QOS SCHEDULING IN IEEE802.16/WIMAX NETWORKS

Mohammad Hamidullah Ahmad

A thesis

in

The Department

of

Computer Science & Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Computer Science

Concordia University

Montréal, Québec, Canada

June 2011

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Mohammad Hamidullah Ahmad

Entitled: QoS Scheduling in IEEE802.16/WiMAX Networks

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________________ Chair
Dr. Y. Yan

______________________________________ Examiner
Dr. L. Narayanan

______________________________________ Examiner
Dr. T. Fancott

______________________________________ Supervisor
Dr. B. Jaumard

Approved by ______________________________________
Chair of Department or Graduate Program Director

____
Dr. Robin A.L. Drew, Dean
Faculty of Engineering and Computer Science

Date ______________________________________

Abstract

QoS Scheduling in IEEE802.16/WiMAX Networks

Mohammad Hamidullah Ahmad

WiMAX/IEEE802.16 is a major broadband wireless access technology that provides an affordable alternative to conventional wired techniques in supporting high-speed data, video and voice services. It is quite challenging to provide Quality of Service (QoS) in wireless networks, which guarantees different service classes meeting specific requirements. In our current work, we have presented a bandwidth scheduling algorithm to facilitate end-to-end QoS in IEEE802.16/WiMAX network. A differentiated service (DiffServ) approach is adopted, which serves QoS provisioning by differentiating multiple traffic classes into individual service flows. There exists a number of well-known traffic scheduling algorithms for scheduling wireless traffic. However, a few of them can differentiate services and, in the literature, attempts have been made to use different scheduling algorithms for different services to meet the service class constraints. This approach of having different algorithms dealing different traffic classes leads to increased scheduler design complexity and eventual challenging implementation. Furthermore, the majority of these available algorithms assume the preexistence of Call Admission Control and thus describes the scheduling and resource allocation algorithm without having any interactivity with Admission Control mechanism. This approach is unrealistic since scheduling and admission control are strongly related in QoS provisioning. In this thesis, a novel scheduling algorithm has been proposed in combination with a Call Admission policy for WiMAX OFDMA/TDD systems, utilizing dynamic preference function that determines transmission order and bandwidth allocation on sub-carrier basis. The dynamic preference function takes into account the required throughput of a request, the current data rate, the expiry of the remaining bits and the signal strength to dynamically determine the best subcarrier through the usage of adaptive modulation and coding. The QoS requests get allocation in the order of the best subcarrier available; that is, the highest priority QoS requests are allotted the best subcarriers with greater bandwidth capacity in an OFDMA symbol, whereas the lowest priority requests get allocation of lower quality subcarriers, if required more than one of them to maintain the required data rate. Consequently, in our method the lower priority QoS requests also have a chance of being served in every round of scheduler execution. Thus our algorithm allows fairness for all classes of requests, also maintaining priorities for the higher classes of traffic. System performance is evaluated using simulations. Numerical results show that the proposed scheduling algorithm is efficiently able to handle multi-class QoS in WiMAX uplink systems with fair allocation of available resources among the admitted requests.

Acknowledgments

I would like to thank the Almighty God for His divine guidance and blessings bestowed upon me throughout my life and particularly during this time that have made this thesis possible.

I am grateful to my supervisor, Dr. Brigitte Jaumard who provided guidelines on how to select the research topic and progress with the research work. Her understanding, encouragement, personal guidance, and great efforts to explain things clearly and simply to me, have provided a good basis for this thesis work.

I also wish to thank all the faculty members and staff of the Computer Science Department. I am grateful to the Faculty of Engineering and Computer Science, Concordia University for supporting this thesis work.

My special gratitude goes to my friends and family. Without their encouragement and mental support it would have been impossible for me to continue and complete this long journey.

Finally, I would like to dedicate my thesis to my parents, who have always been an immense source of support and encouragement in a way that none can ever be.

Contents

List of Figures

List of Tables

Acronyms

AMC	Adaptive Modulation and Coding
BE	Best Effort
BER	Bit Error Ratio
BR	Bandwidth Request
BS	Base Station
CBR	Constant Bit Rate
CID	Connection Identifier
CS	Convergence Sublayer
CQI	Channel Quality Information
DSA	Dynamic Service Addition
DSC	Dynamic Service Change
DSD	Dynamic Service Deletion
ertPS	Extended Real-Time Polling Services
FDD	Frequency Division Duplex
FFT	Fast Fourier Transform
FTP	File Transfer Protocol
IEEE	Institutes of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
LTE	Long Term Evolution
MAC	Medium Access Control
MPEG	Moving Pictures Experts Group
MR-BS	Multihop Relay Base Station

MS	Mobile Staition
nrtPS	Non-Real-Time Polling Services
OFDMA	Orthogonal Frequency Division Multiple Access
PDU	Payload Data Unit
PMP	Point to Multipoint
PS	Physical Slot
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase-Shift Keying
RSSI	Receive Signal Strength Indicator
rtPS	Real Time Polling Services
SFID	Service Flow Identifier
SNR	Signal-to-Noise Ratio
SS	Subscriber Station
SSID	Subscriber Station Identification (MAC address)
SUI	Stanford University Interim Models
TDD	Time-Division Duplex
TDM	Time-Division Multiplexing
UGS	Unsolicited Grant Services
VoIP	Voice over IP
WiMAX	Worldwide Interoperability for Microwave Access

Chapter 1

Introduction

1.1 Motivation

Envisioned to provide Broadband Wireless Access (BWA) in Wireless Metropolitan Area Networks (Wireless MAN), IEEE 802.16 standard has now emerged into a technology to realize what could be called 'pervasive high-speed mobile Internet' [Tec]. The standard has been coined the industry name Worldwide Interoperability for Microwave Access (WiMAX) by vendor interoperability organization, WiMAX forum comprising more than 500 members, and has been used in the literature interchangeably since. The foremost advantage of WiMAX/IEEE802.16 technology is the fast and reduced-cost deployment compared to its counterpart such as digital subscriber lines and cable modems necessitating a wired infrastructure that sometimes becomes inconvenient to build or too costly in suburban or rural areas. The other advantages of WiMAX/IEEE802.16 are variable and high data rates, a large frequency range, strong Quality of Service (QoS) capabilities and a large network coverage.

In the current standard [WiM09] supported Point-to-Multipoint (PMP) mode of transmission, a base station (BS) is responsible for allocating radio resources and transmission schedules for the subscriber stations or mobile stations (SS/MS). In a WiMAX network, the bandwidth allocation policy needs to manage a scarce radio resource that is shared among the SS/MS. A scheduling scheme is also needed at the Medium Access Control (MAC) layer, in order to efficiently coordinate traffic flows with different service class requirements. The standard defines the QoS mechanism, signalling and associated parameter set but leaves the resource allocation strategy, admission control and scheduling policy open for vendor implementation suitably devised for their respective infrastructures. This, along with the fact that channel properties of QoS requests originating from different MS/SS are diverse, has motivated the research community to design an efficient MAC layer scheme that improves system performance taking into consideration the channel condition and traffic request diversity. This will be the focus of my research project.

1.2 Research Challenges

The design of scheduling and resource allocation algorithms is challenging in WiMAX network communications due to the inherent limited capacity and dynamic channel state of wireless communication. The design of an efficient QoS bandwidth scheduling and resource allocation algorithm needs to be concerned with the following features [FL02], [Zha95] [GP99]:

- Efficient channel utilization: One of the foremost requirements is that the algorithm should utilize the channel effectively. This implies that the scheduler should avoid assigning a transmission slot to a request currently having a bad link property.

- Satisfying QoS requirements: The algorithm needs to be able to differentiate certain requirements as necessitated by the classes of service. For example, it must meet minimum data rate requirements or delay bound guarantees for delay sensitive traffic.
- Fairness: The algorithm should be able to fairly assign available resources among the connections. Both short and long term fairness should be provided.
- Implementation complexity: In high-speed networks, scheduling decisions need to be made rapidly; also it should be able to reflect the channel state variation as quickly as possible. It implies that the algorithm must be of low-complexity since it needs to decide fast.
- Scalability: The algorithm needs to be efficiently workable even if the network size increases.

1.3 Thesis Contribution

The principle objective of the thesis is to design an efficient and simple QoS scheduling and bandwidth allocation scheme for multi-class broadband wireless access networks. Here we investigate the strategy of admission control and propose a WiMAX/IEEE802.16 QoS scheduler that allocates available radio resources efficiently to satisfy negotiated QoS requirements for each service request. In our scheme, the QoS parameters that are taken into account are: Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate and Maximum Delay. The key novel contributions of our proposal are:

1. In our framework, we combine resource allocation, admission control and QoS scheduling scheme for the WiMAX supported data types. The major objective of this thesis

is to develop an effective yet simple QoS resource allocation policy for multi-service networks that can be deployed and implemented with less overhead by proposing a preference metric that takes care of data rate and delay constraints of various classes in a single function. We explore the limitation of the original proportional fairness (PF) scheduling scheme, and propose to extend the PF scheme for delay differentiation in IEEE 802.16 networks.

2. We propose a moderately conservative and adaptive call admission control mechanism by considering available bandwidth and the delay constraints that take place in scheduler execution and frame construction. The need for frequent bandwidth reallocation has been minimized by introducing usage statistics over some specific intervals of the day.

3. In our proposal non-realtime (i.e., Non-Real-Time Polling Services (nrtPS) and Best Effort (BE)) requests do not likely starve bandwidth allocation. nrtPS and BE requests act as the 'filler' of the physical slots that remain unused after allocation of higher priority classes such as Unsolicited Grant Service (UGS) and rtPS in each Orthogonal Frequency Division Multiple Access (OFDMA) frame.

While we have the above distinct features, the proposed resource allocation and scheduling algorithm takes advantage of multiuser diversity by utilizing instantaneous channel condition information of different users in the system.

1.4 Thesis Organization

The thesis is organized into six chapters. Chapter 2 gives an overview of WiMAX/IEEE802.16 Physical Layer (PHY) and Medium Access Control (MAC) layer architecture. A literature review of the related works is presented in Chapter 3. Chapter 4 describes the developed algorithms featuring admission control, scheduling and resource allocation. The analytical and simulation results generated by a customized simulator are presented in Chapter 5. Finally, the conclusions are drawn in Chapter 6.

Chapter 2

Overview of Broadband Wireless Access (BWA)

Broadband Wireless Access (BWA) is a wireless Internet connection technology that supports data, voice and video information at high speeds, typically provided by wired connectivity such as DSL or cable services. It is considered 'broad' because multiple types of services can travel across wide band that the technology uses [Erg09]. The term 'WiBB' is sometimes used in the literature as an acronym for 'Wireless Broad-Band'.

2.1 The Key Technologies

BWA has started with a fixed access in mind to compete with DSL and cable modem to bypass the monopoly of service providers' wire-line infrastructure [Erg09]. However, through time BWA has evolved into 'mobile broadband'. Therefore, there are two major flavors of broadband wireless services. The first type called <u>fixed wireless boradband</u> concerns providing wireless Internet solution at fixed locations, acting as a suitable and cost-effective alternative to land-based technologies, and extending the service to rural and sub-urban areas where wired infrastructure is scarce. Local Multipoint Distribution Service (LMDS) and Multichannel Multipoint Distribution Service (MMDS) are the two key telecommunication technologies that have been used for wireless access over a wide area before the standardization of IEEE 802.16d in June 2004. The second type called mobile broadband, offers the additional functionality of portability, nomadicity, and mobility in wide area networks [AGM07]. In this category, we have evolutions of IEEE 802.16, such as, IEEE 802.16e to 802.16h-2010, the third and fourth generation mobile technologies including Long Term Evolution (LTE) and Ultra Mobile Broadband (UMB). However, only mobile WiMAX addresses both the cellular and fixed wireless broadband Internet.

2.1.1 WiMAX

IEEE802.16 standard, commonly dubbed in the industry as Worldwide Interoperability for Microwave Access (WiMAX), provides MAC Layer QoS support for Broadband Wireless Access networks. The standard has evolved through many phases of revisions, amendments, corrigenda or errata to the current version of 802.16-2009 with the latest amendment, 802.16h-2010. In the following sections, we would like to present a brief history of IEEE802.16 standard, followed by discussions of physical and MAC layer specifications that concern our work.

2.1.2 LTE

The Third Generation Partnership Project (3GPP) addresses next generation technology with Long Time Evolution (LTE) project in order to accommodate ever increasing mobile data usage. LTE air interface is a successor of GSM/EDGE and UMTS/HSxPA network technologies. LTE provides an adaptive resource allocation by flexible coding and modulation (up to 64-QAM) with channel dependent scheduling.

2.1.3 LTE vs. WiMAX

Figure 1, as it appears in [Erg09] summarizes the key cellular technologies that the world has seen. In the following discussion, we are going to compare LTE and WiMAX - the two major wireless broadband technologies.

- Both technologies primarily use some forms of (Orthogonal Frequency Division Multiple Access) OFDMA as the air interface. However, LTE uses OFDMA for downlink and SC-FDMA for the uplink modulations whereas WiMAX uses OFDMA in both directions.

- Both LTE and WiMAX have the support for TDD and FDD spectrum. However, vendor implementation trend is to focus FDD for LTE and TDD for WiMAX.

- Having evolved from generations of mobile technologies, LTE's end-to-end network is made complex with many layers and proprietary protocols. On the other hand, having built from scratch, WiMAX network is a flat all-IP network with IETF-based protocols.

From now onward, our discussion focusses on WiMAX technology.

2.2 A Brief History of WiMAX/IEEE802.16

Following the original publication of IEEE 802.16 in 2001 having only Line of Sight support, two major revisions were made in IEEE 802.16-2004 [80204] commonly called IEEE 802.16d

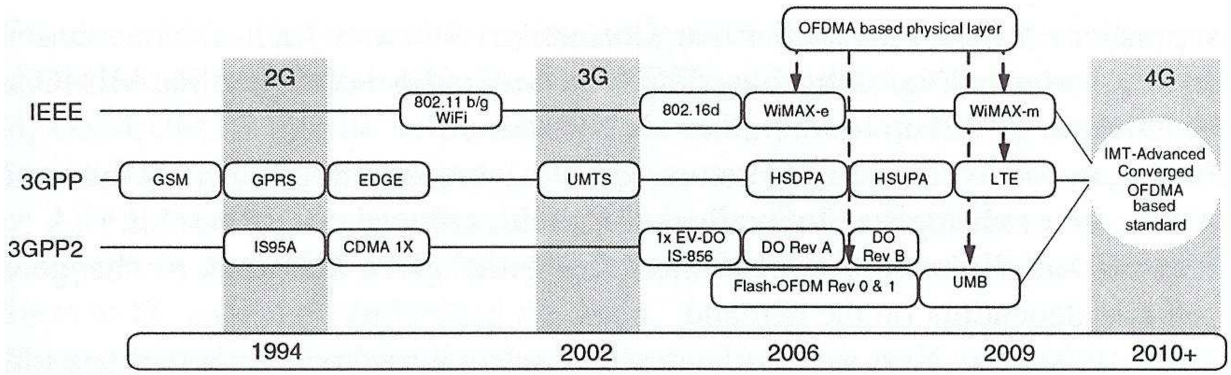

Figure 1: Evolutionary path of cellular technology (reproduced from [Erg09])

and IEEE 802.16-2005 [80205] commonly called IEEE 802.16e, which were the specifications for the fixed and mobile WiMAX respectively. In both IEEE 802.16-2004 and IEEE 802.16-2005, along with the Point to Multipoint (PMP) support (as in Figure 2), there was the optional support for mesh mode (as in Figure 3), where SSs being multiple hops away from the BS could communicate with the BS via intermediate SSs and also could communicate with each other.

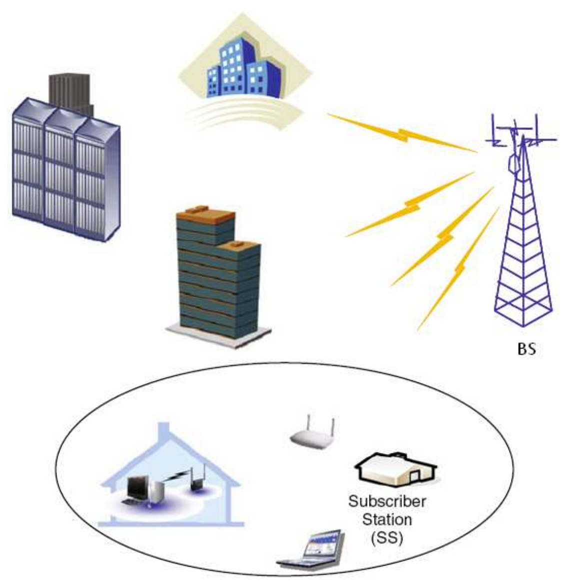

Figure 2: An example of a fixed IEEE 802.16 PMP network (reproduced from [PV10])

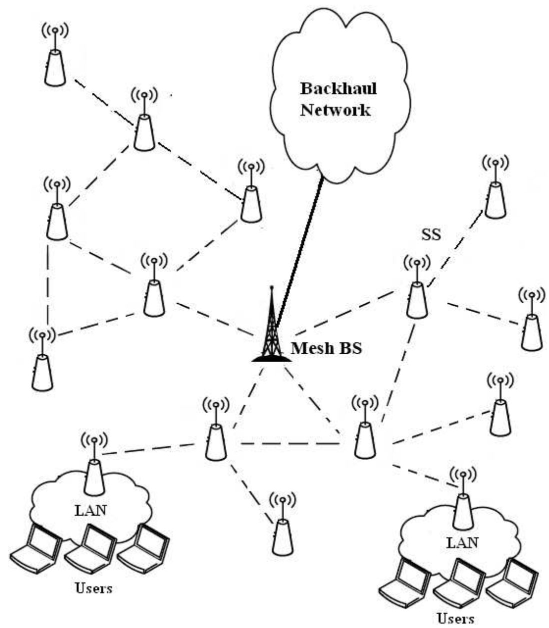

Figure 3: An example of a IEEE 802.16 mesh network

In 2009 the standard committee published a revision of IEEE 802.16-2004. This combines material from IEEE 802.16e-2005, IEEE 802.16-2004/Cor1-2005, IEEE 802.16f-2005, and IEEE Std 802.16g-2007, along with supplimentary maintenance items and improvements to the management information base specifications. However, in this publication of the standard the support for mesh is removed and only PMP support is retained [WiM09]. An amendment, namely IEEE 802.16j-2009 [80209] updates and expands IEEE 802.16, specifying OFDMA physical layer and medium access control layer enhancements to IEEE

802.16-2009 for licensed bands to enable the operation of relay stations. In July 2010, the second amendment to 802.16-2009 is published, termed 802.16h-2010 [80210], which updates and expands the IEEE 802.16, providing specification of improved mechanisms, as policies and MAC enhancements, to enable coexistence among license-exempt systems and to facilitate the coexistence of such systems with primary users.

In our project, we focuss on working with PMP mode of operation as described in the IEEE 802.16-2009, which is the 'current applicable version of IEEE Std 802.16' [80210].

2.3 Packet Transmission Systems

The combination of physical (PHY) layer characteristics that are chosen for a specific network, along with the associated Medium Access control (MAC) layer features, determine the packet transmission systems. The following subsections provide an overview on these specifications.

2.3.1 MAC layer

The MAC layer specifications, among many other things, dictate the node connection topology (PMP or Mesh), the addressing for nodes (48-bit universal MAC address) as well as for links (16-bit Connection Identifier, CID), MAC PDU formats (construction and transmission), ARQ mechanism, QoS provisioning (UGS, rtPS, nrtPS and BE), bandwidth allocation and request mechanisms, MAC support of PHY (Time Division Duplex (TDD) and Frequency Division Duplex (FDD)), contention resolution, network entry, initialization and ranging. In determining the packet transmission system, duplexing plays a vital role in deciding forward and reverse communication channels. In a FDD system, uplink (UL) and downlink (DL) transmissions use different frequencies but are typically simultaneous. In a TDD system, the UL and DL transmissions occur at different times and usually share the same frequency. A TDD frame as in Figure 4, has a fixed duration and contains one DL and one UL subframe. The frame is divided into an integer number of Physical Slots (PSs), which help to partition the bandwidth easily. The TDD framing is adaptive in that the bandwidth allocated to the DL versus the UL can vary.

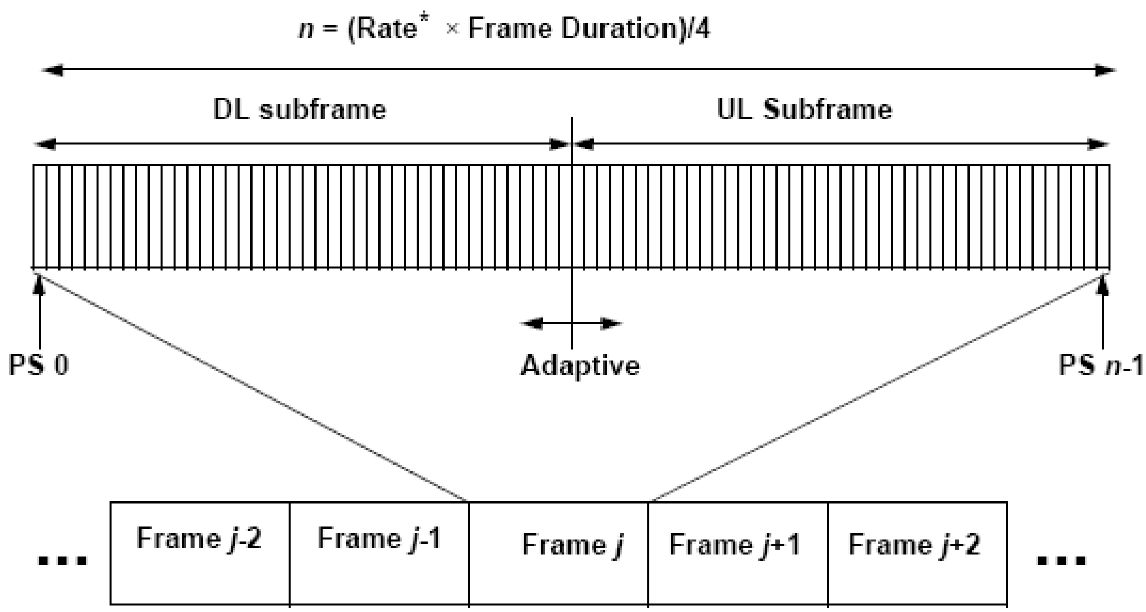

Figure 4: TDD Frame Structure [WiM09]

2.3.2 PHY layer

This layer provides interface between MAC layer and external medium. IEEE802.16-2009 specifies four technologies as described below:

WirelessMAN-SC (Wireless Metropolitan Area Network - Single Carrier)

This is applied for line of sight operations in 10 - 66 GHz frequency band. The uplink (UL) channel is based on a combination of time division multiple access (TDMA) and demand assigned multiple access (DAMA). The downlink (DL) channel is TDM, with the information for each SS multiplexed onto a single stream of data and received by all SSs within the same sector. Data bits are mapped to a Quadrature Phase-Shift Keying (QPSK), 16 Quadrature Amplitude Modulation (QAM), or 64-QAM (optional) signal constellation.

WirelessMAN-OFDM (Wireless Metropolitan Area Network - Orthogonal Frequency Domain Multiplexing)

This is based on OFDM modulation and designed for non line of sight (NLOS) operation in the frequency bands below 11 GHz. The OFDM PHY supports a frame-based transmission. A frame consists of a DL subframe and an UL subframe. A DL subframe consists of only one DL PHY PDU. An UL subframe consists of contention intervals scheduled for initial ranging and bandwidth request purposes and one or multiple UL PHY PDUs, each transmitted from a different SS. After bit interleaving, the data bits are entered serially to the constellation mapper. BPSK, Gray-mapped QPSK, 16-QAM, and 64-QAM are supported, whereas the support of 64-QAM is optional for license-exempt bands.

WirelessMAN-OFDMA (Wireless Metropolitan Area Network - Orthogonal Frequency Domain Multiple Access)

This is also based on OFDM modulation and designed for NLOS operation in the frequency bands below 11 GHz. OFDMA allows some sub-carriers to be assigned to different users. A slot in the OFDMA PHY requires both a time and subchannel dimension and is the minimum possible data allocation unit. OFDMA slots are mapped to subchannels and OFDMA symbols. This technology supports adaptive modulation for every user QPSK, 16QAM, 64QAM and 256QAM.

WirelessHUMAN (Wireless High-speed Unlicensed Metropolitan Area Networks)

This is used for unlicensed MAN and license-exempt bands below 11GHz.

2.3.3 Popular PHY-MAC Configurations

In this section, we will discuss some multiplexing/duplexing configurations that we typically come across in literature. In TDMA, time is divided into time-slots and base station assigns time-slots for each user, allowing access of the medium for the purpose of either transmission or reception in those time slots. In OFDMA, the base station assigns a group of carriers to each user, and the physical carriers can be clustered (localized) or distributed across the channel.

OFDMA/TDD system

An example of an OFDMA frame structure for TDD mode of operation is shown in Figure 5. The horizontal axis indicates the OFDMA symbol number (time domain), and the vertical axis indicates subchannel logical number (frequency domain). A time slot is the basic unit of allocation in the time domain. It consists of some OFDMA symbol duration.

There are two main parts in the frame: downlink subframe and uplink subframe. Each frame begins with some mandatory control overhead: preamble, frame control header (FCH), downlink map (DL-MAP), uplink map (UL-MAP), transmit/receive transition gap (TTG) and receive/transmission transition gap (RTG).

The first portion of uplink subframe consists of ranging information, acknowledgement messages (ACK/NACK), and channel quality information channel (CQICH). The Ranging

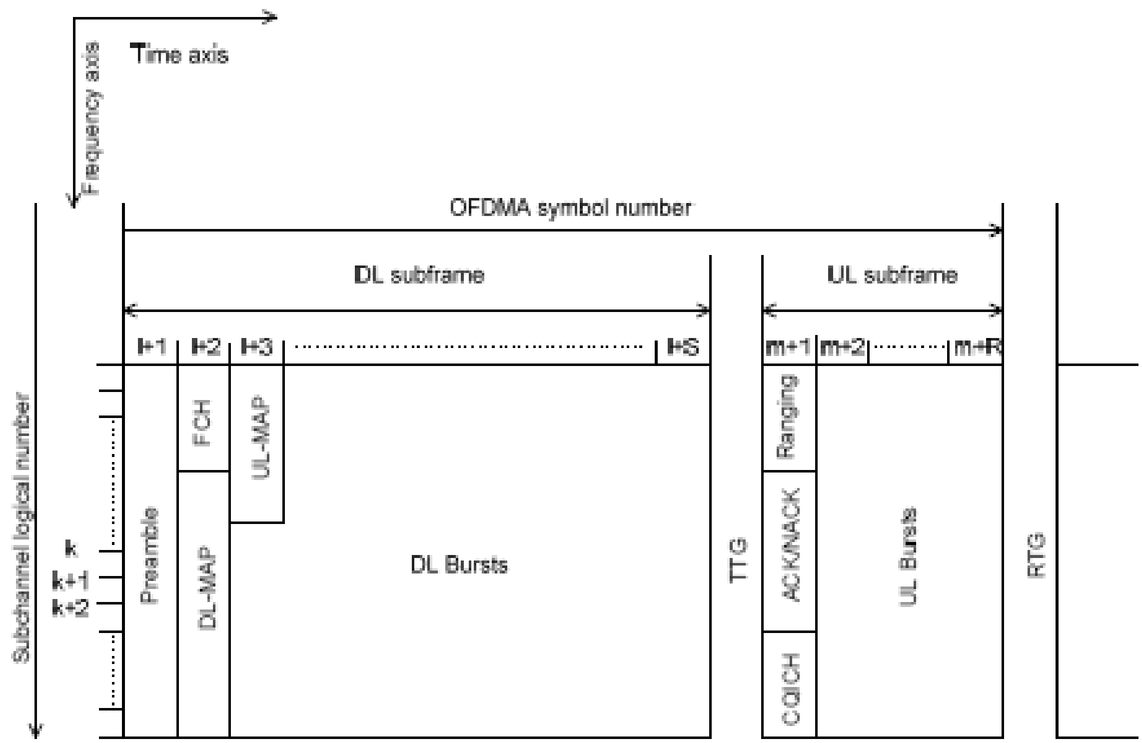

Figure 5: OFDMA/TDD Frame Structure [WiM09]

is used for synchronization and carrier tracking. The CQICH provides fast channel quality feedback. This information is used to construct the next downlink subframe. The roles of ACK/NACK and CQICH messages will play very important role in our SRC algorithm. The automatic repeat request (ARQ) mechanism works based on ACK/NACK message. The BS is informed about channel quality through the CQICH message.

TDMA/TDD system

This duplexing combination is used primarily in mesh mode operation.

CDMA/TDD system

Code Division Multiple Access/With Time-Division Duplexing. This provides the flexibility of the physical layer for accommodating different service types simultaneously and has already been the basis of two 3G standards: TD-CDMA and TD-SCDMA with a synchronous uplink.

2.4 QoS in Wireless Networks

Concurrent provisioning of quality of service (QoS) to a number of end users in a certain network with different requirements is much more challenging a task in wireless media as compared to wired infrastructure. This is due to the inherent highly variable and unpredictable nature of wireless medium that shows time-dependence as well as location dependence. To cope with such issues, QoS in wireless networks is handled at the medium access control (MAC) layer. QoS support in wireless networks makes an effort to meet the end users' varied requirements in terms of data rate, delay, jitter tolerance and reliability of data packets.

Data rate: Measure of how much information can be handled by simultaneous users in a system. This is the basic QoS parameter of concern by most users. The offered data rate is closely related with the channel bandwidth that a particular technology can support. For example, for Line of Sight transmission of WiMAX, with typical channel bandwidth of 25 MHz the raw data rate could be 120Mb/s [WiM09]. The physical-layer pipe between the base station and the enduser terminals, number of active connections and their requirements dictate this feature.

Delay: Measure of end-to-end packet transmission time. Various Qos parameters, packet queuing mechanisms, wireless signal strength and end user mobility characteristics dictate this feature.

Jitter: Measure of delay variation in received packets. Packet buffering and user mobility pattern dictate this feature.

Reliability: Measure of successfully delivered packets without error. The inherent time-wise and location-wise variability of radio waves dictate this feature.

2.4.1 QoS Parameters in IEEE802.16/WiMAX

The QoS metrics of a IEEE802.16/WiMAX network are chiefly parameterized in the standard as Maximum Sustained Traffic Rate (MSR), Maximum traffic burst (MTB), Minimum Reserved Traffic Rate (MRR), Maximum Latency, Tolerated Jitter, Traffic Priority and vendor specific QoS parameters[WiM09].

Maximum Sustained Traffic Rate: Expressed in bits per second, this parameter defines the peak information rate of a specific service flow. it pertains to the service data units (SDUs) at the input to the system.

Maximum traffic burst: This parameter defines the maximum burst size that will be accommodated for a specific service flow. Maximum traffic burst set to zero means no maximum traffic burst reservation is required.

Minimum Reserved Traffic Rate: Expressed in bits per second, this parameter specifies the minimum rate reserved for the service flow. The BS is responsible to satisfy BRs for a connection up to its minimum reserved traffic rate. If less bandwidth than its Minimum reserved traffic rate is requested for a connection, the BS may reallocate the excess reserved bandwidth for other purposes.

Maximum Latency: The value of this parameter defines the maximum interval between the reception of a packet at the Convergence Sublayer of the BS or the SS and the forwarding of the SDU to its Air Interface. This specifies a service commitment.

SDU indicator: The value of this parameter defines whether the SDUs on the service flow are fixed-length or variable-length.

Tolerated Jitter: The value of this parameter defines the maximum delay variation for the connection. This parameter is available in case of a DL or UL service flow.

Traffic Priority: The value of this parameter specifies the priority of associated service flow. This parameter is available in case of a DL or UL service flow, which are associated with any other Uplink Grant Scheduling Types except UGS.

Vendor Specific QoS Parameters: This allows vendors to encode vendor-specific QoS parameters.

2.4.2 QoS Mechanisms in IEEE802.16/WiMAX

Unlike WiFi, IEEE802.16 MAC is connection oriented. For the purposes of mapping to services on SSs and associating varying levels of QoS, all data communications are in the context of a connection. The standard defines QoS mechanism by associating packets that traverse the MAC layer into a service flow. A service flow is identified as a unidirectional flow of packets that has been allotted a particular QoS. The BS and SS provide this QoS according to the QoS parameter set defined for the service flow. Service flows exist in both direction (UL and DL) and typically exist even they are not activated to carry traffic. All service flows have a 32-bit Service Flow ID (SFID); admitted and active service flows also have a 16-bit transport Connection Identifier (CID).

2.4.3 Traffic Classes (Scheduling Services in the Standard)

Scheduling services define the data handling methods supported by the MAC scheduler for data transport on a connection. Five types of service flow are defined in WiMAX/IEEE802.16, and are described below.

- **Unsolicited Grant Services (UGS):** Intended for Constant Bit Rate (CBR) services such as voice applications.

- **Real-Time Polling Services (rtPS):** Intended for services that generate variable size data packets at periodic intervals but certain delay requirements need to be met, such as MPEG video.

- **Non-Real-Time Polling Services (nrtPS):** Intended for delay tolerant services that require variable size data grant burst types on a regular basis such as FTP.

- **Best Effort (BE):** Intended for services that do not require any specific QoS guarantee such as HTTP.

- **Extended rtPS (ertPS):** Intended for services that require to have both the advantages of UGS and rtPS such as VoIP traffic which has active and silent periods, and polling is needed to notify BS of traffic during silent periods.

Table 1 summarizes major QoS parameters for these classes of traffic, which can be quantified to schedule transmission order of admitted service requests in a WiMAX network.

Table 1: Major QoS Parameters

QoS Parameter	UGS	rtPS	ertPS	nrtPS	BE
Max. Sustained Rate	X	X	X	X	X
Min. Reserved Rate	X	X	X	X	
Max. Latency	X	X	X		
Tolerated Jitter	X		X		
Traffic Priority		X		X	
Unsol. Grant Interval	X		X		
Unsol. Polling Interval		X			

A vital component of MAC design is to specify which nodes can transmit when and how this information is propagated through the network. In a centralized single hop scenario specified by the IEEE802.16-2009 PMP mode, solely the BS makes the decision depending on information gathered through various mechanisms, such as: unsolicited bandwidth request, contention based procedures and polling.

Chapter 3

Literature Review

There has been a good number of papers proposing scheduling or resource allocation algorithms reflecting several evolution phases of the WiMAX technology since its formal standardization in 2004. However, a majority of these papers do not consider QoS, some assume single class traffic and a few papers consider service differentiation with multiple classes. These scheduling/resource allocation policies vary whether they pertain to a PMP network or a mesh network, and in a mesh, whether the policy is for centralized or distributed scenarios. There is a further concern if the network is based on multihop (802.16j) standard; the policy adopted needs to take into consideration the type of the relay station, whether it is transparent or non-transparent. Therefore, we see various approaches of scheduling /resource allocation for WiMAX technology. The initial attempt of the current research work was to make an efficiency comparison of centralized and distributed scheduling algorithms in WiMAX mesh mode and find a combined solution that could be effectively deployed in both the mesh scenarios. However, at the advent of the discontinuation of mesh support by the standard's late 2009 edition, we were motivated to work with the QoS scheduling in

WiMAX PMP network. In the following sections, we will attempt to briefly categorize the papers that we have encountered throughout our research endeavor.

3.1 General Scheduling Algorithms

Some researchers have proposed mechanisms to adapt well-known scheduling algorithms within WiMAX/IEEE802.16 framework. **First In First Out (FIFO)** is the simplest form of scheduling to decide upon which queue can be served depending on a node's chronological network entry; however, this method lacks fairness. A little more complicated form of scheduling is **Round Robin (RR)**, where each SS is provided with equal channel resources. It supports fairness since RR allocates resources one by one to all the connections without priority and thus cannot guarantee QoS. A variant of RR, with weight, termed as **Weighted Round Robin (WRR)** is adapted for scheduling in WiMAX networks [CLME06], [SAKH06], [SAH08]. The weights are specified in terms of queue length and packet delay or the number of slots and are adjustable for the throughput and delay requirements [IJT09]. A variant of WRR, termed as **Deficit Round Robin (DRR)** [SV96] that handles variable size packets has also been adapted to WiMAX Scheduling [CAL07], [CWL08]. Alternatively there are schedulers that are specifically proposed for WiMAX. Some of these schedulers are summarized in [BN08] that discusses Temporary Removal Scheduler [BTGK05], Opportunistic Deficit Round Robin Scheduler [RBS06], Frame Registry Tree Scheduler [XPM05] and Adaptive rtPS scheduler [MSJ+06].

3.2 Non-QoS Scheduling Algorithms - Mesh and Relay Based Network

Some researchers have proposed WiMAX specific algorithms. Some of these are channel-condition-aware, which consider the lossy nature of wireless media and some are not, which assume perfect channel condition. A review of these algorithms can be found in [IJT09]. To cite some of the non-QoS mesh network algorithms we have: [TLZL05], [WGIH05] and [YTFM07] discuss concurrent transmission while avoiding interference, [HTLJ06] considers fair allocation of resources, [BN08] discusses maximum signal-to-snterference ratio algorithm, with the drawback that the users always having a small SIR may never be served. The authors of [CTWW07] focuss on spectral reuse in resource allocation, that employs channellevel scheduling for bandwidth adaptation between uplink and downlink subchannels) and link-level scheduling for timeslot allocation among SSs. The authors of [AWA08] propose an exact centralized method with a goal to reduce delay time, and to serve more users in mesh network. Congestion-aware Downlink Scheduling [CWL08], Opportunistic Scheduling Exploiting Multiuser Diversity and Frequency Selectivity [DMR08] consider relay mode of the technology.

3.3 QoS scheduling algorithms - Mesh Network and Relay Based Network

Some works have been done to provide QoS support for traffic in a WiMAX mesh network, such as: End-to-End Bandwidth Reservation [CGLM07], Fair End-to-end Bandwidth Allocation (FEBA) [CAL07], Mesh Election Procedure [CELM07] and Enhanced Election-Based

Transmission [LC08]. For relay networks following can be cited: Scalable QoS Provisioning and Service Node Selection in Relay Based Cellular Networks [JBW08], Admission Control and Interference-Aware Scheduling in Multi-hop WiMAX Networks [GGM07], MAC Performance Evaluation of IEEE 802.16j [ZFYZ08], Performance Evaluation of a WiMAX Multi-Hop Relay System to Support Multicast/Broadcast Service [HYW09] etc.

3.4 QoS scheduling algorithms - PMP Network - Area of Investigation

In this section, we will review the QoS Scheduling for WiMAX PMP networks as well as some of the strategies developed for general OFDMA systems. A few LTE works in the similar network infrastructure will also be mentioned.

Layered Scheduler

Some of the often-cited early works on QoS scheduling in WiMAX PMP infrastructure include [WG03] and [CJW05]. In [WG03] the proposed solution includes QoS support for all types of traffic classes as defined by the standard. The authors in [WG03] have shown the relationship between traffic characteristics and its QoS requirements and the network performance. In [CJW05] a 2-layer service flow management architecture for IEEE 802.16 standard (TDD mode) is proposed. According to the authors, the proposed solution improves the performance of throughput under unbalanced uplink and downlink traffic compared to fixed bandwidth allocation policy. However, both the works introduce complex scheduling of hierarchical algorithms: Priority Queue [WG03] or Deficit Fair Priority Queue (DFPQ) [CJW05] is used at the first layer to distribute the total bandwidth among the service flows in different queues, and at the second layer different schedulers are implemented for servicing different types of traffic, such as, Earliest Deadline First (EDF), Deficit Round Robin (DRR), Weighted Fair Queueing (WFQ), and Worst-case Weighted Fair Queueing (WWFQ) were used to deal with different classes of traffic. This layered scheduling strategy surely adds much overhead in the Base Station and practically unsuitable for OFDM based system.

In [CJG05], the authors propose a two-layer scheduling structure for bandwidth allocation supporting all types of traffic. However, the paper is more focussed in suggesting a QoS architecture rather than elaborating on the scheduling methodology.

In [ZHWW07], Liu et el. proposed a simplified layered scheduling scheme based on Maximum Delay Utility for OFDM networks. In this scheme, the scheduling is divided into two main steps, macro and micro scheduling. In the macro step, the utility functions are defined, and the priorities of each type of service are determined. In the micro step, the scheduling is performed among all users inside a given traffic type determined in macro step. The authors claim that their simplified layered scheduling has much lower computational complexity and almost the same utility values as the standard maximum delay utility scheduling while handling multiple traffic types with diverse QoS requirements.

In [KSC09], the authors proposed a scheduling algorithm that satisfies the QoS requirements of the real-time traffic and maximizes the utility of the non real-time traffic. A step-by-step approach was used to achieve these two objectives with low complexity and traffic class prioritization. A well-known bipartite matching algorithm was adopted for the QoS scheduling of the real-time traffic and a standard gradient scheduling algorithm for the utility maximization scheduling of the non real-time traffic.

Single Class Scheduler

Instead of proposing hierarchical schedulers, some papers focus on scheduling of one service class for a particular application in a WiMAX network. In [LKC04] and [LKC+06] the proposed schedulers only consider realtime traffic. In [YL06] the only service class it deals with is VoIP applications. These policies have limited usage in practical network since in reality a network offers multiple service classes and the scheduler should be able to handle the interaction arising from the coexistence of multiple class traffics.

Uplink Scheduling

The authors in [LAC09] use a customized form of Deficit Round Robin algorithm to provide Uplink scheduling. The proposed algorithm preserves the simplicity and fairness available in the original DRR design and in addition has a low latency bound.

The proposed algorithm in [FF07] supports the four service levels specified by the standard and considers their QoS requirements for scheduling decisions. It uses an approach based on three priority queues, namely, low priority queue, intermediate queue and high priority queue and proposes the scheduling discipline for uplink traffic where QoS is maintained by adhering to the requests' deadline. However, sorting the intermediate and low priority queues and migrating the requests to the high priority queue at each frame's onset is definitely a scheduler overhead. Also the assumption that the number of slots needed to serve the high-priority-queue requests being always less or equal to the number of available slots in the uplink sub-frame is utterly impractical since their might be some requests in the intermediate queue, which would require immediate migration to the high-priority-queue due to near deadline-expiry, thus causing a possible over-flow. The algorithm does not address this issue. Furthermore, the algorithm pre-assumes admission control being met, thus ignoring the effect of the new requests on bandwidth available for the already admitted requests.

Downlink Scheduling

In [RJ09] a downlink scheduling scheme for a multiuser OFDMA/TDD WiMAX network is proposed. They formulate the optimization problem based upon a multichannel proportional fairness with minimum bit rate guarantee. The solution is considered by using the Lagrange multipliers with relaxed constraints. The cross-layer approach integrates the requirement to the solution and schedules packets with finitely backlogged queue consideration.

In [PMl09] Sara et el. propose a design and analysis of a channel-aware Deficit Round Robin (DRR) based scheduling algorithm for downlink traffic delivery in a point-to-multipoint WiMAX network. Their work provide a technique for compensation of channel errors to preserve QoS differentiation and fairness under non-ideal channel conditions. The authors claim this compensation technique under different channel conditions and network loads to be effective.

The authors of [WD10] propose a downlink resource management framework for QoS scheduling in OFDMA based WiMAX systems. Their framework contains a dynamic resource allocation (DRA) module and a connection admission control (CAC) module. A two-level hierarchical scheduler is developed for the DRA module that provides more organized service differentiation among different service classes, and a measurement-based connection admission control strategy is introduced for the CAC module. According to their system-level simulation the proposed framework can work adaptively and efficiently to improve the system performance in terms of high spectral efficiency and low outage probability.

Generalized OFDMA/LTE

In [WJ09], the authors proposed a scheduling scheme with a goal to distribute all available resources among different competing users in OFDMA WiMAX network. However, they supposed that all the time slots within a frame are mapped into the same modulation type, which is not a realistic thing to do since in order to combat fading effect, the subcarriers in an OFDMA frame gets modulated differently.

In [Del10] the author proposes scheduling algorithms that can handle single and multi-class QoS in LTE uplink systems in conjunction with admission control to meet the LTE requirements. This work also include the delay as an important parameter. The major difference lies in the way how this delay parameter is realized in their SC-FDMA LTE compared to our OFDMA/TDD WiMAX system.

In [DJ10b] the authors propose scheduling algorithms for LTE network with single class traffic that, in addition to the channel contiguity constraint, take also into account the end-to-end delay constraint. Their solution chiefly centers around dealing with LTE specific SC-FDMA uplink scheduling requirements, i.e., meeting contiguous resource block constraints.

In [DJ10a] the authors scheduling and resource allocation schemes that deal with QoS requirements in Uplink LTE systems. They investigate the possibility of assigning more than one resource block and its consequences on satisfying stringent QoS requirements in the context of heavy traffic, either in terms of end-to-end delays or of minimum rates. However, their utility function does not include the factor of signal strength variation resulting from mobility and other fading and interference effects. Therefore, it is unclear how the consequent data rate variation would be handled by the modulation and coding scheme involved.

Summary of WiMAX QoS scheduler design issues

In our literature review we have encountered many papers that propose solutions for QoS provisioning in IEEE802.16/WiMAX, however, in many cases, without giving much elaboration on underlying physical layer channel assignment. Also connection admission control mechanism is sometimes ignored. In general, the bridging of the radio resource allocation and the scheduling of traffic needs to be addressed. In our current work we will attempt to combine resource allocation, admission control and QoS scheduling policy. We will also attempt to map various parameters of our algorithm to the ones mentioned in the IEEE802.16-2009 standard.

Chapter 4

QoS Scheduling and Resource Allocation in IEEE802.16 Network

In this chapter we will discuss our proposed scheduling, resource allocation and admission control framework. Before presenting our algorithms, we would first give a description of the OFDMA resource allocation model and channel fading model that have been used in framework.

4.1 OFDMA Resource Allocation Model

In OFDMA, the radio resource is partitioned in both frequency and time domain, resulting in a frequency-time domain resource allocation. We assume the uplink scenario of an OFDMA system with S_U data subcarriers. The time axis is divided into frames of fixed duration, each of which, has two parts: Downlink (DL) subframe and Uplink (UL) subframe to support TDD operation. The split between UL and DL is adaptive and is a system parameter. In each UL subframe, there are T_U timeslots available for uplink transmissions.

Channel coherence in frequency and time can be utilized by grouping G_S subcarriers and G_T timeslots to form an OFDMA slot. An OFDMA slot is the minimum resource unit that can be allocated to a request. Therefore, in each UL subframe, there are $S = S_U/G_S$ subchannels (indexed by s) in frequency domain and $T = T_U/G_T$ timeslots (indexed by t) in time domain, giving a total of $S \times T$ OFDMA slots available in each UL subframe per frame. Each OFDMA slot can be thought of as a unit of transmission opportunity in Frequency-Time domain and can be assigned to different requests and be independently bit and power loaded.

We assume that the system has a bandwidth of B Hz, which is equally distributed among all the subchannels giving each subchannel a bandwidth of $\beta = B/S$. The instantaneous subchannel states are represented by a vector of signal-to-noise ratio values $\gamma_s(t)$. The $\gamma_s(t)$ values are dependent on instantaneous attenuation and transmit power share of subchannel s. Finding maximum information rate $R_{\max}$ of the system is an optimization problem applied on Shannon's capacity equation of individual subchannels and summing them as follows [BGWM07]:

$$R_{\max} = \max \sum_{s \in S} \beta \times \log_2\{1 + \gamma_s(t)\}. \tag{1}$$

Let K be the set of admitted requests. For a given request $k \in K$, we denote by S_k the set of subchannels allocated to request k, by P_k the instantaneous power share of user k, by $P_k^{\max}$ the maximum transmit power, by $\gamma_{s,k}(t)$, the instantaneous SNR value of the subchannel s as allocated to user k at timeslot t, by $r_{s,k}(t)$ the instantaneous channel capacity (the maximum possible data transmission rate) of the subchannel number s if allocated to request k and by $\overline{R}_k$ the request's average throughput. The relation between

$r_{s,k}(t)$ and $\gamma_{s,k}(t)$ is dictated by Shannon's capacity equation [GB06b]:

$$r_{s,k}(t) = \beta \times \log_2\{1 + \gamma_{s,k}(t)\}. \tag{2}$$

Let $\phi_{s,k}(t)$ be a decision variable that indicates the assignment of subchannel s to user k at time point t. Accordingly,

$$\phi_{s,k}(t) = \begin{cases} 1 & \text{if subcarrier } s \text{ is assigned to user } k \text{ at time } t \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We further assume that each of the subchannels can be uniquely assigned to one request at most, while each request can be assigned any number of subchannels during a given time point. This constraint, thus, can be defined as:

$$\sum_{k \in K} \phi_{s,k}(t) \leq 1 \tag{4}$$

for each $s, 1 \leq s \leq S$ and $t, 1 \leq t \leq T$.

The relation between $r_{s,k}(t)$ and $\overline{R}_k$ is:

$$\overline{R}_k = \phi_{s,k}(t) \times \sum_{s \in S_k} r_{s,k}(t). \tag{5}$$

The maximization of the user utility can be written as follows:

$$\max \sum_{k \in K} U(\overline{R}_k | S_k, \gamma_{s,k}(t)) \tag{6}$$

where $U(\overline{R}_k|S_k, \gamma_{s,k}(t))$ is the utility of user k as a function of the throughput $\overline{R}_k$ given the allocation of subchannels $S_k \in S$ to user k having instantaneous SNR values $\gamma_s(k,t)$.

The utility function addresses two issues: Subchannel Allocation: deciding how many subchannels to be allocated to each request and Subchannel Assignment: determining which subchannels to be assigned to each request. The relation between them lies in the fact that when a user is far away from the base station having allocated lower SNR value subchannel(s), the number of subchannels needs to be increased in order to maintain fairness and quality of service.

We need further to define delay and rate constraints. Firstly, we define that each request must not exceed a maximum allowable delay ($\delta_k^{\max}$), otherwise the request would expire and get rejected. The delay constraint is given by:

$$\delta_k(t) \leq \delta^{\max}. \tag{7}$$

Now we further investigate this constraint. Let the time to expire a request's Head of Queue (HoQ) packet be expressed as $W_k(t)$, which is the difference between the maximum allowable delay ($\delta^{\max}$) and HoQ packet delay up till time t, $\delta_k(t)$ i.e.,

$$W_k(t) = \delta^{\max} - \delta_k(t) \tag{8}$$

$W_k(t)$ is an indicator of how close a request is to its deadline expiry. Requests with smallest $W_k(t)$ needs to get resource allocation promptly. Note that the value of $1/W_k(t)$ increases significantly as the time to expire $W_k(t)$ decreases. When the system is heavily loaded with requests, deadline violations start to occur. Therefore, we define a threshold quantity of delay θ_δ. The scheduler checks at start of frame time if a request is due expiry in the next frame when the request has:

$$W_k(t) \leq \theta_\delta. \tag{9}$$

These requests, if any, forms a set, K_θ, which requires immediate allocation in the current frame.

Second constraint is that each request's average throughput must not be smaller than the Minimum Reserved Traffic Rate ($r^{\min}$) as defined by the system.

$$\overline{R}_k \geq r^{\min}. \tag{10}$$

4.2 Channel Fading Model

In our model a single BS serves a number of mobile terminals (or users) from which $k \in K$ requests are generated. The BS receives the feed back channel from all the stations to collect the current channel state. Based on the received channel information via $CQICH$, the BS scheduler decides upon scheduling criteria set forth in the algorithm and selects a request among the active connections to be served in the next physical slot (PS).

If the transmission power of the BS is given by P_t, the receiving power of a mobile user generating request k is given by [CB07]:

$$P_k = |h_k|^2 \times P_t \tag{11}$$

where h_k is the channel gain. The channel gain represents the effects of various physical phenomena such as scattering and absorption of radio waves, shadowing by terrestrial obstacles, and multipath propagation. The channel gain from the BS to the user generating request k can be written as:

$$h_k = \sqrt{c d_k^{-\alpha} S_r} m_k \tag{12}$$

where c is a constant pertaining to the transmission and reception antenna gains, d_k is the distance from the BS to the user with request k, α is the path-loss exponent (estimated to be about 4.0 in typical urban environment), S_r is a random variable for the shadow-fading effect and m_k represents the phasor sum of the multipath components. The shadow fading S_r follows the log-normal distribution with zero-mean and the variance σ_s^2 in decibels in the log-scale. The multipath-fading effect m_k is modeled as the second-order chi square or exponential random variable with a mean of 1.0, representing the Rayleigh fading channel.

As the model assumes a single-cell scenario, the inter-cell interference can be ignored. the SNR of the user generating request k on subcarrier s_k can, therefore, be represented as $\gamma_{s,k} = P_k/P_n$, where P_n is the background-noise power, which includes the thermal noise and other Gaussian interferences. In [CDG02], the median SNR at the cell edge ρ is defined representing the noise level of the wireless environment considered. Since $\rho = cD^{-\alpha}P_t/P_n$, the average SNR of the user generating request k as:

$$\overline{\gamma_{s,k}} = \rho \times (D/d_k)^{\alpha} S_r. \tag{13}$$

With Adaptive Modulation and Coding (AMC) support, WiMAX can vary the modulation and coding rate on a per user basis based on the prevailing Signal to Noise Ratio (SNR) conditions as shown in table 2. The resource allocation algorithm utilizes user's channel

Table 2: Modulation and Coding Schemes in WiMAX/IEEE802.16[WiM09]

Modulation	bits/symbol	Receiver SNR
$BPSK(1/2)$	0.5	3.0
$QPSK(1/2)$	1	6.0
$QPSK(3/4)$	1.5	8.5
$16QAM(1/2)$	2	11.5
$16QAM(3/4)$	3	15.0
$64QAM(2/3)$	4	19.0
$64QAM(3/4)$	4.5	21.0

SNR, $\gamma_{s,k}(t)$ in order to determine which sub-carriers are suitable for it at timeslot t for data transmission. We assume that BS has complete channel quality information (CQI) of all connections, which it does by employing RSSI and CINR signal quality measurements and associated statistics. We further assume that this channel state is not changing over one schedule length.

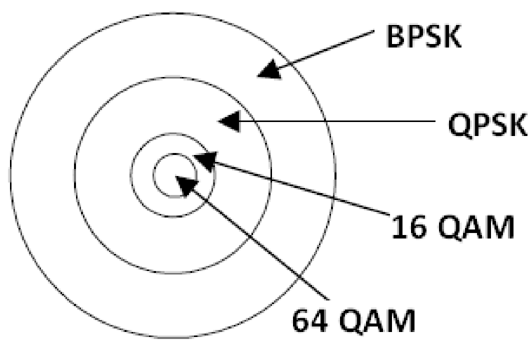

Figure 6: Relative Cell Radii for Adaptive Modulation (reproduced from [LE04])

AMC allows a WiMAX system to adjust the signal modulation scheme depending on the signal to noise ratio (SNR) condition of the radio link. When the radio link is high in quality, the highest modulation scheme is used, giving the system more capacity. During a signal fade, the WiMAX system can shift to a lower modulation scheme to maintain the connection quality and link stability.

4.3 Scheduling and Resource Allocation in WiMAX network

In this section, we propose a heuristic model for the resource allocation problem subject to delay constraints in WiMAX PMP network with OFDMA-TDD operation. We chose OFDMA-TDD packet transmission system from a multitude of other possible variants such SC-OFDM-FDD/TDD, TDMA-TDD/FDD, CDMA-TDD/FDD, OFDMA-FDD since OFDMA provides increased robustness to narrowband interference and impulse noise [SK98] and TDD offers as compared to FDD, less complex transceiver design and more flexibility in changing the UL and DL bandwidth ratio according to the dynamic traffic pattern [tdd01]. We assume that the subscriber stations are stationary or nomadic users with slowly varying channel conditions in a single cell. We further assume that the base station has perfect knowledge of channel information.

4.3.1 Problem Statement

The QoS Scheduling and Resource Allocation with Call Admission Control (QoS-SRC) problem that is addressed in current work can be summarized as follows: *A number of service requests with QoS parameters is given; we need to admit and schedule these requests so that minimum reserved traffic rate and maximum delay requirements of these service flows can be guaranteed as well as the maximum throughout can be ensured under a fairness model.*

4.3.2 Mathematical Formulation

We have S set of subchannels, T timeslots in an OFDM frame, K set of connection requests, indexed respectively by s, t and k. Q_k denotes the length of the queue associated with request k. We assume that $b_{\gamma,\mu}(s)$ is the number of bits per second that Adaptive Modulation and Coding scheme adaptively selects for subchannel s on the basis of received SNR γ using modulation μ.

We would like to define the preference metric which determines the order that the requests get allocation in the OFDMA frames. In doing so, we introduce proportional fairness (PF) to constitute the utility function.

PF was proposed by Qualcomm Company, which was realized in the IS-856 standard for the downlink traffic scheduling (also known as High Data Rate (HDR)) [JPP00]. The essential goals of this packet scheduling scheme are to enhance the system throughput as well as provide fairness among the queues under consideration. A PF scheduler assigns times-lots first to those connections having the highest ratio of current achievable rate to averaged rate. In every frame having each user's average rate been updated and the ratio been calculated, the scheduler keeps on serving the connections in this order as long as there are available slots pending requests. This way the scheduler can take advantage of the temporal variations of the channel by scheduling transmission opportunities to SS/MS during time periods where the SS/MS has the best signal levels. Therefore, Proportional Fairness scheduling is based on one priority function (14):

$$U_k(t) = \frac{r_{s,k}(t)}{\overline{R_k}(t)} \tag{14}$$

where $r_{s,k}(t)$ is the data rate currently achievable by the request k, which is dictated by the channel condition, $\overline{R_k}(t)$ denotes historical average (an exponentially smoothing average) data rate received by request k up to slot t. The request with the highest $U_k(t)$ is served at time slot t, where the average throughput of the request is updated by (15):

$$\overline{R}_k(t+1) = (1 - \frac{1}{T_c})\overline{R}_k(t) + (\frac{1}{T_c})r_{s,k}(t). \tag{15}$$

Otherwise, if the user is not currently receiving service, the queue is updated by (16):

$$\overline{R}_k(t+1) = (1 - \frac{1}{T_c})\overline{R}_k(t) \tag{16}$$

where T_c is the time constant for the moving average, which is an adjustable parameter. In general, T_c is assumed to be 1000 slots in the CDMA-HDR system. The value of T_c dictates the amount of time for which an individual request can be starved. This happens when the request generating MS/SS abruptly moves from a good channel condition to a bad channel condition. This is because the algorithm attempts to serve each request at the peak of its channel environment. Hence, the scheduler will see a drop in channel condition as temporary until the poor channel conditions persist for more than T_c.

The preference metric for a request, Φ_k is given by:

$$\Phi_k = U_k \times \frac{1}{W_k(t)}. \tag{17}$$

4.3.3 Multi-class QoS Scheduling and Resource Allocation Algorithm

We suppose that C different classes of services are provided by the system, which are indexed by c, i.e., $c \in C$ and the system has admitted K_c sets of requests of class c, indexed by k, i.e., $k \in K_c$. We assume that $r_c^{\min}$, $r_c^{\max}$ and $\delta_c^{\max}$ denote respectively the Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate and Maximum Delay associated with the traffic of class c.

In order to incorporate multi-class support, the preference metric needs to be modified along with the related constraints.

The delay constraint:

$$\delta_k(t) \leq \delta_c^{\max}. \tag{18}$$

The rate constraint:

$$r_c^{\max} \geq \overline{R}_k \geq r_c^{\min}. \tag{19}$$

The QoS resource allocation and scheduling scheme will differentiate the delay and rate performance of each QoS request. Considering the priority of traffic types, the modified preference metric is given by:

$$U_k(t) = \frac{r_{s,k}(t)}{\Phi_c \overline{R_k}(t)} \tag{20}$$

where Φ_c denotes priority of classes, $\Phi_{UGS} > \Phi_{rtPS} > \Phi_{nrtPS} > \Phi_{BE}$ but numerically they are reverse.

The resource allocation policy for UGS traffic is set forth in the standard. To accommodate within our multi-class resource allocation scheme we assume to place the UGS requests (K_{CBR}) onto some reserved channels for the duration of the required number of physical slots.

The policy for the multi-class scheduling is outlined in Algorithm (4.1). The algorithm works as follows: The algorithm starts by checking whether there is any request in close proximity of its expiry (lines 8-16). If a request's delay constraint is not met it is discarded and the sender is notified of its outage so that request can be resent (lines 10-11), otherwise the request is allocated with subchannels from the reserved subchannel set without further scrutiny to save time. If there is no request nearing expiry, the algorithm enters into its regular phase (lines 19-30). The priority of allocation is determined by the preference metric: $U_k \times \frac{1}{W_k(t)}$. Allocation of subchannels starts with the requests having the highest value for the preference metric. The parameters U_k and $W_k(t)$ take into account the rate constraint and delay constraint respectively. Requests that have pending traffic and have a high value of $U_k \times \frac{1}{W_k(t)}$ are allotted with the best subchhanels available.

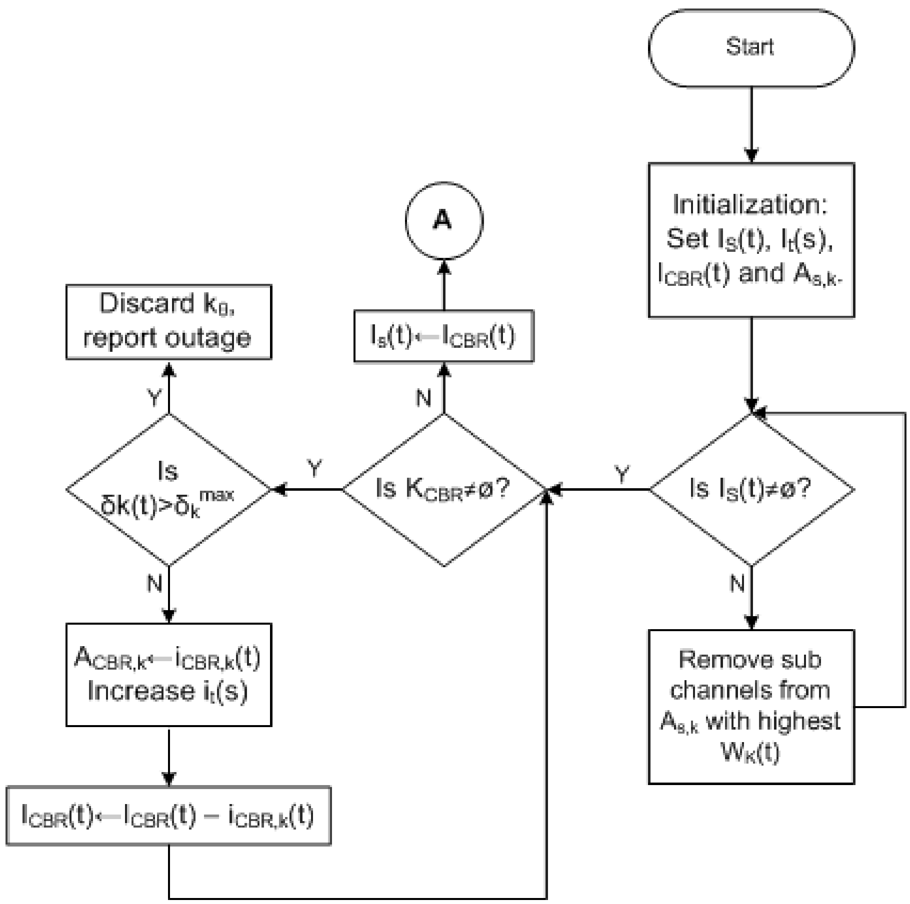

Figure 7: Flowchart for Subchannel Assignment Algorithm (part 1)

The IEEE802.16 standard defines QoS mechanism by associating MAC-layer packets into a service flow. A service flow is identified as a unidirectional flow of packets that has been allotted a particular QoS. Service flows exist in both direction (UL and DL) and

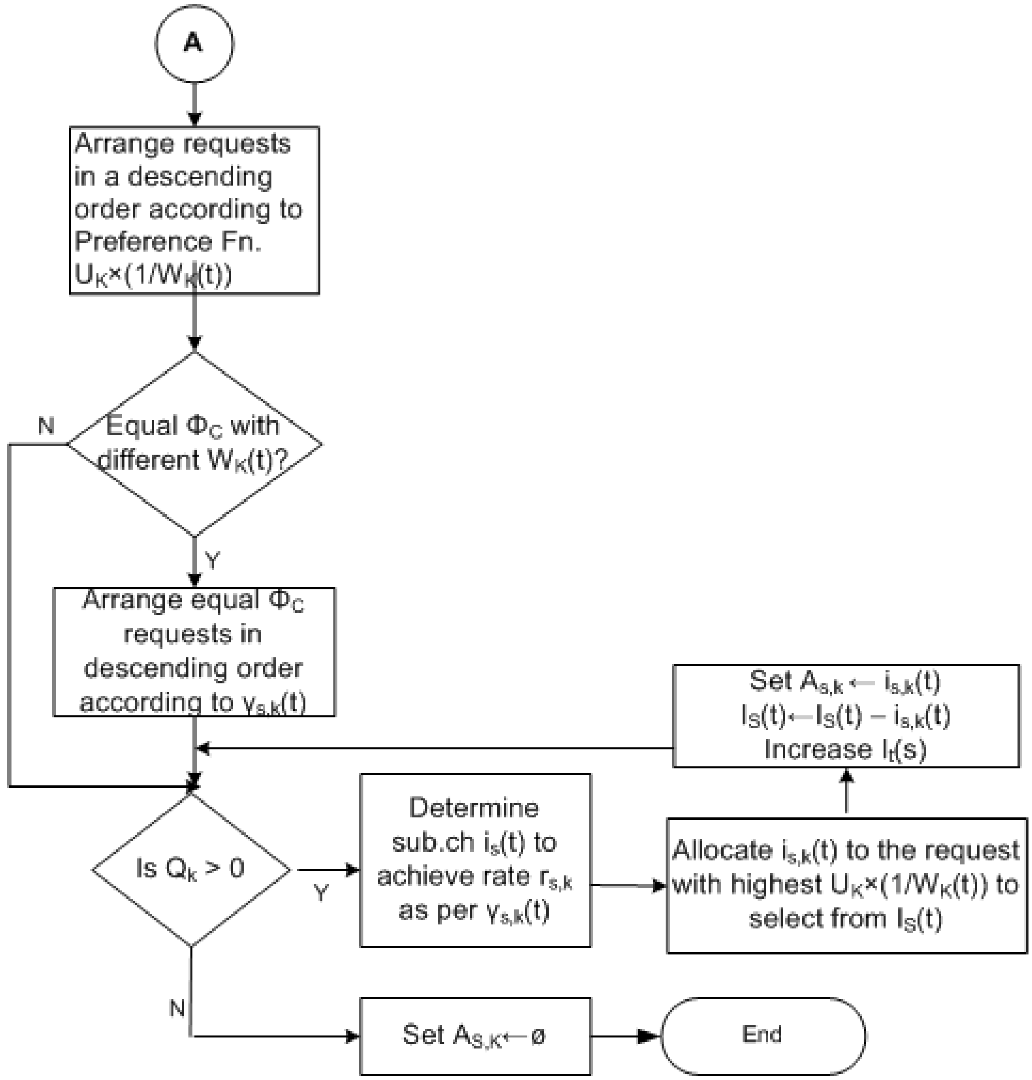

Figure 8: Flowchart for Subchannel Assignment Algorithm (part 2)

typically exist even when they are not activated to carry traffic. All service flows have a 32-bit Service Flow ID (SFID) and active service flows have 16-bit transport Connection Identifier (CID). In OFDMA-TDD, SS/MSs are identified by subcarriers or group of subcarriers and basing on the the QoS requirements, SS/MS's service-flow connections are allotted the time-slots in an OFDMA frame. The mechanism of subcarrier allocation to SS/MSs is determined by the resource allocation policy and occupancy duration of that user's subcarriers in an OFDMA frame is determined by scheduling policy. QoS scheduling support logically resides within the MAC layer of BS and SS/MSs. DL scheduler resides at BS whereas UL scheduler resides both at BS and each SS/MS, where BS UL scheduler allocates UL resources per SS/MS and each SS/MS scheduler allocates the UL grants to its connections. In our scheme we focus on BS scheduler. At the BS each DL connection can be thought to have a packet queue and according to the QoS parameter set and status of the queues, the BS downlink scheduler selects from the DL queues, on a frame basis, the next service data units (SDUs) to be transmitted to the SS/MSs [CLME06]. Alternatively, the queues at the BS can be thought to be queues of requests pertaining to specific QoS types, rather than pertaining to specific connections, since the packet streams (DL and UL) are distinguishable by unique SFIDs and CIDs that are associated with them. In our framework we consider the later approach.

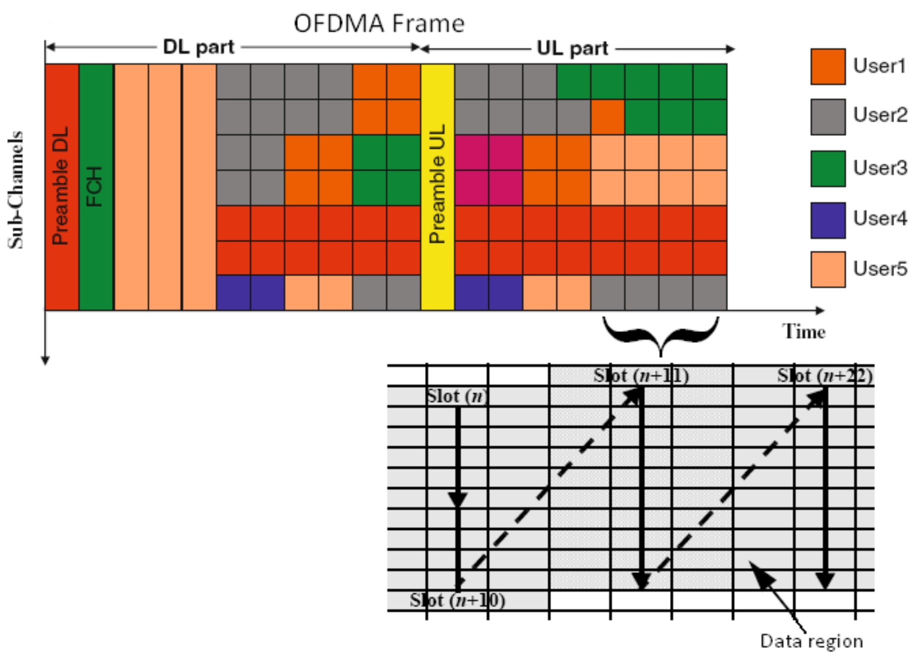

Figure 9: Working of the resource allocation algorithm (adapted from [PV10])

4.4 Admission Control

The **admission control** policy at the base station decides whether it would allow a new connection initiation with QoS request (that has passed authentication and authorization by the security subsystem) basing on the existing network resources. The base station collects all the DSA (Dynamic Service Addition), DSD (Dynamic Service Deletion) and DSC (Dynamic Service Change) requests and accordingly updates estimated available bandwidth. When a subscriber/mobile station requests a service it may either be granted or denied. The denial of service is known as call blocking and its probability is defined as call blocking probability (p_b), which is calculated as the ratio of the number of blocked requests to the total number of requests. The requests that are granted access based on the estimated available system resources and average delay, are called admitted requests. Through successful handover mechanism a mobile user may move from one cell to another. However, if resources are not available for allocating these hand-offs in the new cells or the resource estimation is being ineffective due to signal strength variation, may result in forced discontinuation of service to the user. This is known as call dropping and the probability of such an event is known as call dropping probability or outage probability (p_o), which is calculated as the ratio of the number of requests not fulfilling their throughput and delay requirements to the total number of admitted requests. Given the call blocking and outage probabilities (p_b) and (p_o) respectively, the call completion probability (p_c) is given by [GB06a]:

$$p_c = (1 - p_b)(1 - p_o). \tag{21}$$

Notations Used: We suppose that C different classes of services are provided by the system, which are indexed by c, i.e., $c \in C$ and the system has admitted K^c sets of requests of class c, indexed by k, i.e., $k \in K^c$. We assume that $r^c_{\min}$, $r^c_{\max}$ and $\delta^c_{\max}$ denote respectively the Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate and Maximum Delay associated with the traffic of class c.

4.4.1 Bandwidth Control

The available bandwidth of the system can be specified either in data rate or frequency range. According to the Shannon-Hartley theorem, the data rate of communication is directly proportional to the frequency range of the signal used for the communication. However, in all the analysis that follows, we assume data-rate bandwidth. The admission criterion for a new request (k') is given by:

$$(\sum_{c \in C} \sum_{k \in K^c} r^c_{\min}(k) + BW_{sys}) + r^c_{\min}(k') < BW \tag{22}$$

where BW is a parameter that specifies total capacity of the system. This parameter is assumed to be the bandwidth excluding what is needed for signalling purposes, such as: initial ranging messages, primary management messages etc. BW_{sys} is a 'buffer' bandwidth that can be used by the system to give slightly greater than the minimum rate whenever possible and that also remains available for high priority new incoming requests.

4.4.2 Delay Control

In order to meet the delay requirement of the request, we need to consider the delay incurred in queuing process, which has two components. We assume that the BS and SS/MSs maintain separate queues for different classes of traffic, and there is an elapsed time by

Convergence Specific Sublayer at MAC to enqueue these queues, this contributes to first component of delay. And after a request is queued, a number of OFDM frames are needed to complete the transmission of its requested data blocks and during this time the data remains in the queue, this contributes to another component of delay. We assume that initial queuing delay is negligible compared to scheduling delay. Both these delays are closely related with the data rate asked by requests. The more data rate is asked, the more time is required in queuing and scheduling it, thus contributing to delay. Let $\bar{\delta}^c$ be the average delay that an admitted request of class c undergoes. Following should hold for the admitted requests:

$$\overline{\delta}^c(k) \leq \delta^c_{\max}(k). \tag{23}$$

4.4.3 Relating Bandwidth and Delay

The condition (22) does not represent functional behavior of the system since it does not include delay constraint. We therefore define average throughput of the system, $\overline{R}(T_w)$, which is the amount of data transferred and measured over a time period T_w. This window time period is a multiple of Frame Time T_{frame}. And the average throughput of the admitted requests is denoted by $\overline{R^c}(k, T_w)$. Over any period T_w, following must hold true:

$$\overline{R}(T_w) \leq BW. \tag{24}$$

With the consideration of delay, condition (22) can be re-written in terms of throughput as in (25):

$$
\begin{aligned}
(\sum_{c \in C} \sum_{k \in K^c} r^c_{\min}(k) + BW_{sys}) + r^c_{\min}(k') < \overline{R}(T_w) \\
\Rightarrow \sum_{c \in C} \sum_{k \in K^c} \frac{\bar{\delta}^c(k)}{\delta^c_{\max}(k)} \overline{R}^c(k, T_w) + r^c_{\min}(k', T_w) < \overline{R}(T_w). \quad (25)
\end{aligned}
$$

We further define the minimum usage of resources by each class of traffic to prevent starvation for lower-priority traffics. For this we assume that the BS has a database of usage statistics during prescribed interval of day. This can be used to dynamically set the limit for minimum resource used by different classes of traffic at different times of the day. At a given interval of a day (t), we assume that the estimated bandwidth occupied by ongoing Constant Bit Rate traffic (i.e., UGS), Variable Bit Rate traffic (i.e., ertPS and rtPS) and Non Realtime traffic (i.e., nrtPS and BE) are respectively denoted by: $\overline{R}_{Q_1}(t)$, $\overline{R}_{Q_2}(t)$ and $\overline{R}_{Q_3}(t)$. We generalize this by $\overline{R}_{Q_i}(t)$. If the sum of the estimated bandwidth used by ongoing connections ($\overline{R}_{Q_i}(t)$) and the estimated bandwidth to be used by the incoming service request (k') is larger than a predefined upper threshold $R^{\max}_{Q_i}(t)$, the incoming request is rejected; otherwise, if the request is below the minimum usage threshold $R^{\min}_{Q_i}(t)$, it is accepted with certain probability depending on the estimated bandwidth usage and the connection priority. We denote service-class priority weight by $p_{Q_i} \in [0, 1]$. This parameter helps to prevent unnecessary blocking request due to previous congestion, for an example, if $p_{Q_i} = 90\%$, and the system load is below that point, it will try to admit new requests without checking the priorities. It is assumed that in a relaxed state, all requests have an equal chance of being admitted by the system.

These parameters are checked while admitting a request into the system. For realtime requests, both the rate and delay constraints are checked and for non realtime requests only the rate constraint is verified. The Admission Control Policy is outlined in algorithm 4.2.

{Intialization}
Set $I_s(t) \leftarrow$ Available subchannels at timeslot t $\{I_s(t) \subseteq S\}$
Set $I_t(s) \leftarrow$ Residual timeslots on subchannel s $\{I_t(s) \subseteq T\}$
Set $I_{CBR}(t) \leftarrow$ Reserved subchannels at timeslot t $\{I_{CBR}(t) \subseteq I_s(t)\}$
Set $A_{s,k} \leftarrow$ set of assigned subchannels to connection request k indexed by $a_{s,k}$
Get Q_k for $\forall k$ {get the queue size of connection request k}
if $I_s(t) \neq \emptyset$ **then**
  if $K_{CBR} \neq \emptyset$ **then**
    for each reserved subchannels in $I_{CBR}(t)$, i.e., $i_{CBR} \in I_{CBR}(t)$ **do**
      if $\delta_k(t) > \delta^{max}$ **then**
        Discard request k_θ and report outage
      else
        Set $A_{CBR,k} \leftarrow i_{CBR,k}$(t)
        Increase $i_t(s)$
      end if
      $I_{CBR}(t) \leftarrow I_{CBR}(t) - i_{CBR,k}(t)$ {Remove from $I_{CBR}(t)$ all the subchannels that has been assigned a user}
    end for
  else
    $I_s(t) \leftarrow I_{CBR}(t)$ {Release the reserved subchannels and include them in available subchannel set}
    Calculate $U_k \times \frac{1}{W_k(t)}$ and sort k in descending order, constituting the set K_{ord}
    for each service request, $k \in K_{ord}$ **do**
      if $Q_k > 0$ **then**
        Determine the subchannel $i_s(t)$ for $k \in K_{ord}$ to achieve rate $r_{s,k}$ as per the SNR $\gamma_{s,k}(t)$
        Set $A_{s,k} \leftarrow i_{s,k}(t)$ {i.e., $\phi_{s,k}(t) \leftarrow 1$}
        $I_s(t) \leftarrow I_s(t) - i_{s,k}(t)$
        Increase $i_t(s)$
        if $L(Q_k) < L(PDU)$ **then**
          Update $i_t(s) \leftarrow i_t(s) - \frac{L(Q_k)}{r_{s,k}}$
        else
          Update $i_t(s) \leftarrow i_t(s) - \frac{L(PDU)}{r_{s,k}}$
        end if
      else
        Set $A_{s,k} \leftarrow \phi$ {set $A_{s,k}$ to null}
      end if
    end for
  end if
else
  Remove subchannels from $A_{s,k}$ with highest $W_k(t)$
end if

Algorithm 4.1: Subchannel Assignment Algorithm

Keep statistics of average throughput parameters $\overline{R}(T_w)$, $\overline{R}^c(k, T_w)$, $\overline{R}_{Q_i}(t)$ and average delay $\overline{\delta}^c(k)$ of class c for some multiples of T_w msec;
Initialize $N^c(k')$ (Number of new requests to be admitted by the system)
Categorize new service requests according to Q_i
for each service request, k' **do**
  Collect traffic parameters: Q_i, $r^c_{\min}(k')$, $\delta^c_{\max}(k')$;
  if $r^c_{\min}(k') > \overline{R}(T_w)$ and $\delta^c_{\max}(k') < \overline{\delta}^c(k)$ **then**
    deny request;
  end if
  if $r^c_{\min}(k') < \overline{R}(T_w)$ and $\delta^c_{\max}(k') > \overline{\delta}^c(k)$ but $r^c_{\min}(k') > R^{\max}_{Q_i}(t)$ **then**
    reject request;
  else
    if Request k' satisfies relation (25) and $\delta^c_{\max}(k') > \overline{\delta}^c(k)$ and $r^c_{\min}(k') < R^{\min}_{Q_i}(t)$ **then**
      accept request k' with probability p_{Q_i};
    else
      accept request k' with probability $p_{Q_i} . \frac{\sum_{c\in C}\sum_{k\in K^c}\frac{\overline{\delta}^c(k)}{\delta^c_{\max}(k)}\overline{R}^c(k,T_w)+r^c_{\min}(k',T_w)-R^{\max}_{Q_i}(t)}{R^{\max}_{Q_i}(t)-R^{\min}_{Q_i}(t)}$;
    end if
  end if
  $N^c(k') \leftarrow N^c(k')$ +1;
end for

Algorithm 4.2: Admission Control Policy

Chapter 5

Experimental Results and Analysis

In chapter 4, the new scheduling and resource allocation scheme along with the call admission control policy have been proposed. The performance of these algorithms are further investigated analytically with the help of a custom-built C++ simulator in this chapter. A few free C++ source libraries from Boost [Lib] were used for generating some traffic distribution.

The chapter is organized as follows: The simulation model used in the experiments is presented in Section 5.1. In Sections 5.2 and 5.3 simulations results and analysis are presented respectively. We use the same notations as in chapter 4, and therefore, will not repeat their definitions in this chapter, unless they are new notations.

5.1 Simulation Model

A cellular WiMAX network with a single BS using OFDMA is developed as the system model. The FFT size is set to 512, hence the number of sub-carriers N is 512, of which data sub-carriers is equal to 360. The system bandwidth, B is 5 MHz and number of OFDM symbols in 5 ms frame is equal to 48. The throughput is averaged over a window of 200 OFDM frames. The data rate is upper bounded by Shannon's capacity formula:

$$\overline{R}_k = \sum_{s \in S_k} \phi_{s,k}(B/N) \times log_2\{1 + \overline{\gamma_{s,k}}\} \tag{26}$$

All the Subscriber's Stations are assumed to have independent Rayleigh fading channels. The average received SNR ($\gamma_{s,k}$) of each request (k) from the stations is derived by [CB07], [CDG02]:

$$\overline{\gamma_{s,k}} = \rho \times (D/d_k)^{\alpha} \times S_r \tag{27}$$

where D is the radius of the circle that circumscribes the hexagonal cell, d_k is the distance from the BS to the user generating request k, α is the path-loss exponent (estimated to be about 4.0 in typical urban environment), S_r is a random variable for the shadow-fading effect. The simulation parameters are summerized in table (3)

In each SS, there are four queues corresponding to four traffic classes. However, we assume each SS generates one class of requests at any time, although it can generate requests of any class at different times (i.e., a given user may use different applications). Each queue has an infinite backlog of data.

5.2 Simulation Results: Proper Management of the QoS Requirements

We conducted three sets of experiments that are described next. These experiments deal with successfully admitted requests into the system. Through these experiments we would like to find out if the scheduling and resource algorithm is able to differentiate service classes

Table 3: Simulation Parameters

Parameter Name	Assumed Values
System	OFDMA/TDD
System Bandwidth	5 MHz
FFT Size	512
Frame Duration	5 ms
Downlink-to-Uplink TDD split	1 : 1
Useful Symbol Time (μs)	91.4
Number of OFDM symbol/frame	48
Pathloss Exponent	2
Shadow Fading Effect	6
Shadowing correlation	1.0 for intra-site, 0.5 for inter-site
Median SNR	100
User Distribution	Uniform
Beam Pattern	Omnidirectional
Cell radius	2 km
BS transmit power	20 W
User maximum power	150 mW
Channel Model [Mod]	802.16 SUI-5

Table 4: QoS parameters taken from [AGM07]

Service Type	Rate Budget	Maximum Delay (ms)
$Voice(UGS)$	64 Kbps	100
$Video(rtPS)$	512 kbps	100
$FTPData(nrtPS)$	200 kbps	200
$Web(BE)$	200 kbps	300

and meet the data rate and delay requirements of the QoS traffic under varied offered load scenarios. Each simulation figure shows the mean obtained over fifteen instances.

1. **Experiment 1:** This experiment investigates whether the delay requirements of the realtime traffic is met under low, medium and heavy traffic scenarios. The experiment was conducted with one BS and 70 SSs. There are 5 rtPS requests, 15 nrtPS requests, 15 BE connections and the number of active UGS requests varies from 10 to 35. As can be seen in the Figure (10) the delay of the UGS traffic remained unaffected with UGS load increase. However, there was slight oscillations of rtPS traffic as the higher

Table 5: Traffic Parameters

Type	Traits	Distribution	Parameters
Voice [Bra69](UGS)	ON period	Exponential	Mean = 1.2 s
	OFF period	Exponential	Mean = 1.8 s
	Packet size	Constant	66 bytes
	Inter-arrival time	Constant	20 ms
Video [WD10] (rtPS)	Packet Size	Log-normal	Mean = 4.9 bytes Std. dev. = 0.75 bytes
	Inter-arrival time	Normal	Mean = 33 ms Std. dev. = 10 ms
FTP [FF07] (nrtPS)	Packet size	Exponential	Mean = 512 Kbytes
Web [FF07] (BE)	Packet size	Pareto	Mean = 10558 bytes

priority fixed grant UGS requests increased in the traffic profile. Noticeable that the delay requirements are much below the realtime traffic constraint.

2. **Experiment 2:** This experiment verifies whether or not the increase of UGS/rtPS traffic load has an impact on the QoS level of services with lower priority. This demonstrates that the scheduler was able to provide fixed data grants quite efficiently. The simulation environment includes includes one BS and 91 SSs. In the experiment, there are 6 rtPS connections, 20 nrtPS connections, 20 BE connections, and the number of active UGS connections varies from 5 to 45. As can be seen in the Figure (11), the throughput of the UGS traffic (the realtime voice) remains constant over the experimentation period. The voice data burst is efficiently scheduled with throughput of 64kbps. The usage of adaptive modulation and coding enables the scheduler to efficiently use the radio resources. The throughput of rtPS traffic slightly gets decreased with the increase of UGS traffic, however, lies within the allowable bandwidth requirement for realtime data. The decrease of throughput of non realtime data is well-understood since a system is designed for a specific capacity and accordingly the maximum system bandwidth is restricted by Shannon's equation.

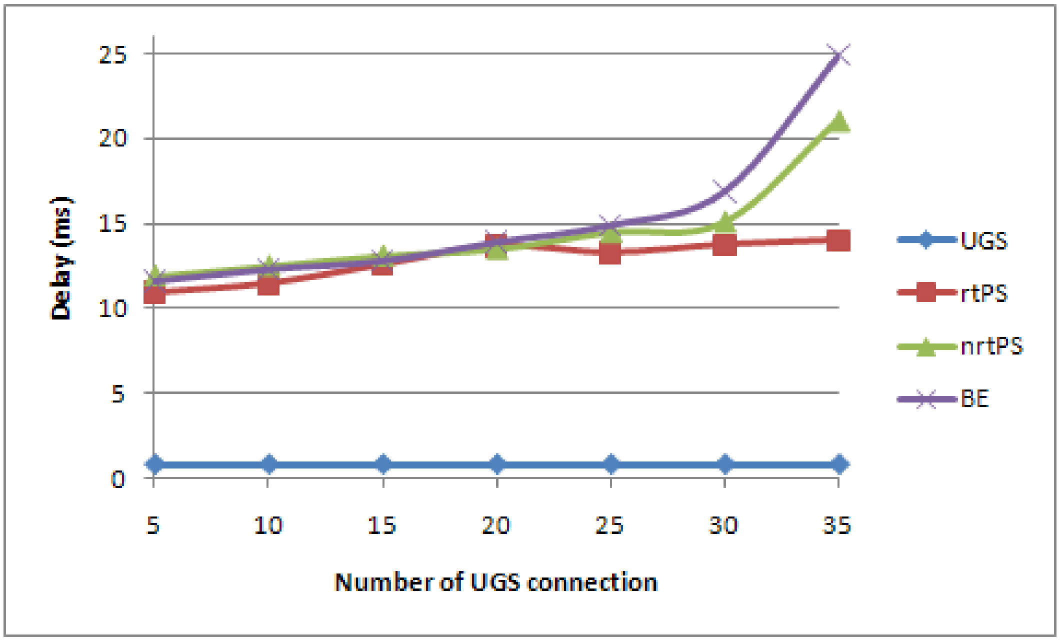

Figure 10: Delay of UGS, rtPS, nrtPS and BE requests with varying UGS load

3. **Experiment 3:** This experiment finds out whether or not the increase of BE traffic load has an impact on the QoS level of services with higher priority (UGS and rtPS). The experiment was conducted with one BS and 62 SSs. There are 12 UGS requests, 5 rtPS requests, 10 nrtPS requests, and the number of active BE requests varies from 10 to 35. As can be seen in the Figure (12), The delay of the UGS traffic remained unaffected with BE load increase. This demonstrates that the scheduler was able to provide fixed data grants quite efficiently. However, there was slight oscillations of rtPS traffic delay.

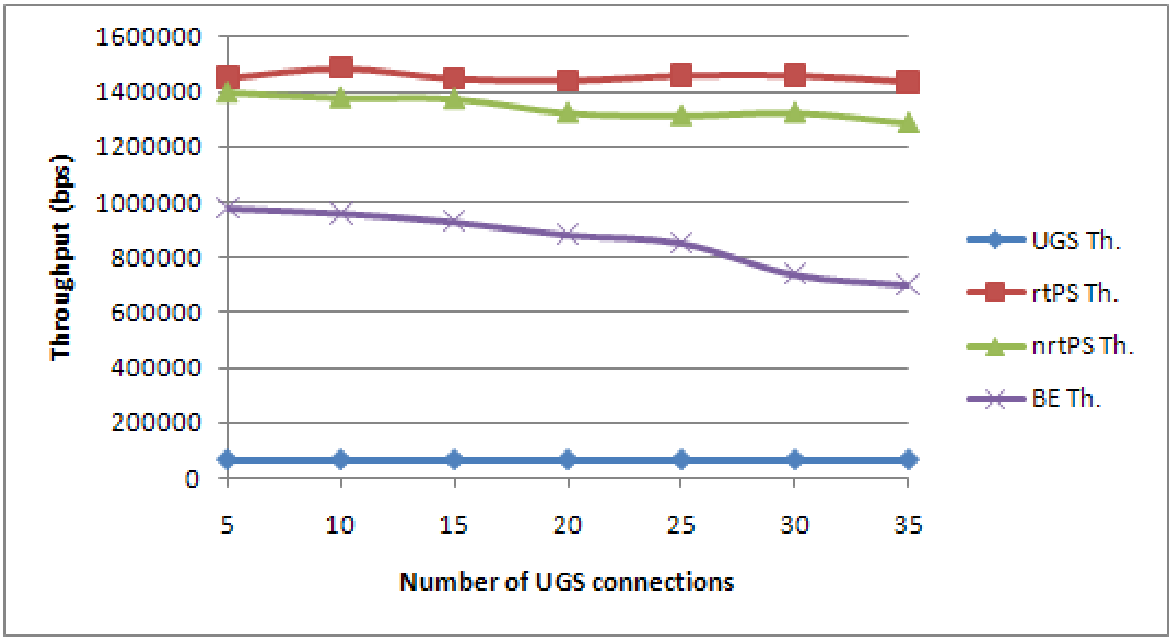

Figure 11: Throughput of UGS, rtPS, nrtPS and BE connections with varying UGS load

5.3 Comparison of Resource Allocation performance with other QoS scheduling schemes

We have chosen to compare our algorithm (algorithm QoS-SRC in Figures (13), (14)) with that of reference [FF07] (algorithm QoS-Comp in Figures (13), (14).

Varying UGS load scenario Both our nrtPS and BE throughput excels in performance compared to [FF07] because of ours efficient usage of Adaptive Modulation and Coding, which the other algorithm did not have the provision of. As can be seen in (13), the BE performance of [FF07] worsens much more sharply with increasing UGS traffic compared to ours.

Varying BE load scenario In a scenario with varying BE traffic, Figure (14), the throughput performance of nrtPS requests is almost similar for both the algorithms. However, we see that the BE throughput decreases sharply after some point, with the increase

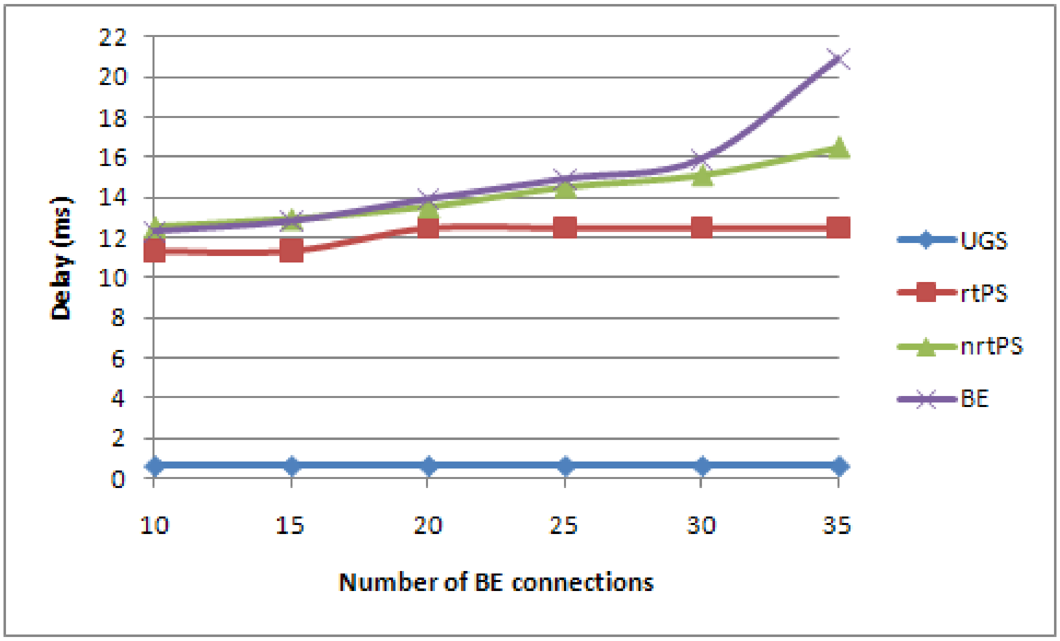

Figure 12: Delay of UGS, rtPS, nrtPS and BE requests with varying BE load

of BE traffic in case of algorithm proposed in [FF07].

Varying rtPS load scenario

In a scenario with varying rtPS traffic, Figure (15), the throughput performance of nrtPS requests is almost similar for both the algorithms. However, we see that the BE throughput decreases sharply (almost linearly) after some point, with the increase of rtPS traffic in case of algorithm proposed in [FF07]. This sharp performance decrease is not desirable in any system.

Comparing short-term fairness

Our algorithm exhibits more fairness in terms of usage of sub-carriers by the requests generated at different distances from the base station as can be seen in Figure (16). This is more evident when we consider the throughput, as depicted in Figure (16). Because of the usage of adaptive modulation, the stations that are closer to the base station receive stronger subcarriers, but they are assigned a fewer of them compared to those at a greater

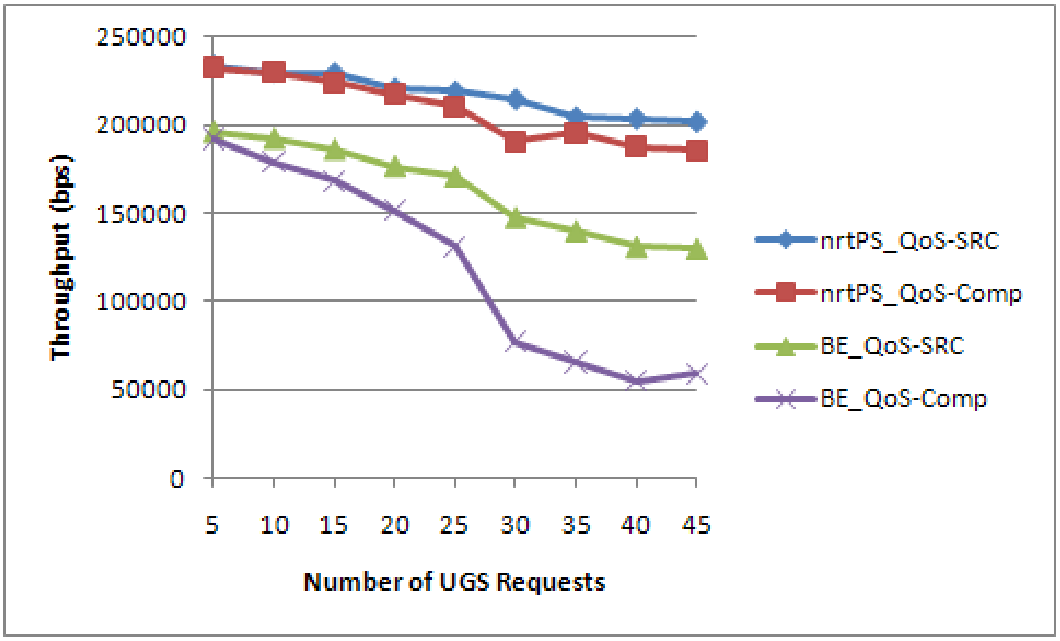

Figure 13: Throughput performance comparison of algorithms QoS-SRC and QoS-Comp, varying UGS load

distance, that receive subcarriers of less capacity. Consequently, the average throughput achieved by the users having the same class of traffic, located at different distances from the base station is fairly even.

For carrying out this experiment, we configured the environment to generate only one type of traffic at a time so that the two algorithms can be compared. The subcarrier and throughput information of 70 SSs generating nrtPS traffic are shown in Figures (16) and (17).

5.4 Comparison of Admission Control policy performance

Since Algorithm ($QoS - Comp$) of [FF07] assumes the preexistence of Call Admission Control (CAC), we make the performance comparison of our CAC scheme with the one presented in [CJG05] (denoted by $CA - Comp$) where the policy only checks the minimum reserved

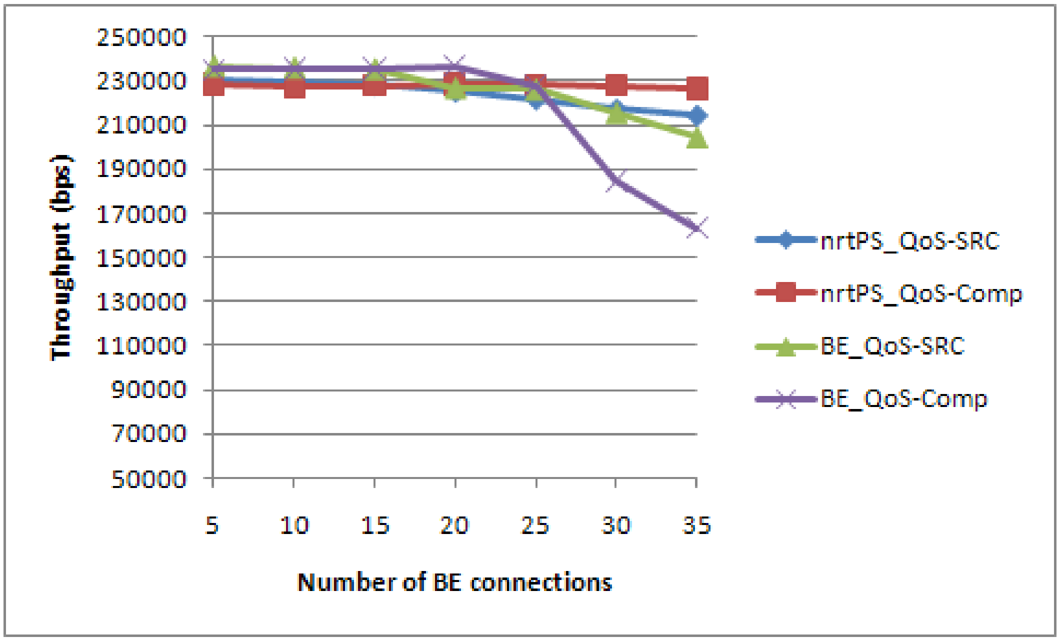

Figure 14: Throughput performance comparison of algorithms QoS-SRC and QoS-Comp, varying BE load

bandwidth of the admitted requests, not considering the delay requirements of the requests. According to [CJG05], the new request admission-deciding function is BW_a, the available bandwidth. BW_a is defined as:

$$BW_a = BW_{total} - \sum_{c \in C} \sum_{k \in K^c} r^c_{\min}(k) \tag{28}$$

Where, BW_{total} is the total capacity of the wireless link. The system continues to admit new requests as long as the value of BW_a is non-zero, i.e., $BW_a > 0$. We assume that the arrival process of the traffic generated to the queue follows a Poisson distribution. Although such a distribution is questionable in order to model the traffic, we have used it here for making the comparison of our algorithm to the referenced one.

Comparing Blocking probabilities: Figure (18) and Figure (19) respectively depict the blocking probabilities for our proposed Admission Control algorithm (CA_{SRC}) and the

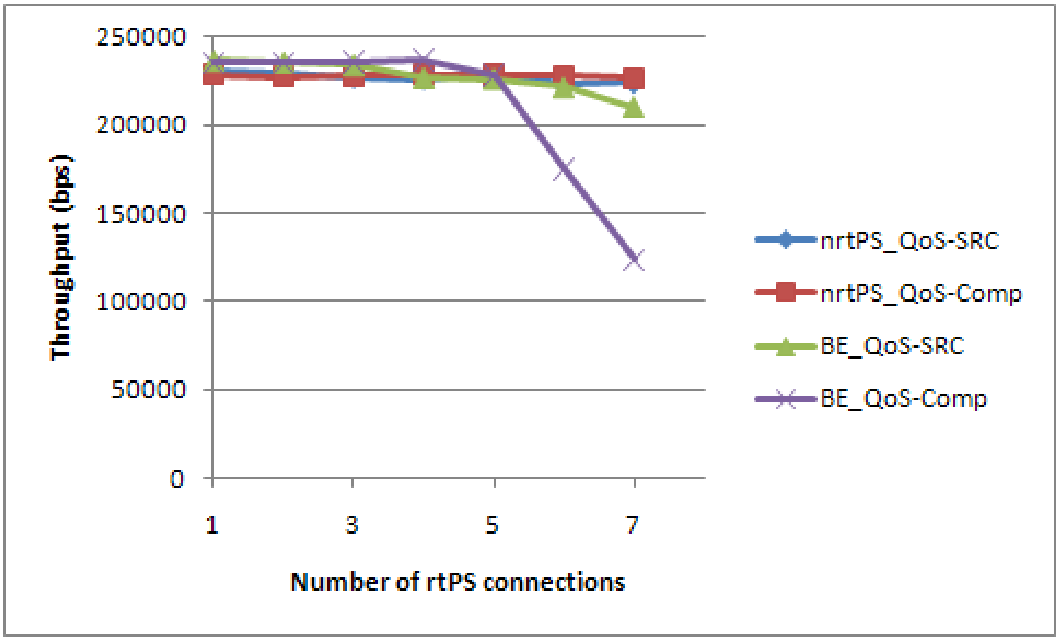

Figure 15: Throughput performance comparison of algorithms QoS-SRC and QoS-Comp, varying rtPS load

algorithm we are comparing with (CA_{Comp}) as a function of the call arrival rate. We observe that the blocking probability increases as the call arrival rate increases. Furthermore, the blocking probability is greater for rtPS and nrtPS classes of traffic since they more demanding in terms of bandwidth compared to UGS traffic.

There are some major differences in Figure (18) and Figure (19). Firstly, as can be seen in figure (19), the BE request is always admitted by CA_{comp} admission policy since for these requests, $r^c_{\min}(k)$ is equal to zero. However, admitting BE connections indefinitely, would cause these non-QoS requests to reach an utterly unusable data rate. Secondly, in figure (18), the UGS blocking probability increases very slowly with the increasing arrival rate due to the reserved channel set as prescribed by the standard. Thirdly, in figure (18), we observe that the blocking probabilities follow the required priority scheme (i.e. $UGS > rtPS > nrtPS$), while, in figure (19), the priority distribution is random.

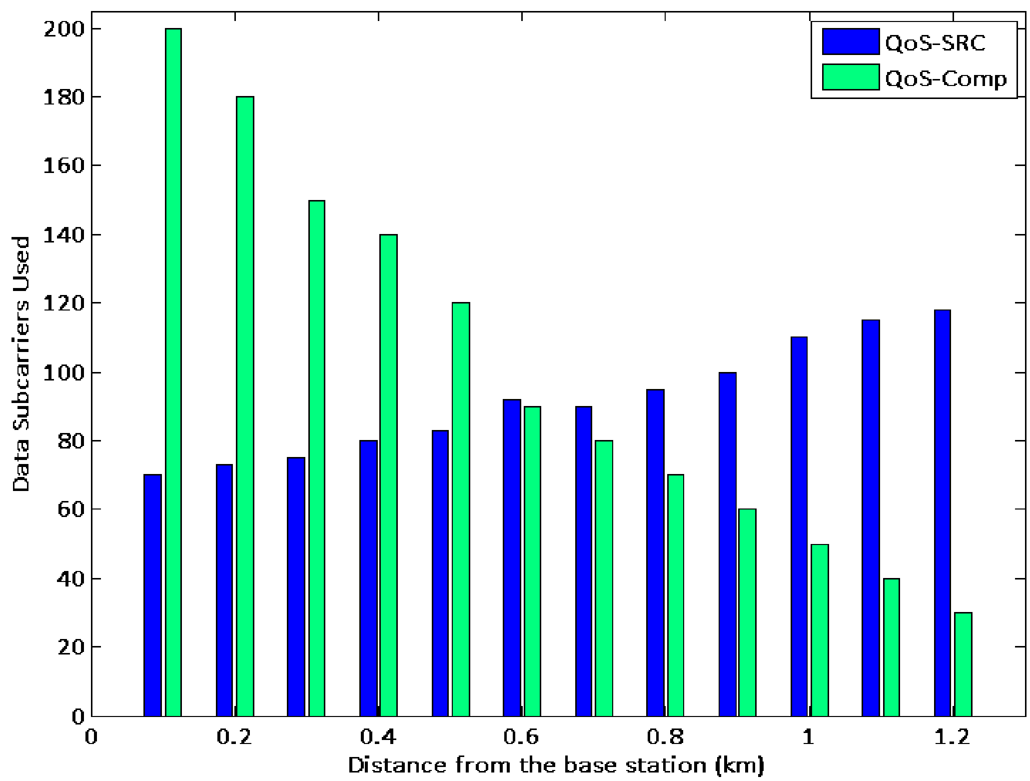

Figure 16: Comparing fairness: average subcarrier allocation vs. distance from the BS

Comparing Outage probability:

As can be seen in figure (20), the outage probabilities are compared. We find that for the requests of the same traffic type the greedy approach described in [CJG05] (indicated by the suffix -comp with the traffic type in the figure) exhibits greater outage probability than ours. When an admitted request gets allocation of sub-carriers, the transmission delay would depend on the current data capacity of the sub-carriers, the request's previous allocation and the remaining bits. The more data rate a request have, the more time it has to stay in the system before suitable sub-carrier allocation, the more likely the request gets closer to its expiry. Since the algorithm CA_{comp} does not consider this delay while admitting, its outage probability gets higher as the requests are more likely to get expired.

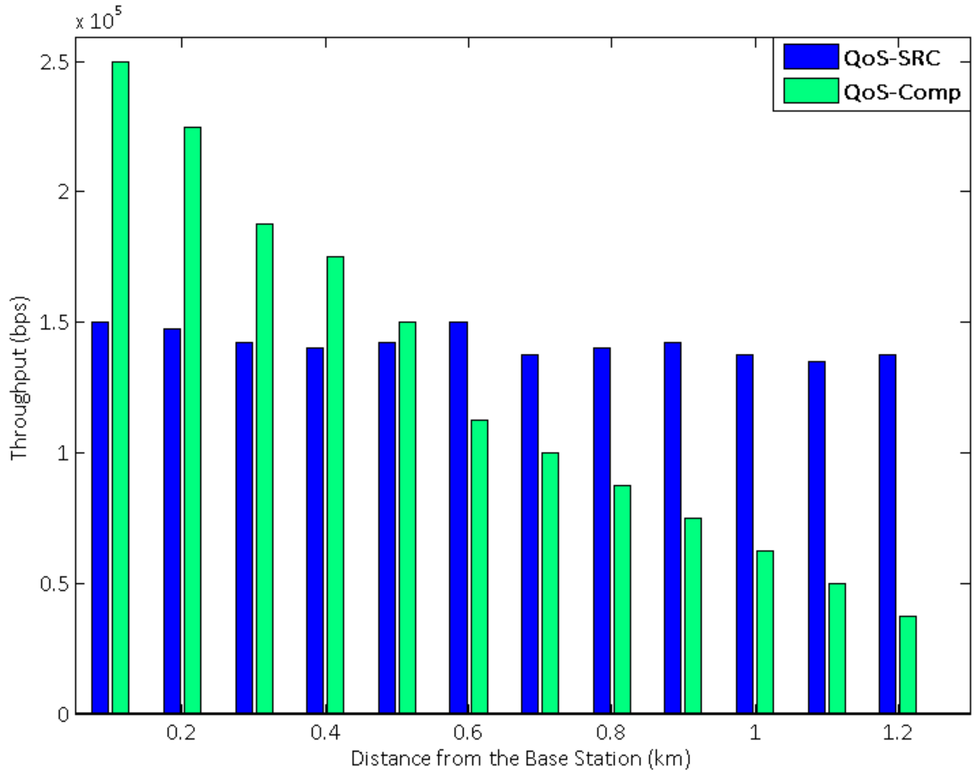

Figure 17: Comparing fairness: average throughput vs. distance from the BS

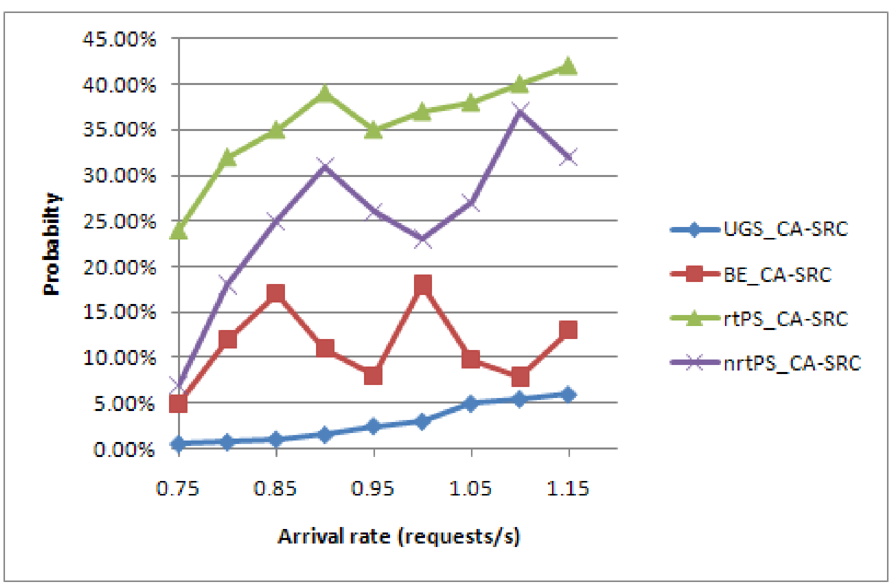

Figure 18: Algorithm CA_{SRC} blocking probabilities vs. arrival rate

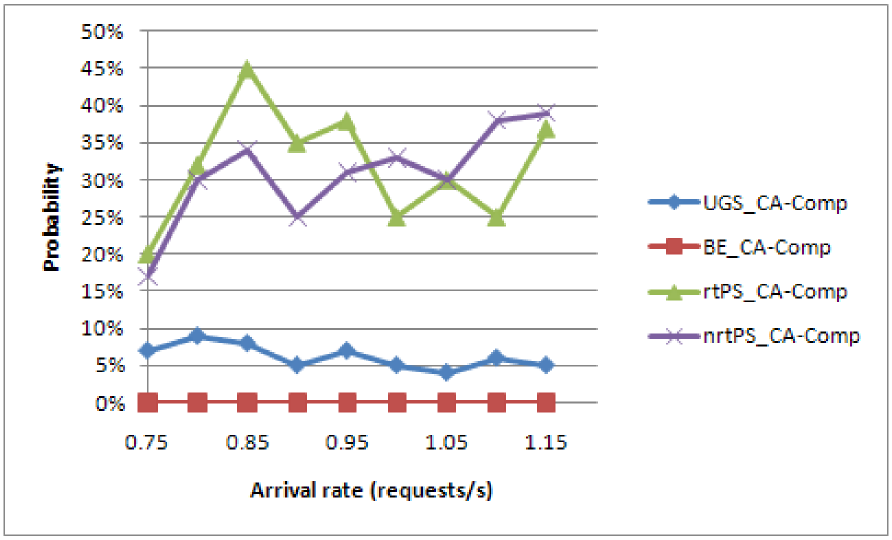

Figure 19: Algorithm CA_{Comp} blocking probabilities vs. arrival rate

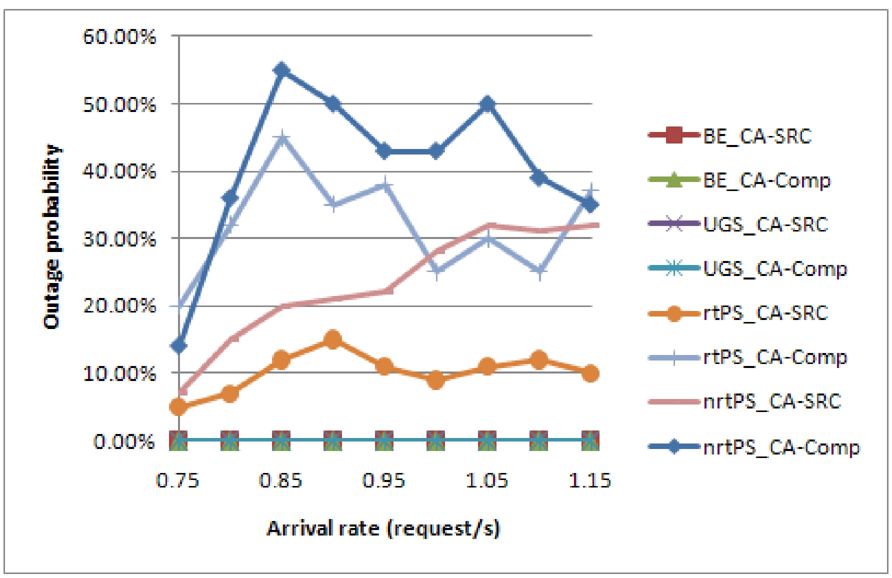

Figure 20: Comparing outage probability

Chapter 6

Conclusion

6.1 Summary of the Thesis

WiMAX/IEEE802.16 network is the perfect candidate to realize the present demand of high speed mobile Internet. However, the success of this network depends not only on the underlying technologies, but also heavily relies on its efficient scheduling algorithms and resource allocation strategies that facilitate end-to-end Quality of Service. Scheduling algorithms determine the transmission opportunities of the participating stations while resource allocation algorithms fairly utilize the system resources among the users. Furthermore, to enhance the capacity of WiMAX network, a critical part is to design admission control mechanism that ensures QoS for the newly admitting requests and minimizes resource outage for the existing users in the system.

Keeping all the above points in mind, a new scheduling and resource allocation with call admission control, SRC algorithm for OFDMA/TDD systems, has been proposed in this thesis. Instead of dealing queues of different classes with different scheduling schemes, one uniform preference function is defined to assign transmission opportunities for each traffic request. This preference metric is conceptually based on the proportional fair scheduling with a modified utility function that makes it simpler to implement. In addition, our algorithm takes into account the QoS satisfaction in terms of average delay and rate budget for each class of QoS request. Furthermore, we associate admission control for the multi-class systems. We have investigated performance of the scheme through extensive simulation. It is observed that the proposed scheme offers a very good performance in the aspects of service differentiation, throughput guarantees and fairness for all types of services.

6.2 Future Work

It is a challenging research topic by itself to provide QoS scheduling for four classes of traffic services in WiMAX network with call admission control mechanism. However, there are a bunch of issues that can be further investigated.

The scheduling algorithm that has been developed here is based on the specification as per IEEE802.16 for single-cell systems. It is necessary to investigate the behavior of our approach in multi-hop relay scenario of IEEE802.16j standard.

In our current work, we have focused on WiMAX uplink systems. The scheme is supposed to be readily extendable to downlink. However, this is worth investigating to find out the performance of the algorithm in a combined admission control strategy for downlink and uplink scenario.

The proposed algorithm performance assumes the perfect channel quality information (CQI) at the base station. However, unreliable channel state measurement may result in wrong decision being made by the algorithm. It is important to study the effects of imperfect CQI, for example, channel estimation error, Doppler shift and feedback delay.

Bibliography

[80204] Part 16: Air interface for broadband wireless access systems. IEEE Computer Society and IEEE Microwave Theory and Techniques Society, 2004.

[80205] Part 16: Air interface for broadband wireless access systems. IEEE Computer Society and IEEE Microwave Theory and Techniques Society, 2005.

[80209] IEEE Standard for Local and Metropolitan Area Networks -Part 16: Air Interface for Broadband Wireless Access Systems Amendment 1: Multiple Relay. IEEE Computer Society and IEEE Microwave Theory and Techniques Society, June 2009.

[80210] IEEE Standard for Local and Metropolitan Area Networks -Part 16: Air Interface for Broadband Wireless Access Systems Amendment 2: Improved Coexistence Mechanisms for License-Exempt Operation. IEEE Computer Society and IEEE Microwave Theory and Techniques Society, July 2010.

[AGM07] J. G. Andrews, A. Ghosh, and R. Muhamed. Fundamentals of WiMAX Understanding Broadband Wireless Networking. Prentice Hall, 2007.

[AWA08] M. Algamali, J. Wang, and M.Y. Abdullah. OBSS: Optimal Base Station Scheduler of IEEE 802.16 Mesh mode. In Wireless Communications, Networking and Mobile Computing, WiCOM, 2008.

[BGWM07] M. Bohge, J. Gross, A. Wolisz, and M. Meyer. Dynamic Resource Allocation in OFDM Systems: An Overview of Cross-layer Optimization Principles and Techniques. In IEEE Network, volume 1, pages 53–59, February 2007.

[BN08] A. Belghith and L. Nuaymi. Comparison of WiMAX Scheduling Algorithms and Proposals for the rtPS QoS Class. In 14th European Wireless Conference, 2008.

[Bra69] P. T. Brady. A model for generating On-Off speech patterns in Two-Way conversations. In Bell System Technical Journal, volume 48, pages 2445–2472, 1969.

[BTGK05] C.F. Ball, F. Treml, X. Gaube, and A. Klein. Performance Analysis of Temporary Removal Scheduling applied to mobile WiMAX Scenarios in Tight Frequency Reuse. In 16th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications, PIMRC, 2005.

[CAL07] C. Cicconetti, I. F. Akyildiz, and L. Lenzini. Bandwidth Balancing in Multi-Channel IEEE 802.16 Wireless Mesh networks. In INFOCOM, 2007.

[CB07] J. Choi and S. Bahk. Cell Throughput Analysis of the Proportional Fair Scheduler in the Single Cell Environment. In IEEE Transactions on Vehicular Technology, volume 56, March 2007.

[CDG02] S. Catreux, P. F. Driessen, and L. J. Greenstein. Data Throughputs Using Multiple-Input Multiple-Output (MIMO) Techniques in a Noise-Limited Cellular Environment. In IEEE Transactions on Wireless Communications, volume 1, April 2002.

[CELM07] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi. Performance Evaluation of the Mesh Election Procedure of IEEE 802.16/WiMAX. 2007.

[CGLM07] C. Cicconetti, V. Gardellin, L. Lenzini, and E. Mingozzi. End-to-End Bandwidth Reservation in IEEE 802.16 Mesh Networks. In Mobile Adhoc and Sensor Systems, MASS, 2007.

[CJG05] J. Chen, W. Jiao, and Q. Guo. An integrated QoS control architecture for IEEE 802.16 broadband wireless access systems. In Global Telecommunications Conference, IEEE GLOBECOM, December 2005.

[CJW05] J. Chen, W. Jiao, and H. Wang. A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode. In IEEE International Conference on Communications, volume 5, pages 3422–3426, 2005.

[CLME06] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund. Quality of Service Support in IEEE 802.16 Networks. In IEEE Network, volume 20, pages 50–55, March.-April. 2006.

[CTWW07] L. W. Chen, Y. C. Tseng, D. W. Wang, and J. J. Wu. Exploiting Spectral Reuse in Resource Allocation, Scheduling, and Routing for IEEE 802.16 Mesh Networks. In IEEE Vehicular Technology Conference, VTC, 2007.

[CWL08] Y. C. Chan, C. L. Wu, and C. Y. Lai. Congestion-Aware Downlink Scheduling for IEEE 802.16j Multi-Hop Relay Networks. In Proceedings of the 4th Annual International Conference on Wireless Internet, WICON, 2008.

[Del10] O. Delgado. Admission Control and Resource Allocation for LTE Uplink Systems. Master's thesis, Concordia University, August 2010.

[DJ10a] O. Delgado and B. Jaumard. Scheduling and Resource Allocation for Multiclass Services in LTE Uplink Systems. In IEEE Wireless and Mobile Computing, Networking and Communications (WiMob), pages 355–360, October 2010.

[DJ10b] O. Delgado and B. Jaumard. Scheduling and Resource Allocation in LTE Uplink with a Delay Requirement. In Communication Networks and Services Research Conference (CNSR), pages 268–275, May 2010.

[DMR08] S. Deb, V. Mhatre, and V. Ramaiyan. WiMAX Relay Networks: Opportunistic Scheduling to Exploit Multiuser Diversity and Frequency Selectivity. In Proceedings of the 14th ACM international conference on Mobile computing and networking, MobiCom, 2008.

[Erg09] M. Ergen. Mobile Braodband Including WiMAX and LTE. Springer, 2009.

[FF07] J. Freitag and N. L. S. Fonseca. Uplink Scheduling with Quality of Service in IEEE 802.16 Networks. In IEEE Global Telecommunications Conference, GLOBECOM, pages 2503–2508, November 2007.

[FL02] H. Fattah and C. Leung. An Overview of Scheduling Algorithms in Wireless Multimedia Networks. In IEEE Wireless Communications, volume 9, pages 76–83, October 2002.

[GB06a] M. Ghaderi and R. Boutaba. Call Admission Control in Mobile Cellular Networks: A Comprehensive Survey. In International Conference on Wireless Communications and Mobile Computing, IWCMC, volume 6, February 2006.

[GB06b] J. Gross and M. Bohge. Dynamic Mechanisms in OFDM Wireless Systems: A Survey on Mathematical and System Engineering Contributions. In TKN Technical Report, May 2006.

[GGM07] D. Ghosh, A. Gupta, and P. Mohapatra. Admission Control and Interference-Aware Scheduling in Multi-hop WiMAX Networks. In IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems, 2007.

[GP99] R. Guerin and V. G. Peris. Quality-of-service in Packet Networks: Basic Mechanisms and Directions. In Computer Networks: The International Journal of Computer and Telecommunications Networking, volume 31, pages 169–189, February 1999.

[HTLJ06] B. Han, F. P. Tso, L. Ling, and W. Jia. Performance Evaluation of Scheduling in IEEE 802.16 Based Wireless Mesh Networks. In IEEE International Conference on Mobile Adhoc and Sensor Systems, MASS, 2006.

[HYW09] C. He, O. Yang, and G. Wang. Performance Evaluation of a WiMAX Multi-Hop Relay System to Support Multicast/Broadcast Service. In IEEE 6th International Conference on Mobile Adhoc and Sensor Systems, MASS, 2009.

[IJT09] C. S. In, R. Jainand, and A. K. Tamimi. Scheduling in IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey. In IEEE Journal on Selected Areas in Communications, volume 27, February 2009.

[JBW08] P. Jiang, J. Bigham, and J. Wu. Scalable QoS Provisioning and Service Node Selection in Relay Based Cellular Networks. In 4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM, 2008.

[JPP00] A. Jalali, R. Padovani, and R. Pankaj. Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireress System. In IEEE Vehicular Technology Conference Proceedings, volume 3, pages 1854–1858, May 2000.

[KSC09] Y. Kim, K. Son, and S. Chong. QoS Scheduling for Heterogeneous Traffic in OFDMA-based Wireless Systems. In IEEE Global Telecommunications Conference, GLOBECOM, pages 1–6, 2009.

[LAC09] E. Laias, I. Awan, and P.M. Chan. Fair and Latency Aware Uplink Scheduler in IEEE 802.16 Using Customized Deficit Round Robin. In International Conference on Advanced Information Networking and Applications Workshops, WAINA, pages 425–432, May 2009.

[LC08] B. H. Lee and C. M. Chen. An Enhanced Election-Based Transmission Timing Mechanism in IEEE 802.16 Mesh Networks. In Asia-Pacific Conference on Communications, APCC, 2008.

[LE04] WiMAX Forum: WiMAXs Technology For LOS and NLOS Environments. 2004.

[Lib] Boost C++ Source Libraries. `http://www.boost.org/`.

[LKC04] H. Lee, T. Kwon, and D. H. Cho. An Efficient Uplink Scheduling Algorithm for VoIP Services in IEEE 802.16 BWA System. volume 5, September 2004.

[LKC+06] H. Lee, T. Kwon, D. H. Cho, G. Lim, and Y. Chang. Performance Analysis of Scheduling Algorithms for VoIP Services in IEEE 802.16e Systems. volume 3, September 2006.

[Mod] WiMAX Channel Model. `http://www.ieee802.org/16/tg3/contrib/802163c-01_29r4.pdf`.

[MSJ+06] R. Mukul, P. Singh, D. Jayaram, D. Das, N. Sreenivasulu, K. Vinay, and A. Ramamoorthly. An Adaptive Bandwidth Request Mechanism for QoS Enhancement in WiMax Real Time Communication. In IFIP International Conference on Wireless and Optical Communications Networks, 2006.

[PMl09] S. Pizzi, A. Molinaro, and A. lera. Channel-Aware Class-Based Scheduling for QoS Support in IEEE 802.16/WiMAX Networks. In IEEE International Conference on Information Networking, ICOIN, 2009.

[PV10] R. Prasad and F. J. Velez. WiMAX Networks : Techno-economic Vision and Challenges. Springer, 2010.

[RBS06] H. K. Rath, A. Bhorkar, and V. Sharma. An Opportunistic DRR (O-DRR) Uplink Scheduling Scheme for IEEE 802.16-based Broadband Wireless Networks. In IETE, International Conference on Next Generation Networks (ICNGN), 2006.

[RJ09] N. Ruangchaijatupon and Y. Ji. Proportional Fairness with Minimum Rate Guarantee Scheduling in a Multiuser OFDMA Wireless Network. In

International Conference on Wireless Communications and Mobile Computing, IWCMC, 2009.

[SAH08] A. Sayenko, O. Alanen, and T. Hamaainen. Scheduling Solution for the IEEE 802.16 Base Station. In International Journal of Computer and Telecommunications Networking, volume 52, 2008.

[SAKH06] A. Sayenko, O. Alanen, J. Karhula, and T. Hamaainen. Ensuring the QoS Requirements in 802.16 Scheduling. In Proc. Int. Workshop Modeling Analysis and Simulation Wireless and Mobile Systems, Spain, 2006.

[SK98] H. Sari and G. Karam. Orthogonal Frequency-Division Multiple Access and its Application to CATV Networks. In European Transactions on Telecommunications (ETT), volume 9, pages 507–516, Nov.-Dec. 1998.

[SV96] M. Shreedhar and G. Varghese. Efficient Fair Queuing Using Deficit Round Robin. In IEEE Transactions on Networking, volume 4, June 1996.

[tdd01] The Advantages and Benefits of TDD Broadband Wireless Access Systems. TDD Coalition White Paper, September 2001.

[Tec] Mobile WiMAX Technology. `http://www.intel.com/technology/wimax/mobile.htm`.

[TLZL05] J. Tao, F. Liu, Z. Zeng, and Z. Lin. Throughput enhancement in wimax mesh networks using concurrent transmission. In Wireless Communications, Networking and Mobile Computing, WiCOM, 2005.

[WD10] H. Wang and L. Dittmann. Downlink Resource Management for QoS Scheduling in IEEE 802.16 WiMAX Networks. In Journal of Computer Communications, volume 33, May 2010.

[WG03] K. Wongthavarawat and A. Ganzz. Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems. In International Journal of Communication Systems, volume 16, pages 81–96, February 2003.

[WGIH05] H.Y. Wei, S. Ganguly, R. Izmailov, and Z.J. Haas. Interference-Aware IEEE 802.16 WiMax Mesh Networks. In IEEE 61st Vehicular Technology Conference, pages 3102–3106, 2005.

[WiM09] Part 16: Air interface for broadband wireless access systems. IEEE Computer Society and IEEE Microwave Theory and Techniques Society, May 2009.

[WJ09] H. Wang and W. Jia. An Optimized Scheduling Scheme in OFDMA WiMax Networks. In International Journal of Communication Systems, volume 23, pages 23–39, July 2009.

[XPM05] S. A. Xergias, N. Passas, and L. Marekos. Flexible Resource Allocation in IEEE 802.16 Wireless Metropolitan Area Networks. In 14th IEEE Workshop on Local and Metropolitan Area Networks, LANMAN, 2005.

[YL06] O. Yang and J. Lu. A New Scheduling and CAC Scheme for Real-Time Video Application in Fixed Wireless Networks. In IEEE Consumer Communications and Networking Conference, IEEE CCNC, pages 303–307, 2006.

[YTFM07] X. Yajing, Li. Z. Tao, Z. X. Fang, and X. H. Min. A bandwidth scheduling algorithm based on minimum interference traffic in mesh mode. In Proceedings of Intl. Conference on Wireless Communications, Networking and Mobile Computing, WiCOM, 2007.

[ZFYZ08] J. Zhang, S. Feng, W. Ye, and H. Zhuang. MAC Performance Evaluation of IEEE 802.16j. In International Symposium on Information Science and Engieering, ISISE, 2008.

[Zha95] H. Zhang. Service Disciplines for Guaranteed Performance Service in Packetswitching Networks. In Proceedings of the IEEE, volume 83, pages 1374–96, October 1995.

[ZHWW07] L. Zhe, Z. He, W. Wu, and X. Wang. A Simplified Layered QoS Scheduling Scheme in OFDM Networks. In IEEE Vehicular Technology Conference, VTC, pages 1842–1846, 2007.