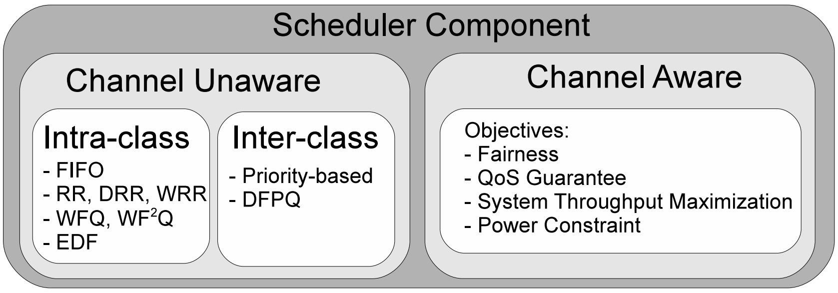


Figure 3.2: Classifications of the Scheduler Component

Intra-class Scheduling organizes the traffic within the queues of each class of service. In this context, So-In, Jain and Tamimi (SO-IN; JAIN; TAMIMI, 2009b) define three groups of scheduling algorithms: (i) Round Robin (RR), (ii) Weighted Fair Queuing (WFQ), and (iii) Delay-based. The first group can be considered the very simple scheduling algorithms, aside from First In First Out (FIFO). However, RR cannot assure QoS for different service classes. Therefore, Weighted Round Robin (WRR) has been applied for IEEE 802.16 scheduling (SAYENKO et al., 2006), (SAYENKO; ALANEN; HAMALAINEN, 2008). The weights can be used to adjust the throughput and delay requirements. In addition, to avoid missing opportunities, variants of RR such as Deficit Round Robin (DRR) or Deficit Weighted Round Robin (DWRR) can be used to serve variable sizes of packets (CICCONETTI et al., 2006). The second group of scheduling algorithms is an approximation of General Processor Sharing (GPS). In the WFQ algorithm each connection has its own FIFO queue and the weight can be dynamically assigned for each queue. Thus, the algorithms based on the WFQ use the virtual time concept (LERA; MOLINARO; PIZZI, 2007). These algorithms keep the delay bound to achieve worst-case fairness property, as the Worst-case Fair Weighted Fair Queueing (WF^{2}Q) algorithm. The last group of scheduling algorithms is designed for real-time services where the delay bound is the primary QoS parameter. The algorithms of this group are usually based on the Earliest Deadline First (EDF) algorithm to serve the connection based on the deadline criteria in the Intra-class scheduling.

Inter-class Scheduling defines the order in which queues of the Intra-class scheduling should be served. For the purpose of guaranteeing QoS to different classes of services, algorithms based on priority have been used for Inter-class scheduling in IEEE 802.16 networks. An interesting research in this context is to define the weights and/or the amount of resources that each class should receive. Queue length can be also used to set the priority level, *e.g.*, more bandwidth is granted to connections with longer queues (NIYATO; HOSSAIN, 2005). The disadvantage of the priority-base algorithms is the occurrence of starvation in some connections of low priority classes of service. This feature can reduce the overall throughput of the network. In this context, new algorithms have been proposed, such as Deficit Fair Priority Queuing (DFPQ) that maintains the maximum allowable bandwidth for each service class (CHEN; JIAO; WANG, 2005).

Channel-Aware Schedulers

According to So-In, Jain and Tamimi (SO-IN; JAIN; TAMIMI, 2009b), it is not valid to consider perfect RF channel conditions, no loss and unlimited power source in the wireless medium with mobility, due to the high variability of RF, such as signal attenuation, fading, interference, and noise. Therefore, in this class the usage of RF conditions in the scheduler component is briefly discussed.

Usually, the channel-aware schedulers provide resources to the MSs with better RF quality to exploit the multiuser diversity. The scheduler component can use the property of multiuser diversity to increase the overall throughput and to support more users. A strategy of the scheduler is to serve the MS with the best RF condition, and not scheduling resources for MSs with high error rate, because the packets would be dropped anyway. Channel-aware schedulers can be classified considering four key objectives: (i) fairness, (ii) QoS guarantee, (iii) system throughput maximization, and (iv) power optimization.

The first objective achieves long term fairness, but cannot guarantee the delay constraint (RUANGCHAIJATUPON; JI, 2008). The second aims at providing throughput and delay guarantees considering a threshold probability (PARK; CHO; BAHK, 2006). On the other hand, the third goal is to maximize the overall throughput, but cannot provide QoS requirements especially delay and unfairness (SINGH; SHARMA, 2006). Finally, the last objective is to minimize the power consumption, but also cannot provide QoS requirements (LIANG; CHEW; KO, 2008).

Comparative Analysis of Scheduler solutions

The aim of this topic is to perform a comparative study about scheduling solutions presented in the IEEE 802.16 literature. Table 3.2 shows the related work according to the classification defined by So-In, Jain and Tamimi (SO-IN; JAIN; TAMIMI, 2009b), *i.e.* channel-aware categories, hierarchical structure using Intra and Inter classes, and QoS guarantees. In this context, to the best of our knowledge, no research addressing all integrated characteristics presented in Table 3.2 has been proposed so far.

The channel-aware scheduler is one of the components designed in the new cross-layer adaptive architecture presented in this thesis to support the high variability of RF channel in IEEE 802.16 networks. Moreover, this new architecture is hierarchical and modular, *i.e.* it is able to support any Intra and Inter-class algorithm to provide guaranteed QoS.

Currently, the proposals on channel-aware schedulers are focused on developing algorithms to optimize the provision of QoS guarantees in IEEE 802.16 networks. However, the scheduler component alone cannot guarantee the QoS requirements for different classes of service because the user's data and BRs must be allocated within the OFDMA frame. Therefore, the guarantees of QoS also are influenced by the allocator component which is discussed in the following subsection.

Table 3.2: Comparasion on Scheduler Proposals

Scheduler proposals	Channel aware	Intra class	Inter class	QoS guarantees
Sayenko *et al.*, 2008		✓		✓
Lera, Molinaro, and Pizzi, 2007		✓		✓
Niyato and Hossain, 2005			✓	✓
Chen, Jiao, and Wang, 2005			✓	
Ruangchaijatupon and Ji, 2008	✓			
Park, Cho, and Bahk 2006	✓			✓

3.1.3 Allocator

The main goal of the allocator component is to select data received from the scheduler component and arrange it within the OFDMA frame. The problem of allocation has been studied in the recent past with the definition of OFDMA multiplexing as a mandatory technique in IEEE 802.16e. The proposals are focused only in the DL subframe allocation where the IEEE 802.16 standard left open the method for allocation of the burst profiles, differently from the allocation method in UL subframe, as described in Subsection 2.3.3.

The earliest work in this area was proposed by Ben-Shimol, Kitroser and Dinitz (BEN-SHIMOL; KITROSER; DINITZ, 2006), where the authors formulated the problem as a two dimensional bin packing to seek the best compromise between the overhead and allocated data. The data within the DL subframe must be allocated using bursts with rectangular shape, and consisting of a set of indivisible sub-bursts, called bin. More recently Cicconetti *et al.* (CICCONETTI et al., 2010) classified the allocator components in two categories: (i) sequential and (ii) non-sequential that are described as follows.

The sequential solutions are those where the data bursts are allocated in the same order as scheduled. In others words, the data bursts received from the scheduler component are allocated from the left to the right, starting in the top of the frame considering the arrival order defined by the scheduling algorithm. Figure 3.3 (a) shows an example of sequential solution. Israeli, Rawitz and Sharon(ISRAELI; RAWITZ; SHARON, 2008) introduced the sequential solution so that the scheduler component can decide which PDUs are most important to allocate in the current DL subframe. Bacioccola *et al.* (BACIOCCOLA et al., 2007) also exploited the sequential solution by proposing the Simple Data Region Allocation (SDRA) and its improved version was proposed by Ohseki, Morita and Inoue (OHSEKI; MORITA; INOUE, 2007), where the data regions are created in a "raster" manner, *i.e.* by filling the DL subframe column by column from the first row and moving downward.

In the non-sequential solutions, the allocator is free to choose the data bursts that best fit in the DL subframe. In Figure 3.3 (b), a simple example of non-sequential solution can be observed. Currently, three main non-sequential proposals have been presented by the scientific community: (i) Micro Scheduling Problem (MiSP), (ii) One Column Striping with non-increasing Area first mapping (OCSA), and (iii) Recursive Tiles and Stripes

(RTS). These proposals are described as follows.

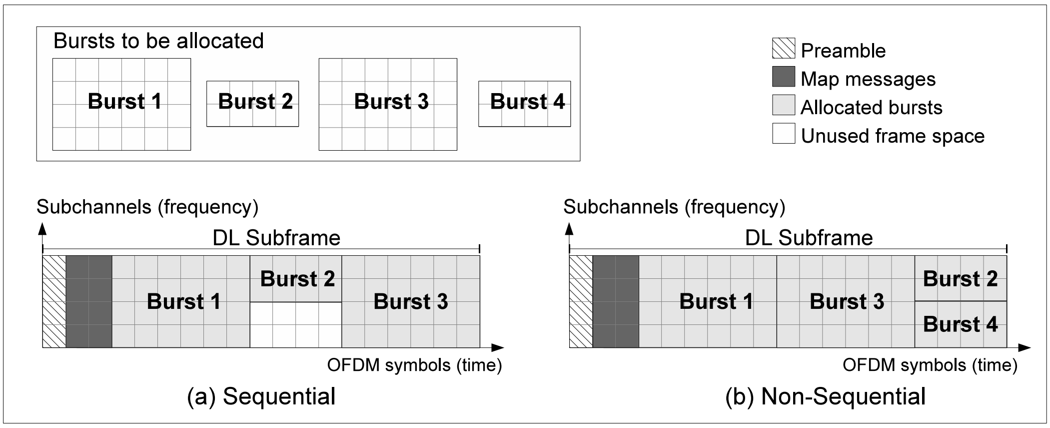


Figure 3.3: Classifications of the Allocator Component

MiSP algorithm (COHEN; KATZIR, 2008) is based on information such as QoS requirements of each connection, RF conditions, and fairness, among others. This information is used by the allocator, that represents each burst profile as a sequence of one or more contiguous rectangles to build the OFDMA DL subframe. This representation is an approximation that returns a suboptimal solution that meets all the constraints of the OFDMA allocation. In order to find a better solution, an "Efficient Dual Approximation for MiSP" was proposed (COHEN; KATZIR, 2010). This approach uses the next-fit bin-packing algorithm. The main idea behind this algorithm is to divide the burst profiles into sets in which all burst profiles are roughly the same size and to place each set in separate columns of the OFDMA DL subframe. The dual approximation returns a solution whose amount of information is at least as high as that of the optimal solution while relaxing some of the constraints.

OCSA algorithm (SO-IN; JAIN; TAMIMI, 2009c) is a two dimensional rectangular burst mapping algorithm that is composed of three steps. First, the set of data bursts to be allocated is sorted in descending order with respect to its size. Second, the largest element that can fit in the unused space of the OFDMA DL subframe is allocated in a horizontal manner. Given the area of a burst, the algorithm determines the possible width-height pair for this burst. The only constraint is that the width of the element must be less than the width of the remaining OFDMA DL subframe. Considering several feasible pairs, it is selected the pair with the smallest width. In the third step, called vertical allocation, if there is space available, the algorithm allocates the largest data burst that fits in the left-over area of the OFDMA DL subframe. However, smaller height and larger width are preferred for the utilization of the space within this current width. This process of filling the vertical strip is repeated until no further allocation will fit vertically in the unused space with the current DL subframe width.

RTS algorithm (CICCONETTI et al., 2010) is composed of two phases. The first one is the iterative selection of a subset of slots (S) used by m PDUs (N_{pdu}) not exceeding L_{DL} slots. In order to prioritize the allocation of PDUs that demand high QoS guarantees, this selection considers an information assigned to each PDU. The second phase is an inner packing algorithm that tries to allocate S (or the largest majority of it) to the bin, returning the best feasible solution found. This is obtained through iterated executions of two basic algorithms, denoted as Tiles and Stripes. Both algorithms determine an allocation of (or part of) N_{pdu} of S to a data region composed of: a rectangular data bin of fixed height (H) and width (Ŵ), and a data column of height (Ĥ), where Ŵ and Ĥ are tentative values for the available bin width and column height, respectively. In the packing algorithm, the authors assume that, for a given packing configuration π of N_{pdu}, a function $\phi(\pi)$ creates the maps and returns the values Ŵ and Ĥ that define the actual data container available for packing π. The packing algorithm checks "a posteriori" if the data region allocation produced by Tiles or Stripes is feasible, and, in such a case, updates the incumbent solutions.

Comparative Analysis of Allocation Solutions

This topic presents a comparative analysis of allocation solutions considering the OFDMA frame. Table 3.3 shows research works about OFDMA allocation in IEEE 802.16 networks associated with sequential approach and static maps overhead. Therefore, it is possible to observe that only two investigations considered static overhead of maps in allocation algorithm, an investigation using sequential approach and another using non-sequential approach. However, the dynamic overhead of maps cannot be disregarded in a QoS architecture for IEEE 802.16 networks. In this context, the new cross-layer adaptive architecture presented in this thesis consider the dynamic overhead of maps in the allocator component. Moreover, this new architecture is modular, *i.e.* it is able to support any allocation algorithm.

Table 3.3: Comparasion on Allocator Proposals

Allocator proposals	**Sequential approach**	**Static overhead**
Shimol, Kitroser, and Dinitz, 2006	✓	✓
Israeli, Rawitz, and Sharon, 2008	✓	
Bacioccola *et al.*, 2007	✓	
Ohseki, Morita, and Onoue, 2007	✓	
Coehn and Katzir, 2010		
So-In, Jain, and Tamimi, 2009		
Cicconetti, 2010		✓

According to the related work previously discussed, it can be observed that many proposals have been presented to design or to optimize specific features of each QoS component. However, the cross-layer integration among the QoS components is fundamental to the performance of the overall QoS architecture in IEEE 802.16 networks. In this context, the cross-layer integration proposals of the components in a QoS architecture are analyzed in the following section.

3.2 Cross-layer Architectures

This section presents the related work about QoS architectures in IEEE 802.16 networks that consider the RF conditions and the OFDMA multiplexing features to design the QoS components. These architectures based on the PHY Layer and MAC Layer are called cross-layer architectures, where the goal is to optimize the communication between the layers to improve the system performance (MSADAA; C?MARA; FILALI, 2010). In this thesis, the related work can be classified in three perspectives according to (a) the usage of the bandwidth, (b) multiplexing technique, and (c) the frame structure, as can be seen in Figure 3.4. Indeed, the classification of the related work on cross-layer architecture proposals published in the IEEE 802.16 literature can be considered the first specific contribution of this thesis. In this context, the three perspectives are discussed in the following.

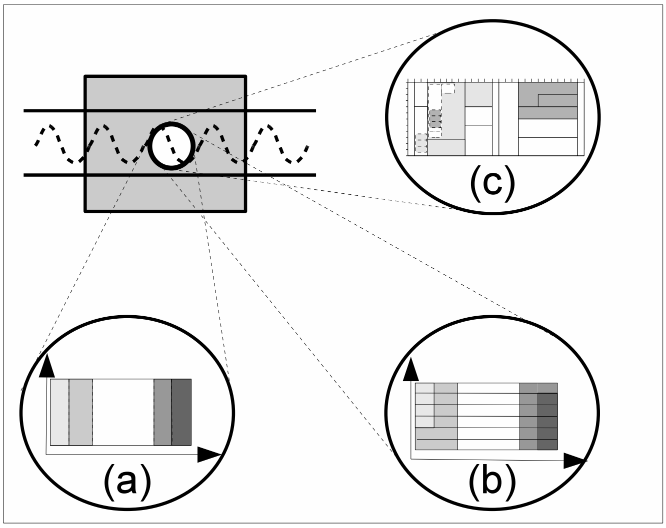


Figure 3.4: Perspective of Cross-layer Architectures

Bandwidth aware is used in proposals that estimate the RF capacity as can be seen in Figure 3.4 (a). Usually, the related work on bandwidth aware considers the MCS, Signal to Interference plus Noise Ratio (SINR), power, and Bit Error Rate (BER) to calculate the RF capacity and consequently to design the QoS components. Lu and Ma (LU; MA, 2010) proposed a RF condition estimator to analyze the impact of the air interface on MAC layer. The cross-layer architecture proposal has two contributions: an opportunistic scheduling algorithm considering real-time physical channel conditions, network traffic and queue conditions, and a CAC algorithm to support QoS in IEEE 802.16 PMP networks. When the system resources become scarce, elastic CAC can automatically and dynamically adjust the bandwidth consumed by the existing connections. Therefore, the QoS requirements of ingoing connections would be admitted as possible. Another work on bandwidth aware was proposed by Triantafyllopoulou *et al.* (TRIANTAFYLLOPOULOU et al., 2007). The cross-layer optimization mechanism takes advantage of the adaptation capabilities of IEEE 802.16 layers. The authors combined the MCS capability of the PHY Layer and the multi-rate feature of the multimedia applications through a cross-layer optimizer at BS and MSs. The optimization process consists in collecting an abstraction of the layer-specific information, such as QoS parameters and RF conditions, and informing the corresponding layers about the required changes.

Multiplexing resource aware is a technique investigated in related work on channel division based systems where the multiplexing resource is the slot composed of OFDM symbols and subchannels as can be seen in Figure 3.4 (b). Yahiya, Beylot and Pujolle (YAHIYA; BEYLOT; PUJOLLE, 2008) proposed a cross-layer scheme with guaranteed QoS for multiuser OFDMA based mobile IEEE 802.16 networks. The scheme defines distinct scheduling priority for each connection on each subchannel that integrates MAC QoS requirements, service flows types and PHY Layer using Channel Quality Indication (CQI). Based on the scheduling mechanism combined with subchannel allocation scheme, efficient QoS guarantees were achieved in terms of maximum delay requirement for real-time service flows and MRTR for non real-time service flows sessions. In another research on multiplexing resource aware, Xie, Zhou and Zeng (XIEI; ZHOU; ZENGI, 2008) defined a cross-layer QoS design which consists of CAC and subchannel assignment, taking PHY Layer channel information into account. The CAC checks whether a BS can satisfy the service flow's QoS requirements and decides to accept it or not. Then the scheduler decides which slots will be used to transmit the accepted service flows. All channel information is sent back to the subchannel allocation and MCS algorithm through feedback channels from all MSs. The resource allocation scheme made by the algorithm is forwarded to the OFDMA transmitter. The transmitter then selects different numbers of slots from different users to form an OFDMA symbol.

Frame structure aware is used in proposals that consider the OFDMA frame to design the QoS components of IEEE 802.16 networks, as can be seen in Figure 3.4 (c). The first work on frame structure aware using cross-layer architecture was published by Cohen and Katzir (COHEN; KATZIR, 2008) and updated in 2009 (COHEN; KATZIR, 2010). The authors addressed the OFDMA scheduling problem assuming that the BS determines in advance the PHY-profile to be used for each PDU and the gain obtained by transmitting this PDU using its selected PHY-profile. Then, there are two phases. In the first phase, referred to as Macro scheduling, the BS determines which of the PDUs will be selected for transmission. In the second phase, referred to as Micro scheduling, the scheduler determines how to build the OFDMA frame from the selected PDUs. The Micro scheduling in this thesis is considered to be the allocator component due to the characteristics of this second phase.

3.2.1 Comparative Analysis of Cross-layer Architectures

The classification of the related work on cross-layer architectures shows that there are several strategies to calculate and to estimate the available resources in IEEE 802.16 networks. Table 3.4 presents a comparative analysis of some research works relationship with the perspective for estimation of available resources and QoS components used by cross-layer architectures. To the best of our knowledge, only one work considering the OFDMA frame in the design of a cross-layer architecture has been proposed so far. However, in this work, CAC is not considered to provide QoS.

Table 3.4: Comparison on Cross-layer Architecture Proposals

Cross-layer proposals	**Band. aware**	**Multiplex. aware**	**Frame aware**	**QoS Components**		
				CAC	**Sched.**	**Alloc.**
Lu and Ma, 2010	✓			✓	✓	
Triantafyllopoulou, 2007	✓			✓		
Yahiya *et al.*, 2010		✓			✓	✓*
Xiei, Zhou, and Zengi, 2008		✓		✓	✓	✓*
Coehn and Katzir, 2010			✓		✓	✓

* Only subchannel allocation scheme

Table 3.4 shows that some important features of the IEEE 802.16 technology are ignored in the design of cross-layer architectures, such as maps overhead in the OFDMA frame, overflows in the data flow among QoS components, and overall QoS guarantees, considering all five classes of service defined by the standard. Usually, these features are neglected in the integration of QoS components due to the complexity to address the challenges in this integration. Therefore, the new cross-layer adaptive architecture proposed in this thesis considers the OFDMA frame, as well as the three components to guarantee QoS in IEEE 802.16 networks. In the following section some of these challenges are listed, in order to design a cross-layer architecture.

3.3 Design Challenges for IEEE 802.16 QoS Architecture

In the literature, proposals on each one of the QoS components and also designs of cross-layers architecture for IEEE 802.16 networks can be found. According to the research works, the main challenging problems that arise when trying to develop a cross-layer architecture are:

- To determine the trade-off between an efficient solution and a simple one. The efficient solution would take into account the QoS requirements of the different applications, while simples solution would be implementation-friendly and less time consuming;

- To determine the optimal relationship between fairness and RF utilization. Indeed, giving priority to users with better RF conditions, *e.g.* with lower MCS configuration, would increase the overall RF utilization. Nevertheless, it would be unfair to other users with worse RF conditions, *e.g.* with higher MCS configuration;

- To choose between an optimized solution and a more generic one. The optimized solution targets a specific kind of application, such as real-time, and takes into account its specific needs. On the other hand, the general solution addresses yet efficient and less complex approaches that would consider heterogeneous types of traffic;

- To calculate the performance of the algorithms of each QoS component in an integrated manner in the cross-layer architecture;

- To take advantage of an optimal MCS capability when proposing a new cross-layer adaptive solution;

- To specify an adaptive DL and UL allocation algorithm in the OFDMA frame, in order to make an efficient use of the available resources.

The challenges presented in the literature show the complexity of the proposals related to cross-layer adaptive architectures. Currently, there are open research topics associated with the integration of the three main QoS components based on OFDMA frame. In this context, Charter 4 presents a new cross-layer adaptive architecture for the IEEE 802.16.

3.4 Summary

This chapter presented an overview of the state-of-the-art regarding three main QoS components. Moreover, cross-layer architectures for IEEE 802.16 networks were studied.

First, the discussion about the degradation policies of the services that can be adopted in the CAC component was presented. Second, a review of the definitions associated with the scheduler component was conducted, where the related work about channel-unaware scheduling and channel-aware scheduling were discussed. Third, an analysis of sequential and non-sequential categories of the allocator component was performed. In the sequence, the integration of QoS components using a cross-layer architecture was presented. Analyzing the discussions it was possible to verify that some important features of the IEEE 802.16 technology are usually neglected in the integration of QoS components, leaving many opened questions in this research area. Therefore, this chapter is closed with the discussion about the challenges related to the design of a new cross-layer adaptive architecture.

4 A NEW CROSS-LAYER ADAPTIVE ARCHITECTURE FOR IEEE 802.16 NETWORKS

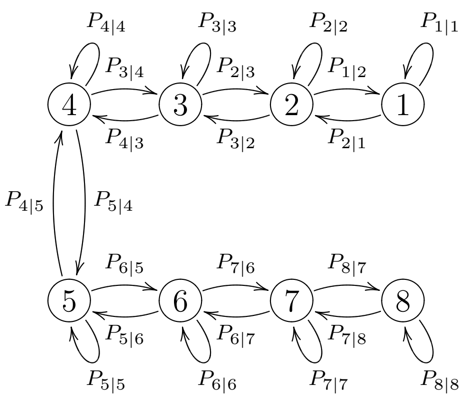
The goal of this chapter is to present a new cross-layer adaptive architecture for IEEE 802.16 networks. The chapter starts presenting the definition of the RF channel model used to inform changes on RF channel conditions to the new cross-layer adaptive architecture. Next, the design of the channel-aware architecture is described, considering an information structure to propose QoS components which are aware of changes on the RF channel and traffic requirements. In addition, a module that integrates the scheduler and allocator components is presented, which takes into account the amount of PDUs and BRs that can be sent from the scheduler to the allocator. Finally, the design of the CAC component is proposed to compose the new cross-layer adaptive architecture.

4.1 Modeling the RF Channel

The design of an efficient cross-layer adaptive architecture for IEEE 802.16 networks requires the QoS components to become aware of the RF channel conditions. The channel-aware requirement is mandated because a bad channel quality may invalidate theoretical fairness and QoS support capabilities of the cross-layer architecture (NI et al., 2010). In this thesis, the RF channel modeling is based on a time-discrete Markov chain composed of eight states, each one representing a different MCS. A graphical representation of the proposed Markov chain and the MCS corresponding to each state of the chain is presented in Figure 4.1.

Each state of the Markov chain has associated transition probabilities. In the design of the model, the probabilities of transition among the states of the Markov chain are parameters that might be set according to the intended scenario. An example of a real scenario is presented by Lakkakorpi, Sayenko and Moilanem (LAKKAKORPI; SAYENKO; MOILANEN, 2008), where a fixed WiMAX network with 55 connections is simulated in the Nokia Research Center. The Markov chain begins with the selection of an initial channel state. Another important parameter that must be informed is the time between evaluations of the channel condition, which is defined in unit of number of frames.

During the evaluation of the RF channel condition, there are three possibilities in terms of transitions: (i) permanence in the same state, when the channel remains in similar propagation conditions, (ii) increase one level of robustness, when the current channel condition is worst than in last evaluation, or (iii) decrease one level of robustness, if the channel condition is better than in the last evaluation. For example, if the current MCS is 16 QAM 3/4, *i.e.* 4 state, and a worst channel condition is detected, the next transmission will apply 16 QAM 1/2, *i.e.* 3 state. Otherwise, considering that the RF channel condition is better than in last evaluation, 64 QAM 1/2, *i.e.* 5 state, should be used. There are two exceptions, which occur when the more robust or the less robust MCS is used. In the former, it is not possible to increase the robustness, while in the latter, it is impossible to decrease the robustness, thus no change on MCS is applied. The transitions among states of the Markov chain used in the performance evaluation of this thesis are presented in Chapter 5.



State	MCS
1	QPSK 1/2
2	QPSK 3/4
3	16 QAM 1/2
4	16 QAM 3/4
5	64 QAM 1/2
6	64 QAM 2/3
7	64 QAM 3/4
8	64 QAM 5/6

Figure 4.1: RF Channel model

The MCS applied to a transmission is used to calculate the throughput of the i^{th} PDU or BR in a given instant of time t. This information may be used by the scheduler component to define the order in which PDUs and BRs will be served by the BS. The throughput of a PDU or BR (τ_i) is calculated according to equation (4.1). The value τ_i is obtained considering the quantity of bits per modulation symbol (M_i), the FEC encoding (K_i), and the symbol duration (T_s).

$$\tau_i = \frac{M_i.K_i}{T_s} \quad [bit/s] \tag{4.1}$$

Considering information from RF channel state, the scheduler component can order the PDUs and BRs within the queues using the QoS requirements of each class of service. The combination of channel state information and QoS requirements compose the channel-aware scheduling for the new cross-layer adaptive architecture of IEEE 802.16 networks. In the next section the proposal of the new channel-aware architecture is presented.

4.2 New Channel-Aware Architecture

The first aim of this thesis is to design a new cross-layer adaptive architecture that is aware of the changes on the RF channel conditions and traffic requirements in IEEE 802.16 networks. In order to reach this aim, the adaptive architecture demands information about the RF channel conditions and of the traffic. In this context, the information used by the QoS components associated with RF channel conditions and traffic transmitted among the BS and MSs is presented in Subsection 4.2.1. Moreover, the design of the channel-aware scheduler component for the BS, which uses this information is discussed in Subsection 4.2.2.

4.2.1 Definition of the Joint Information

The Joint Information is a structure designed to be used in the new cross-layer adaptive architecture that is proposed in this thesis. The Joint Information carries several information regarding the current channel conditions of the IEEE 802.16 network. Moreover, the structure has fields to associate each PDU and BR with its connection, class of service and creation/arrival time. The fields that compose the refereed data structure, as well as units, sizes, and descriptions are presented in Table 4.1.

Table 4.1: Joint Information

Field	Unit	Size	Description
CID	Integer	16	The connection identifier of a PDU or a BR
Data length	Bits	16	Indicates the quantity of bits to be transmitted
Deadline	Millisecond	8	Contains the deadline of realtime PDUs or BRs
MCS	Integer	3	Indicates the Modulation and Coding Scheme to be applied considering the RF conditions
Current throughput	Bits/s	18	Calculated according to Equation 4.1
Class of Service	Integer	3	Class of service assigned to the PDU or BR
Timestamp	Millisecond	32	Timestamp of the creation/arrival of the PDU or BR in BS

The Joint Information is assigned to each PDU or BR on its creation or arrival at the BS. The information contained in the structure is used by QoS components to define the set of PDUs or BRs that will be candidate for allocation during the next available OFDMA frame. Indeed, the scheduler uses the Joint Information to define which PDUs and BRs will be submitted to the allocator. After, a set of fields of the Joint Information is also used by the allocator to define how the PDUs and BRs will be arranged within the next DL subframe. Moreover, the CAC component considers some fields of the Joint Information to decide whether new connections will be admitted or not, as described in Subsection 4.4.

An example of application of the Joint Information is the use in a Linux-based WiMAX BS. In this case, the Joint Information can be associated with the *sk_buff* structure of the Linux Kernel that constitutes the PDU headers or BR headers used by QoS components. The size of each field of the Joint Information is presented, in units of bits, in Table 4.1. The structure of the Joint Information is 12 bytes long, however the *sk_buff* structure already has a timestamp field. Therefore, the Joint Information size really should insert 8 bytes in each PDU or BR. Moreover, it is important to describe that the Joint Information is not transmitted in RF channel, this structure is only used to internal control of QoS components.

One field or a set of fields of the Joint Information can be used to calculate the profit of each PDU or BR. The profit, in the context of network traffic scheduling, is a concept that was introduced by Cohen and Katzir (COHEN; KATZIR, 2008) aiming to associate a gain to the scheduling and allocation of PDUs and BRs in DL transmissions. The authors describe that the assignment of a profit should be based on several information, such as QoS requirements, throughput maximization, fairness, and RF propagation conditions, among others. However, how this information should be used is out of the scope of Cohen and Katzir's work. Then, this thesis introduces the Joint Information to design QoS components aware of the changes on the RF channel in IEEE 802.16 networks. The scheduler component project is presented in the following subsection.

4.2.2 Channel-aware Scheduler Component Design

The design of a new cross-layer adaptive architecture demands channel-aware components. Therefore, in this thesis, the scheduler is the main component used to turn the new architecture aware of RF channel conditions. The scheduler component is projected to assign the Joint Information structure of each PDU and BR to queues, corresponding to each class of service. This process is performed in the Intra-class scheduling, as presented in Subsection 3.1.2. However, there are three fundamental differences on the approach of this thesis, if compared to the related work: (i) the Joint Information structure is well-defined and is applied in both PDU and BR queues, (ii) the channel-aware scheduling component is planned in a hierarchical manner, unlike the classification defined by the So-In, Jain and Tamimi (SO-IN; JAIN; TAMIMI, 2009b), and (iii) each class of service considers different fields of the Joint Information in the arrangement of PDUs and BRs within the queues, due to specific QoS requirements of each class of service.

The priority of each PDU and BR within a queue is calculated using a Weighted Proportional Fair (WPF) algorithm (HOU et al., 2008), that is aware of the RF channel propagation conditions because uses the Joint Information in its logic structure. The choice of WPF is due to this algorithm can improve resource allocation, fairness, and QoS guarantees assigning a weight for each PDU and BR in queues of the Intra-class scheduling. The weight of each PDU and BR used in the WPF algorithm is defined according to a scheduling factor that, in this thesis, is designed for the class of service queues. These scheduling factors are described in the next topic.

Scheduling Factors

The scheduling factors, considering the fields of the Joint Information, are used by WPF algorithm to order the PDUs and BRs inside the queues in Intra-class scheduling, as can be seen in Figure 4.2. To simplify, the figure represents only a group of queues, *i.e.* PDU queues or BR queues. The scheduling factors are used in the five classes of service defined by the IEEE 802.16 standard. However, due to similar features of some classes of service, the scheduling factors are defined in three groups, as follows.

UGS and ertPS classes of service have similar real-time requirements for its applications, *e.g.* VoIP transmissions without silence suppression for UGS services and VoIP transmissions with silence suppression for ertPS services. Thus the same scheduling factor can be used for both classes of service. The φ_{UGS_ertPS} scheduling factor is obtained using equation (4.2), which is based on the information used by Dua and Bambos (DUA; BAMBOS, 2007). In this equation, two fields of the Joint Information that are relevant to VoIP transmissions are combined: the *Deadline* (D_i), and the *Current Throughput* (τ_i) of the i^{th} PDU or BR. The value of the deadline is normalized such as $D_i^{'} = D_i/D_{max}$, where D_{max} is the maximum value for the deadline of UGS or ertPS transmissions, *i.e.* $D_i^{`} \leq D_{max}$. Moreover, the current throughput is normalized with respect to its maximum value ($\tau_{max}$), such as $\tau_i^{`} = \tau_i / \tau_{max}$, where $\tau_i \leq \tau_{max}$. Each one of these normalized values receive a weight, α for the deadline, and $(1 - \alpha)$ for the throughput. Where, α is a constant smoothing factor such as $0 \leq \alpha \leq 1$. In other words, the smoothing factor is a simple weighted average that accounts the current observation ($D_i^{`}$) and the current throughput ($\tau_i^{`}$).

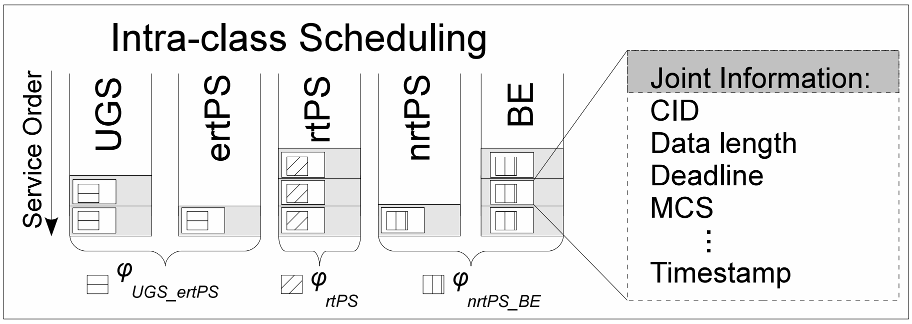


Figure 4.2: Intra-class Scheduling Queues

$$\varphi_{UGS\_ertPS} = \alpha.D_i^{`} + (1 - \alpha).\tau_i^{`} \tag{4.2}$$

The φ_{UGS_ertPS} of each PDU and BR in UGS and ertPS queues is used as weight in the WPF algorithm. Therefore, this algorithm is applied to prioritize the PDUs and BRs considering the smallest scheduling factor of each queue, by computing $argmin_i\{\varphi_{UGS\_ertPS}\}$.

rtPS class of service has specific QoS requirements for its applications, *e.g.* video transmission, thus the scheduling factor is individually defined. In this case, it considers the *Current Throughput* field of the Joint Information for each PDU and BR (τ_i). Moreover, the backlogged information about the throughput ($\overline{\tau}$) can be considered in the scheduling factor of the rtPS service. These information types are typically found in the literature about channel-aware real-time transmissions, as presented by Elsayed and Khattab (ELSAYED; KHATTAB, 2006). Therefore, the historical information about the throughput of the PDU and BR should be calculated by the Intra-class scheduling algorithm. The exponential average of the throughput ($\overline{\tau}_i$) during the transmission in a given instant of time t is calculated in equation (4.3). Where, ρ is also a constant smoothing factor such as $0 \leq \rho \leq 1$.

$$\overline{\tau}_{i(t)} = \rho.\tau_{i(t)} + (1 - \rho).\overline{\tau}_{i(t-1)} \quad [bit/s] \tag{4.3}$$

After calculating $\overline{\tau}$, the scheduling factor for rtPS class (φ_{rtPS}) can be defined using equation (4.4). In the sequence WPF algorithm is applied to prioritize the PDUs and BRs considering the highest scheduling factor this queue, such as $argmax_i\{\varphi_{rtPS}\}$.

$$\varphi_{rtPS} = \frac{\tau_i}{\overline{\tau}} \tag{4.4}$$

nrtPS and BE classes of service have minimal or no QoS requirements guaranteed, therefore these two classes of service also use the same scheduling factor. In nrtPS and BE, the *Current Throughput* of the MS and the *Deadline* of the PDU or BR are not considered. However, the channel state (ϱ_i), *i.e.* the MCS field of the Joint Information such as $1 \leq \varrho \leq 8$, is considered during the scheduling process, because MSs with better propagation conditions have a higher probability of successful transmissions. Therefore, using the scheduling factor presented in equation (4.5) the PDU or BR with the best channel condition is selected by the WPF algorithm in the Intra-class level such as $argmax_i\{\varphi_{nrtPS\_BE}\}$. This approach provides a more efficient use of the network capacity. On the other hand, PDUs and BRs with a bad channel condition should also be served to avoid starvation problem. This situation can be controlled by the CAC component which is described in Section 4.4. The CAC component prevents the occurrence of saturation in the system, since this component ensures the efficient behavior of the scheduling algorithm.

$$\varphi_{nrtPS\_BE} = \varrho_i \tag{4.5}$$

After the Intra-class scheduling, the PDUs or BRs are served according to a prioritization algorithm applied for each class of service in the Inter-class scheduling, *e.g.* UGS, ertPS, rtPS, nrtPS, and BE. In addition, an allocation algorithm arranges the PDUs and BRs within the OFDMA frame.

The design of the new cross-layer adaptive architecture must be modular and flexible to support any Inter-class scheduling and allocation algorithms to provide different QoS requirements according to the network profile. Therefore, the scheduler and allocator components should work in an integrated way to achieve the required QoS guarantees. Otherwise, QoS degradation can take place and affect the overall network performance. In this context, the next section discusses the integration between the scheduler and allocator components in OFDMA-based IEEE 802.16 networks.

4.3 A Module to Integrate Scheduler and Allocator

The scheduler component is responsible for selecting a set of PDUs ($Npdu$) and BRs (Nbr) to be sent to the allocator component. In this thesis, this selecting process has three important features, as previously described in Section 4.2.2: (i) the Intra-class scheduling is channel-aware because it exploits the Joint Information fields, (ii) the PDUs and BRs of each class of service are tagged with a scheduling factor, (iii) the WPF algorithm is applied to select the PDUs and BRs with highest priority according the QoS guarantees of each class of service. Therefore, an integration module needs to be proposed to define a threshold criteria to limit the amount of PDUs and BRs sent from scheduler to allocator, according to the capacity of the OFDMA frame. Sending the correct amount of PDUs and BRs to the allocator is fundamental to exploit the maximum efficiency of the allocation algorithms and to avoid overflow in the allocator component.

The integration module can be observed in Figure 4.3 that shows a view of the new cross-layer adaptive architecture considering the scheduler and allocator components. Two functions compose the integration module: (i) mapping overhead and (ii) threshold criteria. Indeed, before designing the threshold criteria, it is necessary to calculate the mapping overhead of the OFDMA frame. The mapping overhead size cannot be ignored when calculating the amount of the PDUs and BRs to be sent to the allocator. However, some works, as Cohen and Katzir (COHEN; KATZIR, 2008) and So-in *et al.* (SO-IN; JAIN; TAMIMI, 2009c), do not consider this overhead during the allocation process. The equations used to define the mapping overhead, according to the specification of the IEEE 802.16 standard presented in Subsection 2.3.2, are described in the next section.

4.3.1 Mapping Overhead

The size of the DL_{MAP} is proportional to the number of connections currently active in the network and to the MCS, *i.e.* the amount of burst profiles carried in the OFDMA frame (JUYEOP; DONG-HO, 2009). The structure of the DL_{MAP}, presented in Figure 4.4 (a), is composed of a header with fixed length, called DL_{head} and several IE headers, named DL_{IE}. IE is a structure used to inform the number of OFDMA symbols and subchannels allocated for data transmission, as described in Section 2.3.2. A DL_{IE} is composed of multiple CID fields, belonging to connections with an associated MCS level. All the connections belonging to a given IE are part of the same burst profile. In this thesis, the information regarding the CID is called DL_{CID}.

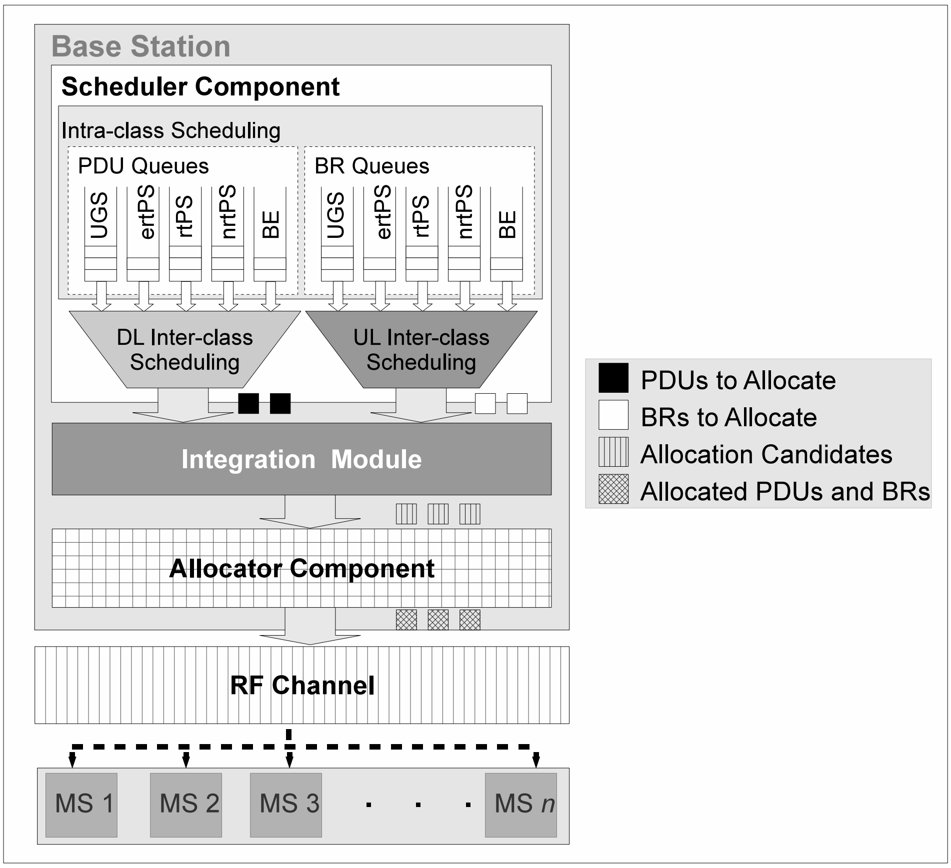


Figure 4.3: A Module to Integrate Scheduler and Allocator

Considering the structure of the DL_{MAP} presented in Figure 4.4 (a), it is possible to define the equation (4.6). This equation permits to obtain the size of DL_{MAP} for a given frame, taking into account that a DL_{IE} is composed of a variable number of DL_{CID}.

$$DL_{MAP} = DL_{head} + \sum_{j=1}^{N} \left(DL_{IE(j)} + \sum_{k=1}^{Cj} DL_{CID(j,k)} \right) \quad [bits] \qquad (4.6)$$

Where, $j \in \{1, 2, ...\ \text{N}\}$ is the index of MCS level, and N is the number of MCS levels, $k \in \{1, 2, ...\ \text{C}\}$ is the number of DL connections scheduled in a DL_{IE}, and C is the number of CIDs. A DL burst is generally formed of a set of $Npdu$ and Nbr, with the same MCS level. Therefore, the number of DL_{IE} in a DL_{MAP} is the number of different MCS levels used for transmitting the scheduled $Npdu$ and Nbr.

On the other hand, the size of the UL_{MAP} depends only on the number of CIDs (JUYEOP; DONG-HO, 2009). Each CID is associated with a burst and its specific MCS, as can be seen in Figure 4.4 (b). This association, called UL_{UIUC} is made through a UL_{IE}, that carries information about resources allocated for uplink transmission. Equation (4.7) defines the manner of calculating the size of the UL_{MAP}.

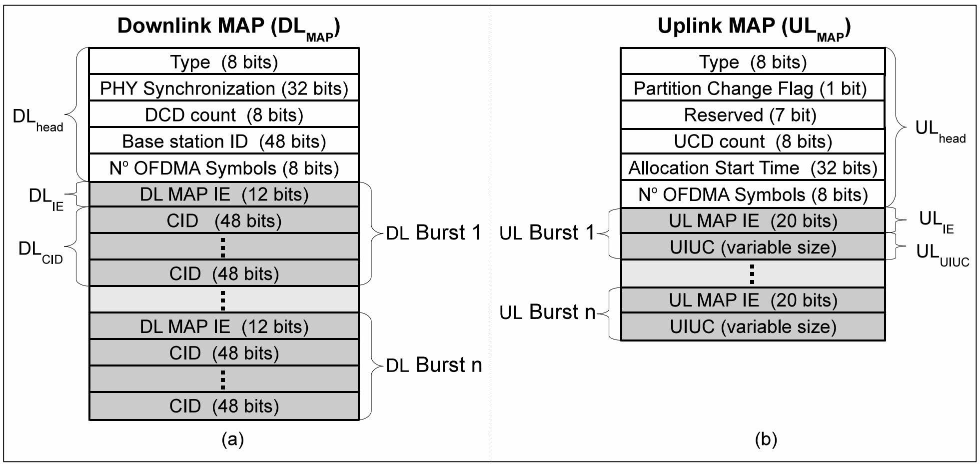


Figure 4.4: Structure of Downlink and Uplink Maps

$$UL_{MAP} = UL_{head} + \sum_{p=1}^{R} \left(UL_{IE(p)} + UL_{UIUC(p)} \right) \quad [bits] \tag{4.7}$$

Where, $p \in \{1, 2, ...$ R$\}$ is the index of resource allocation UL, and R is the number of resource allocated with an Uplink Interval Usage Code (UIUC) association. The UIUC should be used to define the type of uplink access, *i.e.* CDMA bandwidth request, CDMA ranging, CDMA allocation IE, different burst profiles, among others. Therefore, each type of UL_{UIUC} has variable size.

The mapping overhead is composed of other structures, such as DCD and UCD which size is dependent on the quantity of burst profiles transmitted. The organization of DCD and UCD is similar to the organization of DL_{MAP} and UL_{MAP}. Although, these structures are considered in the new cross-layer adaptive architecture, details are not presented to avoid repetitions. After calculating the mapping overhead, it is possible to define a threshold criteria designed along with the integration module between scheduler and allocator components, as presented in the following subsection.

4.3.2 Threshold Criteria

The main goal of the scheduler component is to prioritize PDUs and BRs to be allocated in the OFDMA frame by the allocator component. Indeed, the scheduler component does not define a set of the PDUs and BRs that must be sent to the allocator component, *i.e.* the scheduler only orders the PDUs and BRs prioritizing the QoS requirements. In this context, it is proposed a threshold criteria which is part of the integration module, placed between the scheduler and allocator components. The threshold criteria has two main functions: (i) to convert the mapping overhead, PDUs, and BRs to the allocation unit, and (ii) to select a set of the PDUs and BRs ordered by the scheduler component to send to the allocator component. The first function is important because the scheduler component usually works with bit unit, while the allocator component considers slot unit. The second function is fundamental due to the variability on bursts size, which is caused by both the dynamics of the network traffic requirements and the time variant RF channel conditions.

In order to transmit PDUs and BRs, mapping overhead and data must be inserted into the OFDMA frame. Therefore, the bits of each PDU and BR must be arranged considering the slot as allocation unit. The transmission capacity of the each slot is defined according to the number of OFDMA symbols in each subframe, which is calculated considering the total duration of a symbol and the period of the subframe, as discussed in Subsection 2.2.1. Consequently, the allocation unit of the DL subframe is different of the allocation unit of the UL subframe. Moreover, the allocation unit varies for each frame size.

In mobile IEEE 802.16 networks, the mapping overhead is formed by two kinds of information: static map overhead (Ψ) and dynamic map overhead (Φ) that must be considered in the computation of the threshold criteria. Therefore, these maps overhead are defined as follows.

Ψ is composed of Preamble, FCH, and DL and UL map headers, as presented in equation (4.8). The static map overhead has fixed length, because its value is dependent of network parameters that do not change dynamically.

$$\Psi = Preamble + FCH + DL_{head} + UL_{head} \quad [bits] \tag{4.8}$$

On the other hand, Φ varies in time. This variation happens due to the MCS level and to the number of connections scheduled in a DL subframe, as defined in equation (4.6). Moreover, the variation is affected by the amount of resources allocated in an UL subframe, as presented in equation (4.7). The DL_{head} and UL_{head} are considered in Ψ, thus these maps must be disregarded in Φ calculation. In addition, the channel descriptors of DL and UL are computed, as can be seen in equation (4.9).

$$\Phi = DL_{MAP} - DL_{head} + UL_{MAP} - UL_{head} + \sum_{j=1}^{N} \left(DCD_{(j)} + UCD_{(j)}\right) \quad [bits] \tag{4.9}$$

After calculating the static and dynamic overhead, the threshold criteria must convert the mapping overhead, PDUs, and BRs to the allocation unit, *i.e.*, slots. The number of slots (N_{slots}) associated with the amount of data that should be transmitted in a subframe is defined in equation (4.10). The N_{slots} value is obtained considering the quantity of bits per modulation symbols (M) and the FEC encoding (K) for each m^{th} and s^{th} slots of the subframe, which is able to carry mapping control information, PDUs, and BRs. Moreover, the calculation of N_{slots} must consider the amount of symbols per slots (S_{slot}) and the number of subcarries per subchannel (N_{sc}).

$$N_{slots} = \left\lceil \frac{(\Psi + \Phi).M_m.k_m + \sum_{s=1}^{N_{slot}} \left(\sum_{r=0}^{N_{pdu(s)}} PDU_{(s,r)} + \sum_{l=0}^{N_{br(s)}} BR_{(s,l)}\right).M_s.K_s}{S_{slot}.N_{sc}} \right\rceil \tag{4.10}$$

Where, m and $s \in$ {1, 2, ... N_{slot}} are the indexes of slot in the OFDMA subframe, r $\in$ {0, 1, ... $N_{pdu(s)}$} is the index of a PDU in a slot s, and $l \in$ {0, 1, ... $N_{br(s)}$} is the index of a BR also in a slot s.

After converting the amount of bits to the number of slots of the subframe, the threshold criteria must satisfy inequation (4.11), that selects a set of PDUs and BRs ordered by the scheduler component to be allocated considering the size of the subframe (L) in units of slots. The allocator component, receives the set of PDUs and BRs, and is responsible for assembling the frame to be transmitted. The PDUs and BRs that are not allocated due to restrictions imposed by the arrangement algorithm in the OFDMA frame are reinserted in the first position of scheduling queues to be served in the next frame, *i.e.* the integration module assures that the QoS requirements are regarded.

$$N_{slots} \leq L + \Lambda \tag{4.11}$$

An adjustment factor (Λ) is used to allow variations on the amount of PDUs and BRs in slots unit to be sent from the scheduler to the allocator. In other words, this factor is used to grant a bigger diversity of bursts size to the arrangement algorithm and to enable a better allocation efficiency. In Figure 4.5 an example of the Λ that shows the behavior of the threshold criteria considering the OFDMA DL subframe capacity can be seen. In Figure 4.5 (a), the threshold criteria considers Λ equal to the DL subframe capacity ($\Lambda = 0\%$). In this case, the arrangement algorithm does not use all available resource due to little diversity of bursts size. On the other hand, in Figure 4.5 (b) the available resource is better used when the arrangement algorithm receives a bigger diversity of bursts size ($\Lambda = 10\%$). The generalization of this example, *i.e.*, the complete representation of the integration between scheduling and allocation components is discussed in the next subsection.

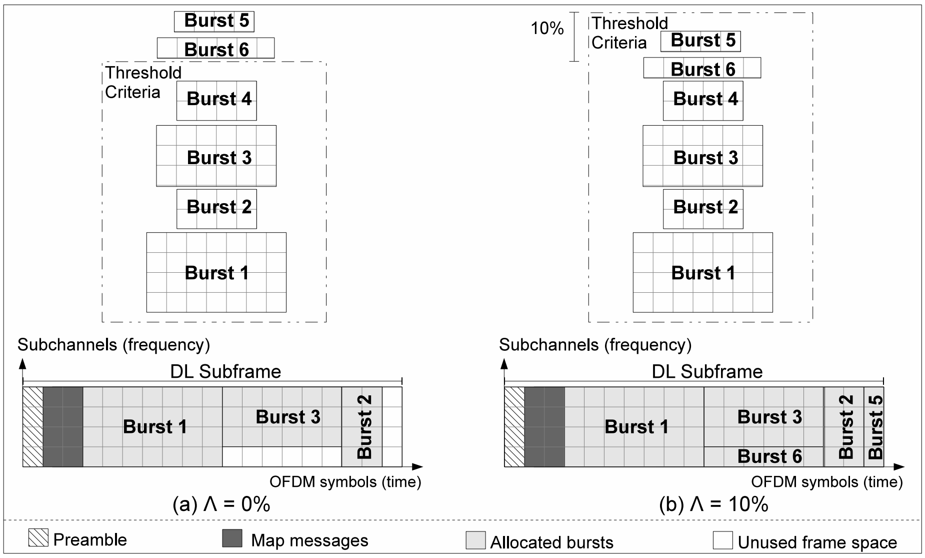


Figure 4.5: Example of Adjustment Factors

4.3.3 Modeling of the Integration between Scheduler and Allocator

The goal of this subsection is to present a model that represents the flow of PDUs and BRs in the new cross-layer adaptive architecture considering the scheduler and allocator components. The flowchart, that can be observed in Figure 4.6, shows the processing in the MAC layer of PDUs and BRs, transmitted from the BS to the MSs, in IEEE 802.16 networks, *e.g.* 5ms long OFDMA frame transmission.

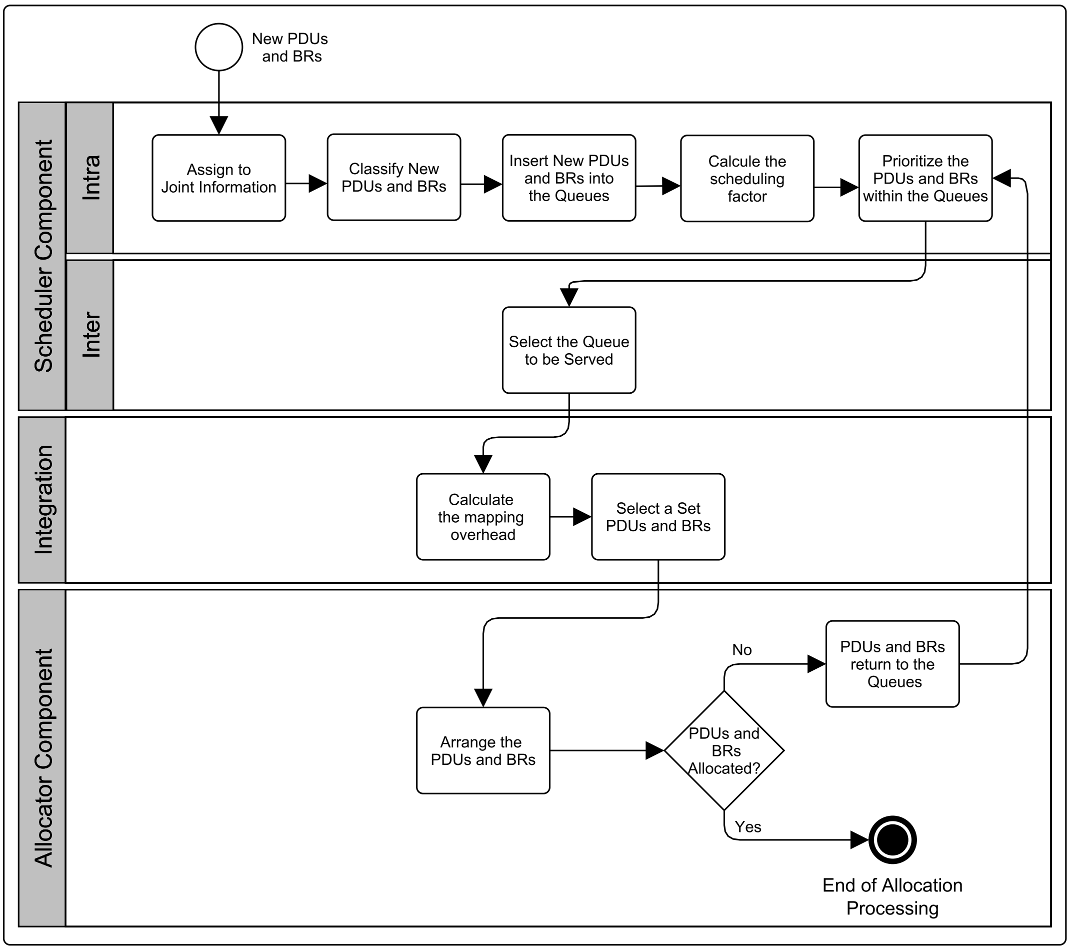


Figure 4.6: Flow of PDUs and BRs in the Cross-layer Adaptive Architecture

The processing of a frame in the MAC layer begins with the scheduler component analyzing the new PDUs and BRs which arrived in the BS. The Joint Information structure is assigned to each new PDU and BR, and then they are classified according to their classes of service to be inserted into the PDU queues and BR queues, respectively. After, the scheduling factor is calculated to each PDU and BR and an Intra-class scheduling algorithm is applied to prioritize the PDUs and BRs within the queues. To select the queue that must be served according to the service priority, an Inter-class scheduling algorithm is applied. These initial phases are performed by the scheduler component, however the next phases that calculate the mapping overhead of the OFDMA frame and select a set of PDUs and BRs to be sent to the allocator component are implemented by the integration module. Finally, the allocator component, through an allocation algorithm, decides which PDUs and BRs will be arranged within the OFDMA frame. This decision is performed aiming to use the largest amount available resource, *i.e.* to avoid unused spaces due to assembling of the OFDMA frame. The PDUs and BRs that are not allocated in the current frame are reinserted in the beginning of queues to be served in the next frame, assuring the QoS requirements of each class of service.

The scheduler and allocator are fundamental components to guarantee QoS in the new cross-layer adaptive architecture. However, these two components alone are not able to provide strict QoS requirements. For example, some connection requests may be denied to guarantee a good performance of the system when the network resources are saturated. The CAC is the component responsible for this function in connection-oriented networks, as IEEE 802.16. Therefore, the design of a CAC component able to be integrated into the new cross-layer adaptive architecture is presented in the next section.

4.4 CAC Component of the New Cross-layer Adaptive Architecture

The function of the CAC component is to decide whether to accept or not new connections. This decision is performed using a CAC algorithm that needs to receive accurate information about the usage of the network resources. The CAC component must consider the OFDMA multiplexing technique, the frame structure, the mapping overhead, the traffic requirements of a new connection, as well as the performance of the scheduler and allocator components. Therefore, the CAC component demands an infrastructure that provides information to support the decision of accepting or denying an incoming connection. In the following subsection the design of the Information Infrastructure is presented.

4.4.1 Information Infrastructure to the CAC Component

The Information Infrastructure to support the CAC component can be seen in Figure 4.7 representing the outgoing (f_{out}) and the incoming (f_{in}) OFDMA subframes in the new cross-layer adaptive architecture. The modules presented in this section are numerated according to Figure 4.7 to help in the description of the Information Infrastructure (1) to the CAC Component (2). This infrastructure must manage three information types: (i) the connections, (ii) the RF channel conditions, and (iii) the data scheduled and allocated. In order to provide accurate information to the CAC component (2), the infrastructure proposed in this thesis has a module called Raw Data (3), which is composed of three data structures that carry each information type. These data structures are presented as follows.

Traffic Description Information (3.a) is the data structure that refers to the ongoing connections in the IEEE 802.16 network. Therefore, the traffic description fields for each connection are considered in this structure, as can be observed in Table 4.2.

Table 4.2: Traffic Description Information

CID	SFID	Class of Service	MSTR (bps)	MRTR (bps)	ML (ms)	JT (ms)	Deadline (ms)	Timestamp (ms)
13	15	ertPS	124928	4096	70	30	150	50763
...	...	...	...	...	...	...	...	...

The Traffic Description Information (3.a) provides the QoS requirements of a connection, such as MSRT, MRTR, ML, JT, and Deadline. These QoS requirements must be guaranteed while a connection is active in the network. Moreover, the CID, SFID, Class of Service, and timestamp are necessary to identify a given connection.

RF Channel Information (3.b) is composed of feedback information about the RF channel conditions, as can be seen in Table 4.3. These information types can be collected frame-by-frame or in specific time intervals. The timestamp field records the time when the collect is performed.

Table 4.3: RF Channel Information

MSID	Reception Power (dB)	MCS (Burst Profile)	Timestamp (ms)
32	16.4	4 (16-QAM 3/4)	50763
...	...	...	...

The information is received from the RF Channel Information Handler (4) and is stored in the RF Channel Information (3.b) data structure during an interval. The RF channel condition refers to a transmission channel between a MS and the BS, modeled according to Section 4.1. The RF channel is identified by the Mobile Station Identifier (MSID) field. The MCS and the reception power of each RF channel are associated with the connections and its PDUs. This association between RF channel conditions and the connections is performed by timestamp fields in both Traffic Description Information (3.a) and RF Channel Information (3.b) data structures.

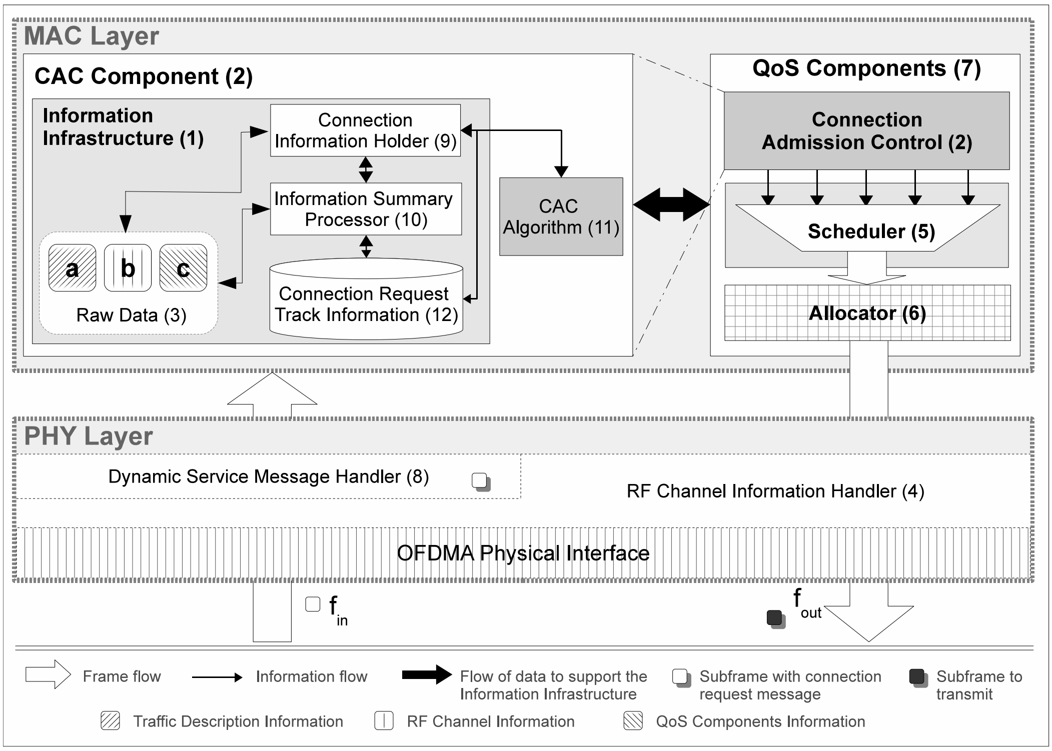


Figure 4.7: Information Infrastruture to the CAC Component

QoS Components Information (3.c) provides data about the Scheduler (5) and Allocator (6) in each frame transmitted by QoS Components (7). For example, this data structure stores the amount of bits transmitted in DL and UL subframes, as can be seen in Table 4.4. Moreover, the average delay and average jitter to serve the PDUs and BRs in each frame is also computed. The timestamp field represents the frame transmission time, and is used to its identification.

The Raw Data (3) stores accurate information in the three data structures previously described. However, these data structures should be managed to provide useful information to support the CAC component (2) decisions. Initially, the Dynamic Service Message

Handler (8) captures the register messages of connection requests in each UL subframe received by the BS. DL register messages do not need to be filtered in the OFDMA frame because they are created by the BS itself. The register messages are sent to the Connection Information Holder (9). There are three types of messages: DSA to add a connection, DSC to change ongoing connection, and DSD to delete active connections, as describe in Subsection 2.3.3.

Table 4.4: QoS Components Information

DL Data (bits)	**UL Data (bits)**	**DL Delay (ms)**	**UL Delay (ms)**	**DL Jitter (ms)**	**UL Jitter (ms)**	**Timestamp (ms)**
602660	371330	88	113	24	33	50772
...	...	...	...	...	...	...

The Connection Information Holder (9) provides five basic functions to the infrastructure of the CAC component (2): (i) holds the information of one or more connection requests during the decision process, (ii) requests to the Information Summary Processor (10) the data necessary to the CAC Algorithm (11) can take the decisions, (iii) sends connection requests and the result received from the Information Summary Processor (10) to the CAC Algorithm (11) to preform the decision process, (iv) forwards the accepted connection requests to the Raw Data (3) in Traffic Description Information (3.a) structure or forwards the denied connection requests to the Connection Request Track Information (12), (v) moves the ongoing connections from Traffic Description Information structure (3.a) to the Connection Request Track Information (12) upon the reception of a DSD register message. Moreover, the Connection Information Holder (9) should manage the DSC register messages.

The Information Summary Processor (10) computes accurate information about the usage of the network resources. This information is used by the CAC Algorithm (11) to support the decision of accepting of denying a connection. Moreover, the information is associated with the main QoS requirements guaranteed by the network, *e.g.* maximum throughput, minimum delay, jitter, among others. After the decision processing, the accepted connection request is temporary stored in the Traffic Description Information (3.a) structure of the Raw Data (3), while the connection is active. On the other hand, the denied connection request is stored in the Connection Request Track Information (12) to be used in the performance analysis of the CAC component (2). A model that represents a general view of the data flow in the Information Infrastructure (1) of the CAC component (2) is discussed in the next subsection.

4.4.2 Modeling of the Information Infrastructure to the CAC Component

The goal of this section is to present the actions performed by modules of the Information Infrastructure of the CAC Component. Therefore, the modules of the Information Infrastructure follow the same numerical representation of Figure 4.7. The data flow modeling in the Information Infrastructure (1) of the CAC Component (2) can be observed in Figure 4.8. The flowchart shows the processing of a register message in the OFDMA frame and the handling of the connection request by the Information Infrastructure (1) of the CAC Component (2). Without loss of generality, this flowchart does not consider DSC and DSD register messages once the goal of this subsection is to present the actions of the infrastructure modules for a connection request, *i.e.* DSA register messages.

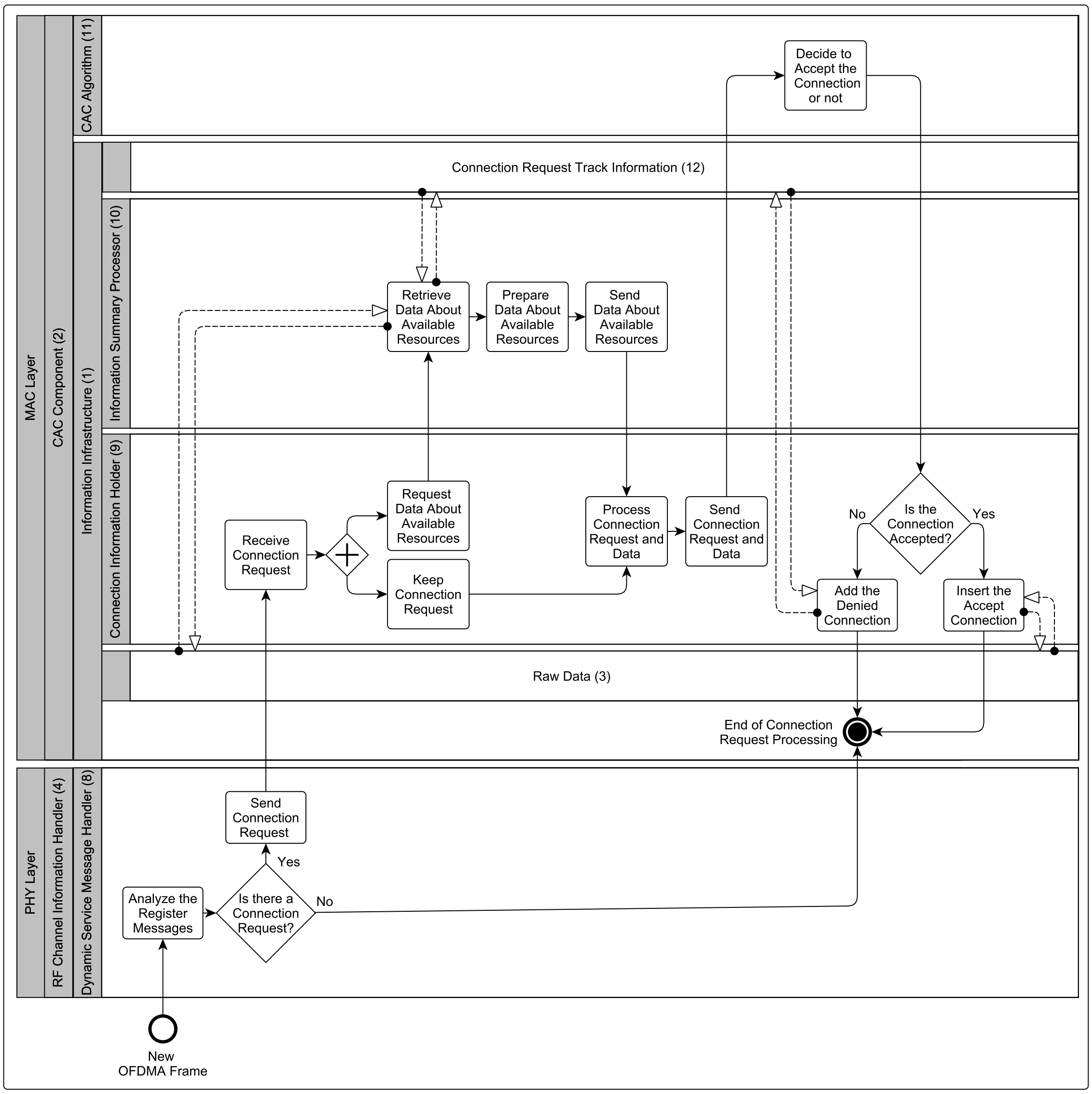


Figure 4.8: Flowchart of Information Infrastructure to the CAC Component

Initially, a new OFDMA frame arrives at the PHY Layer and the Dynamic Service Message Handler (8) analyzes whether there are register messages inside the frame. When the frame carries a DSA register message, *i.e.* a connection request, this message is filtered and sent to the Information Infrastructure (1). In this point, the connection request processing begins at the MAC Layer. The module which receives the connection request in the infrastructure is the Connection Information Holder (9). This module keeps the connection request while it exchanges data about the usage of the network resources with the Information Summary Processor (10). The accurate information over the usage of the network resources are summarized analyzing the Raw Data (3) and the Connection Request Track Information (12) in accordance with the QoS requirements previously defined. The analysis performed to calculate the usage of the network resources is discussed in Subsection 4.4.3.

The Information Summary Processor (10) computes accurate information about the usage of the network resources. This information is used by the CAC Algorithm (11) to support the decision of accepting of denying a connection. Moreover, the information is associated with the main QoS requirements guaranteed by the network, *e.g.* maximum throughput, minimum delay, jitter, among others.

After these steps, the Connection Information Holder (9) sends the connection request and the summarized information regarding the usage of network resources to the CAC Algorithm (11). The function of CAC Algorithm (11) is to decide whether to accept or not the connection request considering the set of data received from the Connection Information Holder (9). The decision is sent from the CAC Algorithm (11) to the Connection Information Holder (9) which inserts the accept connection into the Raw Data (3) or adds the denied connection in the Connection Request Track Information (12).

The data flow processing in the Information Infrastructure previously presented is performed frame-by-frame. In this context, the next section presents the mathematical analysis to calculate the accurate usage of the network resources in the new cross-layer adaptive architecture for IEEE 802.16 networks.

4.4.3 Accurate Information to the CAC Component

The three main QoS requirements according to the Traffic Description Information of the Raw Data are throughput associated with MSTR and MRTR, delay related to ML; Deadline, and timestamp; and jitter (JT). Therefore, the Information Summary Processor is the module of the Information Infrastructure responsible for analyzing the accurate usage of these requirements in the network. The throughput, delay, and jitter values used in each frame are stored in the QoS Components Information of the Raw Data and are considered to calculate the accurate information to the CAC component. In this context, this subsection is composed of three topics about: (i) throughput, (ii) delay, and (iii) jitter that present the mathematical analysis regarding to the usage and availability of the network resources to the CAC component.

Throughput Analysis

First, the throughput analysis must calculate the accurate amount of bits transmitted in each OFDMA subframe (F_{bits}). Taking into account equation (4.10), where the static (Ψ) and dynamic (Φ) mapping overhead, PDUs, and BRs are considered. The calculation of the F_{bits} can be seen in equation (4.12) that considers the modulation symbols (M) and the FEC encoding (K) for each m^{th} and s^{th} slots of the subframe.

$$F_{bits} = (\Psi + \Phi).M_m.k_m + \sum_{s=1}^{N_{slot}} \left(\sum_{r=0}^{N_{pdu(s)}} PDU_{(s,r)} + \sum_{l=0}^{N_{br(s)}} BR_{(s,l)} \right) .M_s.K_s \quad [bits] \tag{4.12}$$

After obtaining F_{bits}, the accurate throughput, *i.e.* the used capacity (C_{used}) of the network, can be calculated in a given time interval t. Where, the duration of a frame is represented by T_f, $f \in \{1, 2, ... \text{F}\}$ is the index of the frame, and F is the number of frames transmitted in an interval t. In other word, C_{used} is calculated considering the accurate historical throughput of F subframes, as can be seen in equation (4.13).

$$C_{used} = \frac{\sum_{f=1}^{F} F_{bits(q)}}{T_f.F} \quad [bit/s] \tag{4.13}$$

The throughput analysis must also estimates the total capacity of the RF channel

(C_{tot}). However, before it is necessary to define which MSC should be used for this estimative. According to Section 3.2, this thesis considers a realistic estimative, which is aware of the frame structure by considering two approaches. First, the average MCSs in each OFDMA subframe is considered, assuring the representation of channel variability. For example, an optimist case, *i.e.* using MCS equal to 64 QAM 5/6, and a pessimist case, *i.e.* considering MCS equal to QPSK 1/2, are not realistic approaches. Moreover, they do not represent the channel variability. Second, the historical average of MCSs for F is considered subframes. In other words, the historical average of MCS is used to estimate C_{tot}. The calculation of the historical average of MSC ($\overline{MCS}$) can be seen in equation (4.14).

$$\overline{MCS} = \left\lceil \frac{\sum_{f=1}^{F} \left(\frac{\sum_{j=1}^{DL_{IE(f)}} burst(f,j)}{DL_{IE(f)}} \right)}{F} \right\rceil \tag{4.14}$$

Where, $j \in \{1, 2, ...\ DL_{IE}\}$ is the index of MSC for a burst profile of a given subframe f and DL_{IE} is the amount of burst profiles in a OFDMA subframe f. Therefore, based on Both *et al.* (BOTH et al., 2008) C_{tot} can be obtained considering the total available bandwidth (W), a sampling factor (q), the number of OFDM subcarriers total (N_{fft}) and used (N_{used}), a guard interval, also called Cyclic Prefix (CP), and $\overline{MCS}$. The calculation of C_{tot} can be seen in equation (4.15).

$$C_{tot} = \lfloor W.q \rfloor . \frac{N_{used}}{N_{fft}} . \frac{1}{1+CP} . \overline{MCS} \quad [bit/s] \tag{4.15}$$

Finally, the available capability ($C_{available}$) can be calculated according to equation (4.16). Where, $0 \leq \varepsilon \leq 1$ represents the efficiency of scheduler and allocator components. For example, the amount of mapping overhead and data arranged into OFDMA subframe is related with the efficiency of the allocation algorithm, *i.e.* the variability on bursts size, which is caused both by the dynamics of the network traffic requirements and the time variant RF channel conditions.

$$C_{available} = C_{tot}.\varepsilon - C_{used} \quad [bit/s] \tag{4.16}$$

The efficiency of scheduler and allocator components integrated is presented in Chapter 5. In addition, the second QoS requirement is presented in the next topic.

Delay Analysis

The second accurate information essential to the CAC component is the residence time of PDUs and BRs inside of the new cross-layer adaptive architecture, *i.e* the delay generated by the QoS architecture in the WiMAX BS. Therefore, the delay is obtained considering the creation or arrival time (T_c) of each PDU and BR in the QoS architecture, as well as the system up time (T_{sys}), such as $DLY = T_{sys} - T_c$, where $T_c \leq T_{sys}$. T_c is related to the timestamp field of the Joint Information structure associated with each PDU and BR.

After the analysis of the delay of each PDU and BR it is possible to calculate the average delay for a specific connection ($\overline{DLY_c}$). Equation 4.17 presents as the $\overline{DLY_c}$ is obtained.

$$\overline{DLY_c} = \frac{\sum_{i=1}^{N_{data}} DLY_{(c,i)}}{N_{data}} \quad [ms] \tag{4.17}$$

Where, $i \in \{1, 2, ... N_{data}\}$ is the index of a PDU or BR and $N_{data} = N_{pdu} + N_{br}$ for a given connection (c). In similar manner, the association between the PDU and BR with a connection is using the CID field of the Joint Information. After the delay analysis, it is possible to perform the jitter analysis as presented in the next topic.

Jitter Analysis

The last accurate information sent to the CAC component is the variation of the delay of PDUs and BRs, called jitter. The variation of each PDU and BR is calculated such as $JTR = |DLY_{(i+1)} - DLY_{(i)}|$. Therefore, the average jitter for a determined connection $\overline{JTR_c}$ is obtained according to the equation (4.18).

$$\overline{JTR_c} = \frac{\sum_{i=1}^{N_{data}} JTR_{(c,i)}}{N_{data}} \quad [ms] \tag{4.18}$$

The Information Infrastructure provides to the CAC Algorithm accurate information about two fundamental perspectives: (i) the usage of QoS metrics in the network and (ii) the QoS requirements requested by new connections. Based on these information types, the CAC Algorithm decides whether to accept or not new connections. Therefore, in this approach the CAC algorithm can be simple and efficient to support the QoS requirements according to the characteristics of each network. The performance evaluation of the CAC algorithms, as well as of the new cross-layer adaptive architecture is presented in the next chapter.

4.5 Summary

This chapter presented the proposal of a new cross-layer adaptive architecture for IEEE 802.16 networks. First, the RF channel model using a time-discrete Markov chain was presented. The aim of this model was to represent the different MCS levels that are supported in transmissions among MSs and the BS to consider time-variant changes of the RF channel conditions. Second, the design of the new channel-aware architecture was presented. The focus of this design was to allow the QoS components to be aware of changes in the RF channel using the proposed Joint Information structure. Third, the project of a module to integrate scheduler and allocator components was discussed, considering the mapping overhead of the OFDMA frame and a threshold criteria to define the amount of PDUs and BRs that can be sent from the scheduler to the allocator. This chapter was closed with the design of an unified CAC component in the new cross-layer adaptive architecture. In order to reach this unification an Information Infrastructure to support the CAC component was presented.

5 EVALUATION OF THE PROPOSED ARCHITECTURE

The aim of this chapter is to present the evaluation of the new cross-layer adaptive architecture. The chapter starts with the description of the simulation design considering the simulator model, the physical aspects of the simulation scenario, and both traffic and connection request models used in the evaluation of the proposed architecture. After, the results are discussed to analyze the performance of the integration among CAC, scheduler, and allocator components.

5.1 Simulation Design

The simulator used in this thesis was designed with two high-level computing languages: Labview and Matlab. The first was used as core language, *i.e.* the programmable logic of components, as well as the integration among them were developed using the Labview. The choice as core language for developing the simulator was due to the faster programming, friendly user interface, and good support to debugging code. The second language was applied to design the traffic and connection request models because the Matlab has an excellent support for probabilistic and statistical requirements. In the following subsection the simulator model is presented.

5.1.1 Simulator Model

The simulation environment is composed of three levels as can seen in Figure 5.1. The first level regards to the configuration of the simulation environment itself. The second level refers to the activities of the MSs. The third level involves the three components of the new cross-layer adaptive architecture in the WiMAX BS.

Initially, the simulation tool reads the configuration parameters, *e.g.* simulation time, frame duration, amount of connection requests for each class of services, and others. In parallel the PHY parameters are assigned, as well as the traffic and connection request models are initialized using the Matlab language. Aiming to a better organization, the PHY parameters is discussed in Subsection 5.1.2.

After the configuration of the simulation environment, the connection requests from MSs are sent to the BS. Moreover, bandwidth requests are sent to the BS regarding to each accept connection. In this context, without loss of generality of a cellular system, each MS can only have one assigned connection, which is not used for QoS prioritization purposes. Thus, the prioritization is based on the class of service and not on MSs. The traffic and connection request models of the class of services are presented in Subsection 5.1.3.

Finally, the three components of the new cross-layer adaptive architecture are simulated in the WiMAX BS. The PDUs belonging to connections accepted by the CAC component are created using the traffic model that represents the behavior of each class of service. In the next step, PDUs and BRs are sent to the scheduler component which is integrated with the allocator component, as presented in Section 4.3. PDUs and BRs that were not allocated in the OFDMA frame return to the scheduler component as discussed in Subsection 4.3.2. Moreover, feedback information in each frame is sent to the CAC component to be used in the decision process of accepting or not new connections.

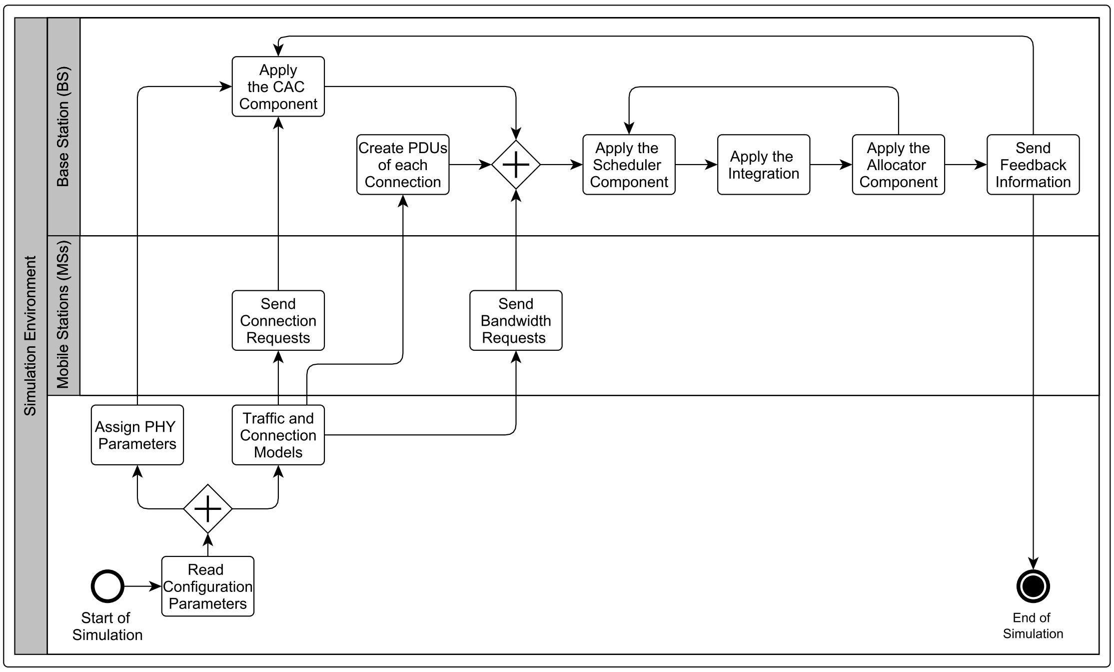


Figure 5.1: Flowchart of the Simulation Environment

5.1.2 Simulation Scenario

In order to analyze the behavior of the new cross-layer adaptive architecture, it is necessary to characterize the physical aspects of the IEEE 802.16 networks. These aspects are similar for all simulation scenarios. In this context, this subsection is divided in two topics. The first topic presents the transition probabilities of the RF channel model proposed in this thesis. The second topic refers to physical parameters used to the performance evaluation of the proposed architecture.

Transition Probabilities of the RF Channel Model

The aim of this topic is to present the transition probabilities among the Markov chain states defined in Section 4.1 to model the RF channel conditions. The transition probabilities defined in Table 5.1 are an adaptation of the scenario presented by Lakkakorpi, Sayenko and Moilanem (LAKKAKORPI; SAYENKO; MOILANEN, 2008), where a fixed WiMAX network with 55 connections is simulated by the Nokia Research Center. The transition from a state to another is performed in each frame.

The adaptation of probabilistic values in the RF channel model performed in this thesis aims to adjust the fixed model to a mobile scenario. In other word, only the probabilistic values were changed and no new state or transition was inserted in the RF channel model. The state sojourn times were decreased and the transition values among the states were increased, allowing a scenario with higher mobility than one proposed by Lakkakorpi, Sayenko, and Moilanem (LAKKAKORPI; SAYENKO; MOILANEN, 2008).

Table 5.1: Transition Probabilities

MCS	QPSK 1/2	QPSK 3/4	16 QAM 1/2	16 QAM 3/4	64 QAM 1/2	64 QAM2 2/3	64 QAM 3/4	64 QAM 5/6
QPSK 1/2	0.5	0.5	0	0	0	0	0	0
QPSK 3/4	0.33	0.34	0.33	0	0	0	0	0
16 QAM 1/2	0	0.33	0.34	0.33	0	0	0	0
16 QAM 3/4	0	0	0.33	0.34	0.33	0	0	0
64 QAM 1/2	0	0	0	0.33	0.34	0.33	0	0
64 QAM 2/3	0	0	0	0	0.33	0.34	0.33	0
64 QAM 3/4	0	0	0	0	0	0.33	0.34	0.33
64 QAM 5/6	0	0	0	0	0	0	0.5	0.5

After the transition values have been adapted to the mobile scenario, the physical parameters must be defined in the simulation environment. In this context, the next topic discusses these simulation parameters.

Physical Parameters

The parameters considered in the simulations are based on typical values found in the literature (WANG et al., 2008) and are summarized in Table 5.2. The mobile WiMAX allows almost any available spectrum bandwidth to be used. Allowed channel bandwidths vary from 1.25 MHz to 28 MHz. A RF channel bandwidth of 10 MHz TDD system with 5ms frame duration and PUSC subchannelization mode is considered to simulate mobile WiMAX networks in this thesis. These are the default values recommended by the mobile WiMAX forum system evaluation methodology and are also common values used in practice (WiMAX Forum, 2008) (SO-IN; JAIN; TAMIMI, 2009a).

Table 5.2: Physical Parameters

Parameters	Values
Channel bandwidth	10 MHz
Frame duration	5ms
FFT size (subcarrier)	1024
FFT data (subcarrier)	720
OFDMA subchannels (DL:UL)	30:35
Subframe symbols (DL:UL)	29:18
PUSC zone in slots (DL:UL)	420:210
Cyclic prefix	1/8
Ranging channel	12 slots

The Fast Fourier Transform (FFT) size of 1024 with 720 subcarriers for data transmission has been chosen to reflect the mandatory implementation to guarantee devices interoperability. Moreover, considering the description in Section 2.2.1, the group of 28 subcarriers is called a DL subchannel leading to 30 DL subchannels. In the same way, 24 subcarriers form a UL subchannel resulting 35 UL subchannels in a subframe.

Considering the values previously described, the DL subframe is composed of 29 symbols and the UL subframe is formed by 18 symbols. Therefore, the DL subframe consists of a total of $((29-1)/2).30 = 420$ DL slots, where a symbol is discounted due to preamble and the division by 2 represents the amount of symbols in a DL slot. The UL subframe has capacity of $((18-3)/3).35 = 175$ UL slots, where 3 symbols are discounted because of ranging channel and the division by 3 represents the amount of symbols in a UL slot. The ranging channel in UL subframe is fixed in 35 slots, where each one can transport 30 different CDMA codes (YOU; KIM; KIM, 2005). Finally, a cyclic prefix configuration of 1/8 is considered since this is the only mandatory value for mobile WiMAX systems (WiMAX Forum, 2008).

QoS Components Parameters

The new cross-layer adaptive architecture is modular and flexible to support any algorithm in the design of its QoS components. Therefore, the parameters defined to configure the CAC, scheduler, and allocator components are specific for each algorithm. The WPF algorithm is considered in the Intra-class scheduling to define the order of PDUs and BRs within the queues service. The choice of WPF is because this algorithm can improve resource allocation, fairness, and QoS guarantees by assigning a weight for each PDU and BR in the queues of the Intra-class scheduling (HOU et al., 2008).

In the Inter-class scheduling, Proportional Fair (PF) (SHREEDHAR; VARGHESE, 1996) and DRR algorithms are considered to select which queue must be the next one to be served. The service order used in the simulation scenario is: ertPS, rtPS, and BE to reproduce the traffic demand of a cellular system. In this context, it is considered a deadline equal to 155ms for ertPS and a deadline equal to 100ms for rtPS, as presented in Table 5.3. These values were based on the performance study of scheduling algorithms by Ali, Dhrona and Hassanein (ALI; DHRONA; HASSANEIN, 2009). Moreover, a confidence interval of 95% is guaranteed in the results discussed in Subsection 5.1.3.

Table 5.3: QoS Parameters

Parameters	**Values**
Queues priority	ertPS > rtPS > BE
Deadline for ertPS	155ms
Deadline for rtPS	100ms
Confidence interval	95%

In the allocator component, MiSP, OCSA, and RTS algorithms are chosen to arrange PDUs and BRs within the OFDMA frame. In the RTS algorithm it is considered the maximum of 10 iterations and the minimum acceptable parameter equal 0.5, once these values presented the best performance in the analysis conducted by Cicconetti, *et al.* (CICCONETTI et al., 2010). MiSP and OCSA algorithms demand no generic parameters for the configuration scenario.

The RTS algorithm defines a function $\phi(\pi)$ to calculate the mapping overhead for control information, but do not specify how this function should be implemented. MiSP and OCSA algorithms do not consider the mapping overhead. Therefore, the equations defined in Subsection 4.3.1 are used to calculate the mapping overhead in the three algorithms to analyze the performance of allocation in the new cross-layer adaptive architecture.

In the CAC component, very simple CAC algorithms admit new connections based on the occupancy of the bandwidth allocated to each WiMAX class of service (ZHU et al., 2008); as long as a class can accommodate the traffic of new connections, these new connections are admitted. Recent investigations, however, go a step further and propose more sophisticated CAC algorithms which decisions also take into account dynamic traffic requirements, *e.g.*, maximum latency and tolerated jitter. In this context, two CAC algorithms using a policy without degradation of service are defined considering the accurate information described in Subsection 4.4.3.

The first one, called BW_CAC, is depicted in Algorithm 1, which major goal is to accept all connection requests while there is available capacity *i.e.* available space in the frames. The CAC component calculates the available capacity and returns this information to the BW_CAC algorithm using $C_{available}(\,)$ function, which is part of the Information Summary Processor of the CAC component and is defined in equation 4.15. In this context, the new cross-layer adaptive architecture allows the usage of simple, modular, and efficient CAC algorithms. Moreover, to accept or not a connection, the BW_CAC algorithm considers the capacity related to QoS parameters, *e.g.* MSTR and MRTR, for each class of service.

Algorithm 1 BW_CAC

Require: $MSTR\_ertPS$
Require: $MSTR\_rtPS$
Require: $MRTR\_BE$
Ensure: $C_{available}(\,)$ returns the amount of bits that can be allocated in DL subframe
if $class\_of\_service = "ertPS"$ **then**
  if $(C_{available}(\,) - MSTR\_ertPS) > 0$ **then**
    Accept the connection request
  else
    Reject the connection request
  end if
else if $class\_of\_service = "rtPS"$ **then**
  if $(C_{available}(\,) - MSTR\_rtPS) > 0$ **then**
    Accept the connection request
  else
    Reject the connection request
  end if
else
  if $(C_{available}(\,) - MRTR\_BE) > 0$ **then**
    Accept the connection request
  else
    Reject the connection request
  end if
end if

The second one, called QoS_CAC, is presented in Algorithm 2, where the acceptance policy is based on the conformance of capacity, delay, and jitter QoS requirements of the specific class of service of the incoming connection request. For example, in the case of an incoming connection belonging to ertPS class of service, it will be accepted only if there is available capacity, and both the delay and jitter requirements can be guaranteed. The delay requirement of the ertPS class of service is associated with the $\overline{DLY}$ function defined in equation 4.17. In the same way, the jitter requirement is related with the $\overline{JTR}$

function defined in equation 4.18. These functions are also part of the Information Summary Processor of the CAC component and return the average values of delay and jitter, respectively.

Algorithm 2 QoS_CAC

Require: $MSTR\_ertPS$
Require: ML_ertPS
Require: JT_ertPS
Require: $MSTR\_rtPS$
Require: JT_rtPS
Require: $MRTR\_BE$
Ensure: $C_{available}(\ )$ returns the amount of bits that can be allocated in DL subframe
Ensure: $\overline{DLY}(class\_of\_service)$ returns the average delay of a specific class of service
Ensure: $\overline{JTR}(class\_of\_service)$ returns the average jitter of a specific class of service

if $class\_of\_service = "ertPS"$ **then**
  if $(C_{available}(\ ) - MSTR\_ertPS) > 0)$ AND $(\overline{DLY}(ertPS) \leq ML\_ertPS)$ AND $(\overline{JTR}(ertPS) \leq JT\_ertPS)$ **then**
    Accept the connection request
  else
    Reject the connection request
  end if
else if $class\_of\_service = "rtPS"$ **then**
  if $(C_{available}(\ ) - MSTR\_rtPS) > 0)$ AND $(\overline{JTR}(rtPS) \leq JT\_rtPS)$ **then**
    Accept the connection request
  else
    Reject the connection request
  end if
else
  if $(C_{available}(\ ) - MRTR\_BE) > 0$ **then**
    Accept the connection request
  else
    Reject the connection request
  end if
end if

The parameters defined for the algorithms of the QoS components are not sufficient to configure the simulation environment. The modeling of the traffic and connection requests also are fundamentals for the design of an accurate simulation scenario. The next subsection presents details on traffic and connection request models used in the simulation scenarios.

5.1.3 Traffic and Connection Request Models

Three traffic and connection request models are developed to associate the features of ertPS, rtPS, and BE classes of service with respectively VoIP, Video Clip, and HTTP applications. Such association between the classes of service and the applications is based on the work proposed by So-In, *et al.* (SO-IN; JAIN; TAMIMI, 2009a). Furthermore, the traffic and connection request models used to describe the behavior of these applications are based on the *System Evaluation Methodology* document, published by the WiMAX Forum (WiMAX Forum, 2008). Initially, the traffic models for each application are described in next three topics. The connection request models are presented in the fourth topic.

VoIP Traffic Model

The parameters of Adaptive Multi Rate (AMR) codec are used to model the VoIP traffic, *i.e.*, it is considered a voice traffic with silence suppression. The parameters of the VoIP traffic model are presented in Table 5.4.

The VoIP application behaves like an ON/OFF source model. In other words, PDUs

Table 5.4: Codec AMR parameters

Parameters	Values
Talk spurt length (On)	Exponential, μ = 1026ms
Silence length (Off)	Exponential, μ = 1171ms
Transmission Time Interval	20ms
Total MAC PDU size during a talk spurt (ON)	MAC header (6 bytes) + compressed RTP/UDP/IP header (3 bytes) + voice packet (33 bytes) = 42 bytes

are transmitted during ON period and only comfort noise is sent during OFF period. Estepa *et al.* (ESTEPA; ESTEPA; VOZMEDIANO, 2004), using experimental measurements, proved that ON and OFF periods of AMR codec follow an exponential probability distribution, with average values of 1026ms to the ON period and 1171ms to the OFF period. Moreover, the VoIP traffic model considers the transmission time interval of 20ms and the headers of transport protocols, as Real-time Transport Protocol (RTP), User Datagram Protocol (UDP), and IP, are compressed.

Video Clip Traffic Model

A Video Clip application is characterized as a short video, usually of musical or advertising nature. In this traffic, aspects as video codec and display dimension influence in the file size of the video transmitted between BS and MSs. The parameters considered for the Video Clip traffic model are presented in Table 5.5.

Table 5.5: Video Clip traffic parameters

Parameters	Values
Video Clip length	Truncated exponential Mean 15s, Max = 60s
Video codec	MPEG-4
Display size	176 x 144
Mean compressed frame size	2.725 Kbytes

The parameters and values used in the traffic model may change from one video trace to another. The values suggested by the WiMAX Forum (WiMAX Forum, 2008) are considered. The duration of the Video Clip is provided by the truncated exponential distribution, with average duration of 15s and maximum period of 60s. The MPEG-4 codec is chosen due to its wide usage and compression capability. Moreover, it is considered the generation of 25 frames per second and the display dimension of 176 x 144 pixels. The Video Clip traffic is modeled with the transmission of files, encapsulated within FTP protocol. In this case, the period between the transmission of two different videos behaves like an exponential distribution, with average of 180s.

HTTP Traffic Model

The HTTP traffic is modeled using typical sessions of web browsing. In this context, each web page consists of a number of objects, such as a main page and embedded objects. The session is divided into ON/OFF periods representing web page downloads and the reading times. The download of the web page and each one of the constituent objects is represented by the ON period, while the parsing time and protocol overhead are represented by the OFF periods. The reading time is the period between the access of two web pages and includes the time for the user to read all or parts of the web page.

The parameters of the HTTP traffic are presented in Table 5.6, where "SD" means the Standard Deviation value, "Min" represents the Minimum value, and "Max" is the Maximum value. Moreover, the values of the Probability Density Functions (PDF) are shown, considering the recommendations of WiMAX Forum (WiMAX Forum, 2008).

Table 5.6: HTTP traffic parameters

Componet	**Distribution**	**Parameters**	**PDF**
Main page size	Truncated Lognormal	Mean = 10710 bytes SD = 25032 bytes Min = 100 bytes Max = 2 Mbytes	$\sigma = 1.37$ $\mu = 8.37$
Embedded object size	Truncated Lognormal	Mean = 7758 bytes SD = 126168 bytes Min = 50 bytes Max = 2 Mbytes	$\sigma = 2.36$ $\mu = 6.17$
Number of embedded objects page	Truncated Pareto	Mean = 5.64 Max = 53	$\sigma = 1.1$ $\mu = 55$
Reading time	Exponential	Mean = 30s	$\mu = 0.033$
Parsing time	Exponential	Mean = 0.13s	$\mu = 7.69$

The traffic models are used when a connection request is accepted by the CAC algorithm. Therefore, the connection request descriptors also must be modeled. In this context, the next topic presents the connection request models.

Connection Request Models

The traffic descriptors of connection requests are received by the BS and must be characterized in the simulator. Consequently, three connection request models are defined to associate the characteristics of ertPS, rtPS, and BE classes of service with VoIP, Video Clip, and Web applications, respectively. Such association is based on the work of So-In, *et al.* (SO-IN; JAIN; TAMIMI, 2009a). Therefore, the values of all QoS parameters that compose the traffic descriptor, *i.e.* MRTR, MSTR, ML, and JT, are summarized in Table 5.7.

The QoS parameters MSTR and MRTR are modeled using the Truncated Lognormal distribution with $\sigma = 1.37$ and $\mu = 8.37$ to the PDF values due to the characteristic of these parameters, *i.e.* bits unit. On the other hand, the ML and JT are modeled using the Truncated Exponential distribution because these parameters are in time units. Moreover, some QoS parameters are not applicable (N/A) for the traffic descriptor of a specific application. After the simulator design has been developed, the performance evaluation of the QoS components was performed. Therefore, the next section discusses the results regarding to the performance of the new cross-layer adaptive architecture.

Table 5.7: QoS Parameter Values

Traffics/ Parameters	**MSTR**	**MRTR**	**ML**	**JT**
VoIP	16.4 Kbps	N/A	Truncated Exponential Mean = 70 ms Min = 5 ms Max = 150 ms	Truncated Exponential Mean = 30 ms Min = 5 ms Max = 50 ms
Video Clip	Truncated Lognormal Max = 2 Mbps	Truncated Lognormal Max = 5 Kbps	N/A	Truncated Exponential Mean = 50 ms Max = 100 ms
Web	Truncated Lognormal Max = 2 Mbps	Truncated Lognormal Max = 10 Kbps	N/A	N/A

5.2 Performance Evaluation

The goal of this section is to analyze the performance of the new cross-layer adaptive architecture taking into account both the traffic requirements and physical characteristics of IEEE 802.16 networks. However, to the best of our knowledge, no architecture considering OFDMA frame structure, as defined in Section 3.2, has been proposed. Indeed, the literature on mobile IEEE 802.16 networks presents that the current research is focused on specific solutions for each QoS component. Moreover, these QoS components are usually evaluated in an individual manner *i.e.* without considering the performance influence of the one component over another. In this context, the performance evaluation of the new cross-layer adaptive architectures in comparison with other architectures is neither fair nor useful.

To analyze the performance of the new cross-layer adaptive architecture, the algorithms and functionalities of QoS components must be evaluated considering the context of the whole architecture. For that reason, this section is organized in three subsection. In the first one the results of the scheduling factors are discussed. In the second one the results about the integration between the scheduler and allocator components are presented. Finally, in the third one the results of the CAC component integrated in the new cross-layer adaptive architecture are analyzed.

5.2.1 Scheduling Factors Analysis

To investigate the scheduler factors, the new cross-layer adaptive architecture was configured in the complete manner, *i.e.* CAC, scheduler, and allocator components were considered. Despite of the results focusing only in the performance of the scheduling factors, in this subsection. In this context, the CAC component been enabled, but no CAC algorithm was used, *i.e.* all connections were accepted. The algorithms used in the Inter-class scheduling and allocation were PF and RTS, respectively. These algorithms were chosen because presented the better performance than others, as describe in the next subsection.

The first analysis on the scheduling factors was performed with the φ_{UGS_ertPS} considering the VoIP traffic in the ertPS queue. The simulation scenario was configured with only one traffic because the aim of this analysis is to observe the smoothing factor α, which sets weights to two QoS metrics: Deadline ($D_i^{`}$) and Throughput ($\tau_i^{`}$), as presented in equation (4.2). The weights used for α were 0.1, 0.5, and 0.9 to investigate the behavior of Intra-class scheduling considering the maximum, average, and minimum values. Moreover, the simulations considered up to 800 connection requests. The amount of connections was limited to 800 to keep the simulation scenario close to the reality.

Considering the previously described, all VoIP traffic was allocated. Therefore, the amount of bytes scheduled was the same using the three values for α. In this case, the φ_{UGS_ertPS} had no effect because all traffic demanded was served. The amount of bytes sent from the scheduler component to the allocator component presented a linear increase in relation to the number of connections. For example, about 400 bytes of VoIP traffic were allocated per frame in a scenario with 100 connections, while 3200 bytes were allocated per frame considering 800 connections. This behavior is due to the characteristic of the VoIP traffic, *i.e.* the packages size are constant as described in Subsection 5.1.3.

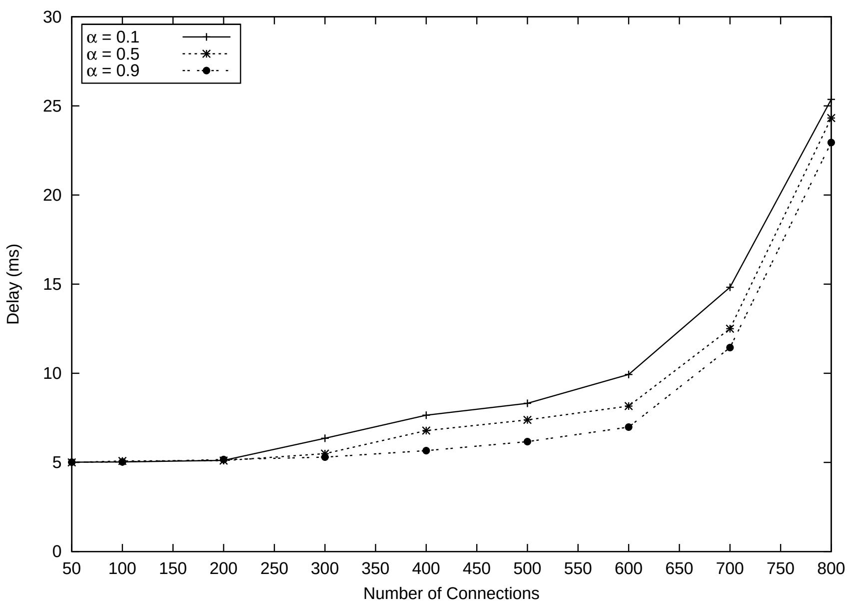


Figure 5.2: VoIP traffic delay for α smoothing

The deadline metric used in the φ_{UGS_ertPS} can be investigated analyzing the delay because the deadline of each PDU or BR is decreased 5ms in each simulation cycle. The analysis of Figure 5.2 shows that the maximum average delay was approximately 25ms for 800 connections, *i.e.* no VoIP PDU was discarded in this simulation scenario. Another interesting observation it is that the smoothing factor α had no effect until 200 connections. This behavior occurs because all VoIP PDUs were served in the first frame.

On the other hand, after 200 connections, the amount of VoIP traffic was higher than the capacity of an OFDMA frame. Therefore, the smoothing factor α began to appear. α equal to 0.9 presented a lower delay than α values of 0.1 and 0.5 because it was applied a higher weigh in the deadline metric than in the throughput metric considering the φ_{UGS_ertPS} scheduling factor.

The investigation of the delay for the VoIP traffic using the φ_{UGS_ertPS} scheduling factor shows that the smoothing factor α is efficient in saturated scenarios. The VoIP traffic is composed of small and constant PDUs, therefore the gain of performance is limited by the amount traffic generated by the VoIP traffic model. The performance of the φ_{UGS_ertPS} scheduling factor must be better in situations where traffic models that demand a high amount of network resources are used.

The φ_{rtPS} scheduling factor was investigated using the same methodology previously described. However, the simulation scenario was configured with only Video traffic in the rtPS queue. The weights used for the smoothing factor ρ were 0.1, 0.5, and 0.9 for the current throughput (τ) and the average throughput ($\overline{\tau}$), as presented in equation 4.3. The average amount of bytes scheduled in an OFDMA frame in relationship with the number of connections is presented in Figure 5.3. The best throughput was found using the smoothing factor ρ equal to 0.1. For example, approximately 6000 bytes were allocated per OFDMA frame considering 800 connections. This behavior occurred due to the higher weight set to $\overline{\tau}$ in detriment to τ. In other words, the best performance was found when the φ_{rtPS} scheduling factor prioritized the historical throughput instead of the current throughput that considers the instantaneous variations of both traffic and physical characteristics. Moreover, this behavior was negligible considering 100 connections because the variability of the traffic and physical characteristics was small in this simulation scenario.

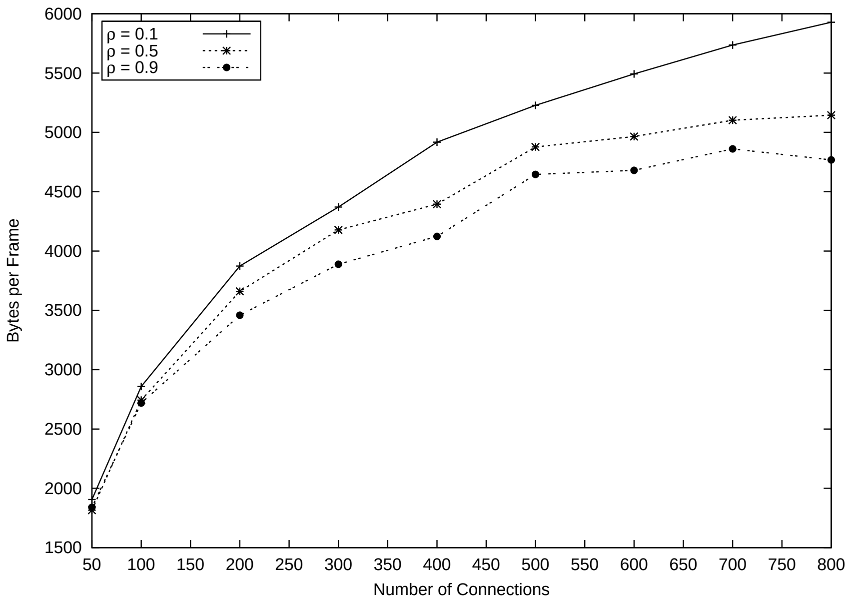


Figure 5.3: Video traffic throughput for ρ smoothing

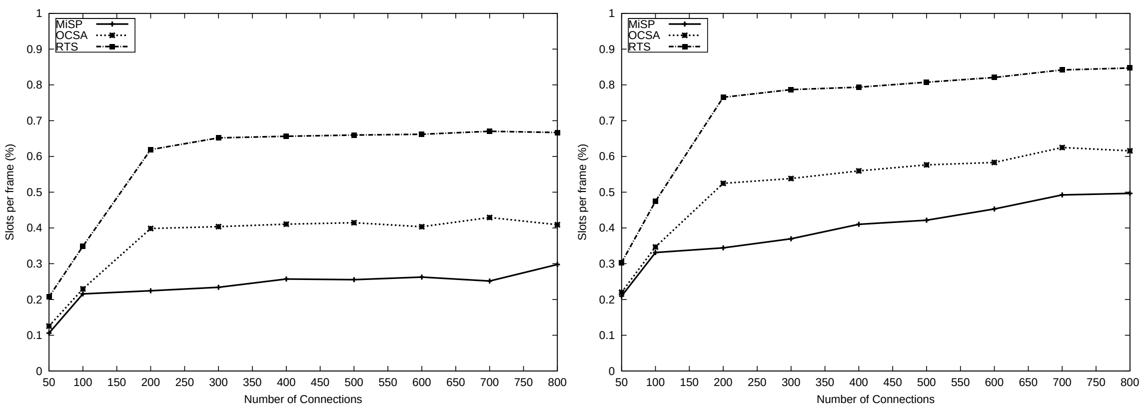
The delay found using the φ_{rtPS} scheduling factor was in average between 20ms and 70ms, *i.e.* no Video PDU was discarded and the QoS requirements were guaranteed by the new cross-layer adaptive architecture. However, the effect of the smoothing factor ρ

was negligible in these simulation scenarios. Indeed, the main goal of the φ_{rtPS} scheduling factor is to improve the throughput of rtPS transmissions, without to compromising the QoS requirements. In this context, the next subsection presents the results of the integration between scheduler and allocator components.

5.2.2 Integration Between Scheduler and Allocator Components Analysis

The aim of this subsection is to analyze the performance of scheduler and allocator components considering the new cross-layer adaptive architecture for IEEE 802.16 networks. It is important to describe that the investigations about allocation algorithms found in the literature analyze these algorithms in an isolated manner, *i.e.* without taking into account the traffic requirements, physical characteristics of IEEE 802.16 networks, and the relationship with the remaining QoS components. In this context, the scheduling and allocation algorithms are investigated, as well as some metrics related to the allocation capacity in the OFDMA frame and to the number of connections. Focusing on the performance analysis of the integration between scheduler and allocator components, the simulation scenarios in this subsection considered α and ρ factors equal to 0.5 for normalization and smoothing values, respectively.

The first investigation of this subsection shows the efficiency of three allocation algorithms: MiSP, OCSA, and RTS in terms of the amount of allocated data within a DL subframe with a maximum capacity of 420 slots. The efficiency of the allocation algorithms can be measured analyzing the relationship between allocated data and the unused space of a DL subframe. The allocation algorithms do not have an efficiency of 100% due to demand for arranging a big variability of bursts size into OFDMA subframe. This variability is associated with both the traffic requirements and physical characteristics of IEEE 802.16 networks, as discussed in section 4.3.2. To analyze this efficiency, a simulation scenario composed of the PF algorithm in the Inter-class scheduling, without considering the adjusting factor (Λ) in threshold criteria for the integration module, as well as VoIP, Video, and HTTP traffics in same proportion of connections was defined.



(a) Without mapping overhead (b) With mapping overhead

Figure 5.4: Performance of the allocation algorithms

In this analysis, RTS algorithm presented the best performance in terms of data allocation, when compared to the MiSP and OCSA algorithms. Figure 5.4 (a) shows that in scenarios with more than 200 connections, RTS algorithm allocated approximately 66% of data, while OCSA algorithm allocated 40% and MiSP algorithm allocated 22%, *i.e.*

RTS algorithm allocated approximately 66% - 40% = 26% more data than OCSA algorithm and allocated around 40% - 22% = 18% more data than MiSP algorithm. Figure 5.4 (a) also shows an interesting behavior when the BS received less than 200 connections. In this case, the amount of traffic generated in the simulation scenario was smaller than the capacity of DL subframe. Thus, the allocation algorithms presented a linear increase until reaching its maximum efficiency. After this point, the amount of data allocated remained constant as the number of connections increased.

The allocation can be classified into two groups: users' traffic, *i.e.* the amount of traffic without considering the mapping overhead, as can be seen in Figure 5.4 (a) and allocated data considering the mapping overhead, as can be observed in Figure 5.4 (b). It is also possible to analyze from Figure 5.4 that the mapping overhead increased the DL subframe usage in about 12%, *e.g.* 78% - 66% = 12% when RTS algorithm has been applied. This increase can be observed in all allocation algorithms and the insertions of mapping overhead did not affect the behavior of algorithms, despite of increasing the DL subframe usage. Thus, it is possible to conclude that the mapping overhead should be considered in the allocation algorithm design. Therefore, the overhead is analyzed in the next investigation.

The overhead is composed of FCH, DCD, UCD, DL-MAP, and UL-MAP structures of the DL subframe. To analyze the overhead nature, the second investigation of this subsection considers the overhead separating the DL and UL mapping. It is possible to observe in Figure 5.5 that UL mapping generated about 4% more overhead than DL mapping, considering an overall overhead of 12%. This behavior occurred because UL transmissions cannot aggregate a set of connections with a same burst profile. These aggregations are not possible because the medium access is shared among the MSs in UL transmissions. Therefore, each UL connection is associated with a burst profile, unlike in the case of DL connections where it is possible to group a set of connections with a same burst profile. In addition, it is import to describe that the overhead of UL connection requests is not considered in this investigation, because the connection requests are allocated in the ranging channel and in the first burst of the UL subframe.

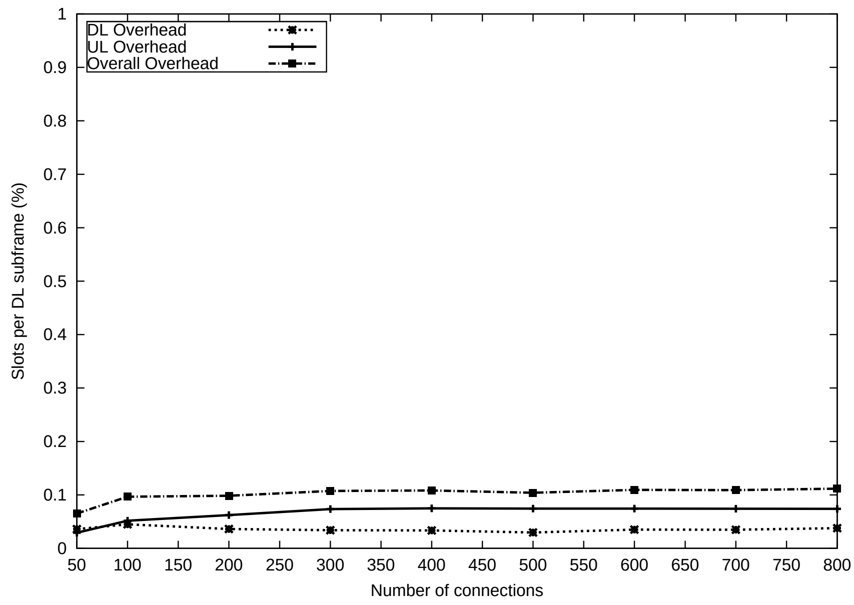
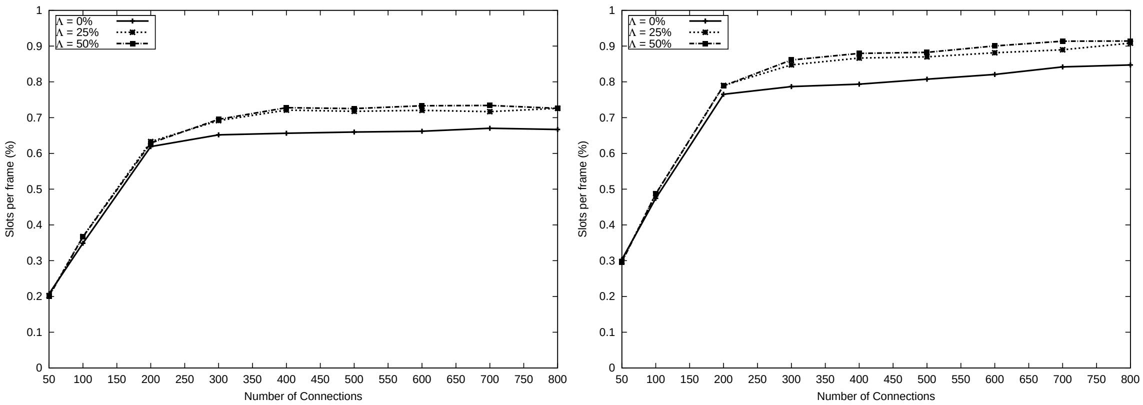


Figure 5.5: Mapping overhead

According to Subsection 4.3.1, the mapping overhead depends on the number of ongoing connections and on the amount of burst profiles. However, due to the frame capacity, *e.g.* 420 slots in the DL subframe, the average amount of connections and burst profiles are constant within OFDMA frames. Consequently the average overhead was also constant, as can be seen in Figure 5.5. This behavior is related to the high diversity of the RF channel and to the number of MSs in the network.

Another interesting feature of the allocator is its capability of feeding back the scheduler with unallocated data, *i.e.* the originally unallocated data can be allocated within an upcoming DL subframe. Although this feature is only natively available in RTS algorithm, it was implemented in the three allocation algorithms analyzed in this thesis, to allow a fair comparison among them. The adjusting factor (Λ) permits to vary the amount of data sent from the scheduler to the allocator, considering the capacity of the DL subframe. Therefore, Λ enables to improve the variability of data to be arranged by the allocation algorithm. In this context, the objective of the third analysis of this subsection is to investigate whether variations on Λ affect the efficiency of the allocation algorithm. The RTS algorithm was chosen for this analysis because it has the best data allocation performance among the algorithms considered in this thesis.

Figure 5.6 shows RTS efficiency considering three Λ values and the same proportion of VoIP, Video, and HTTP traffics. Λ equal to 0% means that the amount of data sent from the scheduler to the allocator corresponds to the capacity of the DL subframe, Λ equal to 25% and 50% indicates that the amount of data sent from the scheduler to the allocator is 25% and 50% higher than the capacity of the DL subframe, respectively. This investigation is important because Λ is used to permit a greater variability of bursts size to the RTS algorithm and to improve the allocation efficiency, as described in Section 4.3.2.



(a) Without mapping overhead

(b) With mapping overhead

Figure 5.6: Adjust factor performance

The variation of Λ did not represent any changes on RTS algorithm's performance in scenarios with less than 200 connections, as can be seen Figure 5.6. This behavior is explained because the amount of traffic generated in this situation is smaller than the capacity of the DL subframe, thus Λ has no effect. Otherwise, when there were more than 200 active connections, the scheduling queues received amounts of data that exceeded the DL subframe's capacity, thus changes in Λ value affected the performance of the RTS algorithm.

It is possible to observe in Figure 5.6 (a) that Λ = 0% had an efficiency of about 66%, while Λ = 25% presented an efficiency of approximately 72%, *i.e.* a gain in the order of 6% due to the higher variability on the size of the bursts sent to the RTS algorithm. However, Λ = 25% e 50% did not present a significant gain because the efficiency is not proportional to the increase on Λ value, since RTS algorithm has a limit of data that can be arranged within the DL subframe. Moreover, the arrangement capacity of the allocation algorithm is related with the traffic and physical characteristics. Another aspect observed in Figure 5.6 (b) is that the mapping overhead did not influence the efficiency of Λ, because the allocation algorithm is not able to distinguish data and mapping overhead.

To investigate deeply the adjust factor performance, another simulation scenario with VoIP predominant traffic and the mapping overhead was performed. The proportion of the traffic was of 70% VoIP, 15% Video, and 15% HTTP. This scenario was called of QoS profile since the predominant traffic demands QoS requirements. The voice-base predominant traffic was considered in this investigation because it is composed of small PDUs that can be better arranged by the allocation algorithm into OFDMA frame. This scenario is typically found in cellular systems. Figure 5.7 shows a gain of about 11% when the allocation algorithm received a greater variability of the data to be arranged within the DL subframe. Furthermore, it is possible to observe that the RTS algorithm used almost all the capacity of the DL subframe in the scenario with 800 connections, *i.e.* the RTS algorithm almost reached the ideal efficiency in the data allocation within OFDMA frame.

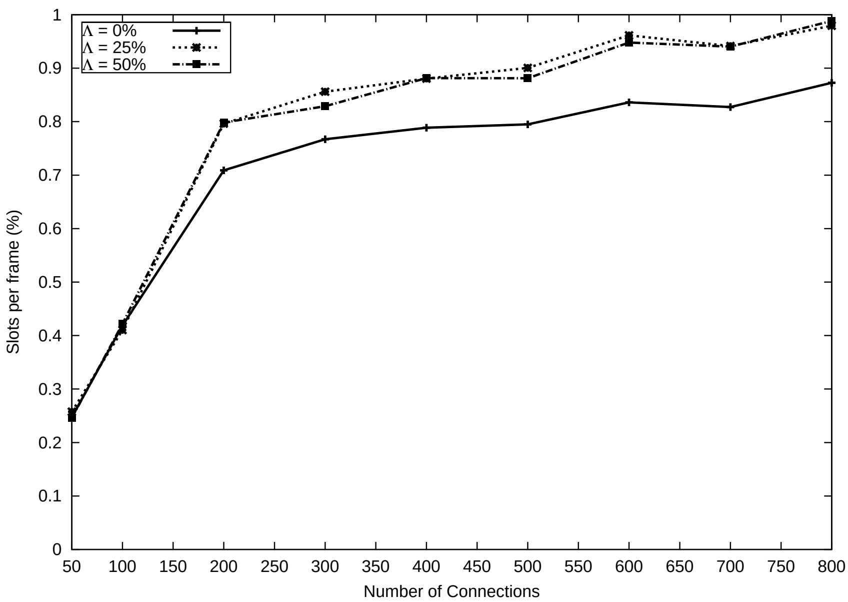
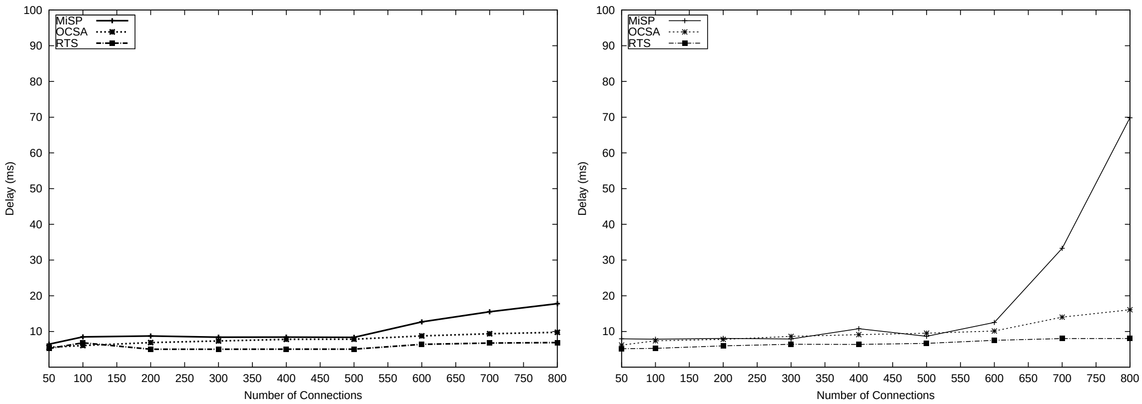


Figure 5.7: VoIP in the QoS profile

The fourth investigation of this subsection focuses on the delay caused by the new cross-layer adaptive architecture in VoIP transmissions, *i.e.* real time traffic. The simulation scenarios for this investigation were defined as VoIP, Video, and HTTP traffics in same proportion of connections. In this context, the PF scheduling algorithm was used to prioritize VoIP traffic in detriment to video and BE traffics. In PF, before serving the next queue, the algorithm assures that all PDUs which are currently in the highest priority queue are served. Afterwards, PF algorithm serves the next queue, considering the priority criteria. The calculation of the delay considers the frame duration, that in this simulation scenarios was equal to 5ms. Therefore, the delay of a PDU is related to the amount of frames transmitted while it is processed by the new cross-layer adaptive architecture.

Figure 5.8 shows that RTS algorithm normally allocated the VoIP PDUs in the first frame after the PDU arrived to the scheduling queue, *i.e.* the typical delay was 5ms. This behavior occurs in two analyzed scenarios, (a) with Λ equal to 0% and (b) Λ equal to 50%. The RTS algorithm is able to guarantee the 5ms delay even when the data sent from the scheduler to the allocator was 50% higher than the capability of the DL subframe. The OCSA algorithm had similar performance in the scenario with Λ factor equal to 0%. On the other hand, OCSA and MiSP algorithms tend to increase the delay after 600 connections in the scenario with Λ factor equal to 50% due to the amount of data allocated. This number of connections indicates a threshold until which the maximum delay efficiency was guaranteed. The delay behavior for the OCSA and MiSP is associated with the amount of data allocated by these algorithms, *i.e.* a small amount of allocated data into an OFDMA frame represented a higher number of frames to transmit the same data. Therefore, the mean of the delay increased in this case.



(a) Λ = 0% (b) Λ = 50%

Figure 5.8: Delay analysis for VoIP traffic

The fifth investigation of this subsection presents the delay of the other real-time class of service, *i.e.* video. This investigation presents the analysis only with Λ equal to 0%, because there was no a significant different in the performance for the video delay using Λ equal 25% and 50%. This behavior is justified due to (i) VoIP PDUs are served first than the video PDUs in the Inter-class scheduling, (ii) VoIP PDUs are smaller than the video PDUs, therefore the VoIP PDUs are easily arranged within OFDMA frame than the video PDUs, and (iii) the VoIP traffic is constant, thus there is a high amount of VoIP PDUs to be allocated.

Figure 5.9 shows that if compared with VoIP transmission, video traffic presented a higher delay, since it had lower priority, as presented in Table 5.3. Another analysis that can be made in Figure 5.9 regards to the constant tendency of the delay for video traffic after 500 connections. This behavior occurred because the maximum deadline of the rtPS queue was equal 100ms, *i.e.* the video PDUs which were not served in 100ms have been discarded by the new cross-layer adaptive architecture. Therefore, the video traffic tended to have a constant average delay when the maximum deadline of rtPS was reached.

The sixth and the last investigation of this subsection is about the efficiency of the scheduling algorithms considering three concomitant classes of traffic and using the RTS as allocation algorithm. Figure 5.10 (a) presents the performance of the PF algorithm. The graph shows that until 200 connections the amount of allocated video and HTTP traffics improved even though the PF scheduling algorithm prioritized VoIP traffic. This behavior occurred because the amount of traffic was smaller than the available resource, *i.e.* the capacity of DL subframe. It is important to describe that the available resources are associated with the efficiency of the allocation algorithm discussed in Figure 5.4. After 200 connections, the amount of the traffic was higher that the available resource associated with the efficiency of the allocation algorithm, therefore the PF algorithm scheduled the traffic with higher priority. In other words, the PF algorithm scheduled VoIP traffic instead video and HTTP traffics, thus the allocation of video and HTTP traffics was reduced. As a consequence of this behavior, after 750 connections the amount of VoIP traffic was higher than HTTP traffic.

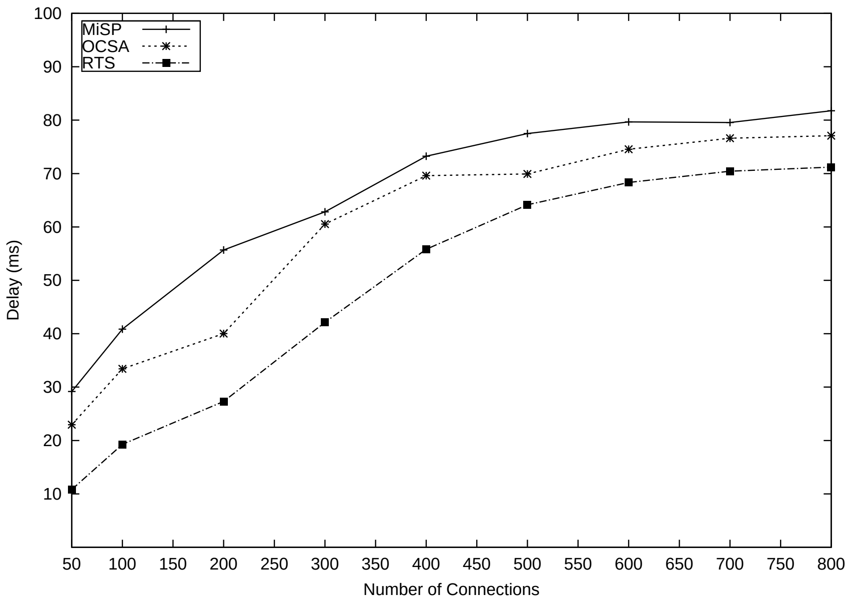
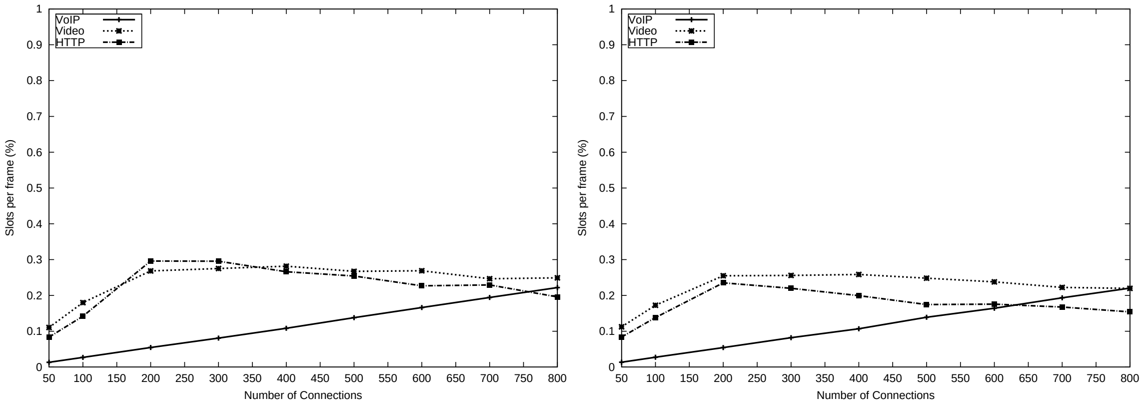


Figure 5.9: Delay analysis for video traffic

The shape of the curves obtained using DRR and PF algorithms is similar. The amount of VoIP traffic allocated in both algorithms was the same, as it can be seen in Figure 5.10 (b). This occurred because VoIP was the first traffic scheduled by DRR algorithm and the length of VoIP PDUs was smaller than the quantum size, which defined the maximum quantity of bytes that can be served for the queue in a given round. The quantum size used was of 4500 bytes in these scenarios, as suggested by Shreedhar and Varghese (SHREEDHAR; VARGHESE, 1996). On the other hand, the maximum allocation of HTTP traffic in DRR algorithm was approximately 10% smaller than using PF algorithm. This behavior happened because HTTP PDUs were typically bigger than the quantum size. Therefore, PF algorithm showed a better performance than DRR algorithm in this scenario, because the PF algorithm allocated a higher amount of the data than the DDR algorithm.

The algorithms of scheduler and allocator components are fundamental to the new cross-layer adaptive architecture to guarantee QoS in the different traffic types. However, these algorithms integrated in scheduler and allocator components alone are not able to provide strict QoS requirements. For example, a HTTP traffic may experience starvation because the available bandwidth can be allocated to traffics with higher priority than the

HTTP traffic. In this context, the CAC component must be also integrated in the new cross-layer adaptive architecture. Therefore, the CAC component results are presented in the next section.



(a) PF algorithm (b) DRR algorithm

Figure 5.10: Efficiency of the scheduling algorithms

5.2.3 CAC Component Analysis

The performance evaluation of the CAC component is organized in two topics. The first one shows how the Information Infrastructure of the CAC component handles the usage and the available network resources. This topic is called of the Accurate Information Monitoring. The second one presents the behavior of two CAC algorithms considering the accurate information received from the Information Infrastructure. This topic is called of the Performance Analysis of the CAC Algorithms. The simulation scenario for these investigations considered the PF algorithm in the Intra-class scheduling and the RTS algorithm in the allocation into OFDMA frame. These algorithms were chosen because presented the better performance as previously discussed.

Accurate Information Monitoring

The analysis of the accurate information monitoring considered a traffic composed of 15% VoIP, 15% Video, 70% HTTP. This proportion was defined to analyze the throughput of the network, since the HTTP traffic is composed of bigger PDUs than the VoIP and Video traffics. Moreover, the investigation considered a screen shot of 10s, *i.e.* 2000 OFDMA frames of 5ms, in a WiMAX network with 800 active connections. Therefore, it is possible to observe in Figure 5.11 three curves which represent the maximum, transmitted, and available throughput.

The maximum throughput was calculated considering 100% of the usage of the capacity of DL subframes and the less robust MCS was configured statically, *i.e* 64 QAM 5/6. In other words, the best RF channel condition was used without considering the variability of the wireless environment. In this context, the maximum throughput can be called of the theoretical throughput. In this condition, the maximum rate reached considering the physical configurations used in the simulation scenario was equal 29.03 Mbps, *e.g.* using a RF channel bandwidth of 10 MHz, TDD system with 5ms frame duration, and PUSC subchannelization mode. More details of the physical configurations can be seen in subsection 5.1.2.

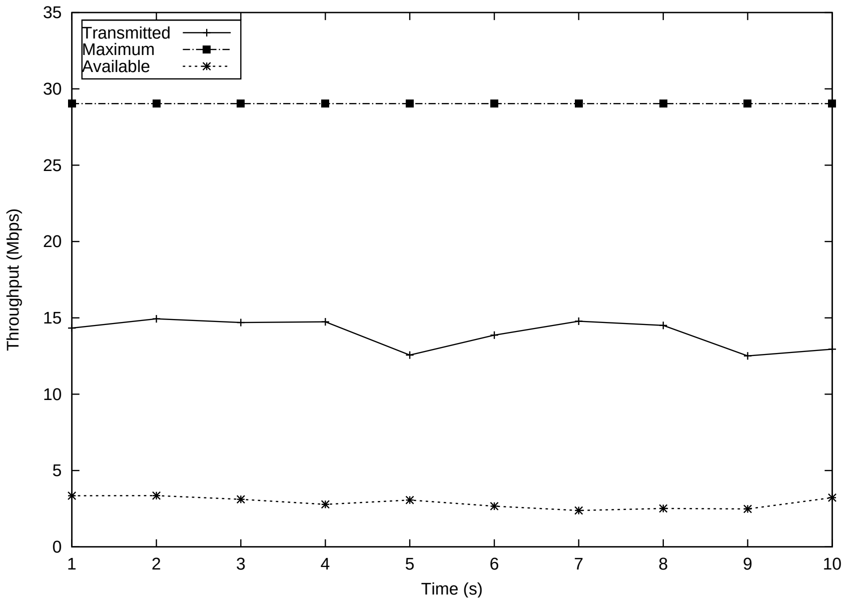


Figure 5.11: Throughput monitoring

Another analysis that can be performed through Figure 5.11 is the transmitted throughput, as defined in equation 4.13. This analysis refers to data allocated in the past, *i.e.* the historical throughput. The difference between the theoretical and transmitted throughput is very significant, about 48%, due to variability of both traffics and RF channel conditions. Therefore, this difference cannot be disregarded in the design of a QoS architecture.

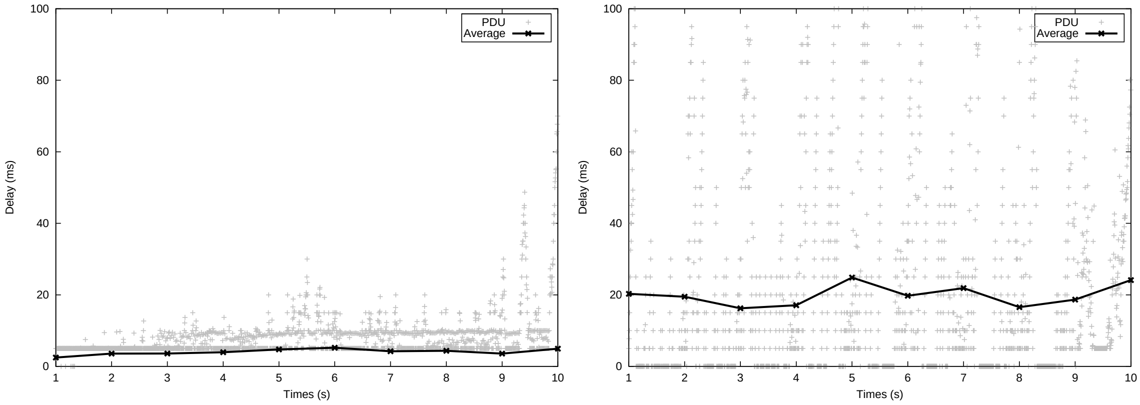
Taking into account the theoretical throughput previously described, it is possible to observe that this approach does not allow the CAC component to report an accurate information to the CAC algorithm. In the same manner, the information about the transmitted throughput must not be used individually in the decision process to accept connection requests, because it is necessary to estimate the MSCs that will be used in the transmission of the next OFDMA frame. In this context, the available throughput was defined in equation 4.16 and can be seen in Figure 5.11. The available throughput was of approximately 4 Mbps in the analyzed scenario. This behavior is related with the average MCS defined in equation 4.14. During the screen shot, the average MCS remained invariable in 16 QAM 3/4. This behavior occurred due to transition probabilities used in the Markov chain.

The second investigation about the accurate information monitoring is related with the delay. Figure 5.12 shows the delay of VoIP and Video traffics considering the same screen shot of 10s. The graphics of the figure have two information types: (i) the accurate delay of 2000 PDUs, (10s using frames of 5ms) and (ii) the average delay of PDUs during the screen shot.

The VoIP delay can be seen in Figure 5.12 (a). It is possible to analyze that the average delay was of approximately 5ms. This behavior means that in average VoIP PDUs were served in the first frame. Moreover, the variability of the delay is low, according to the VoIP traffic model discussed in subsection 5.1.3.

In Figure 5.12 (b), it is possible to observe the Video delay. In this traffic, the average delay was about four frames, *i.e.* 20ms. This occurred because of two distinct features.

The first refers to the Inter-class scheduling of the rtPS class of service. In other works, the Video traffic is served in lower priority than the VoIP traffic. The second regards to the traffic model, since the video has bigger PDUs than the VoIP. Therefore, the arrangement of Video PDUs is worse than of VoIP PDUs into OFDMA frame. Consequently, these two features together cause a high variability in the Video traffic delay.



(a) VoIP traffic (b) Video traffic

Figure 5.12: Delay monitoring

The last analysis regarding the accurate information monitoring refers to the jitter. In Figure 5.13 can be also observed the jitter of VoIP and Video traffics with the same screen shot of 10s. In this simulation scenario the VoIP jitter was practically null, while the Video traffic presented a jitter of approximately 2ms. The behavior of jitter is negligible in this simulation scenario due to the new cross-layer adaptive architecture to be centralized in BS, *i.e.* there is only a clock source. Indeed, the jitter is a characteristic by the undesired deviation of a periodic signal related with different clock sources (NUAYMI, 2007).

This topic investigated the accurate information monitoring that is performed in the new cross-layer adaptive architecture by the Information Infrastructure of the CAC component. This information is sent to the CAC algorithm to decide whether to accept or not new connection requests. The performance analysis of the CAC algorithms is discussed in the next topic.

Performance Analysis of the CAC Algorithms

In this topic, it is presented an evaluation of the impacts caused in the amount of connection requests accepted by two CAC algorithms, *i.e.* BW_CAC and QoS_CAC. The CAC algorithms were investigated considering two different traffic profiles. The first profile had a proportional traffic composed of 15% VoIP, 15% Video, 70% HTTP. This scenario was called of Web profile because the predominant traffic has no QoS requirements. The second scenario was the QoS profile, described previously. Moreover, the configurations of scheduler and allocator components were the same defined in this subsection to evaluate the CAC component. Furthermore, the simulations accommodated up to 800 connection requests.

In Figure 5.14 it is possible to analyze the performance of BW_CAC and QoS_CAC considering the Web profile (a) and QoS profile (b). BW_CAC algorithm accepted all connection requests in both simulation scenarios. Accepting all connection occurred because the BW_CAC algorithm did not guarantee QoS for the accepted connections, since delay and jitter parameters were not taken into account. Another important aspect to be observed in this context is the amount of rejected connection requests when the QoS_CAC algorithm is running. The rejection was of approximately 5% in the Web profile and reached more than 30% when in the QoS profile.

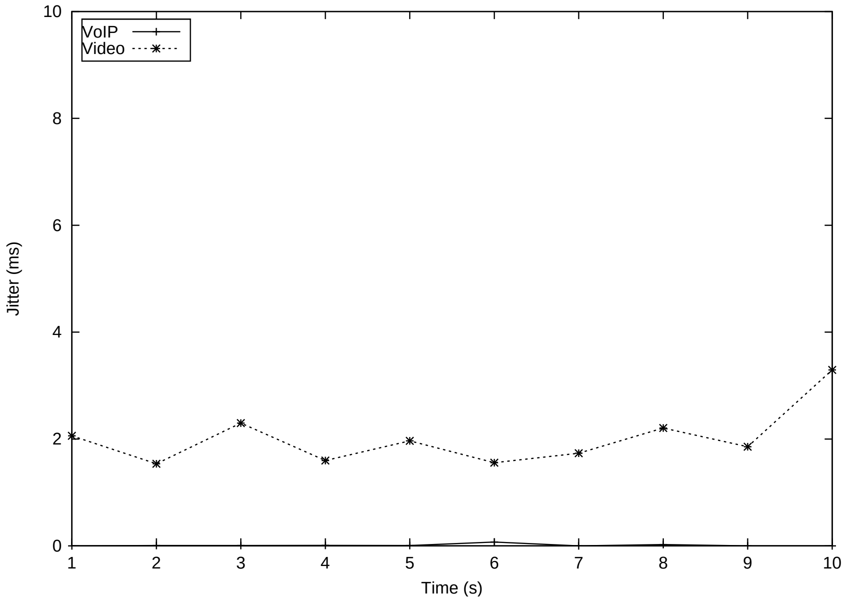


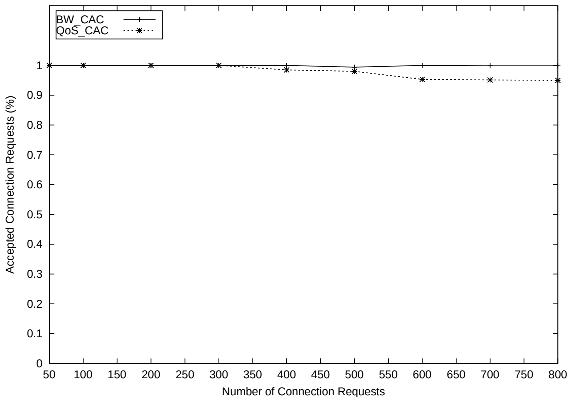
Figure 5.13: Jitter monitoring

Analyzing Figure 5.14, up to 100 connection requests the difference of the traffic profiles did not affect the acceptance of the CAC algorithms. Above that, however, the amount of accepted requests was reduced in the case of the QoS_CAC algorithm because of the large amount of real time connection requests, *i.e.*, 85% of the requests required QoS guarantees (70% VoIP plus 15% Video).

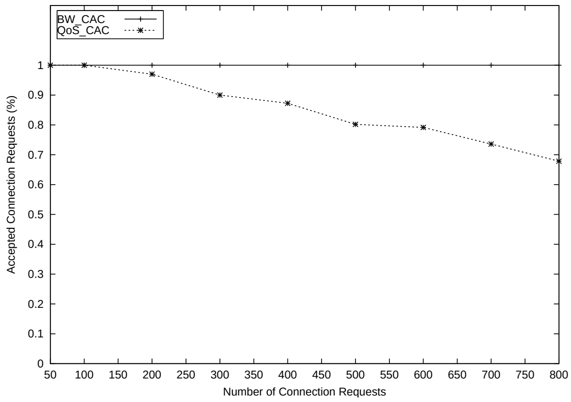
It is possible to conclude that when the predominant traffic was Web, there was not a major need to provide QoS guarantees. Therefore, the QoS aware algorithm should not be used because it reduced the amount of accepted connection requests in about 5%. Nevertheless, when real time traffic was predominant, then the CAC algorithm should be switched to another one able to guarantee QoS for the accepted connections. The proposal to switch the CAC algorithm according with the traffic profile is not inside of the scope this thesis. However, this research work was investigated and can be analyzed in Appendix C.

5.3 Summary

This chapter presented the simulation design and the performance evaluation of the new cross-layer adaptive architecture. First, the simulator model and the simulation scenario were discussed. This scenario was defined considering the physical parameters based on OFDMA, to simulate mobile WiMAX networks. Next, the traffic models for the VoIP, video, HTTP traffics, as well as the connection request models were described. The obtained results were focused on analyzing three aspects: (i) the scheduling factors, (ii) the integration between the scheduler and allocator components, and (iii) the CAC component. Analysis were presented, considering important aspects as, efficiency of scheduling and allocation algorithms, mapping overhead caused by OFDMA frame features, adjusting factor performance, the delay and jitter caused by the new cross-layer adaptive architecture in the real-time traffics, and the evaluation of CAC algorithms in two different traffic profiles.



(a) Web profile



(b) QoS profile

Figure 5.14: Evaluation of CAC algorithms

6 CONCLUSION AND FUTURE WORK

The goals of this chapter are to present the conclusion and future work of this thesis. First, answers for the fundamental questions are presented according to the hypothesis defended. After, the main contributions of this thesis are discussed. Finally, indications on future work related to the results are discussed.

6.1 Answer for the Fundamental Questions

The investigation performed on this thesis showed the study, design, and evaluation of the new cross-layer adaptive architecture to guarantee QoS in WiMAX networks. To drive the directions of this investigation, a hypothesis *"The architecture to guarantee QoS in WiMAX BSs must be modeled using a cross-layer infrastructure able to adapt to the dynamics of traffic requirements as well as to the RF channel conditions"* was defined. The methodology of the work was established considering the resources that the individual QoS components must provide to the new cross-layer adaptive architecture and in the QoS requirements that this architecture must guarantee to the users of IEEE 802.16 networks. Thus, the studies were based on the papers, articles and technical reports published in the literature, in the evaluation methodology published by the WiMAX Forum (WiMAX Forum, 2008), and in the IEEE 802.16 standard (IEEE, 2009).

First, a detailed study on the IEEE 802.16 standard based on OFDMA multiplexing technique was performed. After, the allocator and scheduler components were studied individually, and integrated, as well as the results analyzed using a simulation method. From this preliminary studies, it was possible to understand that the radio resource management should consider the PDUs and BRs of each OFDMA frame, according to the PHY and MAC features of IEEE 802.16 networks, *i.e.* connection-oriented MAC layer, classes of services, frame capacity, arrangement of the data allocated, and mapping overhead. This understanding is different from most of the works proposed in the literature that consider estimated resources, as bandwidth and transmission power. This approach of the literature may lead to overestimation of available network resources. According to the available resources informed by the allocator and scheduler components after integrated, it was possible to design an information infrastructure to the CAC component. This infrastructure is fundamental for the new cross-layer adaptive architecture, because the connection requests and user's applications must be served considering the overall network resources.

To the best of our knowledge, no QoS architecture aware of the OFDMA frame structure, as defined in Section 3.2, and considering the diversity of both traffics and the RF channel conditions has been proposed in the literature. In this context, the performance evaluation of the new cross-layer adaptive architecture with other architectures is not fair and useful. Moreover, single layer architectures were not analyzed because this approach do not regard the diversity previously describe.

The intention behind the definition of the fundamental questions was to help on raising the main points to be analyzed during the investigation of the hypothesis and to establish the way to reach the contributions of this thesis. Therefore, the description below summarizes and highlights the major characteristics of the answers of each fundamental question.

Funtamental question I. *How to provide accurate information to the new cross-layer adaptive architecture?*

Answer. This thesis indicates that two elements are needed to provide accurate information to the new cross-layer adaptive architecture: (i) a Joint Information structure assigned to each PDU and BR that considers the dynamics of network traffic requirements and the RF channel conditions, as well as (ii) an Information Infrastructure used to hold data regarding to the available resources and QoS components.

Fundamental question II. *How to manage the available resources considering the dynamics of network traffic requirements and RF channel conditions?*

Answer. This thesis points that the management of available resources can be done since the QoS components are able to operate in a new cross-layer adaptive architecture using accurate information as their basis. In this sense, the scheduler component here proposed is modeled using scheduler factors that are derived from the Joint information Infrastructure. These scheduler factors are responsible for prioritizing the PDUs and BRs within the scheduling queues. In order to complement the proposed architecture, the CAC component uses the Information Infrastructure to retrieve accurate information provided also by the other QoS components. This way, the CAC component is able to take proper decisions.

Fundamental question III. *What is the impact of integrating adaptive QoS components in WiMAX networks?*

Answer. The main impact caused by the adaptive integration of QoS components is to provide accurate information to algorithms. This information refers to traffic requirements, channel conditions, and QoS components feedbacks that are fundamental to the new cross-layer adaptive architecture do not lead to overestimation of available network resources and guarantees QoS for the user's applications. An example presented in the performance evaluation of this thesis is the relation among the theoretical, transmitted, and available throughput that must be considered by CAC algorithm.

After, answering the fundamental questions that helped on the investigation of the hypothesis of this thesis, the main contributions obtained in this work are highlighted in the next section.

6.2 Main Contributions Obtained

The main contributions of this thesis can be divided into: conceptual and specific ones. Conceptual contribution could be identified due to the investigations of the literature and the experiences gathered during the development this work. In contrast, specific contributions are associated with individual solutions developed for the new cross-layer adaptive architecture. In the following, both contributions are listed.

- Conceptual contribution:
  - Classification of the cross-layer architecture in IEEE 802.16 networks published in the literature in three perspectives according to (i) the usage of the bandwidth, (ii) multiplexing technique, and (iii) frame structure.
- Specific contributions:
  - Specification of the integration among the three main QoS components: allocator, scheduler, and CAC.
  - Definition of the information infrastructure that enables the management of applications traffic requirements, time variant propagation conditions of the RF channel, and QoS components feedbacks.
  - Specification of the CAC component for controlling the admission of connection requests.
  - Identification of the methodology to evaluate a new cross-layer adaptive architecture based on QoS for WiMAX networks.

The contributions previously listed are not exhaustive. Other results obtained in this thesis are the simulation tool and the knowhow of its developed to evaluate the proposed architecture. Using this tool allows to investigate future work on the research topic of this thesis, which are discussed in the next section.

6.3 Future Work

The investigation performed on this thesis leads to the identification of further opportunities of research. These opportunities are listed as follows.

- Propose scheduling policies to the new cross-layer adaptive architecture that prioritize a connection in detriment to others. The IEEE 802.16 standard defines policies of scheduling based on classes of services. However, policies of scheduling that prioritize connections can be an interesting approach for Internet Service Provides (ISPs), *e.g.* classify connections using criteria as diamond, gold, and silver to guarantee different levels of QoS.
- Investigate essential network usage profiles that must be supported by a BS, and the identification of the most suitable physical architecture for the deployment of the self-adapting CAC solution. Furthermore, it is necessary to better understand the thresholds between historical versus current weights in face of very dynamic

and constant changes in the predominant network usage profiles in order to define situations where historical or current weights should be considered for switching the CAC algorithm.

- Design a resource reservation module in the CAC component to support handover process in multi-cell networks. The goal of this future work may be to analyze the behavior of the new cross-layer adaptive architecture in a mesh networks. A good research opportunity in multi-cell scenarios is to guarantee QoS during the handover process.
- Analyze the adaptation of contributions proposed in this thesis to other next generation wireless technologies. For example, the Long Term Evolution (LTE) is a technology similar to IEEE 802.16, since it also uses OFDMA as physical interface, adaptive MCS to adjustment process, and connection-oriented MAC layer. In this context, it is possible to investigate the adaptation of individual solutions proposed in this thesis to LTE networks.

The list of future work discussed previously highlights the major opportunities of research that can be directly derived from the work presented in this thesis. It is important describe that these opportunities can be listed because of the knowhow acquired during the Ph.D program.

REFERENCES

AHMADI, S. An Overview of Next-Generation Mobile WiMAX Technology. **IEEE Communications Magazine**, Piscataway, USA, v.47, n.6, p.84–98, June 2009.

ALI, N. A.; DHRONA, P.; HASSANEIN, H. A performance study of uplink scheduling algorithms in point-to-multipoint WiMAX networks. **Computer Communications**, New York, v.32, n.3, p.511–521, February 2009.

BACIOCCOLA, A. et al. A Downlink Data Region Allocation Algorithm for IEEE 802.16e OFDMA. In: IEEE ICICS - INTERNATIONAL CONFERENCE INFORMATION, COMMUNICATIONS & SIGNAL PROCESSING, Singapore. **Proceedings...**, 2007. p.1–5.

BACIOCCOLA, A. et al. IEEE 802.16: history, status and future trends. **Computer Communications**, New York, v.33, n.2, p.113–123, February 2010.

BEN-SHIMOL, Y.; KITROSER, I.; DINITZ, Y. Two-Dimensional Mapping for Wireless OFDMA Systems. **IEEE Transactions on Broadcasting**, Ottawa, v.52, n.3, p.388–396, September 2006.

BOTH, C. B. et al. Analysis of WiMAX bandwidth allocation mechanism considering physical conditions. In: IEEE ISWCS - INTERNATIONAL SYMPOSIUM ON WIRELESS COMMUNICATION SYSTEMS, Reykjavik. **Proceedings...**, 2008. p.643–647.

CHANDRA, S.; SAHOO, A. An Efficient Call Admission Control for IEEE 802.16 Networks. In: IEEE LANMAN - WORKSHOP ON LOCAL & METROPOLITAN AREA NETWORKS, Princeton, USA. **Proceedings...**, 2007. p.188–193.

CHEN, J.; JIAO, W.; WANG, H. A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode. In: IEEE ICC - INTERNATIONAL COMMUNICATIONS CONFERENCE, Seoul, Korea. **Proceedings...**, 2005. p.3422–3426.

CHIEOCHAN, S.; HOSSAIN, E. Adaptive radio resource allocation in OFDMA systems: a survey of the state-of-the-art approaches. **Wireless Communications and Mobile Computing**, New Jersey, v.9, n.4, p.513–527, October 2008.

CHOI, K. W.; JEON, W. S.; JEONG, D. G. Resource Allocation in OFDMA Wireless Communications Systems Supporting Multimedia Services. **IEEE/ACM Transactions on Networking**, Philadelphia, USA, v.17, n.3, p.926–935, June 2009.

CHOU, S.-F. et al. Performance enhancement of contention-based bandwidth request mechanism in IEEE 802.16 WiMAX networks. In: IEEE PIMRC - PERSONAL INDOOR AND MOBILE RADIO COMMUNICATIONS, Turkey. **Proceedings...**, 2010. p.1287–1292.

CHUCK, D.; CHEN, K.-Y.; CHANG, J. M. A Comprehensive Analysis of Bandwidth Request Mechanisms in IEEE 802.16 Networks. **IEEE/ACM Transactions on Vehicular Technology**, Waterloo, Canada, v.59, n.4, p.2046–2056, May 2010.

CICCONETTI, C. et al. Quality of service support in IEEE 80216 networks. **IEEE Network**, Edmonton, Canada, v.20, n.2, p.50–55, April 2006.

CICCONETTI, C. et al. Efficient Two-dimensional Data Allocation in IEEE 802.16 OFDMA. In: IEEE INFOCOM - CONFERENCE ON COMPUTER COMMUNICATIONS, San Diego. **Proceedings...**, 2010. p.1–9.

COHEN, R.; KATZIR, L. Computational Analysis and Efficient Algorithms for Micro and Macro OFDMA Scheduling. In: IEEE INFOCOM - CONFERENCE ON COMPUTER COMMUNICATIONS, Phoenix, AZ. **Proceedings...**, 2008. p.1184–1192.

COHEN, R.; KATZIR, L. Computational Analysis and Efficient Algorithms for Micro and Macro OFDMA Downlink Scheduling. **IEEE/ACM Transactions on Networking**, Piscataway, USA, v.18, n.1, p.15–26, February 2010.

DUA, A.; BAMBOS, N. Downlink Wireless Packet Scheduling with Deadlines. **IEEE/ACM Transactions on Mobile Computing**, Piscataway, USA, v.6, n.12, p.1410–1425, December 2007.

ELSAYED, K.; KHATTAB, A. Channel-Aware Earliest Deadline Due Fair Scheduling for Wireless Multimedia Networks. **International Journal Wireless Personal Communications**, Hingham, USA, v.38, n.2, p.233–252, July 2006.

ESTEPA, A.; ESTEPA, R.; VOZMEDIANO, J. A new approach for VoIP traffic characterization. **IEEE Communications Letters**, Santa Clara, USA, v.8, n.10, p.644–646, October 2004.

ETEMAD, K.; WANG, L. Multicast and Broadcast Multimedia Services in Mobile WiMAX Networks. **IEEE Communications Magazine**, Piscataway, USA, v.47, n.10, p.84–91, October 2009.

FERREIRA, A.; CARVALHO, P.; LIMA, S. R. Adaptive Admission Control in a NGN Service Platform. In: IEEE WTS - WIRELESS TELECOMMUNICATIONS SYMPOSIUM, USA. **Proceedings...**, 2010. p.1–6.

GE, Y.; KUO, G.-S. An Efficient Admission Control Scheme for Adaptive Multimedia Services in IEEE 802.16e Networks. In: IEEE VCT - VEHICULAR TECHNOLOGY CONFERENCE, Canada. **Proceedings...**, 2006. p.1–5.

GHAZAL, S.; BEN-OTHMAN, J. Traffic Policing Based on Token Bucket Mechanism for WiMAX Networks. In: IEEE ICC - INTERNATIONAL CONFERENCE COMMUNICATIONS, South Africa. **Proceedings...**, 2010. p.1–6.

HOU, F. et al. Performance Analysis of Weighted Proportional Fairness Scheduling in IEEE 802.16 Networks. In: IEEE ICC - INTERNATIONAL CONFERENCE ON COMMUNICATIONS, Beijing. **Proceedings...**, 2008. p.3452–3456.

HOU, F. et al. A flexible resource allocation and scheduling framework for non-real-time polling service in IEEE 802.16 networks. **IEEE Transactions on Wireless Communications**, Rolla, USA, v.8, n.2, p.766–775, February 2009.

IEEE. **IEEE Standard for Local and Metropolitan Area Networks, Part 16 - Air Interface for Broadband Wireless Access Systems**. IEEE 802.16-2009. Disponível em: <`http://standards.ieee.org/getieee802/download/802.16-2009.pdf`>. Accessed: in March 2010.

ISRAELI, A.; RAWITZ, D.; SHARON, O. On the complexity of sequential rectangle placement in IEEE 802.16/WiMAX systems. In: JOURNAL INFORMATION AND COMPUTATION, Duluth, USA. **Proceedings...**, Academic Press, 2008. p.1334–1345. (Lecture Notes in Computer Science, v.206).

JIANG, C.-H.; TSAI, T.-C. Token bucket based CAC and packet scheduling for IEEE 802.16 broadband wireless access networks. In: IEEE CCNC - CONSUMER COMMUNICATIONS AND NETWORKING CONFERENCE, Las Vegas, USA. **Proceedings...**, 2006. p.183–187.

JUYEOP, K.; DONG-HO, C. Simultaneous transmission of MAP IE and data for minimizing MAC overhead in the IEEE 802.16e OFDMA systems. **IEEE Transactions on Wireless Communications**, Philadelphia, USA, v.8, n.11, p.5431–5435, November 2009.

KIM, W. Mobile WiMax, the leader of the mobile Internet era. **IEEE Communications Magazine**, Piscataway, USA, v.47, n.6, p.10–12, June 2009.

KO, S. et al. Mode Selection-Based Channel Feedback Reduction Schemes for Opportunistic Scheduling in OFDMA Systems. **IEEE Transactions on Wireless Communications**, Rolla, USA, v.9, n.9, p.2842–2852, September 2010.

KUNST, R. et al. On the Impact of Hybrid Errors on Mobile WiMAX Networks. **Computer Networks**, 2011.

KWON, E. et al. Markov Model for Admission Control in the Wireless AMC Networks. **IEICE Transaction Communication**, Tokyo, v.E89-B, n.8, p.2230–2233, September 2006.

LAKKAKORPI, J.; SAYENKO, A.; MOILANEN, J. Comparison of Different Scheduling Algorithms for WiMAX Base Station: deficit round-robin vs. proportional fair vs. weighted deficit round-robin. In: IEEE WCNC) WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE, Las Vegas, USA. **Proceedings...**, 2008. p.1991–1996.

LEE, H.; CHO, D. Reliable Multicast Services Using CDMA codes in IEEE 802.16 OFDMA system. In: IEEE VCT - VEHICULAR TECHNOLOGY CONFERENCE, Stockholm. **Proceedings...**, 2005. p.2349–2354.

LERA, A.; MOLINARO, A.; PIZZI, S. Channel-Aware Scheduling for QoS and Fairness Provisioning in IEEE 802.16/WiMAX Broadband Wireless Access Systems. **IEEE Network**, Edmonton, Canada, v.21, n.5, p.34–41, September 2007.

LI, Q. et al. Advancement of MIMO Technology in WiMAX: from IEEE 802.16d/e/j to 802.16m. **IEEE Communications Magazine**, Piscataway, USA, v.47, n.6, p.84–98, June 2009.

LIANG, Z.; CHEW, Y. H.; KO, C. C. A Linear Programming Solution to Subcarrier, Bit and Power Allocation for Multicell OFDMA Systems. In: IEEE WCNC - WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE, Las Vegas, USA. **Proceedings...**, 2008. p.1273–1278.

LU, J.; MA, M. A cross-layer elastic CAC and holistic opportunistic scheduling for QoS support in WiMAX. **The International Journal of Computer and Telecommunications Networking**, Maryland, USA, v.54, n.7, p.1155–1168, 2010.

MSADAA, I. C.; C?MARA, D.; FILALI, F. Scheduling and CAC in IEEE 802.16 Fixed BWNs: a comprehensive survey and taxonomy. **IEEE Communications Surveys & Tutorials**, Campinas, Brazil, v.12, n.4, p.459–487, 2010.

NI, Q. et al. Performance Analysis of Contention Based Bandwidth Request Mechanisms in WiMAX Networks. **IEEE Systems Journal**, San Antonio, USA, v.4, n.4, p.477–487, 2010.

NIYATO, D.; HOSSAIN, E. Queue-aware uplink bandwidth allocation for polling services in 802.16 broadband wireless networks. In: IEEE GLOBECOM - GLOBAL TELECOMMUNICATIONS CONFERENCE, St. Louis, MO. **Proceedings...**, 2005. p.3702–3706.

NIYATO, D.; HOSSAIN, E. Radio resource management games in wireless networks: an approach to bandwidth allocation and admission control for polling service in IEEE 802.16. **IEEE Wireless Communications**, Piscataway, USA, v.14, n.1, p.27–35, 2007.

WILEY, J.; SONS (Ed.). **WiMAX**: technology for broadband wireless access. Chichester, UK: John Wiley and Sons, 2007.

OHSEKI, T.; MORITA, M.; INOUE, T. Burst Construction and Packet Mapping Scheme for OFDMA Downlinks in IEEE 802.16 Systems. In: IEEE GLOBECOM - GLOBAL TELECOMMUNICATIONS CONFERENCE, Washington, USA. **Proceedings...**, 2007. p.4307–4311.

PAPAPANAGIOTOU, I. et al. A survey on next generation mobile WiMAX networks: objectives, features and technical challenges. **IEEE Communications Surveys & Tutorials**, Campinas, Brazil, v.11, n.4, p.3–18, December 2009.

PARK, W.-H.; CHO, S.; BAHK, S. Scheduler Design for Multiple Traffic Classes in OFDMA Networks. In: IEEE ICC - INTERNATIONAL COMMUNICATIONS CONFERENCE, Istanbul, Turkey. **Proceedings...**, 2006. p.790–795.

RONG, B. et al. Call Admission Control Optimization in WiMAX Networks. **IEEE Transactions on Vehicular Technology**, Piscataway, USA, v.7, n.4, p.2509–2522, July 2008.

RONG, B.; QIAN, Y.; CHEN, H.-H. Adaptive power allocation and call admission control in multiservice WiMAX access networks. **IEEE Wireless Communications**, Piscataway, USA, v.14, n.1, p.14–19, February 2007.

RONG, B.; QIAN, Y.; CHEN, H.-H. Integrated Downlink Resource Management for Multiservice WiMAX Networks. **IEEE Transactions on Mobile Computing**, Los Alamitos, USA, v.6, n.6, p.621–632, June 2007.

RUANGCHAIJATUPON, N.; JI, Y. Simple Proportional Fairness Scheduling for OFDMA Frame-Based Wireless Systems. In: IEEE WCNC - WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE, Las Vegas, USA. **Proceedings...**, 2008. p.1593–1597.

SAYENKO, A.; ALANEN, O.; HAMALAINEN, T. Adaptive Contention Resolution for VoIP Services in the IEEE 802.16 Networks. In: IEEE WoWMoM - INTERNATIONAL SYMPOSIUM ON A WORLD OF WIRELESS, MOBILE AND MULTIMEDIA NETWORKS, Espoo, Finland. **Proceedings...**, 2007. p.1–7.

SAYENKO, A.; ALANEN, O.; HAMALAINEN, T. Scheduling solution for the IEEE 802.16 base station. **Computer Network**, Atlanta, USA, v.52, n.1, p.96–115, January 2008.

SAYENKO, A. et al. Ensuring the QoS requirements in 802.16 scheduling. In: MSWiM - 9TH ACM INTERNATIONAL SYMPOSIUM ON MODELING ANALYSIS AND SIMULATION OF WIRELESS AND MOBILE SYSTEMS, Torremolinos, Spain. **Proceedings...**, 2006. p.108–118.

SEKERCIOGLU, A.; IVANOVICH, M.; YEGIN, A. A survey of MAC based QoS implementations for WiMAX networks. **Computer Communications**, Maryland, USA, v.53, n.14, p.2517–2536, September 2009.

SHE, J. et al. A cross-layer design framework for robust IPTV services over IEEE 802.16 networks. **IEEE Journal on Selected Areas in Communications**, San Diego, USA, v.27, n.2, p.235–245, February 2009.

SHREEDHAR, M.; VARGHESE, G. Efficient fair queuing using deficit round-robin. **ACM Transactions on Networking**, New York, NY, USA, v.4, n.1, p.375–385, June 1996.

SINGH, V.; SHARMA, V. Efficient and fair scheduling of uplink and downlink in IEEE 802.16 OFDMA networks. In: IEEE WCNC - WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE, Las Vegas, USA. **Proceedings...**, 2006. p.984–990.

SO-IN, C.; JAIN, R.; TAMIMI, A.-K. A. OCSA: an algorithm for burst mapping in IEEE 802.16e mobile WiMAX networks. In: APCC) 15TH ASIA-PACIFIC CONFERENCE ON COMMUNICATIONS, Shanghai. **Proceedings...**, 2009. p.52–58.

SO-IN, C.; JAIN, R.; TAMIMI, A.-K. Capacity Evaluation for IEEE 802.16e Mobile WiMAX. **Journal of Computer Systems, Networks, and Communications**, New York, v.2010, n.1, p.1–12, December 2009.

SO-IN, C.; JAIN, R.; TAMIMI, A.-K. Scheduling in IEEE 802.16e Mobile WiMAX Networks: key issues and a survey. **IEEE Journal on Selected Areas in Communications**, Piscataway, USA, v.27, n.2, p.156–171, February 2009.

SYDIR, J.; TAORI, R. An Evolved Cellular System Architecture Incorporating Relay Stations. **IEEE Communications Magazine**, Piscataway, USA, v.47, n.6, p.84–98, June 2009.

TRIANTAFYLLOPOULOU, D. et al. A Heuristic Cross-Layer Mechanism for Real-Time Traffic in IEEE 802.16 Networks. In: IEEE PIMRC - PERSONAL, INDOOR AND MOBILE RADIO COMMUNICATIONS, Athens, Greece. **Proceedings...**, 2007. p.1–6.

VENKATACHALAM, M. et al. Location Services in WiMAX Networks. **IEEE Communications Magazine**, Piscataway, USA, v.47, n.10, p.92–98, October 2009.

WANG, F. et al. Mobile WiMAX systems: performance and evolution. **IEEE Communications Magazine**, One Piscataway, USA, v.46, n.10, p.41–49, October 2008.

WANG, H.; LI, W.; AGRAWAL, D. Dynamic admission control and QoS for 802.16 wireless MAN. In: IEEE WTS - SYMPOSIUM WIRELESS TELECOMMUNICATIONS, California, USA. **Proceedings...**, 2005. p.60–66.

WANG, L. et al. Admission Control for Non-preprovisioned Service Flow in Wireless Metropolitan Area Networks. In: ECUMN - EUROPEAN CONFERENCE ON UNIVERSAL MULTISERVICE NETWORKS, France, Toulouse. **Proceedings...**, 2007. p.243–250.

WiMAX Forum. **WiMAX System Evaluation Methodology version 2.1**. WiMAX Forum. Disponível em: <`http://www.wimaxforum.org/documents`>. Accessed in: March 2010.

XIEI, B.; ZHOU, W.; ZENGI, J. A Novel Cross-Layer Design with QoS Guarantee for WiMAX System. In: ICPCA) INTERNATIONAL CONFERENCE ON PERVASIVE COMPUTING AND APPLICATIONS, Alexandria, Egypt. **Proceedings...**, 2008. p.835–840.

XU, C.; LIANG, Y.-C.; LEON, W. S. Mobile WiMAX systems: performance and evolution. **IEEE Transactions on Wireless Communications**, Rolla, USA, v.7, n.1, p.43–47, January 2008.

YAHIYA, T. A.; BEYLOT, A.-L.; PUJOLLE, G. Cross-Layer Multiservice Scheduling for Mobile WiMAX Systems. In: IEEE WCNC) WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE, Las Vegas, USA. **Proceedings...**, 2008. p.1531–1535.

YANG, O.; LU, J. Admission Control and Scheduling Schemes with QoS Support for Real-time Video Applications. **Journal of Multimedia**, Rochester, USA, v.1, n.2, p.21–29, May 2006.

YANG, Y. et al. Relay Technologies for WiMAX and LTE-Advanced Mobile Systems. **IEEE Communications Magazine**, Piscataway, USA, v.47, n.10, p.100–105, October 2009.

YOU, J.; KIM, K.; KIM, K. Capacity evaluation of the OFDMA-CDMA ranging subsystem in IEEE 802.16-2004. In: IEEE WIMOB) INTERNATIONAL CONFERENCE ON WIRELESS AND MOBILE COMPUTING, NETWORKING AND COMMUNICATIONS, Montreal, Canada. **Proceedings...**, 2005. p.217–222.

ZHU, B. et al. Fair Connection Admission Control Scheme for IEEE 802.16e Systems. In: IEEE WICOM) INTERNATIONAL CONFERENCE ON WIRELESS COMMUNICATIONS, NETWORKING AND MOBILE COMPUTING, Dalian, China. **Proceedings...**, 2008. p.1–4.

APPENDIX A PAPER TO THE ISWCS 2008

In this appendix the paper entitled “Analysis of WiMAX bandwidth allocation mechanism considering physical conditions” is presented. This paper was published in the ISWCS conference and was the first paper regarding to the topics of this thesis. The paper presented an initial analysis of the impact caused by overhead on the bandwidth allocation mechanism of WiMAX networks. The main focus of this paper was to evaluate the number of MSs that can be served, and the influence of variations on MCS configuration. Moreover, it was discussed the importance of considering physical impairments in the process of designing a broadband wireless access network.

- **Title:**
  Analysis of WiMAX bandwidth allocation mechanism considering physical conditions
- **Conference:**
  5th IEEE International Symposium on Wireless Communication Systems (ISWCS 2008)
- **URL:**
  `http://www.iswcs.org/iswcs2008/`
- **Date:**
  October 21-24, 2008
- **Venue:**
  Radisson SAS Saga Hotel, Hagatorgi, 107 Reykjavik, Iceland
- **Digital Object Identifier (DOI):**
  `http://dx.doi.org/10.1109/ISWCS.2008.4726135`

Analysis of WiMAX bandwidth allocation mechanism considering physical conditions

[1,2]Cristiano Bonato Both, [1]Eliane Gobetti de Camargo, [1,3]Rafael Kunst,
[1]Lisandro Zambenedetti Granville, [1]Juergen Rochol
[1]Federal University of Rio Grande do Sul (UFRGS)
[2]University of Santa Cruz do Sul (UNISC)
[3]La Salle University (Unilasalle)
Email: {cbboth, egcamargo, rkunst, granville, juergen}@inf.ufrgs.br

Abstract— **The increasing development of mobile applications has changed the traditional demands of computer networks, which must now provide ubiquitous broadband wireless communication. In this context, IEEE 802.16 standardizes the physical and medium access control layers of the technology, aiming to allow network access in both nomadic and mobile scenarios. In this work we analyze the impact caused by overhead on the bandwidth allocation mechanism of WiMAX networks. Our main focus is to evaluate the number of SSs that can be served, and the influence of adaptive modulation and coding configuration. Finally, we discuss the importance of considering physical impairments in the process of designing a broadband wireless access network.**

I. Introduction

The advent of mobile devices, such as Personal Digital Assistants (PDAs), laptops, and next generation mobile phones, has changed the access profile of computer networks, creating the conditions for the development of ubiquitous applications that need to offer information anytime and anywhere. In this context, Broadband Wireless Access (BWA) networks, such as IEEE 802.16 [1], also known as WiMAX, are refereed as emerging Next-Generation Networks (NGN) able to be used in both nomadic (fixed) and mobile scenarios.

WiMAX networks are susceptible to physical impairments that compromise the resources available for data transmission, leading to performance problems and low availability of network services. Physical impairments affect the propagation conditions, forcing the system to adjust the transmission robustness to consider the changes faced by the wireless channel. The adjustment process is based on Adaptive Modulation and Coding (AMC) offering a trade-off between transmission robustness and data rates. This trade-off leads to a variable quantity of data that can be sent within a transmission opportunity, consequently affecting the bandwidth allocation mechanism. This mechanism defines request and grant procedures in order to permit bandwidth reservation [2]. Request refers to the mechanism that Subscriber Stations (SSs) use to indicate to Base Station (BS) that they need uplink transmission opportunities. Bandwidth allocation mechanism is normally composed by a contention and a polling period. On the other hand, grants refer to the mechanism that allows BS to allocate bandwidth to SSs based on the requests received during the uplink period.

Recent researches have investigated the bandwidth allocation mechanism in WiMAX networks. Lin *et al*. [3] proposed a bandwidth allocation algorithm that considers AMC configuration and the urgency of each bandwidth request. Qin and Kuo [4] proposed an efficient approach to optimize the performance of uplink transmissions in IEEE 802.16e BWA networks. This approach reduces the bandwidth request delay and the control message overhead caused by the bandwidth allocation mechanism in the polling period only. Nuaymi *et al*. [5] analyze upper layers header suppression and compression processes and evaluate their effect in the specific case of Voice over IP (VoIP) transmissions. Some studies [6] [7] [8] have focused on analyzing separately the contention and polling periods without considering the overall overhead added by bandwidth allocation control messages. Thus, there are still opened questions, since issues like the impact caused by the overhead on the number of SSs that can be served by a BS using polling and contention mechanisms were not considered in past research.

Considering the aforementioned scenario, in this paper we investigate the influente caused by overhead on IEEE 802.16 networks. We consider two types of overhead in our analysis: (i) the overhead caused by control messages used for bandwidth allocation, and (ii) the overhead added by redundant informations inserted by AMC. Therefore, we analyze the overhead caused by bandwidth allocation mechanism considering (i) variations on the duration of contention and polling periods, and (ii) time-variant characteristics of the wireless channel which cause changes to AMC configuration in a scenario that considers VoIP transmissions with background traffic. The outcomes of the experiments show the number of SSs that can be served by a BS in the analyzed scenarios. We also show the importance of considering physical impairments in the process of designing a BWA network. Finally we discuss the trade-off between guaranteeing Quality of Service (QoS) and serving a larger number of SSs.

The remainder of this paper is organized as follows. Section II presents background aspects on AMC and bandwidth allocation mechanisms defined for IEEE 802.16 technology. In Section III we discuss the methodology used for the evaluation proposed in this work. In Section IV we present results regarding the overhead caused by bandwidth allocation mechanism and wireless channel propagation conditions. Finally, in Section V we conclude this paper and present directions for future investigations.

II. Background

IEEE 802.16 networks present a group of technological innovations. Two of the most important advances are: (a) adaptive modulation and coding and (b) bandwidth allocation mechanism to provide QoS. The goal of this section is to present technical details about these two advances.

A. Adaptive Modulation and Coding

AMC mechanism provides adaptation according to the physical medium conditions, through variations on coding and modulation parameters. These variations are classified as burst profile, as can be seen in Table I [1].

TABLE I
IEEE 802.16 AMC levels

AMC Modes	Un-coded Size	Coded Size
BPSK-1/2	12	24
QPSK-1/2	24	48
QPSK-3/4	36	48
16-QAM-1/2	48	96
16-QAM-3/4	72	96
64-QAM-2/3	96	144
64-QAM-3/4	108	144

Seven AMC modes are define with the goal to provide a trade-off between robustness and data rates. Using a more robust burst profile, such as BPSK-1/2, data rate decreases, while applying a less robust burst profile, such as 64-QAM-3/4, increases data rates [9]. Second and third column show respectively the size of the un-coded transmission block, and the block size after encoding, being both expressed in bytes. Thus, one can see the redundancy added by ACM mechanism in a data block.

B. Bandwidth Allocation Mechanism

Bandwidth allocation is a procedure for granting and requesting bandwidth to transmit data. In IEEE 802.16 technology data is multiplexed using Orthogonal Frequency Division Multiplexing (OFDM) technique. Before been sent over the wireless medium, data bits are mapped into OFDM symbols. These symbols are transmitted using 256 OFDM subcarriers, of which 192 are used to send data, 56 are used as guard band and 8 are pilot subcarriers used for synchronization and power control. Multiplexed data are distributed through frames. Each frame, in Time Division Duplexing (TDD) mode, consists of a downlink (DL) and an uplink (UL) subframe. In this scenario, DL subframe is transmitted using Time Division Multiplex (TDM), while UL subframe uses Time Division Multiple Access (TDMA), as can be seen in Fig. 1.

The downlink subframe is used by BS for transmitting control messages and downlink data to SSs, this is known as bandwidth grant mechanism. The uplink subframe is shared by SSs for sending bandwidth requests and uplink data to BS, this is known as bandwidth request mechanism.

Fig. 1 shows the structure of DL and UL subframes, which are composed by Information Elements (IE). A donwlink subframe starts with a preamble used for synchronization. It is followed by Frame Control Header (FCH) which is one OFDM symbol long, that contains DL Frame Prefix (DLFP) encoded using BPSK-1/2. The duration of the frame and burst profile shall be specified in DLFP. The next portion of DL subframe is the burst $\#1$, in which control messages are broadcasted. The remaining bursts are used to send data only.

The most important control messages are DL-MAP and UL-MAP. DL-MAP defines the burst start time of each MAC Packet Data Unit (PDU) used to send data. Padding can be added in cases where there is not enough data to complete the subframe duration. UL-MAP contains informations about the start time and the duration of transmission opportunities requested and granted to be transmitted during the next UL subframe. These structures also contain the identifier of the burst profile used during the transmission of each MAC PDU. Finally, DL Channel Descriptor (DCD) and UL Channel Descriptor (UCD) are broadcasted.

Moreover, an UL subframe is composed of three global parts: ranging, contention, and MAC PDU transmission. The UL subframe starts with ranging period which is responsible for specifying an interval to permit that new stations join the network, AMC setup, and power adjustment. The second part is formed by contention slots which allow bandwidth requests. However, if multiple request messages are transmitted at the same time, collisions may occur, and thus the guarantee of resource reservation can not be accomplished. Finally, MAC PDUs are used to transmit data and polling opportunities. In polling scheme, BS first allocates sufficient resources to allow SSs to send bandwidth request messages. Therefore, polling scheme is designed to be used by applications that require higher QoS, since it guarantees bandwidth allocation. Finally. the Tx/Rx Transition Gap (TTG) and Rx/Tx Transmission Gap (RTG) are specified between the DL and UL subframes, and between UL and following DL subframe (in the next frame) to allow SS terminals to turn around from reception to transmission and from transmission to reception, respectively.

III. Physical Layer Characteristics

In order to analyze IEEE 802.16 system behavior, it is necessary to characterize the physical aspects of the system, starting with OFDM symbols which are transmitted through the wireless channel [10]. The duration of these symbols (T_s) is composed of a useful symbol time (T_b) plus a guard interval (T_g), also called Cyclic Prefix (CP). Hence, we calculate both T_b and T_g to obtain the total symbol duration. The value of T_b considers the subcarrier spacing (Δ_f) and is obtained through equation 1.

$$T_b = 1/\Delta_f \tag{1}$$

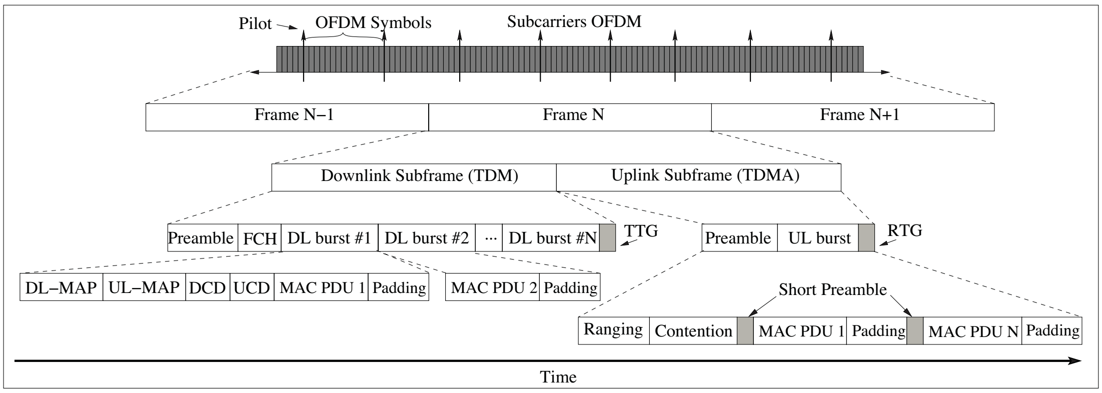


Fig. 1. OFDM frame structure with TDD

The subcarrier spacing is calculated by equation 2, where F_s is the sampling frequency and N_{FFT} is the total number of OFDM subcarriers used, *i.e.*, considering data, guard, and pilot subcarriers.

$$\Delta_f = \frac{F_s}{N_{FFT}} \tag{2}$$

F_s calculation considers the total available bandwidth (B_W), as well as a sampling factor (n), which varies according to the channel bandwidth. The value of F_s is obtained through the application of equation 3.

$$F_s = \left\lfloor \frac{B_W.n}{8000} \right\rfloor .8000 \tag{3}$$

Therefore, replacing Δ_f in equation 1, we have:

$$T_b = \frac{1}{\frac{F_s}{N_{FFT}}} \;\therefore\; T_b = \frac{N_{FFT}}{F_s} \tag{4}$$

Finally, we can replace F_s, in equation 4 to calculate T_b directly, needing only information as the number of subcarriers used, the available channel bandwidth and, consequently, the sampling factor:

$$T_b = \frac{N_{FFT}}{\left\lfloor \frac{BW.n}{8000} \right\rfloor .8000} \tag{5}$$

Another component that affects the assembly of an OFDM symbol is CP (G). The duration of CP is defined by BS and can assume the following values: 1/4, 1/8, 1/16, and 1/32, representing the amount of redundant data that is transmitted in a symbol, in order to provide synchronization and to reduce inter symbol and intra symbol interference. Therefore, T_g is defined in equation 6.

$$T_g = G.T_b \tag{6}$$

Using these information, we can define the parameters used to conduct the experiments proposed in this paper. We consider Point-to-Multipoint (PMP) TDD transmissions, and channel multiplexing using 256 OFDM subcarriers, with 1/16 CP. We simulate a wireless channel with 3.5 MHz of available bandwidth. In this case, the sampling factor is equal to 8/7. Thus, using equation 5 we obtain T_b equal to 64μs, T_g equal to 4μs and, consequently a T_s duration of 68μs.

The amount of data to be associated to each OFDM symbol varies according to AMC mechanism which is composed by the modulation and coding techniques applied to the data before transmission. In this paper, we consider three AMC configurations, to analyze the system performance in situations where wireless propagation conditions are time-variant. Therefore, we analyze AMC configurations of QPSK 1/2, 16-QAM 3/4, and 64-QAM 3/4 in order to study the trade-off between robustness and data rate, brought by AMC. Finally, we variate the frame duration, considering values of 5ms and 20ms, to permit the investigation of the impact of overhead in different scenarios. Table III summarizes the transmission parameters used for measurement purpose.

IV. Overhead Analysis

In this section, we analyze the performance of IEEE 802.16 networks under three main aspects: (i) overall system overhead, (ii) AMC scheme, and (iii) number of SSs supported per BS. These aspects are evaluated considering VoIP communications using G.711 coded for analyzing the behavior of polling mechanism. The remaining of data allocation is modelled as 256 bytes best-effort packets, in order to study contention mechanism. We also vary the duration of the frame and the contention and polling periods. We use a priority queue algorithm in BS, in order to schedule the uplink transmission opportunities.

Fig. 2 shows the maximum number of SSs supported by each BS with variations on AMC configuration, and on the duration of the frame. In this scenario, we consider 50% of

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Physical Layer	OFDM
FFT size	256
FFT used	192
Frame duration	5ms and 20ms
Useful symbol time	$64\mu s$
CP duration	$4\mu s$ (1/16)
OFDM symbol duration	$68\mu s$
OFDM symbols per subframe	36 (5ms) and 148 (20ms)
AMC configurations	QPSK 1/2, 16-QAM 3/4, and 64-QAM 3/4
Confidence Interval	95%

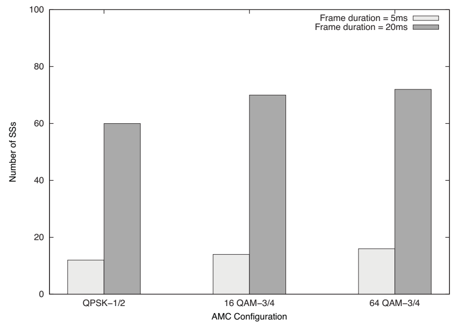


Fig. 2. Maximum number of SSs per BS

VoIP transmissions and 50% of background traffic. Results show that the number of supported SSs is not directly proportional to the frame duration. This behavior is explained because the quantity of OFDM symbols that compose a subframe in a frame duration of 5ms is equal to 36 symbols. Otherwise, if we consider a frame duration of 20ms, this number will be increased to 148 symbols per subframe. For example, if we analyze 16-QAM AMC configuration, we see that a BS can serve up to 70 SSs considering a frame duration of 20ms. This number of SSs represent that each BS can support 14 additional SSs than in situations where it is considered that the number of OFDM symbol per frame is proportional to the frame duration, once in this case it would be expected that 56 SSs could be served instead of 70. This result represents a gain of 20% in the quantity of SSs that can have QoS requirements guaranteed per each BS, showing that considering physical aspects of the network is essential for a proper network design.

We can also observe that if we apply AMC 16-QAM 3/4 in a frame duration equal to 5ms it is possible to grant transmission opportunities for 14 SSs, while increasing the frame duration for 20ms, 70 SSs are served, representing an increase of 80%. The graph also shows the behavior of the network when the wireless channel is affected by physical impairments. As we can see in the graph, when a more robust AMC configuration is applied, the number of SSs that can be served decreases. In quantitative terms, if we consider a frame duration of 20ms the more robust AMC configuration, *i.e.* QPSK 1/2, 60 SSs are supported. This number is increased to 72 SSs in the less robust configuration, *i.e.* 64-QAM 3/4, representing a gain of 17%.

The graph on Fig. 3 illustrates the impact imposed by overhead according to the number of SSs supported by each BS. In this scenario, we consider proportions of 100%, 75%, and 50% for VoIP transmissions with the remaining of the transmission opportunities grantted for background traffic. The AMC configuration considered is 64-QAM 3/4 with a frame duration of 20ms. In order to measure the overhead we consider the maximum number of SSs supported in the network, *i.e.* the number of SSs supported when the less robust AMC configuration, and the maximum frame duration are used.

Analyzing the results, we can see that while there are available resources for allocation, using only VoIP traffic we observe a lower overhead. Initially, while there is a small number of SSs, the overhead is higher. This behavior is observed mainly due to the addition of padding. For example, considering a situation where one BS is responsible for granting opportunities for 8 SSs transmitting VoIP traffic only, the observed overhead is 44,91%. This overhead is divided as follows: 36,41% is due to padding addition, and the remaining 13,5% are caused by control messages inserted by bandwidth allocation mechanism. The need for inserting padding is explained by the fact that there is not sufficient data to fulfill a transmission frame when there is a small number of SSs in the network. Thus, the bandwidth allocation mechanism completes the not used transmission opportunities with padding. The graph shows that the overhead decreases as the number of SSs is increased until it reaches an optimal point, in which the lower overhead is observed. In the mentioned scenario the optimal point is attained with a network configuration of about 30 to 40 SSs per BS. After that point, overhead starts to increase, consequently the resources are allocated until the bandwidth is saturated. The saturation occurs when the entire frame is used for requesting transmission opportunities, and therefore, no data is transmitted.

Finally, we analyze the impact caused by the variation on the period of the frame reserved for polling. We consider three scenarios with respect to polling duration. In the first scenario, BS grants a number of polling opportunities per frame that is equal to the quantity of SSs. The second scenario considers that 50% of SSs receive polling and, in the third, 25% of bandwidth requests based on polling are granted with transmission opportunities per frame. In terms of traffic, we consider only VoIP transmissions, since it uses polling as bandwidth request mechanism. We fixed AMC configuration to obtain measurements considering the maximum number of supported SSs on a frame duration of 20ms. The graph in Fig. 4 shows that granting one polling opportunity per SS is a better

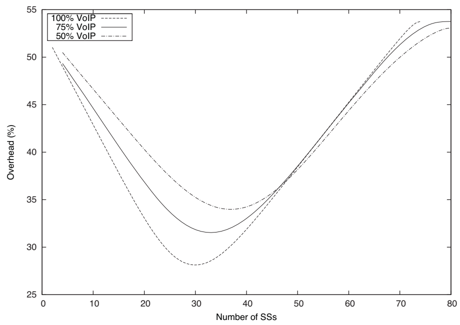


Fig. 3. Impact of number of SSs on overhead

strategy to reduce the overhead while each BS is responsible for serving up to approximately 40 SSs. After that point, it is better to reduce the proportion of polling, because the overhead tends to increase until the networks reaches its saturation point. However, the reduction on the number of polling opportunities diminishes the QoS guarantees, since it is not assured that a given SS will receive a transmission opportunity within a frame. Considering this behavior, the number of SSs supported per BS is variable from 74 with QoS guarantees up to 296, without guaranteeing QoS requirements.

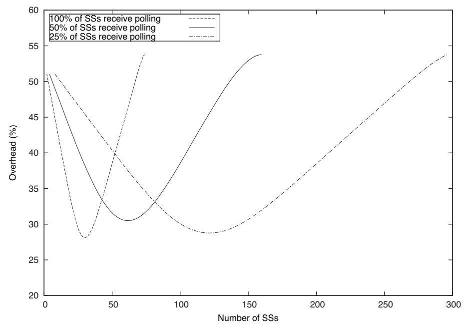


Fig. 4. Impact of polling period duration on overhead

V. Conclusions

In this paper we evaluate the impact caused by overhead on the number of SSs that can be served by each BS in WiMAX networks. In our evaluation, we considere the time-variant physical conditions of the wireless channel. To address this goal we investigate the overhead caused by control messages used for bandwidth allocation, and the overhead added by redundant information inserted by AMC in scenarios considering VoIP transmissions with background traffic.

Analysing the results, we can conclude that the number of supported SSs is not directly proportional to the duration of the frame, due to differences on the quantity of OFDM symbols that can be transmitted within each frame duration. We also show that considering wireless channel propagation conditions is essential for a proper network design, since the overhead caused by AMC is increased when a more robust configuration needs to be used. This characteristic reduces the number of SSs that will have QoS requirements guaranteed. Finally, we discuss the behavior of overhead when the proportion between VoIP and backgroud traffic is changed. In this case, we can conclude that using only VoIP traffic leads to a lower overhead.

Directions for future investigations can involve studies about the impact caused by the delay on bandwidth allocation mechanism. Another aspect to be investigated is the use of other scheduling algorithms in order to provide more restrict QoS guarantees.

Acknowledgements

Authors would like to thank Federal University of Rio Grande do Sul (UFRGS), University of Santa Cruz do Sul (UNISC), Brazilian National Research Center (CNPq), and La Salle University (UNILASALLE) for the financial support.

References

[1] IEEE, "IEEE standard for local and metropolitan area networks - part 16: Air interface for fixed and mobile broadband wireless access systems - IEEE std. 802.16-2005," New York, December 2005.

[2] S.-J. Kim, W.-J. Kim, and Y.-J. Suh, "An efficient bandwidth request mechanism for non-real-time services in ieee 802.16 systems," in *Proceedings of the 2nd International Conference on Communication Systems Software and Middleware - COMSWARE*, vol. 1, January 2007, pp. 1–9.

[3] Y.-N. Lin, C.-W. Wu, Y.-D. Lin, and Y.-C. Lai, "A latency and modulation aware bandwidth allocation algorithm for wimax base stations," in *Proccedings of the WCNC - Wireless Communications and Networking Conference*, vol. 1, 2008, pp. 1408–1413.

[4] Z. Q. nad Geng-Sheng Kuo, "Performance optimization for uplink transmission in ieee 802.16e bwa networks," in *Proccedings of the CCNC - 5th IEEE Consumer Communications and Networking Conference*, vol. 1, 2008, pp. 435–439.

[5] L. Nuaymi, N. Bouida, N. Lahbil, and P. Godlewski, "Headers overhead estimation, header suppression and header compression in wimax," in *Proccedings of the WiMOB - Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, vol. 1, 2007, pp. 17–23.

[6] J. Yan and G.-S. Kuo, "Cross-layer design of optimal contention period for IEEE 802.16 BWA systems," in *Proccedings of the International Conference on Communications - ICC*, vol. 1, 2006, pp. 1807–1812.

[7] J. Zhou, Y. Yang, D. Pang, X. Jin, J. Shi, and Z. Li, "Contention region allocation optimization in IEEE 802.16 OFDMA systems," in *Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*, vol. 1, 2007, pp. 197–200.

[8] A. Sayenko, O. Alanen, and T. Hamalainen, "Adaptive contention resolution for VoIP services in the IEEE 802.16 networks," in *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks - WoWMoM*, vol. 1, 2007, pp. 1–7.

[9] A. Ghosh, D. Wolter, J. Andrews, and R. Chen, "Broadband wireless access with WiMAX/802.16: Current performance benchmarks and future potential," *IEEE Communications Magazine*, pp. 129–136, February 2005.

[10] C. Wang and W. Xu, "A new class of generative models for burst-error characterization in digital wireless channels," *IEEE Transactions on Communications*, vol. 55, no. 3, pp. 453–462, 2007.

APPENDIX B PAPER TO THE SBRC 2011

In this appendix the paper entitled "Uma Análise do Overhead dos Mapas em Redes Metropolitanas Sem Fio Baseadas em OFDMA" is presented. This paper discussed some results regarding to this thesis through a deep analysis of the overhead generated by the map messages. The analysis showed the behavior of the network in several scenarios that consider variations in the amount of data transmitted and in the quantity of mobile stations. The variable conditions of the wireless channel in DL and UL transmissions was also considered. Results obtained through simulations showed to be crucial for the design and implementation of the components that compose a QoS architecture for wireless metropolitan networks, such as connection admission control, scheduler, and allocator.

- **Title:**
  Uma Análise do Overhead dos Mapas em Redes Metropolitanas Sem Fio Baseadas em OFDMA
- **Conference:**
  29th Brazilian Symposium on Computer Networks and Distributed Systems (SBRC 2011)
- **URL:**
  `http://sbrc2011.facom.ufms.br/`
- **Date:**
  May 30-June 3, 2011

Uma Análise do *Overhead* dos Mapas em Redes Metropolitanas Sem Fio Baseadas em OFDMA

Cristiano Bonato Both[1,2], Rafael Kunst[1,2], Lisandro Zambenedetti Granville[1], Juergen Rochol[1]

[1]Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brasil
[2]Universidade de Santa Cruz do Sul (UNISC)
Santa Cruz do Sul – RS – Brasil

{cbboth, rkunst, granville, juergen}@inf.ufrgs.br

Abstract. *The main techniques applied to provide quality of service guarantees in metropolitan area networks are OFDMA and MCS. OFDMA organizes the information within a bidimensional frame considering the time and frequency domains while MCS defines the level of robustness of a transmission. The localization of the data within the frame are defined in map messages transmitted in each frame. Thus, the main contribution this paper is to present a deep analysis of the overhead generated by these map messages. This analysis presents the behavior of the network in several scenarios that consider variations in the amount of data transmitted and in the quantity of mobile stations. Furthermore, we consider the variability of the wireless channel.*

Resumo. *As principais técnicas utilizadas para prover garantias de qualidade de serviço em redes sem fio metropolitanas são OFDMA e MCS. A técnica OFDMA organiza as informações nos domínios tempo e frequência, formando um quadro bidimensional, enquanto a técnica MCS define o nível de robustez da transmissão. Nesse contexto, para os dados serem dispostos no quadro, deve-se utilizar informações de mapas que indicam a localização destes dados. Assim, a principal contribuição deste artigo é apresentar uma profunda análise relativa ao overhead causado pelos mapas no quadro OFDMA. Essa análise apresenta o comportamento da rede em diversos cenários, variando a quantidade de dados transmitidos e a quantidade de estações móveis, bem como considerando a diversidade do meio sem fio.*

1. Introdução

As futuras tecnologias de redes de informação deverão prover infraestrutura de acesso de banda larga em diversos contextos, como para as cidades inteligentes e para a ”Internet das coisas” [Andreini et al. 2010]. Esse acesso de banda larga deverá suportar mobilidade e altas taxas de transmissão a um baixo custo de implementação. Além disso, essas redes devem oferecer garantias de qualidade aos diferentes tipos de serviços multimídia, para que dessa forma as aplicações possam oferecer qualidade de experiência (QoE - *Quality of Experience*) aos seus usuários [Bernardo et al. 2009]. Atualmente, as opções de redes que atendem esses requisitos são as tecnologias sem fio locais, como WiFi (IEEE 802.11), redes metropolitanas WiMAX (IEEE 802.16), sistemas celulares de terceira geração (3G) e o seu sucessor tecnológico conhecido como LTE (*Long Term Evolution*). Dentre essas opções, tanto WiMAX, quanto LTE são considerados precursores das futuras redes de quarta geração (4G) [Mosyagin 2010].

As redes sem fio com abrangência metropolitana, ou futuras redes 4G, têm como característica comum no nível físico a utilização das técnicas OFDMA (*Orthogonal Frequency-Division Multiple Access*) e MCS (*Modulation and Coding Scheme*). A técnica de multiplexação OFDMA atua nos domínios de tempo e de frequência, permitindo que múltiplos usuários compartilhem o acesso ao canal de Rádio Frequência (RF) simultaneamente [Ben-Shimol et al. 2006]. Esse compartilhamento é possível através da designação de um conjunto de subcanais para cada usuário, possibilitando uma maior eficiência no uso da largura de banda do canal. Por sua vez, a técnica MCS indica o tipo de modulação e a razão de codificação utilizada para minimizar os efeitos da variabilidade do canal de RF em transmissões metropolitanas sem fio. Sendo assim, a camada MAC (*Medium Access Control*) deve ser projetada para suportar essas características. Por exemplo, o quadro OFDMA é representado como uma matriz bidimensional, na qual os eixos x e y representam os domínios de tempo e frequência, respectivamente.

A camada MAC das redes metropolitanas sem fio, que utilizam o quadro OFDMA, deve possuir no mínimo três componentes, projetados como parte de uma arquitetura para garantir Qualidade de Serviço (QoS) às aplicações. Esses componentes são chamados de Controle de Admissão de Conexão (CAC), escalonador e alocador. O aceite ou rejeição de uma conexão, considerando os recursos disponíveis na rede, como por exemplo largura de banda, é de responsabilidade do CAC [Msadaa et al. 2010]. A classificação do tráfego dos usuários em ordem prioritária, de acordo com as garantias de QoS, é realizada pelo escalonador [So-In et al. 2009b], enquanto a organização desses tráfegos em um quadro bidimensional é realizada pelo alocador [Cicconetti et al. 2010]. Entretanto, esses componentes não podem ser projetados considerando somente o tráfego dos usuários, pois é preciso adicionar informações de mapas que indicam a disposição dos dados dentro do quadro OFDMA [Sekercioglu et al. 2009]. Essas informações são chamadas de *overhead*, pois apesar de utilizarem espaço no quadro não são dados úteis para os usuários.

A literatura atual sobre redes metropolitanas sem fio, que consideram o quadro OFDMA, possui uma grande quantidade de trabalhos que apresentam algoritmos de CAC [Msadaa et al. 2010] [Bashar e Ding 2009], de escalonamento [Wang et al. 2008] [So-In et al. 2009b] e de alocação [Cicconetti et al. 2010] [Cohen e Katzir 2009]. Entretanto, nenhum destes trabalhos considera o *overhead* causada pelos mapas em duas transmissões, *downlink* (DL) e *uplink* (UL), isto é, sobrecarga total da tecnologia. Os trabalhos que abordam essas informações consideram o *overhead* com valores fixos [Cicconetti et al. 2010] [So-In et al. 2009a], ou ainda, somente o *overhead* dos mapas DL ou UL são consideradas. Nesse contexto, a principal contribuição deste trabalho é realizar uma profunda análise relativa ao *overhead* causado pelos mapas no quadro OFDMA. Essa análise objetiva apresentar o comportamento da rede em diversos cenários, variando a quantidade de dados transmitidos e a quantidade de conexões ativas na rede, bem como considerando a variabilidade do meio sem fio, tanto em transmissões DL, quanto UL. Os resultados apresentados são fundamentais para o desenvolvimento de componentes de CAC, escalonamento e alocação para o projeto de arquiteturas que provem QoS para redes sem fio metropolitanas.

O restante deste artigo está organizado da seguinte forma. Na Seção 2, descreve-se a fundamentação teórica sobre as características das redes metropolitanas sem fio, além disso, são apresentados os trabalhos relacionados sobre o *overhead* causado pelos mapas no quadro OFDMA. Na Seção 3, é apresentada a implementação do simulador que considera a diversidade das redes metropolitanas sem fio e os componentes da camada MAC. Na Seção 4, são discutidos os resultados obtidos e, finalmente, na Seção 5, apresenta-se conclusões e aponta-se direções para futuras investigações.

2. Fundamentação Teórica

As redes metropolitanas sem fio possuem funcionalidades semelhantes as redes sem fio convencionais. Entretanto, operam em grandes áreas de cobertura, com altas taxas de transmissão (tipicamente superiores a 100 Mbit/s), provendo garantias de QoS para diferentes tipos de tráfego de usuários e suporte a mobilidade com altas velocidades, por exemplo 120 Km/h. Nesta seção são descritas as principais fundamentações teóricas relativas as características dessas redes, bem como os trabalhos relacionados.

2.1. Características das Redes Metropolitanas Sem Fio

Uma das principais técnicas utilizadas em redes metropolitanas sem fio é a multiplexação de canal OFDMA. A principal característica dessa técnica é o suporte a múltiplos usuários em um mesmo símbolo OFDM (*Orthogonal Frequency-Division Multiplexing*). Tal característica é obtida através da divisão do canal sem fio em subcanais que são formados por conjuntos de subportadoras. Esse processo de multiplexação é conhecido como subcanalização ou zonas de permutação. As tecnologias WiMAX e LTE especificam alguns métodos de subcanalização, mas apenas o método PUSC (*Partially Used Subchannelization*) é de implementação obrigatória.

O processo de subcanalização torna complexo o mecanismo de alocação de banda, pois as alocações devem ser realizadas considerando os domínios de tempo (símbolos OFDM) e de frequência (subcanais), utilizando uma unidade de alocação mínima chamada *slot* [Camargo et al. 2009]. Por exemplo, nas redes WiMAX, em uma transmissão DL cada *slot* é composto por 2 símbolos OFDM e 1 subcanal, sendo que cada subcanal utiliza 28 subportadoras de dados. Já na transmissão UL, cada *slot* é formado por 3 símbolos OFDM que utilizam 24 subportadoras de dados [So-In et al. 2009a]. O tipo de subcanalização utilizado determina a estrutura do quadro OFDMA, que pode ser observada na Figura 1.

O quadro OFDMA pode ser representado como uma matriz bidimensional, na qual os eixos x e y representam os domínios de tempo e frequência, respectivamente. A construção dos subquadros deve considerar, principalmente, as características de acesso ao meio. No subquadro DL apenas a BS (*Base Station*) transmite informações para todas as MSs (*Mobile Station*), iniciando com um preâmbulo de sincronização que é transmitido em todos os subcanais, tendo a duração de 1 símbolo. No segundo símbolo inicia a transmissão do FCH (*Frame Control Header*), utilizando 4 subcanais, como pode ser observado na Figura 1. O FCH contém informações sobre a localização e duração da mensagem DL-MAP. Esse mapa possui estruturas de dados chamadas de IEs (*Information Element*) que carregam informações sobre a localização, duração, quantidade de conexões e configuração MCS das rajadas DL. Uma configuração MCS indica o tipo de modulação

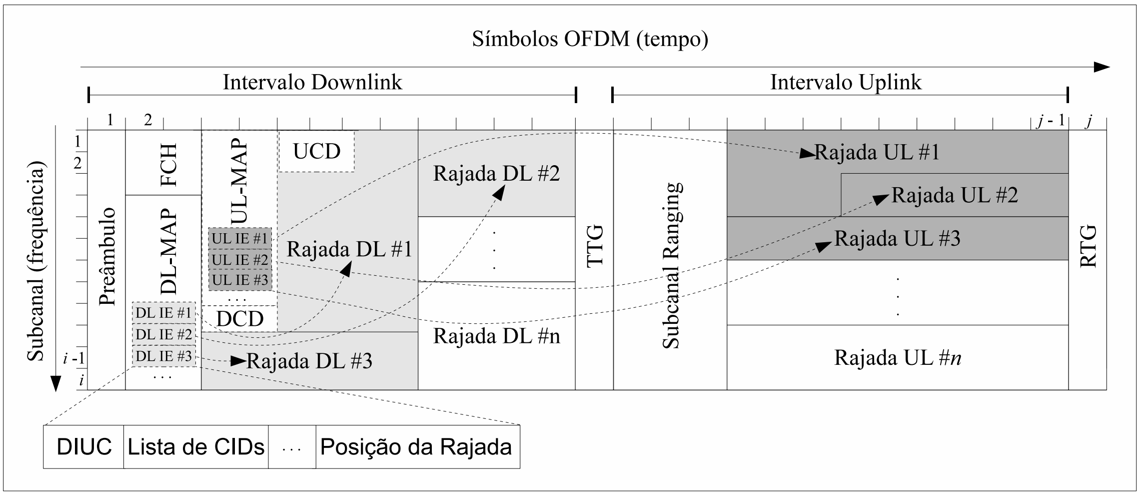


Figura 1. Estrutura do quadro OFDMA

e a razão de codificação utilizada na transmissão de uma determinada rajada de dados. Em outras palavras, o agrupamento das conexões com o mesmo MCS em um quadro OFDMA forma uma rajada DL. Além disso, informações sobre o canal de RF são transmitidas na mensagem chamada DCD (*Donwlink Channel Descriptor*).

Já, no subquadro UL, o meio de transmissão é compartilhado por todas as MSs que desejam transmitir informações para a BS. Por isso, o acesso das MSs ao subquadro UL precisa ser coordenado pela BS, através de um mecanismo de concessão de oportunidades de transmissão. Quando um subquadro UL é alocado, a primeira rajada DL do quadro OFDMA deve conter um mapa chamado UL-MAP e um descritor do canal denominado UCD (*Uplink Channel Descriptor*). Esse mapa contém informações referentes as rajadas UL, seguindo uma estrutura semelhante a do DL-MAP. Cada rajada UL é formada por informações enviadas por apenas uma MS. Essas informações podem conter dados e nesse caso a rajada UL representa uma oportunidade de transmissão de dados concedida a uma determinada MS. Por outro lado, essa informação pode ser uma requisição de banda e dessa forma a rajada UL representa uma oportunidade de requisição de banda concedida para uma determinada MS. Além das rajadas, um subquadro UL contém um canal de *Ranging*, que é um período compartilhado entre todas as MSs, utilizado para o envio de requisições de banda, como no exemplo da Figura 1.

Quando não houver informação suficiente para preencher completamente os subquadros DL ou UL, os mesmos serão preenchidos com *padding*. Além disso, entre os subquadros DL e UL existe um TTG *(Transmit/Receive Transition Gap)* alocado para permitir que a BS alterne do modo de transmissão para o de recepção. No final do subquadro UL, é alocado um RTG *(Receive/Transmit Transition Gap)* para que a BS retorne ao modo de transmissão.

As informações contidas nos mapas são fundamentais para a disposição dos dados dos usuários no quadro OFDMA. Entretanto, devido a estrutura do quadro ser bidimensional, essas informações, chamadas de *overhead* neste trabalho, devem ser consideradas para o projeto dos algoritmos de CAC, escalonamento e alocação das futuras redes metropolitanas sem fio. Por exemplo, na Tabela 1 são apresentados os tamanhos dos mapas para o quadro OFDMA discutido na Figura 1 da tecnologia WiMAX. Essa tabela resume as informações contidas no padrão IEEE 802.16 [IEEE 2009].

Tabela 1. Overhead of the OFDMA Frame

<table>
<tr><th>Mapas</th><th>Fixa (bits)</th><th colspan="2">Variável (bits)</th></tr>
<tr><td>FCH</td><td>96</td><td colspan="2">-</td></tr>
<tr><td>DCD</td><td>24</td><td colspan="2">24.N_{R-DL}</td></tr>
<tr><td>UCD</td><td>48</td><td colspan="2">24.N_{R-UL}</td></tr>
<tr><td>DL-MAP</td><td>104</td><td colspan="2">(12.DL_{IE}) + (48.N_C)</td></tr>
<tr><td rowspan="3">UL-MAP</td><td rowspan="3">64</td><td rowspan="3">(20.UL_{IE}) +</td><td>(32.N_B)</td></tr>
<tr><td>(40 .N_A)</td></tr>
<tr><td>(10.N_D)</td></tr>
</table>

Os mapas formados por IEs podem ser divididos em duas partes: fixa e variável, com excessão do FCH que possui tamanho fixo de 96 bits. Os descritores de canal DCD e UCD são relacionados com o número de rajadas de dados, N_{R-DL} e N_{R-UL}, respectivamente. Ou seja, a cada nova rajada de dados alocada no quadro OFDMA, 24 bits de *overhead* são adicionados. Já, o tamanho do mapa DL-MAP está associado a quantidade de IEs no DL (DL_{IE}) e a quantidade de conexões (N_C) em cada rajada. Sendo assim, a medida que o número de conexões com o mesmo MCS aumenta, o tamanho do DL-MAP é acrescido em 48 bits. Por fim, o tamanho das informações do UL-MAP está relacionado a quantidade de IEs no UL (UL_{IE}) e com o tipo de informação transmitida, podendo ser requisições de banda CDMA (*Code Division Multiple Access*), representadas neste artigo como N_B, informações sobre alocações CDMA, definidas como N_A ou dados (bits) enviados das MSs para a BS, indicada por N_D. Neste contexto, onde o aumento das informações dos mapas possui uma grande variabilidade, é fundamental analisar o *overhead* das informações dos mapas para projetos de algoritmos de CAC, escalonamento e alocação. A seguir, são descritos alguns trabalhos que apresentam estudos parciais sobre a análise do *overhead* dos mapas em redes metropolitanas sem fio baseadas no quadro OFDMA.

2.2. Trabalhos Relacionados

Os três principais componentes em uma arquitetura para garantir QoS em redes metropolitanas sem fio baseadas em OFDMA são normalmente projetados sem considerar o *overhead*. Msadaa, Câmara e Filali [Msadaa et al. 2010] apresentam um levantamento completo sobre o estado da arte no que diz respeito aos componentes de CAC e escalonamento, analisando diversos trabalhos que não consideram o *overhead* em suas propostas. Já, Cohen e Katzir [Cohen e Katzir 2009] exploram a integração entre os componentes escalonador e alocador, sem considerar a alocação dos mapas no quadro OFDMA.

Por sua vez, a abordagem típica apresentada nos trabalhos relacionados ao cálculo do *overhead* não considera os três componentes. Por exemplo, Jin *et. al.* [Jin et al. 2008] estudam o *overhead* considerando somente transmissões DL utilizando o método de multiplexação OFDMA, sem considerar uma arquitetura de QoS. Por outro lado, Tykhomyrov *et. al.* [Tykhomyrov et al. 2010] propõem uma otimização nos mapas, a fim de reduzir o *overhead*. Para isso, é apresentado um comparativo entre o *overhead* DL com e sem a utilização de compressão do cabeçalho. Os resultados mostram que a utilização de compressão não representa ganhos significativos em termos de vazão, pois segundo o padrão IEEE 802.16 [IEEE 2009], que especificada as redes WiMAX, a compressão é aplicada somente na parte fixa do *overhead*. Sendo assim, neste trabalho, os estudos apresentados não consideram o emprego de compressão nos cabeçalhos dos mapas.

O fato dos trabalhos relacionados não considerarem o cálculo do *overhead* no projeto dos três componentes utilizados para provimento de QoS pode causar alguns problemas no funcionamento da arquitetura, os quais são investigados neste trabalho. Por exemplo, caso o componente CAC não considere a existência do *overhead*, a quantidade de recursos, isto é, largura de banda disponível pode ser superestimado, uma vez que o *overhead* não é descontado do total de recursos. Por outro lado, se o *overhead* for desconsiderado no escalonador, este tende a sobrecarregar o alocador, enviando mais dados do que o alocador é capaz de organizar no quadro. Por consequência, se o alocador não contabilizar o *overhead*, o cálculo do espaço disponível para alocação de dados no quadro pode ser superestimado. Sendo assim, neste trabalho são considerados os três principais componentes de uma arquitetura para provimento de QoS em redes metropolitanas sem fio para o cálculo do *overhead*.

3. Infraestrutura para Análise do *Overhead* dos Mapas

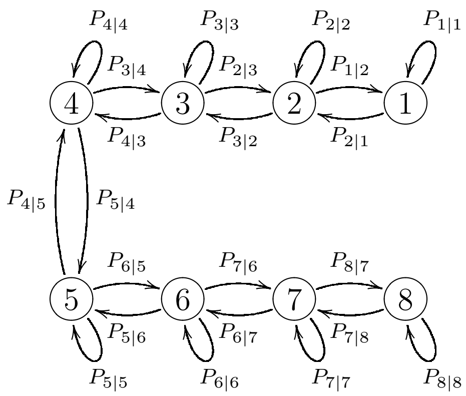
Com o objetivo de analisar o *overhead* dos mapas do quadro OFDMA para redes metropolitanas sem fio, desenvolveu-se um simulador composto de dois módulos. O primeiro módulo, descrito na Subseção 3.1, tem como objetivo representar a diversidade do canal de RF. Já, o segundo módulo, apresentado na Subseção 3.2, possui os componentes da camada MAC, que consideram as técnicas de multiplexação OFDMA e MCS.

3.1. Modelo de Canal de RF

O modelo do canal de RF utilizado no simulador foi projetado através de uma cadeia de Markov de tempo discreto, composta por oito estados. Cada estado representa uma configuração MCS, isto é, considera um canal de RF com uma determinada modulação e taxa de codificação. A informação de qual MSC está sendo utilizado em uma transmissão é fundamental para o cálculo do *overhead* do quadro OFDMA. Os estados são divididos em dois grupos: (i) o conjunto de estados 1-4 é utilizado em canais com uma condição de transmissão ruim, enquanto (ii) os estados 5-8 são usados em casos que o canal apresenta uma boa condição de transmissão. A representação gráfica da cadeia de Markov proposta e os MCSs correspondentes a cada estado da cadeia são apresentados na Figura 2.

Cada estado da cadeia de Markov tem probabilidades de transição associadas. Na concepção do modelo de simulação, as probabilidades de transição entre os estados da cadeia de Markov são parâmetros que podem ser definidos de acordo com o cenário de simulação pretendido. Esses parâmetros são informados considerando probabilidades de transição para os estados da cadeia. Outro parâmetro importante que deve ser informado é a espera entre avaliações da necessidade de transição entre os estado do canal, que é definida em número de quadros.

Durante a simulação da condição do canal, existem três possibilidades em termos de transições: (i) a permanência no mesmo estado, quando o canal mantém a mesma



Estado	MCS	Canal
1	QPSK 1/2	Ruim
2	QPSK 3/4	
3	16QAM 1/2	
4	16QAM 3/4	
5	64QAM 1/2	Bom
6	64QAM 2/3	
7	64QAM 3/4	
8	64QAM 5/6	

Figura 2. Modelo do Canal de RF

condição de transmissão, (ii) aumento de um nível de robustez, quando a condição do canal atual é pior do que na última transmissão, ou (iii) reduzir um nível de robustez, se a condição do canal é melhor do que na última transmissão. Por exemplo, se o MCS atual é de 16 QAM 3/4 e uma condição pior de canal é detectada, na próxima transmissão será aplicada a configuração 16 QAM 1/2. Caso contrário, considerando que a condição atual do canal é melhor do que na última transmissão, deve ser usado 64 QAM 1/2. Existem duas exceções que ocorrem quando o MCS mais e menos robusto são utilizados. No primeiro caso, não é possível aumentar a robustez, enquanto no segundo, é impossível reduzir a robustez.

O modelo do canal de RF descrito nesta seção é utilizado na simulação dos três principais componentes para garantir QoS nas tecnologias para redes metropolitanas sem fio. A seguir são apresentados esses componentes que fazem parte da camada MAC de uma BS.

3.2. Componentes da Camada MAC

Os componentes da camada MAC da arquitetura para garantir QoS devem trabalhar de forma interligada para garantir o bom funcionamento da rede metropolitana sem fio. Sendo assim, na simulação considerou-se a interligação entre esses componentes, conforme ilustrado na Figura 3. O primeiro componente da arquitetura é o CAC, com a finalidade de definir quais conexões serão aceitas e quais serão negadas, através da análise da quantidade de recursos disponíveis para garantir a QoS para as aplicações. Na literatura, existem diversos tipos de CAC, sendo que o escolhido para a simulação foi o CAC baseado em análise da largura de banda disponível, por ser a abordagem mais comumente empregada [Msadaa et al. 2010].

O segundo componente considerado na simulação da arquitetura é o escalonador. O objetivo deste componente é selecionar qual fila deve ser atendida, de acordo com os requisitos de QoS de cada tipo de tráfego dos usuários. Neste trabalho, é simulado o algoritmo *Proportional Fair* (PF) [Potter e Zukerman 2001], que associa uma prioridade a cada fila, a fim de definir qual fila deve ser atendida. A primeira fila a ser atendida

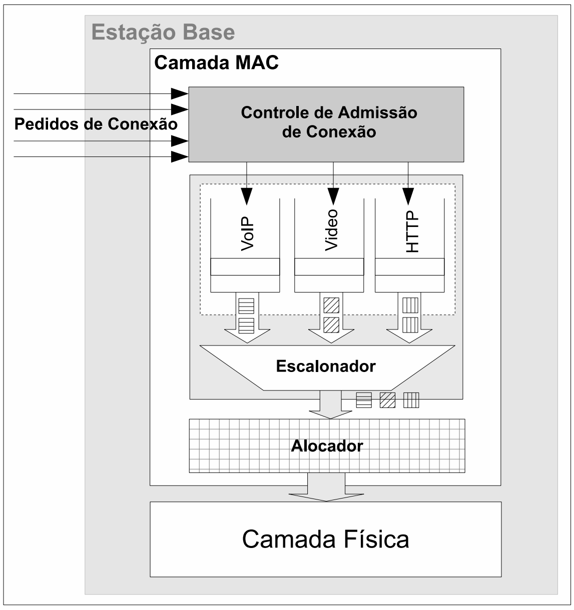


Figura 3. Componentes da Camada MAC

é sempre a de maior prioridade, a qual é servida até que todos os quadros armazenados nela sejam atendidos. Após, a próxima fila na ordem de prioridade é servida. A cada quadro servido em uma fila de menor prioridade, a fila de mais alta prioridade é verificada novamente para garantir que caso exista um novo quadro, esse seja atendido considerando o critério de priorização. Neste trabalho, o algoritmo PF é utilizado para priorizar a fila de voz sobre IP (VoIP) em relação as filas de video e HTTP. Isto se justifica porque uma rede metropolitana sem fio pode atender a tráfego de sistemas celulares e, portanto, a priorização do tráfego de voz é fundamental.

O terceiro componente simulado é o alocador, que é responsável por organizar o tráfego dos usuários no quadro OFDMA. O alocador simulado neste trabalho foi o *Recursive Tiles and Stripes* (RTS), por se mostrar a proposta mais eficiente dentre as atualmente publicadas [Cicconetti et al. 2010]. O funcionamento deste alocador é organizado em duas fases. A primeira fase, chamada de *Tiles* consiste na tentativa de organização verical dos dados dentro do quadro OFDMA. Por outro lado, a segunda fase, conhecida como *Stripes* busca alocar horizontalmente os dados no quadro. Uma quantidade parametrizável de iterações é realizada e a melhor solução é selecionada, isto é, a solução que for capaz de alocar a maior quantidade de dados dentro do quadro OFDMA. Nesse algoritmo é apresentada a necessidade de uma função ϕ que calcula o *overhead* causado pelos mapas. Entretanto, a implementação dessa função não é especificada no algoritmo RTS.

Sendo assim, neste trabalho, é proposto o Algoritmo 1 que tem a finalidade de calcular o valor da função ϕ. Este algoritmo recebe como parâmetros os conjuntos de rajadas DL e UL, bem como os conjuntos de alocações CDMA, de requisições de banda, de conexões atendidas e de rajadas de dados UL. Além disso, define-se uma função, chamada de γ neste trabalho, que calcula o *overhead* fixo do quadro, de acordo com as informações apresentadas na Tabela 1. O resultado obtido através da aplicação da função γ é somado ao *overhead* variável, calculado pelo algoritmo proposto, a fim de obter o *overhead* total, que corresponde ao valor de ϕ.

Algoritmo 1 Algoritmo para cálculo da função ϕ

Parâmetros: $R-DL=\{R-DL_1, R-DL_2, ..., R-DL_n\}$: conjunto de rajadas DL
Parâmetros: $R-UL=\{R-UL_1, R-UL_2, ..., R-UL_n\}$: conjunto de rajadas UL
Parâmetros: $A=\{A_1, A_2, ..., A_n\}$: conjunto de alocações CDMA
Parâmetros: $B=\{B_1, B_2, ..., B_n\}$: conjunto de requisições de banda
Parâmetros: $C=\{C_1, C_2, ..., C_n\}$: conjunto de conexões em um quadro
Parâmetros: $D=\{D_1, D_2, ..., D_n\}$: conjunto das rajadas de dados UL
Parâmetros: $DL_{IE}=\{D_{IE_1}, D_{IE_2}, ..., D_{IE_n}\}$: conjunto de IEs no subquadro DL
Parâmetros: $UL_{IE}=\{U_{IE_1}, U_{IE_2}, ..., U_{IE_n}\}$: conjunto de IEs no subquadro UL

$N_{R-DL}=|R-DL|$
$N_{R-UL}=|R-UL|$
$N_A=|A|$
$N_B=|B|$
$N_C=|C|$
$N_D=|D|$
$N_{DL_{IE}}=|DL_{IE}|$
$N_{UL_{IE}}=|UL_{IE}|$
γ é somatório do tamanho fixo dos Mapas, conforme Tabela 1
$\phi=\gamma$
para todo novo quadro a ser transmitido **faça**
    para i de 1 ate N_{R-DL} passo 1 **faça**
        $\phi+=24$ //cálculo do *overhead* do DCD
        $\phi+=12.N_{DL_{IE}}+48.N_C$ //cálculo do *overhead* do DL-MAP
    fim para
    para i de 1 ate N_{R-UL} passo 1 **faça**
        $\phi+=24$ //cálculo do *overhead* do UCD
        $\phi+=20.UL_{IE}+(40.N_A+32.N_B+10.N_D)$ //cálculo do *overhead* do UL-MAP
    fim para
fim para

O valor retornado pela função ϕ é disponibilizado aos três componentes que compõem a arquitetura para provimento de QoS em redes metropolitanas sem fio. Com isso, é possível que o projeto desses componentes considere informações acuradas sobre o *overhead*, permitindo um melhor funcionamento da rede como um todo. Na próxima seção é apresentada uma análise profunda do impacto causado pelo *overhead* gerado pelos mapas.

4. Resultados

O cálculo do *overhead* considerando os componentes da camada MAC da arquitetura para provimento de QoS em redes metropolitanas sem fio é analisado através de simulações, implementadas usando a linguagem de programação Labview e Matlab integradas. Os parâmetros utilizados nas simulações são apresentados na Tabela 4 e refletem as configurações recomendadas pelo documento que descreve a metodologia de avaliação das redes WiMAX [WiMAX Forum 2008]. Este documento possui uma estrutura que permite simular o comportamento de sistemas celulares, como por exemplo a tecnologia LTE.

Tabela 2. Parâmetros de Simulação

Parâmetro	**Valor**
Largura de banda	10 MHz
Subportadoras total	1024
Subportadoras de dados por subcanal DL	28
Subportadoras de dados por subcanal UL	24
Duração do quadro	5 ms
Dimensão do subquadro DL (Subcanais x Símbolos)	14 x 30 *slots*
Dimensão do subquadro UL (Subcanais x Símbolos)	6 x 15 *slots*
Intervalo de confiança	95%

A primeira análise realizada neste trabalho refere-se a quantidade global de *overhead* no quadro OFDMA. O *overhead* apresentado no gráfico é causado pelos mapas e descritores de canais gerados devido a transmissões DL e UL. Entretanto, como os mapas e descritores de canal estão localizados no subquadro DL, a análise do *overhead* em relação a quantidade de dados alocados deve ser realizada considerando somente este subquadro. Neste contexto, o eixo y da Figura 4 representa a capacidade normalizada do subquadro DL, em relação a um subquadro DL contendo 420 slots de capacidade, em um quadro de 5 ms. Por sua vez, no eixo x da figura é apresentado o número de conexões ativas na rede, ou seja, a quantidade de conexões aceitas pelo mecanismo de CAC.

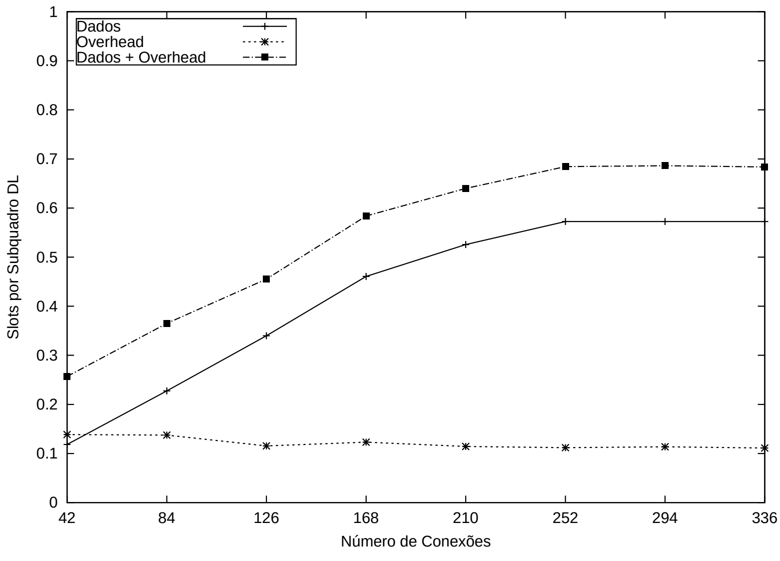


Figura 4. *Overhead* Global

Na Figura 4 pode-se observar que o *overhead* global utiliza em média 13% da capacidade do subquadro DL. Esse comportamento é independente da quantidade de conexões aceitas pelo algoritmo de CAC na BS. Isto se justifica porque o *overhead* é causado pela quantidade de conexões, quantidade de rajadas e pedidos de requisições UL em um quadro OFDMA e não em relação ao número total de conexões aceitas pela BS. Pode-se perceber também que quando há poucas conexões na rede, por exemplo, 84, a quantidade de tráfego dos usuários é menor que a quantidade de *overhead*, devido a baixa demanda por transmissões de dados por parte das MSs. Já, a partir de 252 conexões, o desempenho do algoritmo de alocação RTS não permite arranjar todos os dados no subquadro OFDMA, conforme apresentado por Cicconetti *et al.* [Cicconetti et al. 2010]. Nessas duas situações existe um espaço do subquadro OFDMA que não é utilizado para alocação de dados e *overhead*, o qual é preenchido com *padding*. Quando existem poucas conexões, *padding* é adicionado devido a existência de pouco tráfego de usuários. Por outro lado, quando existem muitas conexões, a inserção de *padding* se deve à ineficiência do algoritmo de alocação.

A segunda análise apresentada neste trabalho é referente ao comportamento do *overhead* gerado por transmissões DL e UL. Na Figura 5 é analisado o número de conexões em relação a capacidade normalizada do subquadro DL, sendo possível observar que o *overhead* UL utiliza em média cerca de 10% do subquadro, enquanto o *overhead* DL ocupa aproximadamente 3% da capacidade. A redução do *overhead* DL é justificada, devido a transmissão DL poder associar um conjunto de conexões a uma mesma rajada, o que não é possível em transmissões ou requisições UL, por causa do compartilhamento de acesso ao meio entre as MSs. Sendo assim, nas transmissões UL cada conexão será associada a uma rajada.

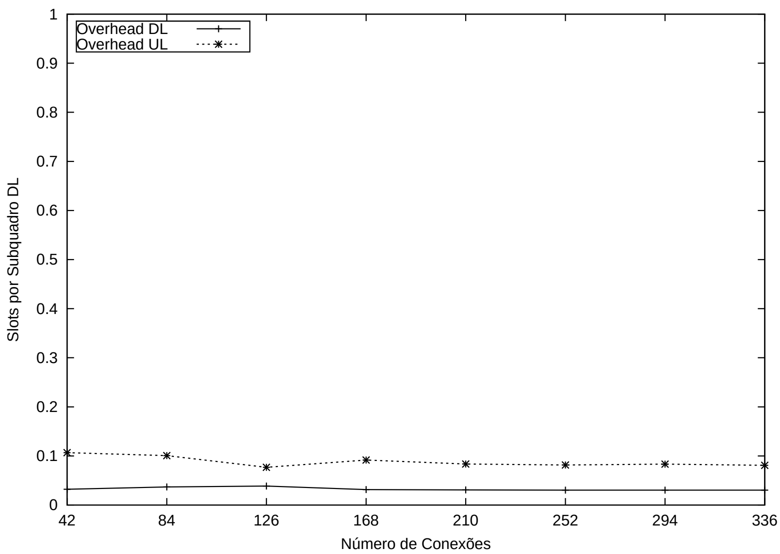


Figura 5. *Overhead* causado pelos Mapas

O objetivo da Figura 6 é aprofundar a análise do *overhead* UL, através dos tipos de mensagens que o compõe. É interessate ressaltar que o *overhead* dos dados UL é transmitido no subquadro DL, enquanto o *overhead* das requisições UL é transmitido no subquadro UL. Assim, a Figura 6 apresenta o percentual do *overhead* de dados UL considerando a capacidade do subquadro DL igual a 420 slots. Por outro lado, o percentual do *overhead* de requisições UL é calculado considerando a capacidade do subquadro UL de 90 slots. O *overhead* de dados UL é de aproximadamente de 8% da capacidade do subquadro DL, pois esse depende da quantidade de rajadas que são associadas com um conjunto de conexões e MCSs. Já, o *overhead* das requisições UL apresenta um crescimento até 126 conexões e depois se mantém em aproximadamente 54%. Esse comportamento se justifica, uma vez que até 126 conexões, o subquadro UL possui espaço disponível para as requisições de banda UL e após essa quantidade de conexões ocorre um saturamento no subquadro UL. Maiores detalhes sobre o impacto causado por essa sobrecarga nas requisições de banda no subquadro UL podem ser observados no trabalho de Camargo *et al.* [Camargo et al. 2009].

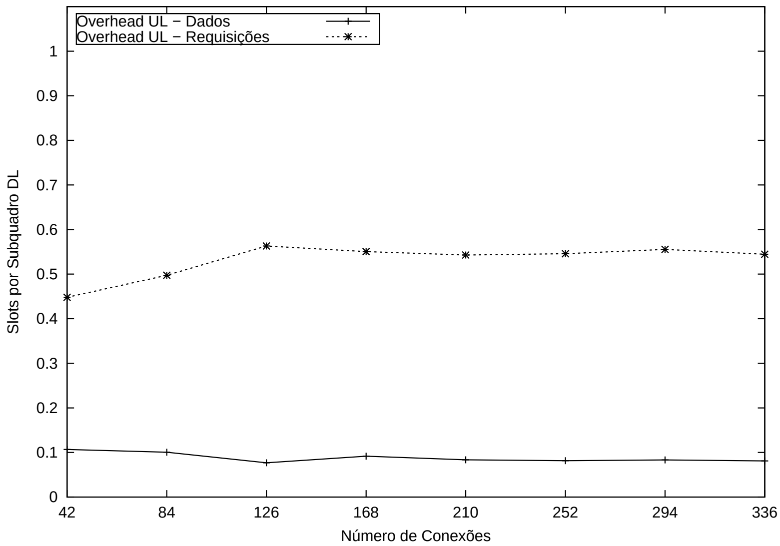


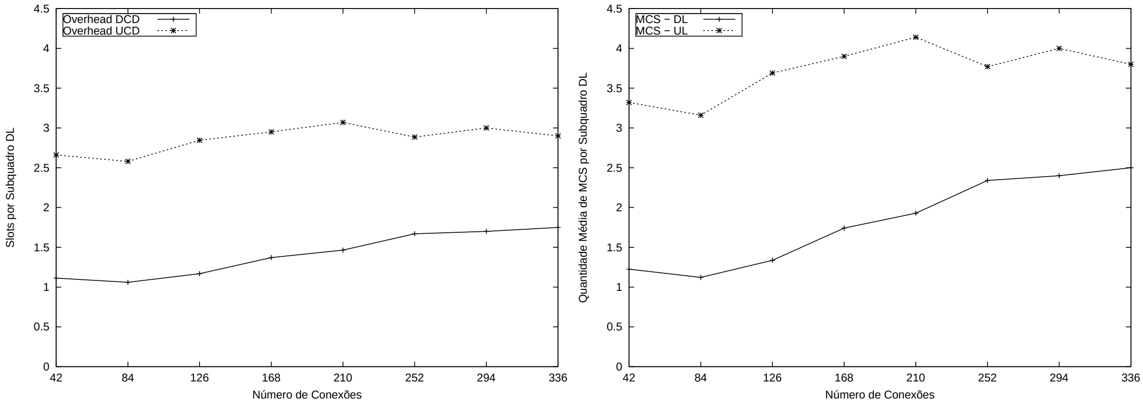
Figura 6. Análise do *Overhead* UL

A última análise realizada neste trabalho refere-se ao *overhead* causado pelos descritores de canais. Por isso, na Figura 7 são mostrados dois gráficos que se complementam a fim de analisar separadamente o *overhead* causado pelo DCD e UCD. Sendo assim, na Figura 7 (a) é analisado o número de conexões em relação a capacidade do subquadro DL em *slots*, enquanto na Figura 7 (b) o eixo y refere-se à quantidade média de MCS por subquadro. Na Figura 7 (a) pode-se perceber que a capacidade do subquadro DL ocupada pelos descritores de canal é inferior a 3 *slots*. Por esse motivo é apresentado o valor absoluto, uma vez que essa quantidade de *slots* torna-se insignificante percentualmente se comparada aos 420 *slots* disponíveis.

Na Figura 7 (a) percebe-se que o *overhead* causado pelo UCD é superior ao gerado pelo DCD. Isto é justificado através da análise da Figura 7 (b), pois a quantidade média de MCS no subquadro UL é superior a do subquadro DL. Analisando os dois gráficos conjuntamente, percebe-se que existe uma relação diretamente proporcional entre a quantidade de MCSs e o tamanho dos descritores. Além disso, pode-se concluir que a quantidade de *overhead* DL é influenciada exclusivamente pela quantidade de MCSs que não pode superar 8 diferentes configurações de modulação e codificação, visto que uma rajada deve comportar todas as conexões com o mesmo MCS. Por outro lado, no subquadro UL, cada rajada é associada a uma conexão com seu respectivo MCS e, por isso, uma quantidade maior de *overhead* é observada no UCD em comparação do DCD.

5. Conclusões e Trabalhos Futuros

Neste artigo, foi realizada uma profunda análise do *overhead* global em uma arquitetura para provimento de QoS em redes metropolitanas sem fio baseada no quadro



(a) (b)

Figura 7. *Overhead* causado pelos descritores

OFDMA. A maior contribuição do trabalho foi a definição da forma de cálculo do tamanho dos mapas e descritores de canal, bem como a sua variabilidade considerando as condições do meio sem fio, o número de conexões ativas e a quantidade de transmissões DL e UL. Essa contribuição é fundamental para apoiar o desenvolvimento de projeto dos três principais componentes da camada MAC, que são: CAC, escalonador e alocador.

Os resultados mostram o *overhead* global de uma rede metropolitana sem fio considerando o quadro OFDMA. Além da visão global, que afeta o desempenho do componentes da camada MAC, foi realizado um estudo detalhado das causas do *overhead*. Por exemplo, foi detalhada a composição do *overhead* DL e UL, através da análise do comportamento dos mapas e descritores de canal. Com os resultados obtidos pode-se concluir que os projetos de componentes para a camada MAC de redes metropolitanas sem fio não podem desconsiderar o efeito causado pelo *overhead* no desempenho dos componentes.

Direções para futuras investigações devem envolver variações na configuração da rede metropolitana sem fio, como alterção no tamanho do quadro, número de subportadoras e largura de banda do canal sem fio. Além disso, é possível explorar as futuras técnicas de multiplexação do canal de RF.

Agradecimentos

Os autores gostariam de agradecer ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNpQ), à Universidade Federal do Rio Grande do Sul (UFRGS) e à Universidade de Santa Cruz do Sul (UNISC) pelo apoio financeiro.

Referências

Andreini, F., Crisciani, F., Cicconetti, C., and Mambrini, R. (2010). Context-aware location in the internet of things. In *IEEE Globecom - Workshop on Network of the Future*, pages 300–304.

Bashar, S. and Ding, Z. (2009). Admission control and resource allocation in a heterogeneous OFDMA wireless network. *IEEE Transactions on Wireless Communications*, 8(8):4200–4210.

Ben-Shimol, Y., Kitroser, I., and Dinitz, Y. (2006). Two-dimensional mapping for wireless OFDMA systems. *IEEE Transactions on Broadcasting*, 52(3):388–396.

Bernardo, V., Sousa, B., and Curado, M. (2009). Voip over wimax: Quality of experience evaluation. In *IEEE Symposium Computers and Communications*, pages 42–47.

Camargo, E., Both, C., Kunst, R., Granville, L., and Rochol, J. (2009). Uma arquitetura de escalonamento hierárquica para transmissões uplink em redes WiMAX baseadas em OFDMA. In *Simpósio Brasileiro de Redes de Computadores*, pages 525–538.

Cicconetti, C., Lenzini, L., Lodi, A., Martello, S., Mingozzi, E., and Monaci, M. (2010). Efficient two-dimensional data allocation in IEEE 802.16 OFDMA. In *IEEE INFOCOM - Conference on Computer Communications*, page 1.

Cohen, R. and Katzir, L. (2009). Computational analysis and efficient algorithms for Micro and Macro OFDMA downlink scheduling. *IEEE Transactions on Networking*.

IEEE (2009). IEEE standard for local and metropolitan area networks, part 16 - air interface for broadband wireless access systems.

Jin, X., Zhou, J., Hu, J., Shi, J., Sun, Y., and Dutkiewicz, E. (2008). An efficient downlink data mapping algorithm for IEEE802.16e OFDMA systems. In *IEEE GLOBECOM - Global Communications Conference*.

Mosyagin, J. (2010). Using 4G wireless technology in the car. In *12th International Conference on Transparent Optical Networks*, page 1.

Msadaa, I. C., Câmara, D., and Filali, F. (2010). Scheduling and CAC in IEEE 802.16 fixed BWNs: A comprehensive survey and taxonomy. *IEEE Communications Surveys & Tutorials*, 12(4):459–487.

Potter, P. G. and Zukerman, M. (2001). Analysis of a discrete multipriority queueing system involving a central shared processor serving many local queues. *IEEE Journal on Selected Areas in Communications*, 9(2):194–202.

Sekercioglu, Y. A., Ivanovich, M., and Yegin, A. (2009). A survey of MAC based QoS implementations for WiMAX networks. *Computer Networks*, 53(14):2517–2536.

So-In, C., Jain, R., and Tamimi, A.-K. (2009a). Capacity evaluation for IEEE 802.16e mobile WiMAX. *Journal of Computer Systems, Networks, and Communications*, 2010(1):1–12.

So-In, C., Jain, R., and Tamimi, A.-K. (2009b). Scheduling in IEEE 802.16e mobile WiMAX networks: Key issues and a survey. *IEEE Journal on Selected Areas in Communications*, 27(2):156–171.

Tykhomyrov, V., Sayenko, A., Puchko, O., and Hamalainen, T. (2010). Decreasing the map overhead in the IEEE802.16 OFDMA systems. In *European Wireless Conference*.

Wang, F., Ghosh, A., Sankaran, C., Fleming, P. J., Hsieh, F., and Benes, S. J. (2008). Mobile WiMAX systems: Performance and evolution. *IEEE Communications Magazine*, 46(10):41–49.

WiMAX Forum (2008). WiMAX system evaluation methodology v 2.1.

APPENDIX C JOURNAL TO THE JNCA 2011

In this appendix the journal entitled "A Self-Adapting Connection Admission Control Solution for Mobile WiMAX: Enabling Dynamic Switching of Admission Control Algorithms Based on Predominant Network Usage Profiles” is presented. This journal refers to join of two research topics. The first one considers the scope of this thesis, while the second one regards the thesis Ph.D of Clarissa Cassales Marquezan. Therefore, the join these two research topics introduced a self adapting CAC design able to consider the variable characteristics of the network. The designed incorporates self-* properties to autonomously select the most appropriate CAC algorithm, according to changes on the characteristics of the network. The article presented an architecture to the self-adapting CAC system; two case studies that can benefit from the proposed design; and, the advantages and challenges of the proposal. The magazine reviewers have suggested to add an evaluation of the CAC system proposed. Currently, this paper is under review.

- **Title:**
  Self-adapting Mobile WiMAX networks: Rethinking the Design of Connection Admission Control System
- **Magazine:**
  Journal of Network and Computer Applications (JNCA)
- **URL:**
  `http://journals.elsevier.com/10848045/journal-of-network-and-computer-applications/`
- **Publisher:**
  Elsevier

A Self-Adapting Connection Admission Control Solution for Mobile WiMAX: Enabling Dynamic Switching of Admission Control Algorithms Based on Predominant Network Usage Profiles

Cristiano Bonato Both*, Clarissa Cassales Marquezan, Rafael Kunst, Lisandro Zambenedetti Granville, Juergen Rochol

Institute of Informatics – Federal University of Rio Grande do Sul (UFRGS)
Av. Bento Gonçalves, 9500 – Porto Alegre, Brazil

Abstract

WiMAX is a connection-oriented wireless network that provides QoS in metropolitan broadband communications. One important component in WiMAX QoS provisioning and management is the Connection Admission Control (CAC), which must be aware of the network conditions (*e.g.*, user traffic demands and physical aspects). In our research, we define the association between a particular user traffic demand and a specific physical condition as a *network usage profile*. State-of-the-art proposals focus on optimizing CAC algorithms considering a single network usage profile; the adaptation of CAC algorithms when the predominant network usage profile changes is partially or fully neglected. In this article, we introduce a self-adapting CAC solution that, using a library of CAC algorithms, is able to switch the running algorithm according to the current network usage profile. The evaluation results, obtained through simulations, demonstrate that our self-adapting CAC solution is able to detect the changes on the predominant network usage profile. In addition, the results show how much different profiles can impact on the efficiency of CAC algorithms, thus confirming the need of switching the running CAC algorithm so that QoS can be guaranteed for the ongoing connections.

Keywords: Connection Admission Control, Mobile WiMAX, Self-*

*Corresponding author
Email addresses: cbboth@inf.ufrgs.br (Cristiano Bonato Both), clarissa@inf.ufrgs.br (Clarissa Cassales Marquezan), rkunst@inf.ufrgs.br (Rafael Kunst), granville@inf.ufrgs.br (Lisandro Zambenedetti Granville), juergen@inf.ufrgs.br (Juergen Rochol)

1. Introduction

Worldwide Interoperability for Microwave Access (WiMAX) has been considered a cost-effective alternative for metropolitan broadband Internet access when wired solutions are not geographically or economically viable [1]. WiMAX follows the specifications of the IEEE 802.16 standard [2] and, in practice, the two terms have become synonymous. IEEE 802.16 defines five classes of service [2] as part of its strategy to support Quality of Service (QoS). A key component of WiMAX QoS provisioning solutions is Connection Admission Control (CAC), which is responsible for preventing network overload by rejecting the admission of new connections that exceed network capacity. Although the IEEE 802.16 standard assumes the existence of CAC, it does not define how it should actually be designed or implemented, which opens research opportunities in academia and industry.

CAC solutions employ, inside each WiMAX base station, algorithms to decide whether new connections requested by mobile stations can be admitted or not [3]. Very simple CAC algorithms admit new connections based on the occupancy of the bandwidth allocated to each WiMAX class of service [4]; as long as a class can accommodate the traffic of new connections, these new connections are admitted. Recent investigations, however, go a step further and propose more sophisticated CAC algorithms which decisions also take into account dynamic traffic requirements (*e.g.*, maximum latency and tolerated jitter) and wireless physical aspects (*e.g.*, signal power and channel bandwidth). In this article, we call these the *vertical aspects* that guide the decisions of a CAC solution.

State-of-the-art solutions have taken a number of approaches to address the vertical aspects of CAC. The statistical CAC mechanism proposed by Yu *et al.* [5] considers channel bandwidth over two classes of service. Bashar and Ding [6]'s adaptive and cross-layer CAC solution also operates over two classes of service but adds consideration of the Signal-to-Noise Ratio (SNR) and signal power along with channel bandwidth. Ghazalt *et al.* [7] account for bandwidth over two classes of service but use a probabilistic and self-configuring CAC algorithm that employs a non-conventional fuzzy-based admission control method for downlink connections.

All of these proposed CAC solutions have been designed assuming specific networking scenario conditions, such as user traffic demands and network physical aspects. In WiMAX networks, these conditions typically change because of time-varying traffic requirements and signal propagation. In our research, we define the association between a particular user traffic demand and a specific physical aspect as a *network usage profile*. Over time, a WiMAX network may present a number of usage profiles, although at a given point in time there will be only one *predominant network usage profile*.

Current CAC solutions address a single specific network usage profile; once user demands or physical aspects change (*i.e.*, the network usage profile changes), the particular CAC solution being employed is no longer suitable to the new conditions, *i.e.*, it is not able to self-adapt in favor of a more appropriate behavior aligned to the new predominant network usage profile. In this article, we refer to as the *horizontal aspects* of a CAC solution those that are related to the dynamic management of different usage profiles. Current CAC solutions, despite of properly addressing the vertical aspects, partially or fully neglect the horizontal ones.

Aiming to jointly address both vertical and horizontal aspects, in this article we rethink the design of the traditional WiMAX CAC. We present a novel CAC solution where properties widely referred to as self-* [8] are incorporated into the CAC architecture, thereby enabling autonomous selection of the most appropriate CAC algorithm according to the predominant network usage profile. Our proposal introduces, for example, some new components into traditional CAC architectures: (a) a repository of CAC algorithms from which the most appropriate one can be chosen; (b) an information infrastructure that keeps track of scheduler feedback, traffic requirements, and physical aspects; and (c) algorithms devoted to adapting CAC decisions to different network usage profiles.

We implemented a simulation environment that focused on evaluating the behavior of the proposed solution considering dynamic variations on the network usage profile. The results highlighted the identification of transition between network usage profiles, demonstrated how the change on network usage profiles impacts the acceptance rate of CAC algorithms, and endorsed the need of adapting the type of running CAC algorithm to a suitable predominant network usage profile, so that QoS can be guaranteed for the ongoing connections.

The remainder of this article is organized as follows. In Section 2 we provide the context of this research with a review of WiMAX and CAC. Our proposed self-adapting CAC solution is introduced in Section 3, and its evaluation is presented in Section 4. Finally, in Section 5 we conclude this article presenting final remarks and future work.

2. Background

WiMAX defines connection-oriented wireless networks where each new connection must be first admitted by a WiMAX base station [2]. To decide whether or not to admit a new connection, the base station must analyze information exchanged with mobile stations. The IEEE 802.16 standard defines Ranging and Register messages that provide the information needed by the CAC component to accept or not a network connection request. As shown in Figure 1, Ranging and Register messages are exchanged between the mobile station and the base station through the ranging manager and the connection manager. Details about these messages, the information provided to the CAC component, and the characteristics of such component are presented as follows.

2.1. WiMAX Messages for the CAC Component

Ranging messages are exchanged continuously between the mobile station and the base station. These messages are not associated with any connection request in particular, but with the physical conditions of the wireless channel used by the mobile station. Ranging messages exchange has two phases. In the initial phase, transmission power and time synchronization are negotiated between the mobile and base station, and the mobile station receives a basic Connection IDentification (CID). Periodic ranging begins immediately after initial negotiation between mobile and base station. In this phase the mobile receives a primary CID to exchange two types of messages: a Ranging Request (RNG-REQ) and a Ranging Response (RNG-RSP). These messages are used to handle with time-varying conditions of the wireless channel between mobile and base station. In addition, based on the content of these messages, it is possible to derive other physical information. For example, Signal to Noise plus Interference Ratio (SNIR) is calculated based on the signal power information inside RNG-REQ messages.

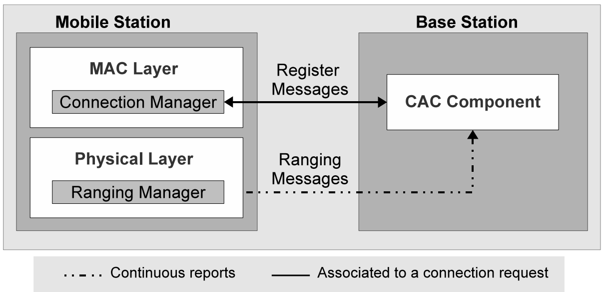


Figure 1: WiMAX messages related to CAC System

Register messages are directly related to requests to admit a new connection. Before establishing a new connection, the mobile station needs to request a registration from the base station. The registration is granted after the exchange of a REG-REQ and a REG-RSP. Register messages enclose traffic descriptors that contain information, such as QoS requirements, associated with the user demands of the new connection. Moreover, these messages can encapsulate three types of actions to manage the QoS requirements of the connections: Dynamic Service Addition (DSA), Dynamic Service Change (DSC), and Dynamic Service Deletion (DSD). Based on these actions, the mobile station can request modifications on its QoS requirements to the base station, to handle changes on traffic conditions.

2.2. Providing Information to the CAC Component

As previously presented, ranging and register messages carry information that is used by the CAC component. Indeed, this information can be used as input parameters for the decision-making process of admitting new connections. Table 1 shows information commonly retrieved from the physical and traffic descriptors carried by the ranging and register messages used by the CAC component.

Table 1: Common Parameters for the CAC Component

Type	Parameters
Physical Descriptor	Power Level
	Available Bandwidth
	Modulation and Coding Scheme (MCS)
	Signal to Noise plus Interference Ratio
Traffic Descriptor	Class of Service
	Minimum Reserved Traffic Rate (MRTR)
	Maximum Sustained Traffic Rate (MSTR)
	Tolerated Jitter (TJ)
	Maximum Latency (ML)

The most important parameter retrieved from the register message is the class of service. The CAC component analyzes all other traffic parameters according to the class of service indicated by the mobile station. Today, WiMAX networks support five classes of services: Unsolicited Grant Service (UGS), extended real-time Polling Service (ertPS), real-time Polling Service (rtPS), non-real time Polling Service (nrtPS), and Best Effort (BE). These classes of service reflect the heterogeneity of network applications.

Current CAC investigations use different combinations of parameters. For example, Lee and Kim [9] use UGS and rtPS classes of service, considering MRTR for each class, MCS, and SNIR. Qin *et al.* [10] employ only one class of service (rtPS), but they consider additional physical and traffic parameters, *i.e.*, MRTR, MSTR, power level, SNIR, available bandwidth, handoff, and MCS. In these works, the parameters used in the CAC component reflect its characteristics. In the next subsection, these characteristics are summarized.

2.3. Characteristics of CAC Component

Before accepting a new connection request, the CAC component must verify whether the available resources would be sufficient to accommodate both the ongoing and the new incoming connections. According to Msadaa, Câmara and Filali [3], to take such decision, two degradation policies can be adopted when no resources are available for the new connections: (i) a flexible policy that reduces the allocation of resources to existing connections in order to manage the requirements of new ones; or (ii) a conservative no-degradation policy that maintains the QoS provided for ongoing connections by simply rejecting new connection requests.

The main goal of policies with degradation of service is to allow the CAC component to admit new connections by reducing the amount of resources provided to the ongoing connections. These policies fall into three categories: (i) service degradation, which decreases the bandwidth assigned to existing connections having lower priority than the priority of the new connection request; (ii) bandwidth borrowing, which decreases the bandwidth assigned to rate-adaptive connections during network congestion; and (iii) bandwidth stealing, in which connection requests are decided based on analysis of a connection's token rate and bucket size. In contrast, when a policy without degradation of services is applied, a new connection request is accepted or rejected based on two requirements: (i) the need to guarantee QoS for the new connection (*e.g.*, real time services should have guarantees of bandwidth, delay, and jitter) and (ii) whether the QoS of ongoing connections should be maintained or not.

Regardless of the type of policy applied, the core of a CAC component is the algorithm that analyzes the parameters, calculates probabilities (*e.g.*, Connection Blocking Probability (CBP), Connection Dropping Probability (CDP), and Handoff Failure Probability (HFP)), and decides whether a new connection should be admitted or not [11]. The type and number of parameters and probabilities used by the algorithm should be related to the objectives for running CAC. Currently, most CAC component for WiMAX networks support the execution of a single network usage profile. This transforms a CAC into a static solution that addresses the vertical aspects discussed before, but which fails to properly handle the dynamic management of horizontal aspects of CAC solutions.

3. Proposal

The main objective of the self-adapting CAC design is to handle both vertical and horizontal aspects associated with WiMAX networks. To achieve this objective, the networking conditions (user demands and physical aspects) that characterize different network usage profiles must be constantly observed at the base station. Whenever necessary, the CAC solution must self-adapt to enhance connection admission decisions. The architecture of our CAC solution is depicted in Figure 2.

Our self-adapting CAC is composed of two levels, operating at the Medium Access Control (MAC) and Physical layers of a base station. The *Running Level* is responsible for ordinary tasks, such as processing incoming frames and providing information to CAC decisions. The *Adaptation Level* performs self-management tasks, such as analyzing networking conditions to eventually switch the running CAC algorithm. These two levels of execution are detailed in the next subsections.

3.1. Running Level

Two components form the running level: *Information Infrastructure* and *CAC*. Each component hosts modules. The *Information Infrastructure* component encloses the *Frame Processing* and *Data Management* modules. The *CAC* component, in its turn, is composed of the *CAC Algorithm* and the *Change Controller* modules.

All incoming frames received in the *Physical Layer* are forwarded to the *Running Level* and processed by the *Information Infrastructure*. Received frames first pass through the *Frame Processing* module, which is responsible for both

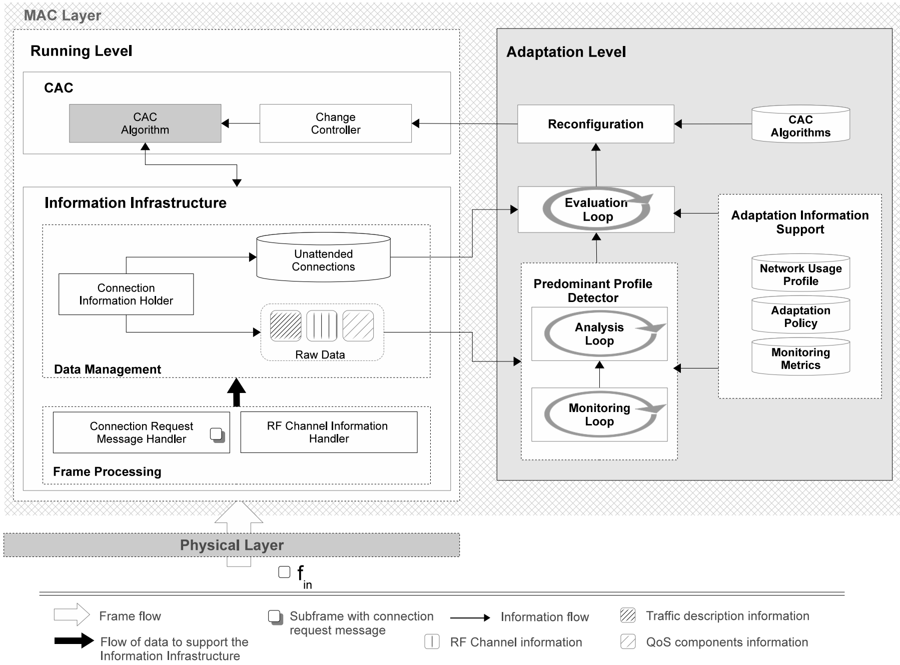


Figure 2: Architecture of Self-adapting CAC Design

identifying connection requests and obtaining Radio Frequency (RF) channel conditions.

Connection request messages, which are processed by the *Connection Request Message Handler*, carry a traffic descriptor that informs the QoS requirements of the incoming connection. The fields of a traffic descriptor are described in Table 2. The *RF Channel Information Handler* is responsible for computing the RF channel conditions of the WiMAX network by processing all incoming frames, regardless they carry a connection request or not. Optionally, the *RF Channel Information Handler* can check the channel conditions by sampling frames in time interval, instead of doing so frame-by-frame. Every connection request carrying QoS requirements detected by the *Connection Request Message Handler*, and every record of the channel conditions computed by the *RF Channel Information Handler* are timestamped so that both requested QoS and channel conditions in a specific moment in time can be related to each other.

After the aforementioned processing activities, data from each frame is forwarded to the *Data Management* module inside the *Information Infrastructure* component. If the received frame contains RF channel feedback information from mobile stations, then the *RF Channel Information Handler* stores the associated data in the *Raw Data* element of the *Data Management* module. If the received frame contains a connection request, then the *Connection Request*

Message Handler forwards such request to the *Connection Information Holder* element inside the *Data Management* module.

Table 2: Traffic Descriptor Fields

Field	Description
CID	Unique identifier of a connection
SFID	Identifies different flows of data carried by a given connection
CoS	Indicates the class of service of a given connection
MSTR	Maximum sustained traffic rate of a connection
MRTR	Maximum reserved traffic rate of a connection
ML	Maximum latency supported by a connection
JT	Jitter limit of a connection
Deadline	Deadline of a VoIP connection
Timestamp	Time of connection arrival

The *Connection Information Holder* provides three basic functions associated with two phases of the CAC decision-making process. The first function takes place before the CAC decision, when the *Connection Information Holder* forwards the connection requests to the CAC algorithm and in parallel holds the information of connection requests until the decision process is finished. The other two functions take place after the CAC decision, when the *Connection Information Holder* receives the CAC decision, and either stores information regarding the accepted connections into the *Raw Data* element or stores information related to the denied connections in the *Unattended Connections* base.

The data stored in the elements *Raw Data* and *Unattended Connections* is used for two purposes. It will be used by the elements of the *Adaptation Level* to identify changes in the predominant network usage profile and eventually trigger adaptations, as explained later in this article; and optionally it can be used by the *CAC Algorithm* in the *Running Level* to provide the decision-making process of such algorithm with more accurate information related to the vertical aspects of CAC solutions.

In our architecture, the *CAC Algorithm* is not tightly coupled with the *Information Infrastructure* component, but it can be designed to benefit from such component. The goal is to enable the compatibility with previous developed CAC algorithms and, at the same time, provide better tools for the design of new ones. Although the *Adaptation Level* is responsible for identifying the need of switching the running CAC algorithm, we believe that it is necessary to define in the *Running Level* an element to materialize the changing process. This element is called *Change Controller* and is designed to cache incoming connection requests while the process of switching the running CAC algorithm is undergoing. In summary, the *Running Level* takes care of the vertical aspects, while the horizontal aspects are delegated to the *Adaptation Level*, which is described as follows.

3.2. Adaptation Level

The objective of the adaptation level is to identify alternations on predominant network usage profile and self-adapt the behavior of the CAC so the admission of connections becomes aligned to the predominant network usage profile. In our architecture, this adaptation is implemented by the following components depicted in Figure 2: *Adaptation Information Support*, *Predominant Profile Detector*, *Evaluation Loop*, *Reconfiguration*, and *CAC Algorithms*. In this article, we focus on the details of the first two components of our architecture.

3.2.1. Adaptation Information Support

The *Adaptation Information Support* is one of the most important components of the *Adaptation Level* because it contains the directives that drive the monitoring, analysis, and evaluation process of the self-adapting mechanism. These directives are in fact data defined by the network administrator that are stored in three modules. The information related to the network usage profiles supported by the base station is stored in the *Network Usage Profile* base. Policies are used to guide the decision of adapting the CAC solution, and they are stored in the *Adaptation Policy*. Finally, monitoring metrics are specifications describing which information should be monitored and which kind of metric is associated to such information. These specifications are defined by the administrator and stored in the *Monitoring Metric* module.

There is an explicit relationship among the *Network Usage Profile*, the *Adaptation Policy*, and the *Monitoring Metrics*, as illustrated in Figure 3. For each network usage profile supported, there must be an adaptation policy and a set of monitoring metrics explicitly associated with the conditions of the adaptation policy (as illustrated in Figure 3). For example, the conditions of the adaptation policy "AP_A" are related to the percentage of BE, rtPS, and ertPS connection requests in a given an interval of time "t"[1]. In this way, it is necessary to define monitoring metrics that will present, within time "t", the total number of connection requests ("MM_1"), the number of BE, ertPS, and rtPS connection requests (respectively, "MM_2", "MM_3", and "MM_4"). One monitoring metric can be associated with more than one adaptation policy (*i.e.*, a network usage profile). For instance, in Figure 3 the monitoring metrics "MM_1" and "MM_4" are associated with both "NUP_Web" and "NUP_VoIP" network usage profiles, without meaning that they must be computed twice, *i.e.*, once the value of the metric is computed it is used in the evaluation of the adaptation policy conditions of all network usage profiles associated with such a metric, as it will be further discussed in the next section.

In our approach, two fields specify the monitoring metric: the type of information (*e.g.*, throughput, each type of class of service, number of handoff requests) and the type of statistics used to generate the monitoring metric (*e.g.*, sum, average, last value). The supported types of information and statistics

[1]The interval of time "t" is actually the interval between monitoring cycles of the Predominant Profile Detector module.

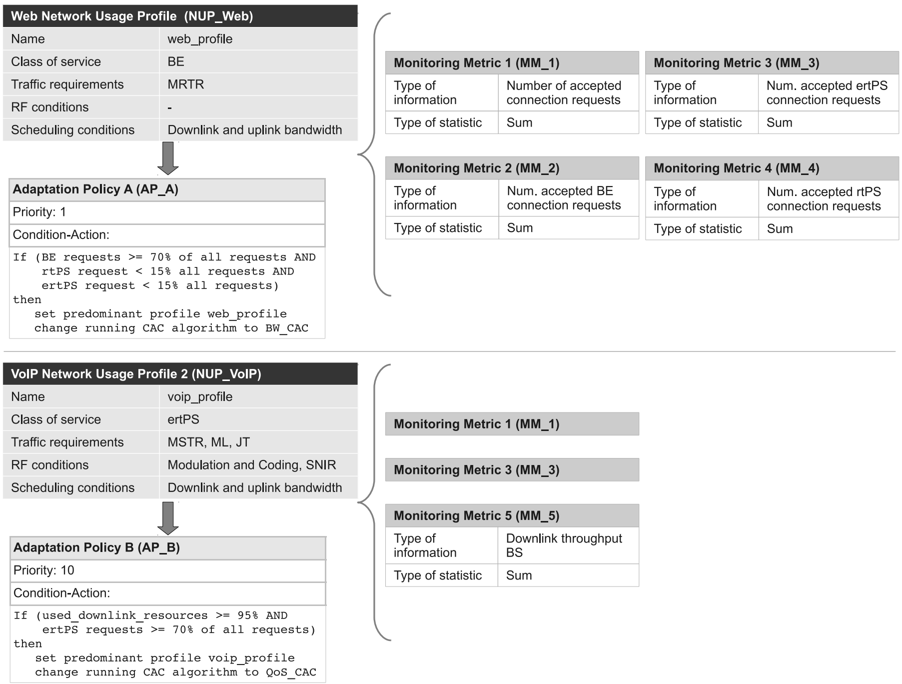


Figure 3: Examples of data associated with the Adaptation Information Support

are directly associated with the information stored in the *Raw Data* element of the *Data Management* module of the *Information Infrastructure* component in the *Running Level*. For example, in Figure 3, the monitoring metric "MM_1" is associated with "Traffic description information" inside the *Raw Data* element (Figure 2), while "MM_5" is related to the "QoS components information" also inside the *Raw Data* element (Figure 2).

3.2.2. Predominant Profile Detector

Based on the monitoring metric specifications, adaptation policies, and actual information retrieved from the *Raw Data* element, as illustrated in Figure 4 the *Predominant Profile Detector* component is able to monitor and identify the updated predominant network usage profile. To achieve these two actions, two modules are defined and depicted as follows.

The *Monitoring Loop*, detailed in Algorithm 1 uses a set of monitoring metrics ($MonMetrics$) to search information in the *Raw Data* and calculate (using for this the function *calc*()) the average value of the monitoring metric (line 7 of Algorithm 1) within an interval t of a monitoring loop. In our approach, both current and historical information are considered in order to generate the average value of each monitoring metric within the cycle t. In addition, we defined

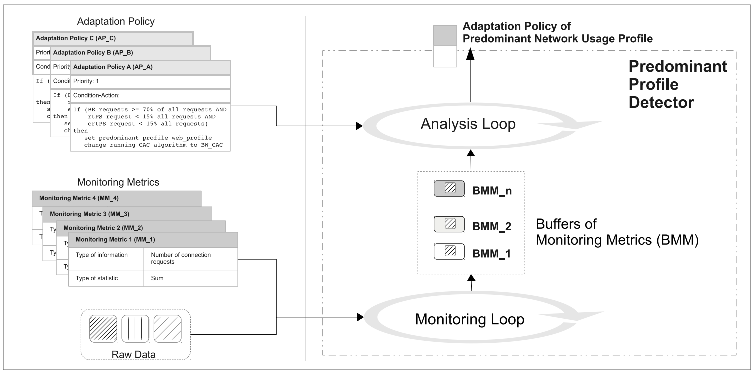


Figure 4: Overview on the Predominant Profile Detector module

weights that adjust how much emphasis the current and historical monitored information have on the monitoring process. Line 7 of Algorithm 1 depicts how the weight ρ is related to the current information (generated by $calc(MM\_i)$) and the historical data ($BMM\_j.value$). Current and historical information are merged to generate the average value of the monitoring process to reduce the chances of starting adapting processes because of an instantaneous, punctual change on the network usage profile. Finally, the average value of each metric is stored in its respective monitoring metric buffer ($BufferMM$ represents the set of those buffers BMM_j, as presented in Algorithm 1).

Algorithm 1 - Monitoring Loop of Predominant Profile Detector Module

Require: $MonMetric = \{MM_1, MM_2, ..., MM_i\}$, where MM_i is $[type, statistic]$
Require: $BufferMM = \{BMM_1, BMM_2, ..., BMM_j\}$, where BMM_j is $[type, value]$
Require: $i = j$, where $i > 0$
Require: ρ, where $0 \leq \rho \leq 1$
Ensure: $Calc(MM)$ function that returns the value of a monitoring metric

1: **loop**
2:     **for all** $MM_i \in MonMetric$ **do**
3:         **if** $init_i$ **then**
4:             $BMM_j.value \leftarrow calc(MM_i)$
5:             $init_i \leftarrow false$
6:         **else**
7:             $BMM_j.value \leftarrow (\rho.calc(MM_i) + ((1-\rho).BMM_j.value))/2$
8:         **end if**
9:     **end for**
10:     Wait for the next monitoring cycle t
11: **end loop**

The *Analysis* component, presented in Algorithm 2, uses the information from $BufferMM$ and the adaptation policies ($AdaptPolicy$) in order to identify matches among the values of the monitoring metrics (stored in $BufferMM$) and the conditions of the adaptation policies ($AP\_i.Conditions$), as presented in line 3 of Algorithm 2. In essence, the match between conditions and the monitoring metrics denotes the identification of a potential set of predominant network usage profile (stored at $PredominantProfileSet$ in line 4 of Algorithm 2). We consider that eventually more than one match could be found during the analysis cycle y. For this reason, we check the priority of the adaptation policies in the $PredominantProfileSet$ (lines 7 - 14 of Algorithm 2) and select the adaptation policy with the highest priority, *i.e.*, the predominant network usage profile during the analyzed cycle y, and inform the Evaluation component of the Predominant Profile Detector module.

Algorithm 2 - Analysis Loop of Predominant Profile Detector Module

Require: $AdaptPolicy = \{AP_1, AP_2, ..., AP_i\}$, where AP_i is $[Priority, Conditions, Actions]$
Require: $PredominantProfileSet \subseteq AdaptPolicy$
Require: $BufferMM = \{BMM_1, BMM_2, ..., BMM_k\}$
Ensure: $match()$ function that returns the match between the monitoring metrics and conditions of an adaptation policy
Ensure: $check\_priority()$ function that returns the adaptation policy with highest priority

1: **loop**
2:   **for all** $AP_i \in AdaptPolicy$ **do**
3:     **if** $match(BufferMM, AP_i.Conditions)$ **then**
4:       $PredominantProfileSet := AP_i$
5:     **end if**
6:   **end for**
7:   **if** $|PredominantProfileSet|! = 0$ **then**
8:     **if** $|PredominantProfileSet| > 1$ **then**
9:       $predominant\_profile := check\_priority(PredominantProfileSet)$
10:     **else**
11:       $predominant\_profile := PredominantProfileSet$
12:     **end if**
13:     Inform to the Evaluation Module about $predominant\_profile$
14:   **end if**
15:   Wait for the next analysis cycle y
16: **end loop**

3.2.3. Further Components

It is not the focus of this paper to discuss the details of the *Evaluation Loop*, *Reconfiguration*, and *Change Controller* components of the *Adaptation Level*. Nevertheless, this section presents their brief description as follows.

The tasks of identifying changes on predominant profiles and verifying the viability of adapting the CAC solution are actually performed by the *Evaluation Loop* component. To determine whether an adaptation is required, the *Evaluation Loop* component checks the reliability of the chosen predominant profile by looking for evidences on the CAC decisions stored in the *Unattended Connections* (Figure 2). Finally, it is also responsibility of this module to evaluate the costs involved to execute the adaptation and, in the end, the adaptation is performed only if there is a favorable cost-effective tradeoff.

In the case when an adaptation is finally required, the *Reconfiguration* component enforces the new predominant network usage profile. Thus, the *Running Level* is prepared to switch the running CAC algorithm by the one associated with the new predominant network usage profile. The new algorithm is retrieved from the pool of *CAC Algorithms*. The Change Controller, as discussed before, must avoid the interruption of admitting or not connections. The switching process between the running and the new CAC algorithms is key to the overall performance of the self-adapting CAC solution. In next section we present and discuss results regarding to the evaluation of the Predominant Profile Detector module of the proposed self-adapting CAC design.

4. Evaluation

The self-adapting CAC solution proposed in this article deals with changes of the predominant network usage profile. Because vertical aspects, as discussed before, are largely observed in the literature, in this section we focus on evaluating the horizontal aspects associated with CAC solutions. The evaluation here conducted has two major objectives: demonstrate the monitoring capacity of the Predominant Profile Detector component from our self-adapting CAC proposal, and demonstrate how distinct CAC strategies and different predominant network usage profile impact the WiMAX base station performance in terms of connection request acceptance. In order to understand the evaluation, we first describe the design of the simulation environment in Section 4.1 and then the analysis of the results in Section 4.2.

4.1. Simulation Environment

In the description of the simulation environment, there are three main groups of information that are important: types of CAC algorithms, physical parameters and types of connection requests model, and types of network usage profiles. The types of CAC algorithms are described together with the design of the simulator in Section 4.1.1, and those types are important for the evaluation of static CAC strategies versus varying network usage profiles. The physical parameters of the simulation and types of connection requests model used to simulate the traffic inside the simulation environment are described in Section 4.1.2. Finally, the types of network usage profiles determine the distribution of the number of connection requests arrival per connection request model within the simulation time as depicted in Section 4.1.3.

4.1.1. Design of the Simulator

The self-adapting CAC proposed in this article is part of a QoS-enable architecture designed to operate in a WiMAX base station (which is out of the scope of this paper, and under review in another article). In order to obtain simulation results, this architecture, which is also composed of Scheduler and Allocator components, as proposed by Cohen [12], was implemented in the simulator. In the proposed simulator design, Scheduler component is responsible for two functions. The first one consists in ordering the data of each connection within the class of service queues. The second one aims to select in which order the queues must be served according to the priority of each class of service. The Allocator component, in turn, aims to arrange the data within the frames.

The first function of the scheduler uses the Weighted Proportional Fair (WPF) algorithm [13], because this algorithm can improve resource allocation, fairness, and QoS guarantees, by assigning a weight for data in the queues. The second function, uses the Proportional Fair (PF) algorithm [14] in order to serve ertPS, rtPS, and BE classes of service in this order of priority to reproduce the traffic demand of a cellular system. Moreover, the implemented allocation algorithm was Recursive Tiles and Stripes (RTS) because it presents the best performance if compared with other allocation algorithms found in the literature [15]. Further information about these components of the architecture are described in articles which are currently under review.

Although the Scheduler and Allocator are crucial components to provide QoS, towards the goals of this article, we focus only on analyzing the behavior of the CAC component. Therefore, we defined two simple algorithms to be used in the evaluation of the impacts caused by a static/fixed CAC algorithms when the network usage profile changes. The first one is called BW_CAC and is depicted in Algorithm 3, whose major goal is to accept all connection requests while there are available resources, *i.e.* bandwidth in the downlink frame. The second is presented in Algorithm 4 and is called QoS_CAC, whose acceptance criteria is based on the conformance of QoS requirements (bandwidth, delay, and jitter) for the specific class of service of the incoming connection request. For example, in the case of an incoming connection belonging to ertPS class of service, it will be accepted only if there is available bandwidth, and both the delay and jitter requirements can be guaranteed.

Algorithm 3 BW_CAC Algorithm

Require: $MSTR\_ertPS$
Require: $MSTR\_rtPS$
Require: $MRTR\_BE$
Ensure: $total\_dl\_res()$ returns the total amount of bits supported by the downlink subframe
Ensure: $used\_dl\_res()$ returns the amount of bits already used in the downlink subframe
if $class\_of\_service = "ertPS"$ **then**
    if $(used\_dl\_res() + MSTR\_ertPS) \leqslant total\_dl\_res()$ **then**
        Accept the connection request
    else
        Reject the connection request
    end if
else if $class\_of\_service = "rtPS"$ **then**
    if $(used\_dl\_res() + MSTR\_rtPS) \leqslant total\_dl\_res()$ **then**
        Accept the connection request
    else
        Reject the connection request
    end if
else
    if $(used\_dl\_res() + MSTR\_BE) \leqslant total\_dl\_res()$ **then**
        Accept the connection request
    else
        Reject the connection request
    end if
end if

Algorithm 4 QoS_CAC Algorithm

Require: $MSTR\_ertPS$
Require: ML_ertPS
Require: JT_ertPS
Require: $MSTR\_rtPS$
Require: JT_rtPS
Require: $MRTR\_BE$
Ensure: $total\_dl\_res()$ returns the total amount of bits supported by the downlink frame
Ensure: $used\_dl\_res()$ returns the amount of bits already used in the downlink subframe
Ensure: $current\_jitter(class\_of\_service)$ returns the average jitter of a specific class of service
Ensure: $current\_delay(class\_of\_service)$ returns the average delay of a specific class of service

if $class\_of\_service = "ertPS"$ **then**
    if $((used\_dl\_res() + MSTR\_ertPS) \leqslant total\_dl\_res())$ AND $(current\_delay(ertPS) \leqslant ML(ertPS))$ AND $(current\_jitter(ertPS) \leqslant JT_ertPS)$ **then**
        Accept the connection request
    else
        Reject the connection request
    end if
else if $class\_of\_service = "rtPS"$ **then**
    if $((used\_dl\_res() + MSTR\_rtPS) \leqslant total\_dl\_res())$ AND $(current\_jitter(rtPS) \leqslant JT_rtPS)$ **then**
        Accept the connection request
    else
        Reject the connection request
    end if
else
    if $(used\_dl\_res() + MRTR\_BE) \leqslant total\_dl\_res()$ **then**
        Accept the connection request
    else
        Reject the connection request
    end if
end if

4.1.2. Physical Parameters and Connection Request Model

The parameters associated with the physical aspects of IEEE 802.16 networks considered in our simulations are summarized in Table 3. These parameters are based on typical values found in the literature [16] and in a simulation methodology suggested by the WiMAX forum [17].

Table 3: Physical Simulation Parameters

Parameters	Values
Channel bandwidth	10 MHz
FFT total size (subcarrier)	1024
FFT data (subcarrier)	720
Subframe symbols (DL:UL)	29:18
OFDMA subchannels (DL:UL)	30:6
PUSC zone in slots (DL:UL)	420:90
Frame duration	5ms

In addition, the connection requests that arrives to the CAC component are modeled in order to simulate the connections belonging to the three classes of services considered in this article. Consequently, three connection request models are defined to associate the characteristics of ertPS, rtPS, and BE classes of service with VoIP, Video Clip, and Web applications, respectively. Such association is based on the work presented by So-In, *et al.* [18]. The values of all QoS parameters that compose the traffic descriptor, *i.e.* MRTR, MSTR, ML, and JT, are summarized in Table 4.

Table 4: QoS Parameters Values

	MSTR	**MRTR**	ML	JT
VoIP	16.4Kbps	N/A	Truncated Exponential Mean = 70ms Min = 5ms Max = 150ms	Truncated Exponential Mean = 30ms Min = 5ms Max = 50ms
Video Clip	Truncated Lognormal Max = 2Mbps	Truncated Lognormal Max = 5Kbps	N/A	Truncated Exponential Mean = 50ms Max = 100ms
Web	Truncated Lognormal Max = 2Mbps	Truncated Lognormal Max = 10Kbps	N/A	N/A

The QoS parameters MSTR and MRTR are modeled using the Truncated Lognormal distribution with $\sigma = 1.37$ and $\mu = 8.37$ to the PDF values due to the characteristic these parameters, *i.e.* bits unit. On the other hand, ML and JT are modeled using the Truncated Exponential distribution because these parameters are in temporal units. Moreover, some QoS parameters are not applicable (N/A) for the traffic descriptor of a specific application.

4.1.3. Simulated Testbed Scenario

Based on the connection requests models we defined a simulation testbed with two network usage profiles. As previously discussed, the simulation of network usage profiles relies on the definition of a distribution that indicates the amount of connection requests for each connection request model that will be generated within the duration of the simulation.

One network usage profile is called *Web Network Usage Profile* (NUP_Web), and its behavior is characterized by distribution of the connection requests per connection model, as follows: 70% of Web connection requests and 15% for both VoIP and Video. This profile represents the typical Internet traffic [19]. The monitoring metrics and adaptation policy related to this profile are the same ones illustrated in Figure 3. The second network usage profile is called *VoIP Network Usage Profile* (NUP_VoIP) [20], and it is characterized by the following distribution of connection requests: 70% of VoIP and 15% for both Video and Web traffic connection models. The monitoring metrics and adaptation policy of this profile are the same ones illustrate in Figure 3.

4.2. Result Analysis

We assume a single-cell WiMAX network and analyze snapshots of the system performance. The combination of the network usage profiles with the different CAC algorithms is explored in our simulation and detailed in the next sections. In addition, all simulations have been carried out to guarantee a confidence interval of 95%.

4.2.1. Monitoring Capability of Predominant Profile Detector

The objective of this section is to demonstrate the monitoring capabilities of the Monitoring Loop module from the Predominant Profile Detector component (illustrated in Figure 4), and the importance of the monitoring strategies for further detection of network usage profile changes. To achieve this goal we simulated, in the initial 10s, a total of 800 connection requests using for this the distribution of connection requests according to the VoIP network usage profile (NPU_VoIP); and in the final 10s, a total of 800 connection requests but now following the Web network usage profile (NUP_Web).

The curves in Figure 5 depict the actual value of the observed monitoring metrics within each monitoring cycle (where $t = 2$) when the BW_CAC algorithm is employed. The observed metrics are: "MM_2" (*i.e.*, BE accepted connections named in the graphics "Web"), "MM_3" (*i.e.*, ertPS accepted connections named in the graphics "VoIP"), and "MM_4" (*i.e.*, rtPS accepted connections named in the graphics "Video Clip").

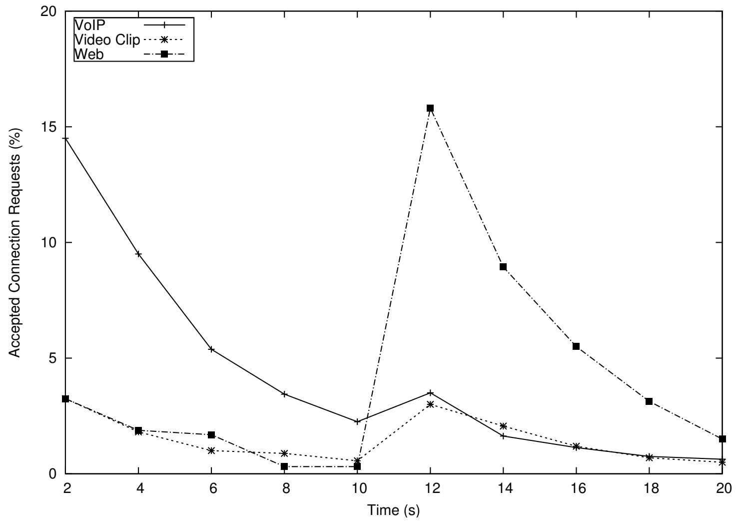
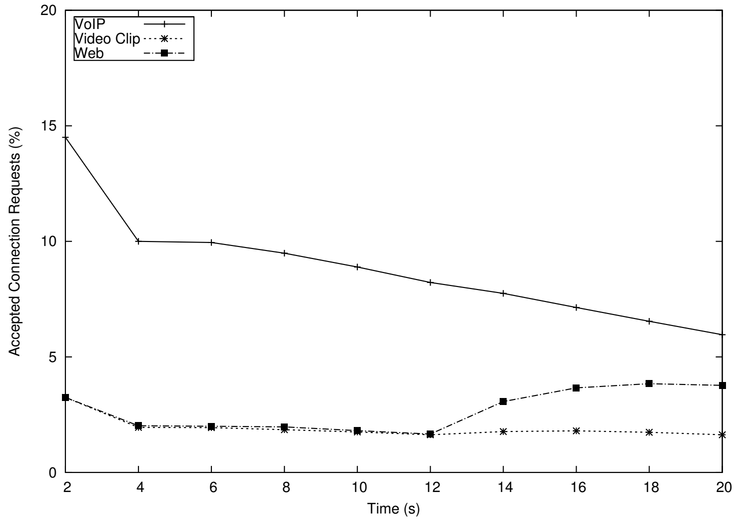
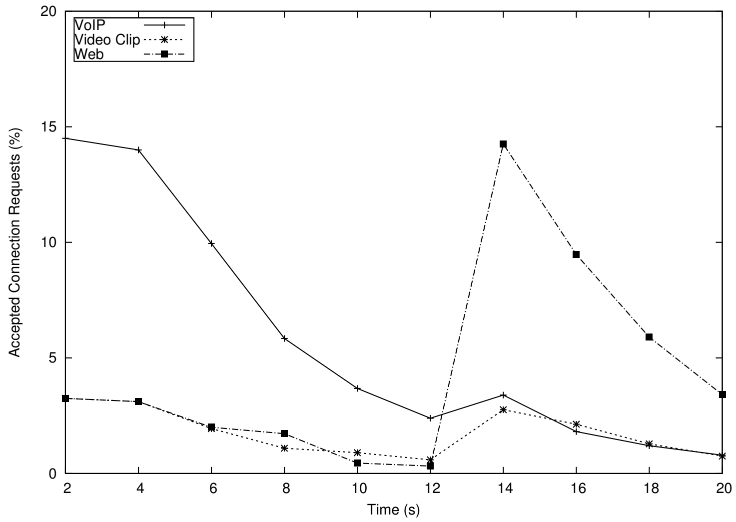


Figure 5: Monitoring Metrics Value under BW_CAC algoritm employment

Figure 6 presents the output value of the monitoring loop after processing each observed monitoring metric according to the execution of Algorithm 1, *i.e.*, the average between historical and current values of the monitoring metrics according to the smoothing factor ρ. For the curves in Figure 6, we used the extremes cases, when the historical values play the major role, as illustrated in Figure 6(a), and when the current values are the most important, as illustrated in Figure 6(b).



(a) Output with $\rho = 0.9$ for historical data



(b) Output with $\rho = 0.9$ for current data

Figure 6: Output values of monitoring loop using BW_CAC

Using as example the VoIP monitoring metrics and output, respectively in Figure 5 and Figure 6(a), we can observe that an excessive attribution of weight for the historical data can create situations that might compromise the detection of changes on network usage profile. For example, at the simulation time $t = 14$, according to Figure 6(a), the highest percentage of accepted connections belongs to the VoIP connection request model (approximately 8%). However, Figure 5, shows that the current highest percentage of accepted connections belongs to the Web connection request model. In this sense, the use of higher weight in the current information, as illustrated in Figure 6(b), can keep the approximation to the current behavior.

We believe that it is important to understanding how much historical versus current values influence the monitoring processes that will drive CAC self-adapting actions. We believe that in the case of CAC vertical aspects the current values typically should have higher weights so that punctual and dynamic changes in the RF channel conditions could be detected and the parameters of the CAC adapted. However, in the context of realizing the behavior of user's demand (*i.e.*, network usage profiles), historical values gain importance, and the simulations here presented are the start point for this better understanding.

4.2.2. CAC Strategies x Predominant Network Usage Profiles

In this subsection we present an evaluation of the impacts caused in the amount of connection requests accepted by a CAC algorithm when the network usage profile changes. We examine the details of the connection request acceptance associated with each CAC algorithm (*i.e.*, BW_CAC and QoS_CAC) for

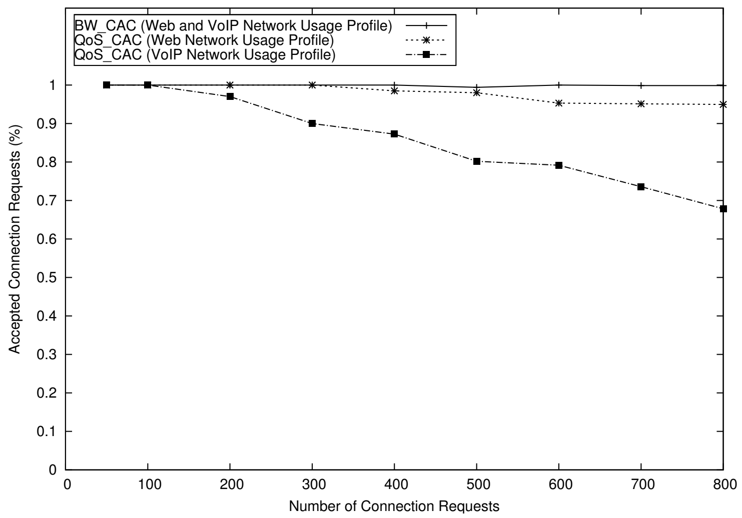


Figure 7: Overall performance of the CAC algorithms

each one of the network usage profiles here considered (VoIP and Web network usage profiles). In Figure 7, we present the overall efficiency of BW_CAC and QoS_CAC algorithms considering VoIP and Web network usage profiles in a scenario that accommodates up to 800 connection requests.

We observe that the BW_CAC algorithm accepts all connection requests in VoIP and Web network usage profiles. Accepting all connection occurs because the BW_CAC algorithm does not guarantee QoS for the accepted connections, since delay and jitter parameters are not taken into account. Another important aspect to be observed in this context is the amount of rejected connection requests when the QoS_CAC algorithm is running. The rejection is of approximately 5% in the Web network usage profile and reaches more than 30% when the network usage profile is VoIP.

We recognize, in Figure 7, that up to 100 connection requests the difference of the network usage profiles does not affect the acceptance of the CAC algorithms. Above that, however, the amount of accepted requests is reduced in the case of the QoS_CAC algorithm because of the large amount of real time connection requests, *i.e.*, 85% of the requests require QoS guarantees (70% VoIP plus 15% Video Clip).

We can also conclude that when the predominant traffic is Web (*i.e.*, there is not a major need to provide QoS guarantees), the QoS aware algorithm should not be used because it reduces the amount of accepted connection requests in about 5%. Nevertheless, when real time traffic is predominant, then the CAC algorithm should be switched to another one able to guarantee QoS for the accepted connections.

5. Conclusions and Future Work

The proposed self-adapting CAC solution is able to adapt to the heterogeneity of WiMAX networks by considering both the vertical and horizontal aspects defined in this article. The major benefits introduced by our approach are: the capacity to dynamically identify changes on the user demands by managing the alternation of network usage profiles; a differentiation between algorithms of running level (CAC operational tasks) and adaptation level (management tasks related to the network usage profiles); and the support to switch the running CAC algorithm without manual intervention.

The results, obtained through simulations, emphasize the need of self-adapting CAC in respect to the horizontal aspects, *i.e.*, the changes on the predominance of different network usage profiles. We demonstrated with the simulations that our self-adapting CAC solution is able to detect the changes on the current profile. We also observed how much the predominant network usage profile can impact the efficiency of CAC algorithms.

As future work, we plan to investigate essential network usage profiles that must be supported by a base station, and the identification of the most suitable physical architecture for the deployment of the self-adapting CAC solution. Furthermore, it is necessary to better understand the thresholds between historical x current weights in face of very dynamic and constant changes in the predominant network usage profiles in order to define situation where historical or current weights should be considered for switching the CAC algorithm.

References

[1] Y. A. Sekercioglu, M. Ivanovich, A. Yegin, A survey of MAC based QoS implementations for WiMAX networks, Computer Networks 53 (14) (2009) 2517–2536.

[2] IEEE standard for local and metropolitan area networks - part 16: Air interface for fixed and mobile broadband wireless access systems - IEEE std. 802.16-2005 (December 2005).

[3] I. C. Msadaa, D. Câmara, F. Filali, Scheduling and CAC in IEEE 802.16 fixed BWNs: A comprehensive survey and taxonomy, IEEE Communications Surveys & Tutorials 12 (4) (2010) 459–487. doi:10.1109/SURV.2010.033010.00038.

[4] B. Zhu, K. Xue, H. Lu, P. Hong, Fair connection admission control scheme for IEEE 802.16e systems, Vol. 1, IEEE, 2008, pp. 1–4. doi:10.1109/WiCom.2008.736.

[5] K. Yu, X. Wang, S. Sun, L. Zhang, X. Wu, A statistical connection admission control mechanism for multiservice IEEE 802.16 network, in: Proceedings of IEEE 69th Vehicular Technology Conference, 2009, pp. 1–5. doi:10.1109/VETECS.2009.5073347.

[6] S. Bashar, Z. Ding, Admission control and resource allocation in a heterogeneous OFDMA wireless network, IEEE Transactions on Wireless Communications 8 (8) (2009) 4200–4210. doi:10.1109/TWC.2009.080844.

[7] S. Ghazalt, Y. H. Aout, J. B. Othman, F. Nait-Abdesselam, Applying a self-configuring admission control algorithm in a new QoS architecture for IEEE 802.16 networks, in: Proceedings of IEEE Symposium on Computers and Communications, 2008, pp. 1023–1028. doi:10.1109/ISCC.2008.4625606.

[8] N. Samaan, A. Karmouch, Towards autonomic network management: an analysis of current and future research directions, IEEE Communications Surveys & Tutorials 11 (3) (2009) 22–36. doi:10.1109/SURV.2009.090303.

[9] J. Y. Lee, K. B. Kim, Statistical connection admission control for mobile WiMAX systems, in: Proceedings of IEEE Wireless Communications and Networking Conference, 2008, pp. 2003–2008. doi:10.1109/WCNC.2008.356.

[10] C. Qin, G. Yu, Z. Zhang, H. Jia, A. Huang, Power reservation-based admission control scheme for IEEE 802.16e OFDMA systems, in: Proceedings of IEEE Wireless Communications and Networking Conference, 2007, pp. 1831–1835. doi:10.1109/WCNC.2007.344.

[11] M. H. Ahmed, Call admission control in wireless networks: a comprehensive survey, IEEE Communications Surveys & Tutorials 7 (1) (2005) 49–68.

[12] R. Cohen, L. Katzir, Computational analysis and efficient algorithms for Micro and Macro OFDMA downlink scheduling, IEEE/ACM Transactions on Networking 18 (1) (2010) 15–26. doi:10.1109/TNET.2009.2022937.

[13] F. Hou, J. She, P.-H. Ho, X. Shen, Performance analysis of weighted proportional fairness scheduling in IEEE 802.16 networks, in: Proceedings. IEEE ICC - International Conference on Communications, Beijing, 2008, pp. 3452–3456. doi:10.1109/ICC.2008.649.

[14] M. Shreedhar, G. Varghese, Efficient fair queuing using deficit round-robin, ACM Transactions on Networking 4 (1) (1996) 375–385. doi:10.1145/217382.217453.

[15] C. Cicconetti, L. Lenzini, A. Lodi, S. Martello, E. Mingozzi, M. Monaci, Efficient two-dimensional data allocation in IEEE 802.16 OFDMA, in: Proceedings. IEEE INFOCOM - Conference on Computer Communications, San Diego, 2010, pp. 1–9. doi:10.1109/INFCOM.2010.5462040.

[16] F. Wang, A. Ghosh, C. Sankaran, P. Fleming, F. Hsieh, S. Benes, Mobile WiMAX systems: performance and evolution, IEEE Communications Magazine 46 (10) (2008) 41–49. doi:10.1109/MCOM.2008.4644118.

[17] WiMAX Forum, WiMAX system evaluation methodology version 2.1, WiMAX Forum. in http://www.cse.wustl.edu/~jain/wimax/ftp/wimax_system_evaluation_methodology_v2_1.pdf. Accessed in August 2011. (July 2008).

[18] C. So-In, R. Jain, A.-K. Tamimi, Capacity evaluation for IEEE 802.16e mobile WiMAX, Journal of Computer Systems, Networks, and Communications 2010 (1) (2009) 1–12. doi:doi:10.1155/2010/279807.

[19] H. Schulze, K. Mochalski, Internet study 2008/2009, IPOQUE. in http://www.ipoque.com/userfiles/file/ipoque-Internet-Study-08-09.pdf. Accessed in April 2011. (July 2009).

[20] A. Estepa, R. Estepa, J. Vozmediano, A new approach for VoIP traffic characterization, IEEE Communications Letters 8 (10) (2004) 644–646. doi:doi:10.1109/LCOMM.2004.835318.