Dynamic Bandwidth Allocation in WiMAX Broadband Networks

By

Ovidiu Jianu

Supervisor: Dr. Hassan Naser

Submitted in Partial fulfillment of the requirement
for the Degree of Masters of Science in
Electrical and Computer Engineering
at
Lakehead University

Thunder Bay, Ontario, Canada 2011

Abstract

Wireless communication networks have experienced remarkable growth in recent years and are becoming attractive alternative to cable and DSL in providing last mile network access. Candidate wireless technologies should be cost-competitive, offer broadband wireless access, support portable and mobile operations with integrated voice, video and data services. Worldwide Interoperability for Microwave Access (WiMAX) is emerging as the leading wireless access technology that provides cost-competitive, ubiquitous broadband wireless access and Quality of Service (QoS) capabilities.

WiMAX refers to the interoperable implementations of the IEEE 802.16 family of wireless networks standards ratified by the WiMAX Forum. The IEEE 802.16 standard provides specifications for the Medium Access Control (MAC) and Physical (PHY) layers. WiMAX uses Orthogonal Frequency Division Multiple Access (OFDMA) at the Physical level to achieve multiple access. The presence of multiple users means that uplink resources in the form of *slots* have to be shared among the users. Slots are the minimum data allocation unit in OFDMA and are defined in two dimensions: one in time (OFDMA symbol number) and one in frequency (subcarrier). Since users share the uplink resources, this can result in collision of data signals originating from different subscriber stations.

We have developed two centralized bandwidth allocation algorithms for the uplink traffic to resolve contention for bandwidth and determine the transmission order of users, while maintaining fairness and QoS capabilities. We have developed an in-house simulation program in C++ to evaluate the performance of the proposed algorithms in terms of network throughput, packet delay and packet loss ratio.

The first algorithm, called Fairness Assured Scheduling Algorithm (FASA), employs a credit pooling technique to allocate slots among subscriber stations in a fair manner according to channel conditions. Based on channel conditions, WiMAX standard defines

seven Signal to Noise Ratio (SNR) levels resulting in seven different transmission rates at which subscriber stations can transmit data. We assign a zone to each of the seven different transmission rates. Allocating equal number of slots to subscriber stations irrespective of channel conditions, results in unequal data transmission rates among subscriber stations. This is due to the fact that different subscriber stations located in different zones experience dissimilar transmission bit rates per slot. To ensure approximately equal transmission rates among subscriber stations we introduce a normalization parameter to assign slots to subscriber stations based on channel conditions. In this way, FASA distributes unequal number of slots to subscriber stations, but guarantees approximately equal subscriber station transmission rates to satisfy service level agreements. Simulation results will show that FASA guarantees almost identical subscriber stations throughput. When the network is fully loaded, FASA overall throughput is approximately 50% of the WiMAX maximum theoretical throughout. This is attributed to the fact that users experience different channel conditions, which limits the achievable transmission rate. However, under best channel conditions, FASA achieves an overall throughput of about 93% of the maximum WiMAX theoretical throughput. The slight difference can be attributed to the slot packing problem. The slot packing problem is concerned with packing variable size data packets into fixed size slots. If the data packets are not exact multiple of the slot size then some slots may not be full, resulting in waste of bandwidth.

The second algorithm, called Fairness Assured with Quality of Service (FASAQ), employs a credit pooling technique to allocate slots among subscriber stations' classes of service in a fair manner. Various classes of service are transmitted based on their QoS requirements, transmission history and their channel conditions. A weighted round robin arbitration mechanism is employed in FASAQ to determine the priority of users by their long term transmission records. In FASAQ, we also employ the concept of zoning to assign slots to subscriber stations based on their channel conditions. Similar to FASA, a zone is assigned to each of the seven different transmission rates corresponding to the seven WiMAX defined Signal to Noise Ratio levels.

To ensure approximately equal transmission rates among subscriber stations, we used the normalization parameter introduced in FASA to assign slots to subscriber stations based on channel conditions. In this way, FASAQ distributes unequal number of slots to subscriber stations, but guarantees approximately equal subscriber station transmission rates to satisfy service level agreements. When the normalization parameter is not employed, subscriber stations receive equal number of slots, which does not translate into equal share of bandwidth among subscriber stations as promised. This is due to the fact that based on channel conditions, slots have different transmission bit rates, and so an equal distribution of slots results in an unequal transmission rate among subscriber stations. We used the name Quality of Service Assured Algorithm (QASA) to describe the case in which the normalization parameter is not employed.

Simulation results will show that FASAQ is able to satisfy QoS requirements and guarantees almost identical throughput among subscriber stations regardless of channel conditions experienced by the subscriber stations. Simulation results will also show that when zoning is not employed, subscriber stations will not receive their fair share of bandwidth as promised. Subscriber stations with poor channel conditions will experience lower throughput than subscriber stations experiencing good channel conditions. Simulation results will evaluate and compare the performance of FASAQ and QASA in terms of network throughput, packet delay and packet loss ratio.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Hassan Naser for his support and guidance over the past two years. Without his continuous feedback this thesis would not have been possible.

I would also like to thank Dr. Atoofian and Dr. Shami for being the readers of this thesis and for their insightful comments and suggestions.

Last but not least, I want to thank my loving fiancé and my wonderful parents and sister for being patient with me during my long absence, and for their constant words of encouragement.

Table of Contents

List of Figures

List of Tables

List of Symbols

k	Pareto shape parameter
b	Pareto location parameter
ρ	Network load
f_k	OFDM subcarrier frequency
R_s	OFDM modulation symbol rate
S_α	Average queue size
S_{max}	*M*aximum queue threshold
S_{min}	Minimum queue threshold
P	Packet drop probability
Pd	Packet maximum drop probability
N	Number of SSs in the system
M	Number of CoSs
$RTT^{(n)}$	Round trip time to SS n
N_{total}	Total number of slots
C_p	Capacity of the WiMAX system
F	Buffer size in bytes
L_{req}	Request Message Length
L_{grt}	Grant Message Length
r_{CBR}	Rate of the CBR traffic
T_{max}	*M*aximum cycle time
$Tsch$	Algorithms scheduling time
T_u	Useful symbol period
T_g	Guard Time
T_s	Symbol duration
F_s	Sampling frequency
Δf	Subcarrier spacing

β	Mean power
U_j	Size of CoS j credit pool
V_i	Size of SS i credit pool

List of Abbreviations

A/D	Analog to Digital
AMC	Adaptive Modulation and Coding
ATM	Asynchronous Transfer Mode
BE	Best Effort
BER	Bit Error Rate
BPSK	Binary Phase Shift Keying
BS	Base Station
BW	Nominal Bandwidth
BWA	Broadband Wireless Access
CBR	Constant Bit Rate
CINR	Carrier to Interference and Noise Ratio
CP	Cyclic Prefix
CPS	Common Part Sublayer
CS	Convergence Sublayer
D/A	Digital to Analog
DL	Downlink
DFT	Discrete Fourier Transform
DHCP	Dynamic Host Configuration Protocol
DRR	Deficit Round Robin
DSL	Digital Subscriber Line
EDF	Earliest Deadline First
ertPS	extended real-time Polling Service
FDD	Frequency Division Duplex
FDM	Frequency Division Multiplexing
FFT	Fast Fourier Transform
GPC	Grant per Connection
GPSS	Grant per SS

HiperMAN	High performance radio Metropolitan Area Network
IFFT	Inverse Fast Fourier Transform
IEEE	Institute of Electrical and Electronic Engineers
ISI	Intersymbol Interference
LAN	Local Area Network
LOS	Line-of-Sight
MAC	Medium Access Control
MAN	Metropolitan Area Network
NLOS	Non-Line-of-Sight
nrtPS	non real-time Polling Service
OFDM	Orthogonal Frequency Division Multiplex
OFDMA	Orthogonal Frequency-Division Multiple Access
PAN	Personal Area Network
PDF	Probability Density function
PFS	Proportional Fair Scheduler
PHY	Physical Layer
PMP	Point to Multipoint
P/S	Parallel to Serial
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RLC	Radio Link Control
RR	Round Robin
RSSI	Received Signal Strength Indicator
rtPS	real-time Polling Service
RV	Random Variable
SAP	Service Access Point
SC	Single Carrier
SDMA	Spatial Division Multiple Access

SDU	Service Data Unit
SINR	Signal to Interference Noise Ratio
SLA	Service Level Agreement
S/P	Serial to Parallel
SS	Subscriber Station
TFTP	Trivial File Transfer Protocol
TDD	Time Division Duplex
UCD	Uplink Channel Descriptor
UGS	Unsolicited Grant Service
UL	Uplink
VoIP	Voice over Internet Protocol
WAN	Wide Area Network
WFQ	Weighted Fair Queuing
WiMAX	Worldwide Interoperability Microwave Access
WirelessHUMAN	Wireless High Speed Unlicensed Metro Area Network
WLAN	Wireless Local Area Network
WRED	Weighted Random Early Detect
WRR	Weighted Round Robin
WSSO	Weighted Subscriber Station Scheduling Order

Chapter 1 Introduction

1.1 Motivation of Research

The motivation behind this master thesis stems from the need to address the technical challenges faced by network operators as they try to cope with ever increasing user demands for high data rates, mobility and connectivity. The 21st century technologically savvy users expect to access information anytime, anywhere and on any device. One solution to satisfy high data rates and mobility is the deployment of wireless technologies. New wireless technologies have enabled the development of a large number of applications which changed the way we work, live and play. As the number and versatility of wireless applications is increasing, so is the number of people using wireless devices on a daily basis.

According to the International Communication Union [1], the number of worldwide mobile users has increased significantly in the past decade, reaching 5.3 billion mobile subscriptions at the end of 2010, with much of the growth coming from the developing countries. With the proliferation of laptops and introduction of new mobile devices, such as smart phones and tablets, the number of people accessing the mobile internet is also growing rapidly. It is expected that within five years, the mobile Web access will overtake the PC as the most popular way to get on the Web. At the current growth rate, in just over two years there will be over 1 billion mobile internet users worldwide [1].

These remarkable growth statistics, combined with future bandwidth-intensive applications, will definitely overwhelm the wireless communication networks and will send wireless network operators scrambling to create new technical solutions. There have been continued efforts to deliver ubiquitous broadband wireless access by developing and deploying new broadband wireless access technologies.

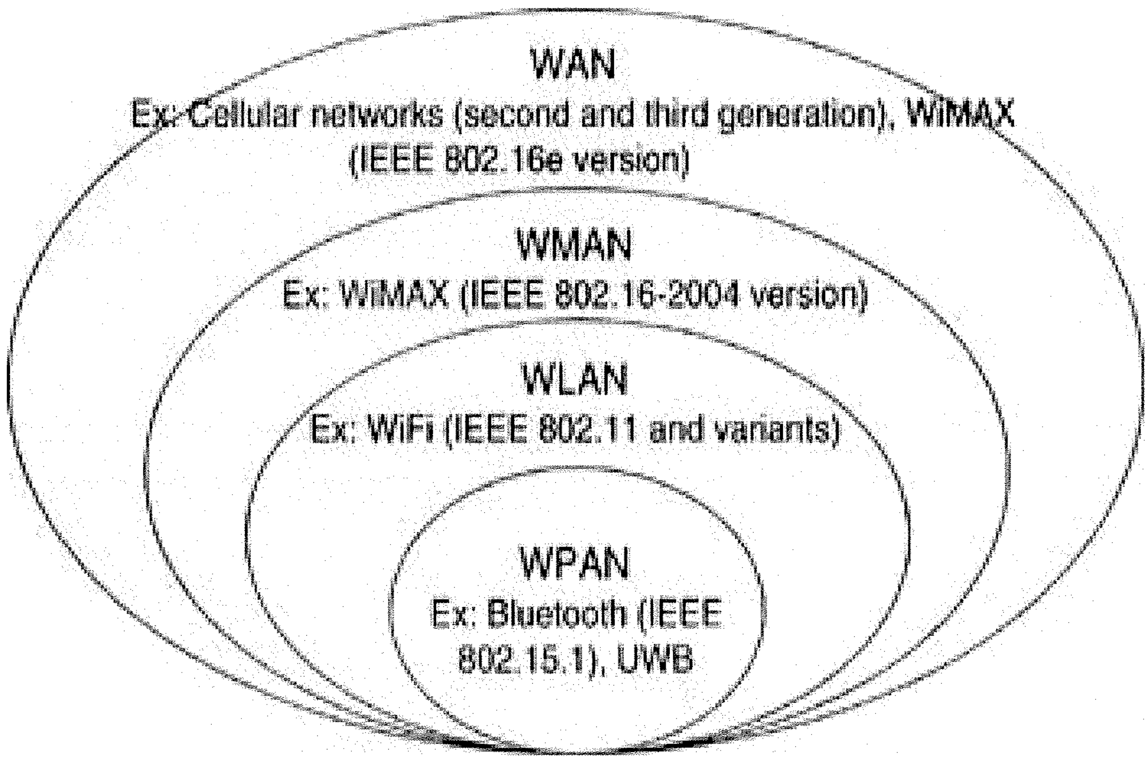

Figure 1.1: Illustration of Network Types (Source: [11])

Figure 1.1 illustrates wireless network categories and the most well-known technologies for each type of network. Wireless access technologies can be broadly characterized by coverage area, mobility, complexity and data rates supported as follows:

- Personal Area Network (PAN) is used for communication among data devices close to one person. An example of PAN technology is Bluetooth [2], whose coverage area is limited to a few meters with no mobility.

- Local Area Network (LAN) is a wireless or wired communication network which covers a small area such as a house or office, with limited mobility. Wi-Fi is a wireless local area network (WLAN) based on the Institute of Electrical and Electronics Engineers (IEEE) 802.11 standards [3].

- Metropolitan Area Network (MAN) is a data network that connects a number of LANs or a group of stationary/mobile users distributed over a large area, such as an airport terminal or a neighbourhood. An example of a MAN is Worldwide Interoperability for Microwave Access (WiMAX) based on the IEEE 802.16 standards [4].

- Wide Area Network (WAN) is a data network covering a relatively large geographical area by interconnecting a set of nodes, hosts, or LANs. Examples of WANs include the Internet [5], cellular networks, and mobile WiMAX.

In this thesis we are interested in wireless access technologies developed to address the last-mile bottleneck by providing an alternative to cable and DSL. The last-mile is the final leg of delivering connectivity from the service provider to the customer. Candidate technologies should be cost-competitive, offer broadband wireless access (BWA), support fixed, portable and mobile operations with integrated voice, video and data services. Worldwide Interoperability for Microwave Access (WiMAX) is emerging as the leading wireless access technology that provides cost-competitive, ubiquitous broadband wireless access and Quality of Service (QoS) capabilities. WiMAX is a family of telecommunications protocols ratified by the WiMAX Forum®, which is an industry-led, not-for-profit organization that certifies and promotes the compatibility and interoperability of broadband wireless products based upon IEEE 802.16 family of broadband wireless networks standard [6]. The IEEE 802.16 Air Interface for Broadband Wireless Access Systems provides specifications for the Physical (PHY) and Medium Access Control (MAC) layers.

WiMAX is providing fixed, nomadic, portable and eventually, mobile wireless broadband connectivity without the need for direct line-of-sight (LOS) communication with a base station. WiMAX forum has published three licensed spectrum bands (2.3GHz, 2.5 GHz and 3.5 GHz) that can offer theoretical rates of up to 75 Mbps with coverage from a single tower of up to 50 km. However, practical considerations such as terrain, weather, and buildings limit the range to about 10 km. WiMAX supports a large number of channel bandwidths ranging from 5MHz to 20MHz and radio frame size of 5ms [7].

IEEE 802.16 standard can support a variety of multiple access schemes such as WirelessMAN-SC, Orthogonal Frequency Division Multiplexing (OFDM) and Orthogonal Frequency Division Multiple Access (OFDMA), which allow the users to share the radio resources. However, OFDMA is the most widely used WiMAX access scheme and will be the only scheme discussed in detail in this thesis. OFDMA is a technique for transmitting large amounts of digital data over a radio wave. The technology works by splitting the radio signal into multiple smaller sub-signals or subcarriers, which are then transmitted simultaneously at different frequencies to the receiver. User generated digital data modulates the subcarriers using any of the four digital modulating techniques: BPSK, QPSK, 16-QAM and 64-QAM. In an OFDMA system, resources are available in the time domain by means of OFDMA symbols and in the frequency domain by means of subcarriers. This two dimensional partition of resources creates a resource matrix with time as the horizontal axis and frequency as the vertical axis.

WiMAX defines a *slot* as the smallest unit of Physical layer resource and hence the smallest element of the resource matrix. A slot is also represented in two dimensions: one in time (time symbols) and one in frequency (subcarriers). WiMAX users experience variable channel conditions. As such, besides carrying data, slots also carry reference signals called *pilots* that convey information about the channel conditions. The pilots carry information such as carrier to interference and noise ratio (CINR) or Received Signal Strength Indicator (RSSI), which is used by the base station and users to determine modulation and coding rates employed during transmission.

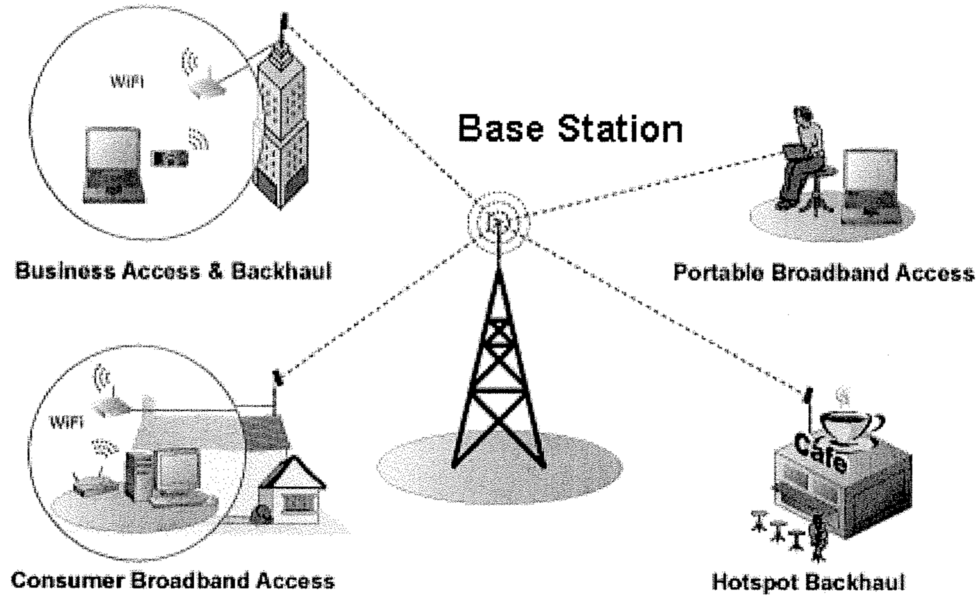

Figure 1.2: Typical WiMAX Architecture (Source: [39])

WiMAX is a broadband wireless access network that delivers data between a base station (BS) and Subscriber Stations (SS) located at customer premises as shown in Figure 1.2. Communication occurs in two directions: from BS to SS is called downlink and from SS to BS is called uplink. With respect to duplexing, which refers to bi-directional communication between two devices, WiMAX supports both Time Division Duplex (TDD) and Frequency Division Duplex (FDD). In TDD systems, a common carrier is shared between the uplink and downlink, the resource being switched in time. In FDD systems, one frequency band is allocated for the uplink and another for the downlink.

WiMAX is a point-to-multipoint (PMP) architecture in the downlink, meaning that the BS broadcasts identical data to all subscribers. Each subscriber is responsible for determining which data is intended for it, while ignoring all other data. In the uplink, WiMAX employs a multipoint-to-point architecture, meaning that the subscriber stations share the wireless capacity and resources, which can result in collisions of data signals from different subscriber stations. There is no direct communication between subscriber stations, and as such they cannot detect possible uplink collisions. As a result, a Medium Access Control (MAC) mechanism is required to mediate access between contending subscriber stations. This process involves allocating bandwidth among users and determining their transmission order and is usually performed by a scheduler located at the base station.

WiMAX supports a multitude of applications such as video and audio streaming, online gaming, Voice over IP (VoIP) and video conferencing, which demand different Quality of Service (QoS) requirements. Quality of Service means different things to different end users, but in general is defined as the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow. Some commonly used QoS parameters are bandwidth, latency, jitter and reliability. It is crucial that the QoS requirements of the users are satisfied when designing any uplink scheduling algorithms. WiMAX specifies a suite of QoS options and a limited number of techniques for users to request bandwidth resources in the uplink. However, WiMAX forum left the details of the implementation of the uplink scheduling algorithms open to vendors and network providers. It is our belief that proprietary solutions are expected to start growing soon, and scheduling algorithms are going to make the difference among the competitors.

1.2 Objectives and contribution

Designing a centralized WiMAX scheduling algorithm for uplink traffic is a challenging task. The goal of this thesis is to design uplink traffic algorithms that partition the uplink bandwidth among users in a fair, efficient and cost effective way, regardless of the user's position and channel conditions. Any scheduling algorithm must possess a number of desirable qualities and should be evaluated based on a number of criteria as follows:

- **Fairness**: The algorithm should ensure that a reasonable level of fairness is maintained among users. Users should be prevented from monopolizing the available resources and the scheduler should maintain a balance with respect to resources allocated to users. In a wireless network such as the one presented in this thesis, fairness is even more important since users experience different channel conditions depending on their spatial location.

- **Complexity**: The algorithm should be simple both conceptually and at the component level. A simple algorithm will require fewer components and will allow for cost efficient implementation on a large scale. Conceptually, the algorithm should be easy to understand, debug and reconfigure by network operators whenever changes need to be made in response to any network state variations.

- **Flexibility**: The algorithm should be able to accommodate users with different QoS requirements and meet the minimum requirements of those users. The algorithm should be able to respond to any unpredictable traffic fluctuations in the network and to changes in user channel quality.

- **Scalability**: The algorithm should continue to work efficiently even as the number of scheduled users varies.

In this thesis, we design two dynamic bandwidth allocation algorithms for the uplink traffic in WiMAX to distribute the slots among users, while satisfying the above mentioned criteria. In the first algorithm, called *Fairness Assured Scheduling Algorithm (FASA)*, the slots are distributed to the users employing priority based scheduling and using a credit pooling technique with request and grant method of arbitration. Based on the channel conditions, the algorithm will grant slots to users such that each user will transmit information at comparable transmission rates. This prevents users with good channel quality from monopolizing the network resources at the expense of users who experience poor channel conditions. The second algorithm, called *Fairness Assured Scheduling Algorithm with Quality of Service (FASAQ)*, also addresses the issue of channel variability by normalizing the user's transmission rates to predefined values. In addition, the algorithm will introduce Quality of Service capabilities to better reflect the classes of traffic supported by WiMAX. A weighted round robin arbitration mechanism is employed in FASAQ to determine the priority in which users can send information based on their long-term transmission records. We assess the performance of the algorithms using an in-house built simulation, implemented in C++, which analyzes network throughput, average delay and average packet loss ratio.

1.3 Thesis Outline

The rest of this thesis is organized as follows. In Chapter 2, we provide a description of WiMAX, with focus on the Physical Layer (PHY) and Medium Access Control (MAC) layer protocols. We give background information on the digital modulation and multiplexing schemes supported by WiMAX. Emphasis will be placed on Orthogonal Frequency Division Multiple Access (OFDMA), which is the scheme used in this thesis. Following the theoretical background, Chapter 3 provides an overview of the scheduling algorithms that exist in the literature. We look at scheduling algorithms designed specifically for WiMAX and algorithms employed in wired systems which can be employed in the wireless world. Chapter 4 details the system model, transmission cycle, architecture of the scheduling algorithm and the uplink channel model. In Chapter 5 we provide a detail description of our algorithms. Chapter 6 will detail the simulation setup and analyze the simulation results. We conclude with Chapter 7, which summarizes this thesis and provides any recommendation for future work.

Chapter 2 WiMAX Theoretical Background

In this chapter we examine the characteristics of the Physical (PHY) and Medium Access Control (MAC) Layers as specified in the IEEE 802.16-2009 standard, which are also applicable to WiMAX protocol as defined by WiMAX Forum. More specifically, we will provide background on PHY layer concepts such as channel coding and modulation techniques, uplink access schemes, subchannelization and uplink physical resources. With regards to the MAC layer we will introduce the reader to the three sublayers of the MAC layer, the concept of Quality of Service (QoS) and bandwidth allocation.

2.1 WiMAX Physical Layer

WiMAX Physical Layer (PHY) covers a wide array of concepts and technologies. Some of those concepts such as downlink access schemes and multi-antenna transmission schemes are not applicable to the work presented in this thesis and as such were not discussed. The aim of this section is to introduce the basic concepts used in this thesis that are needed for a minimum understanding of WiMAX.

The WiMAX Physical Layer has several roles: establishes the physical connection between receiver and transmitter; it is responsible for the transfer of data; it defines the types of signal used, the type of modulation and demodulation employed; and the transmission power. WiMAX is a broadband wireless access technology, meaning that data is transmitted at high speed on the air interface through radio electromagnetic waves using a given frequency. The IEEE 802.16 Physical layer covers a frequency spectrum ranging from 2 GHz to 66 GHz. This large spectrum range is divided into Line of Sight (LOS) environments for frequency bands greater than 11 GHz and non-Line of Sight (NLOS) environments for 2-11 GHz frequency range [4]. Table 2.1 outlines the various air interface variants supported by the IEEE 802.16 standard.

Designation	Applicability	Duplexing
WirelessMAN-SC	10–66 GHz	TDD, FDD
Fixed WirelessMAN-OFDM	Below 11 GHz licensed bands	TDD, FDD
Fixed WirelessMAN-OFDMA	Below 11 GHz licensed bands	TDD, FDD
WirelessHUMAN™	Below 11 GHz license-exempt bands	TDD

Table 2.1: Air Interface Nomenclature (Source: [4])

- **Wireless Metropolitan Area Network-Single Carrier (WirelessMAN-SC)** is targeted for operation in the 10–66 GHz frequency band. It is designed with a high degree of flexibility in order to allow service providers the ability to optimize system deployments with respect to cell planning, cost, radio capabilities, services, and capacity [4].

- **Wireless Metropolitan Area Network-Orthogonal Frequency Division Multiplexing (WirelessMAN-OFDM)** is based on OFDM technology designed mainly for fixed SSs, where the SSs are deployed in residential areas and businesses. It is used for licensed bands below 11 GHz and can support TDD and FDD frame structures. The OFDM PHY can only support sub-channelization in the uplink [4].

- **Wireless Metropolitan Area Network-Orthogonal Frequency Division Multiple Access (WirelessMAN-OFDMA)** is based on OFDMA technology and offers sub-channelization in both uplink and downlink. It is used for licensed bands below 11 GHz and can support TDD and FDD frame structures [4].

- **Wireless High Speed Unlicensed Metro Area Network (WirelessHUMAN)** specification of the PHY layer is similar to the OFDM based layer except it is focused on Unlicensed National Information Infrastructure (UNII) devices and other unlicensed bands. This specification can only support TDD frame structures [4].

In this thesis we will focus on the WirelessMAN-OFDMA FDD profile with licensed bands below 11GHz since OFDMA technology can allow multiple users access on the same channel.

Since the radio frequency spectrum is a scarce resource, wireless systems must be designed such that they allow simultaneous communication of as many users as possible. In order to allow for flexible and efficient spectrum usage, WiMAX supports duplexing, which separates the spectrum into uplink and downlink transmission. WiMAX supports two duplexing techniques: time division duplexing (TDD) and frequency division duplexing (FDD). In TDD, uplink data is sent at times that are different from downlink transmission times, but using the same frequency channel. In FDD, uplink and downlink data is sent on two different frequency bands, usually separated by a frequency gap [8]. In this thesis we consider FDD duplexing.

2.1.1 Modulation formats

WiMAX is a digital technology in which data bits are mapped to signal waveforms that can be transmitted over an analog channel. Telecommunications systems use a number of digital modulation techniques in which an analogue signal is modulated with a digital sequence in order to transport this sequence over a given medium such as radio waves. Modulation techniques are employed in order to change the physical characteristics of the sinusoidal carrier, such as the frequency, phase or amplitude, or a combination of some of these. Four modulations techniques are supported by the WiMAX protocol: BPSK, QPSK, 16-QAM and 64-QAM [4]. We need a basic understanding of modulation techniques to derive the transmission rates of the users in the uplink direction. We will introduce each modulation technique in the following sections.

2.1.1.1 Binary Phase Shift Keying (BPSK)

Binary Phase Shift Keying (BPSK) is a digital modulation scheme that conveys data by changing, or modulating, the phase of a reference signal, also known as the carrier wave. In other words, sinusoidal carrier waveforms are used to transmit coded bits (0s and 1s) over the air. Before the sequence of bits is transmitted, they are passed through a symbol mapper, which maps each bit of the sequence to a baseband symbol. The baseband symbol is +1 if the corresponding bit is 1 and −1 if the corresponding bit is 0 [9]. Two possible symbols are produced that are real numbers and are positioned on the real axis only as shown in Figure 2.1, which represents the constellation diagram for BPSK. There are no imaginary components in these symbols. Also note that these two symbols differ only in their phase by 180^0 and not in their magnitude.

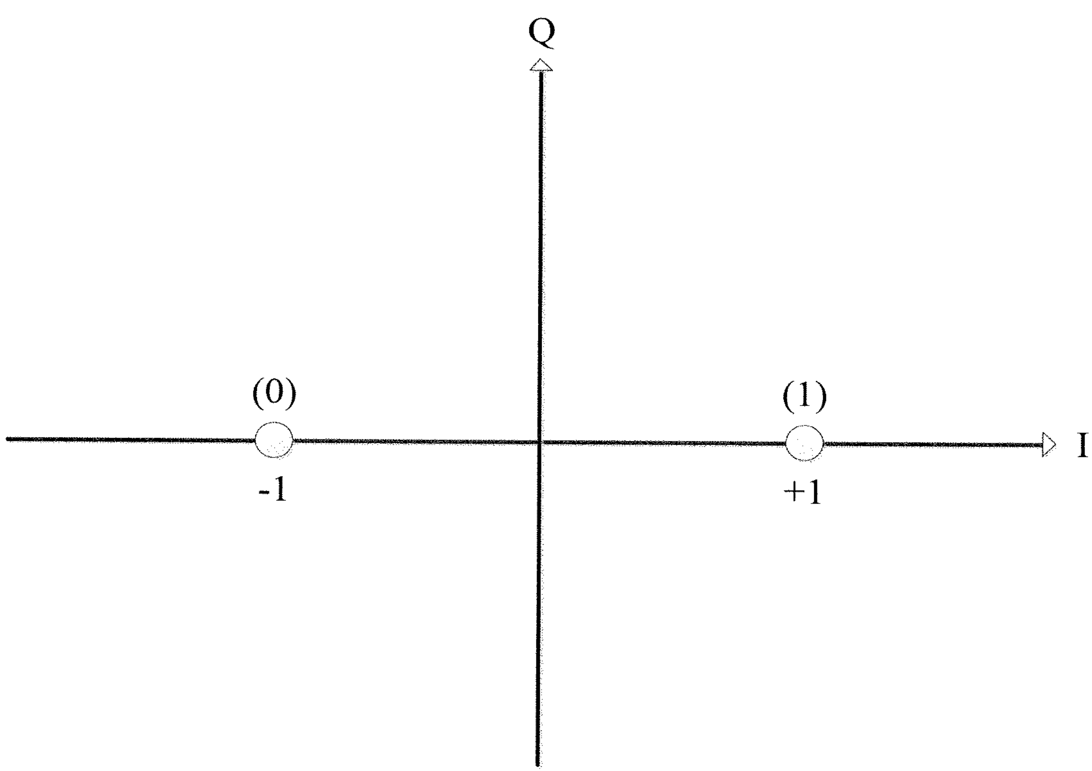

Figure 2.1: BPSK Constellation

Since there are only two symbols in the symbol constellation, each symbol can only carry one bit of data. The baseband symbol is then multiplied by a waveform of the form cos(2πft) that is transmitted over the air. Figure 2.2 shows what happens when a bit stream of 1101 is used to modulate the carrier. If the data bit is a 1, then a positive cosine waveform is transmitted; if the data bit is a 0, then a negative cosine waveform is transmitted.

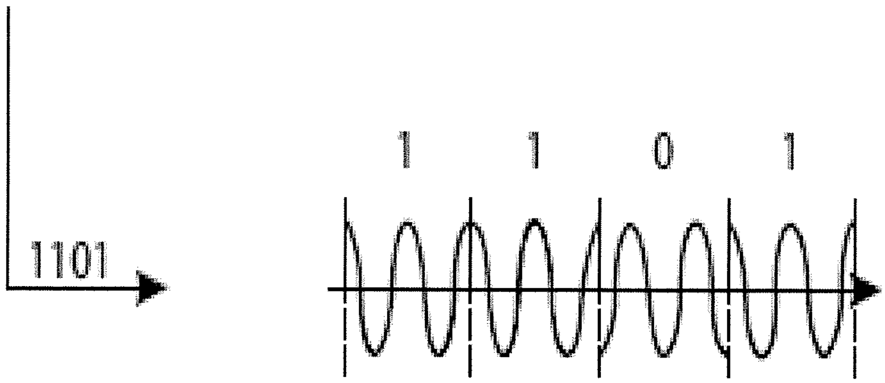

Figure 2.2: Time Domain Waveform for BPSK

2.1.1.2 Quadrature Phase Shift Keying (QPSK)

Similar to BPSK, Quadrature Phase Shift Keying (QPSK) is a digital modulation technique in which a sinusoidal carrier is varied in phase while keeping the amplitude and frequency constant. If one adds two additional symbols along the imaginary axis of Figure 2.1, then there would be a total of four symbols in the constellation. Figure 2.3 shows the four symbols, which are {+1, -1, +j, −j}. These four symbols are real numbers and imaginary numbers. The four symbols again differ only in their phase, not in their magnitude. All symbols have a magnitude of one, but adjacent symbols now have a phase difference of 90°. Because there are now four symbols in the constellation, each symbol can carry two data bits at a time. Figure 2.3 also shows a possible way to assign groups of two bits to the symbols.

Figure 2.4 shows what happens when a bit stream of 1101 is used to modulate the carrier. The baseband bit stream of 1101 goes into a symbol mapper, which now maps a group of two bits to a symbol as opposed to one bit to a symbol in BPSK. The corresponding symbol is -1 if the bits are 11, $+j$ if the corresponding bits are 01 and so on as shown in Figure 2.3. The symbol -1 is more completely and correctly written as $-1 + 0j$ in complex form, meaning that the mapper outputs -1 on the real line and 0 on the imaginary line during the corresponding symbol period.

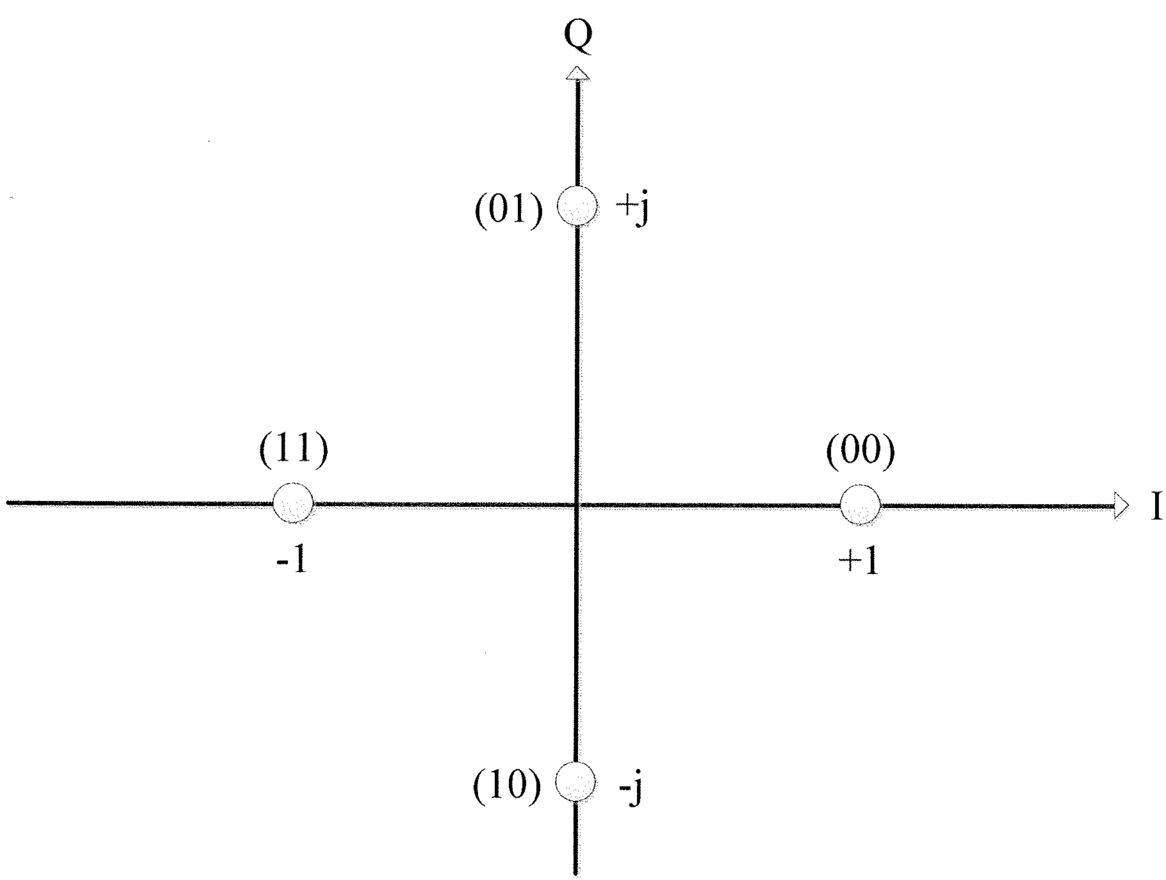

Figure 2.3: QPSK Constellation

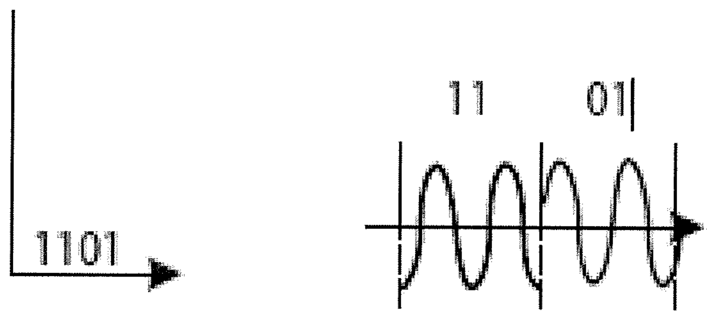

Figure 2.4: Time Domain Waveform for QPSK

Similarly, for the symbol $+j$ we can write it as $0 + j$ in complex form meaning the mapper outputs 0 on the real line and +1 on the imaginary line. This modulation scheme can transmit twice the amount of data in a given bandwidth compared with BPSK, but is less immune to interference and noise. Since more bits per symbol means higher data rates, using increasingly higher order constellations enables more bits to be transmitted using a given symbol.

2.1.1.3 Quadrature Amplitude Modulation: 16-QAM and 64-QAM

In the previous section we saw that QPSK can transmit two bits per symbol. We can transmit even more bits per symbol by increasing the number of symbols in the symbol constellation, thus increasing the spectral efficiency. In this section we will expand on the BPSK and QPSK schemes to show how we can transmit more information in each symbol interval. One way to achieve this is to employ Quadrature Amplitude Modulation (QAM). QAM is a digital modulation technique that changes the amplitude of two sinusoidal carriers, which are out of phase by $\pi/2$.

Figure 2.5 shows the rectangular constellation of 16-QAM in which each of the 16 symbols can carry four data bits at a time. A larger constellation of 64-QAM can be constructed the same way, with each of the 64 symbols carrying six data bits at a time. In general, the number of bits b carried by a symbol is

$$b = log_2 M \tag{2.1}$$

where M is the total number of symbols in the constellation. For 64-QAM there are 64 symbols ($M=64$) in the constellation, so each symbol can carry six bits ($b=6$) at a time. Also note that for all these modulation schemes (QPSK, 16-QAM, and 64-QAM), the bits assigned to a symbol differ from those of an adjacent symbol by at most one bit. This is done to facilitate error correction in digital transmission, which is based on the Grey coding technique [9].

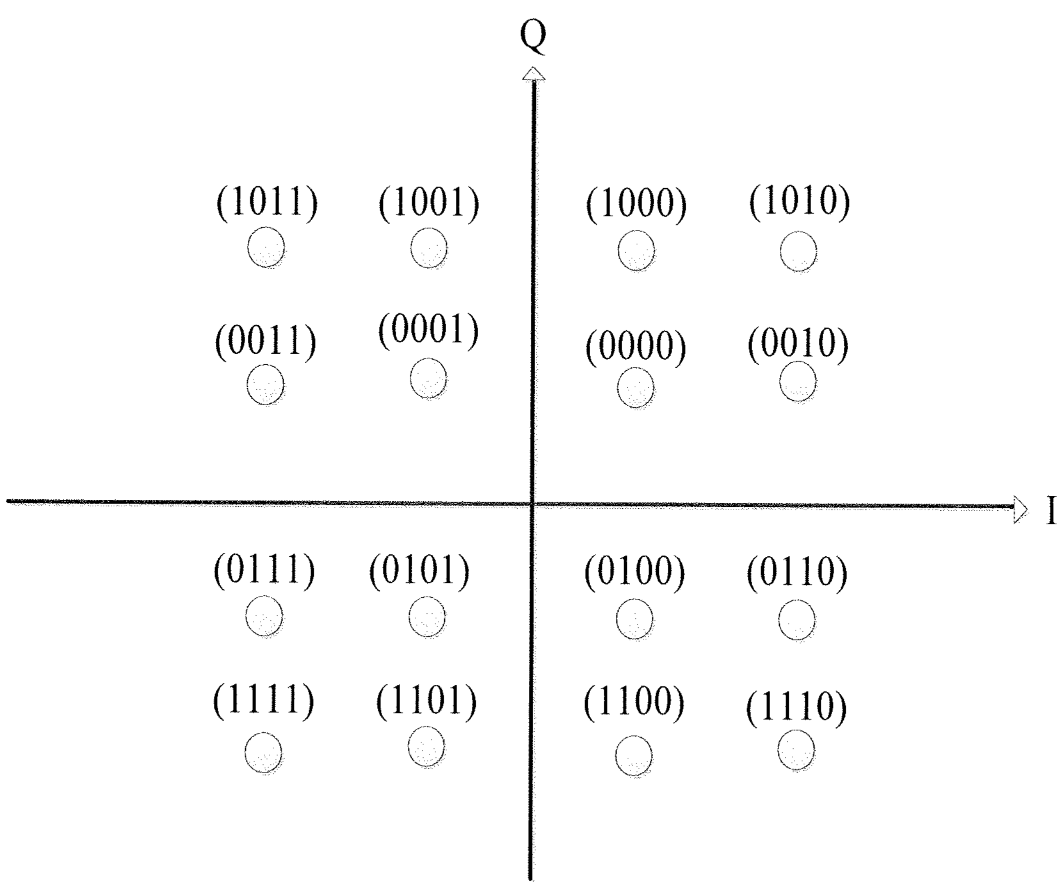

Figure 2.5: 16-QAM Constellation

The increased bandwidth efficiency achieved by employing higher order modulation is done at the cost of higher transmitter power. Also, increasingly higher order constellations are more spectrum efficient but less robust. This is because a higher order modulation means symbols are closer together, so there is a higher chance for the demodulator to misinterpret the symbol. Nevertheless, depending on channel conditions, all of the above modulation techniques are supported by WiMAX, in the uplink and downlink direction.

In this section we briefly showed how data bits are modulated onto carriers using multiple modulating techniques. In the next section, we will show how these carriers are transmitted over the air to maximize throughput and minimize the effect of dispersion and multipath propagation.

2.1.2 Multiplexing Techniques

In order to cope with an increasing demand for broadband wireless access, wireless network providers are deploying systems that can support high-speed data rates. These high data rates are possible by using higher signal bandwidths, which translates into shorter symbol time. Signals suffer from multipath propagation and echoes from surrounding objects, which can cause some of the signals to arrive at the destination in a time-delayed fashion. When the symbol time becomes small as compared to the channel delay, the delayed version of the symbol may interfere with some of the subsequent symbols [10].

To address the inter-symbol interference (ISI), the symbol time is increased by reducing the symbol rate, which is achieved by dividing the high-rate data stream into multiple low-rate substreams. This translates into dividing a wideband carrier into smaller narrowband subcarriers with lower bandwidth than the original carrier. The process is known as Orthogonal Frequency Division Multiplexing (OFDM). OFDM is a method of multiplexing where a high-rate data stream is divided into multiple low-rate substreams, which are then simultaneously transmitted over multiple subcarriers. Data carried by the subcarriers is sent in such a way that the subcarriers do not interfere with one another in frequency [9].

If more than one user needs to access the air interface at the same time, different groups of subcarriers are assigned to different users using the multi user version of OFDM, called Orthogonal Frequency Division Multiple Access (OFDMA). OFDMA multiplexing scheme is the primary technology used in WiMAX. To get a better understanding of OFDMA, we first introduce the conventional frequency division multiplexing (FDM), which forms the basis of both OFDM and OFDMA.

2.1.2.1 Conventional FDM

When the useful bandwidth of the transmission medium exceeds the required bandwidth of signals to be transmitted we can employ a multiplexing technique known as Frequency Division Multiplexing (FDM). Figure 2.6 shows the transmitter of a digital FDM system. A high rate stream of data symbols, produced by the symbol mapper, enters the serial to parallel converter at a rate of R_s modulation symbols per second (sps). The converter splits this high rate stream into K parallel streams of lower rate R_s/K. These streams pass through a digital to analog (D/A) converter so they can be transmitted over the air. Each stream is then modulated by its own complex sinusoid $exp(-j2\pi f_k t,)$, where f_k is the subcarrier frequency assigned to each low-rate substream. All the K subcarriers are summed and the composite signal is transmitted over the air as shown in Figure 2.7.

Besides combating intersymbol interference, another advantage obtained from this method is individual per subcarrier modulation and coding, which is necessary when not all subcarriers experience the same channel conditions. However, FDM has a couple of drawbacks as well. In Figure 2.7, we note that the transmitted signal has guard bands placed between the subcarriers to minimize interference between adjacent subcarriers.

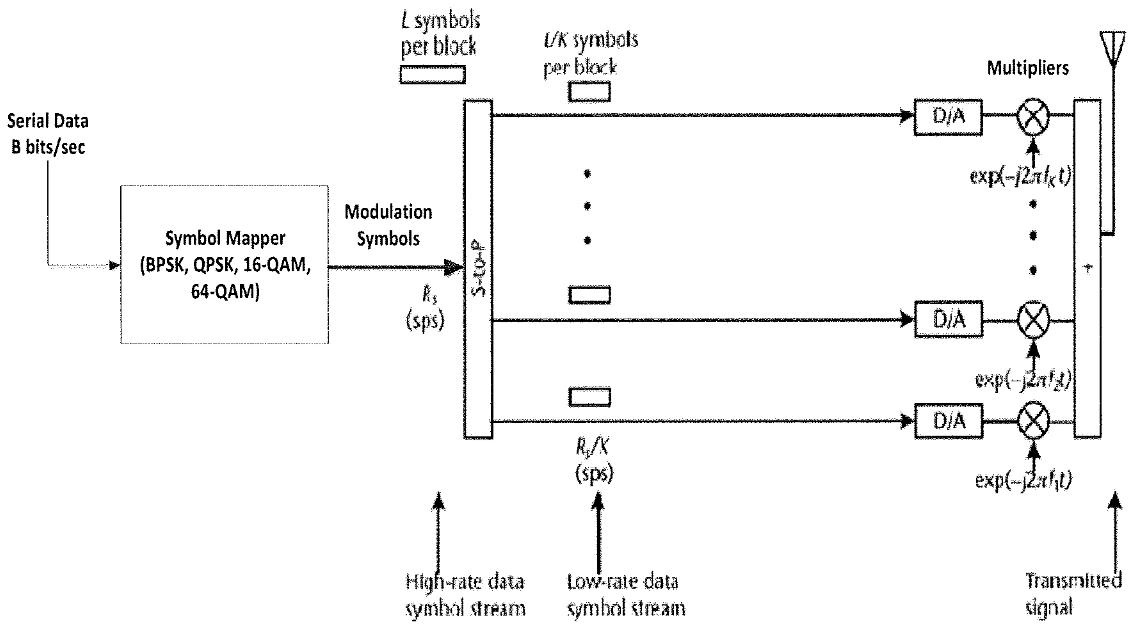

Figure 2.6: Conventional FDM Transmitter (Source: [9])

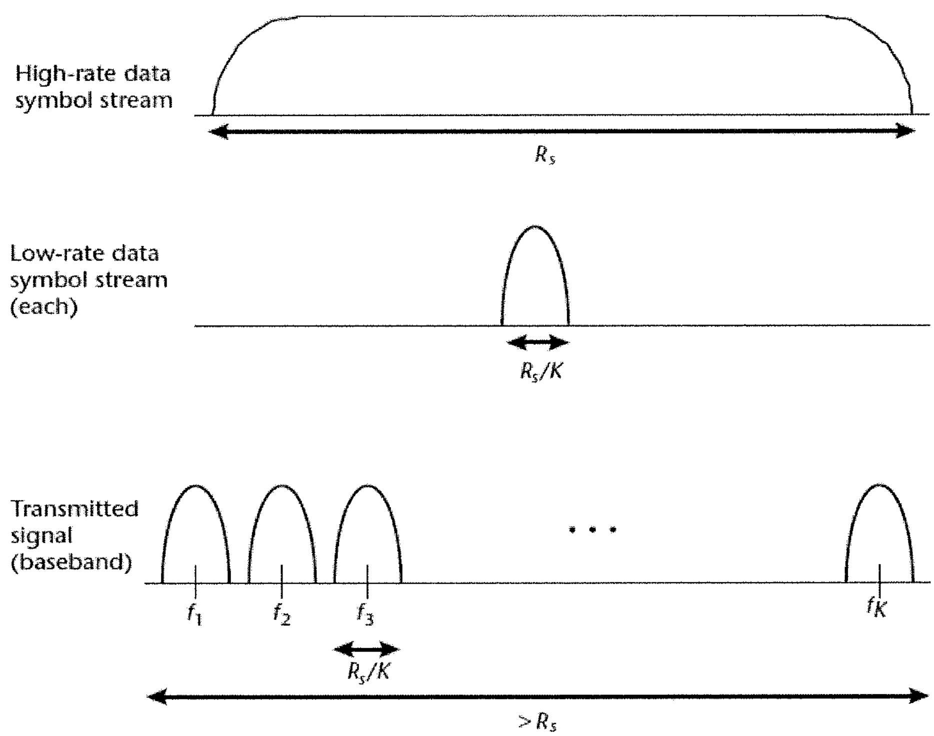

Figure 2.7: The Spectrum of FDM Signal (Source: [9])

Since the guard bands carry no data, this adds to the bandwidth requirements without increasing the data throughput. Second, each substream needs its separate digital to analog converter and modulator. This adds to the complexity and cost of the system. One way to overcome the FDM shortcomings is to employ OFDM.

2.1.2.2 OFDM Basics

Similar to FDM, the goal of OFDM is to use a large number of subcarriers spread in the allocated bandwidth, with each subcarrier being modulated by a proportionally lower data rate that would be the case in a single carrier scenario. As mentioned in the previous section, the use of a large number of subcarriers is necessary to limit the effects of multipath propagation that affects the transmission rate in a non-line of sight environment. One drawback of FDM was the presence of guard bands which is not bandwidth efficient. OFDM solves this problem by using Fourier analysis. More specific, OFDM uses Inverse Fast Fourier Transform (IFFT) at the transmitter side and Fast Fourier Transform (FFT) to recover the signal at the receiver side [11].

The subcarriers generated in this way do not need separate individual multipliers for each subcarrier and there is no need for guard bands. However, the subcarriers need to be orthogonal to each other over the duration of the symbol. A carrier is orthogonal with respect to another carrier if its sideband nulls (in frequency) fall at the main lobe frequency of the next carrier so that the next carrier is unaffected by the presence of the first carrier [9]. If T_s is the symbol time during which one particular carrier is present as shown in Figure 2.8, then for orthogonality to hold true the next carrier frequency should be separated from the first by $1/T_s$. If the first carrier is at frequency f_0, the next carrier should be at f_0+1/T_s and so on, obtaining the expression for the carrier frequencies given by $f_n = f_0+n/T_s$, where n is the number of subcarriers.

Figure 2.9 shows the process by which an OFDM symbol is generated. A stream of high-rate data bits B bits/sec is passed through a symbol mapper to obtain a high-rate stream of modulation symbols. The high rate stream of modulation symbols is running

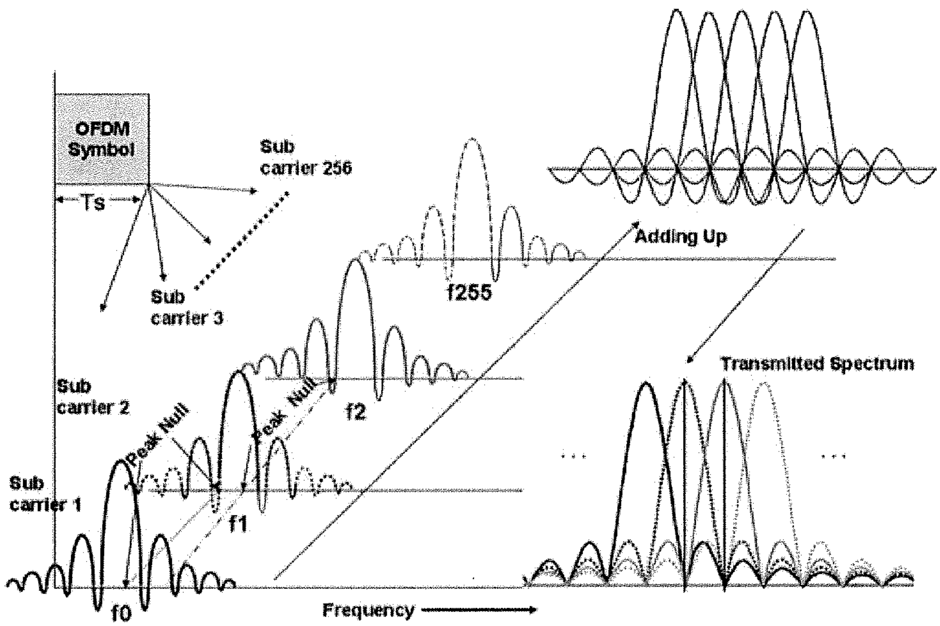

Figure 2.8: Orthogonal Subcarriers in OFDM (Source: [9])

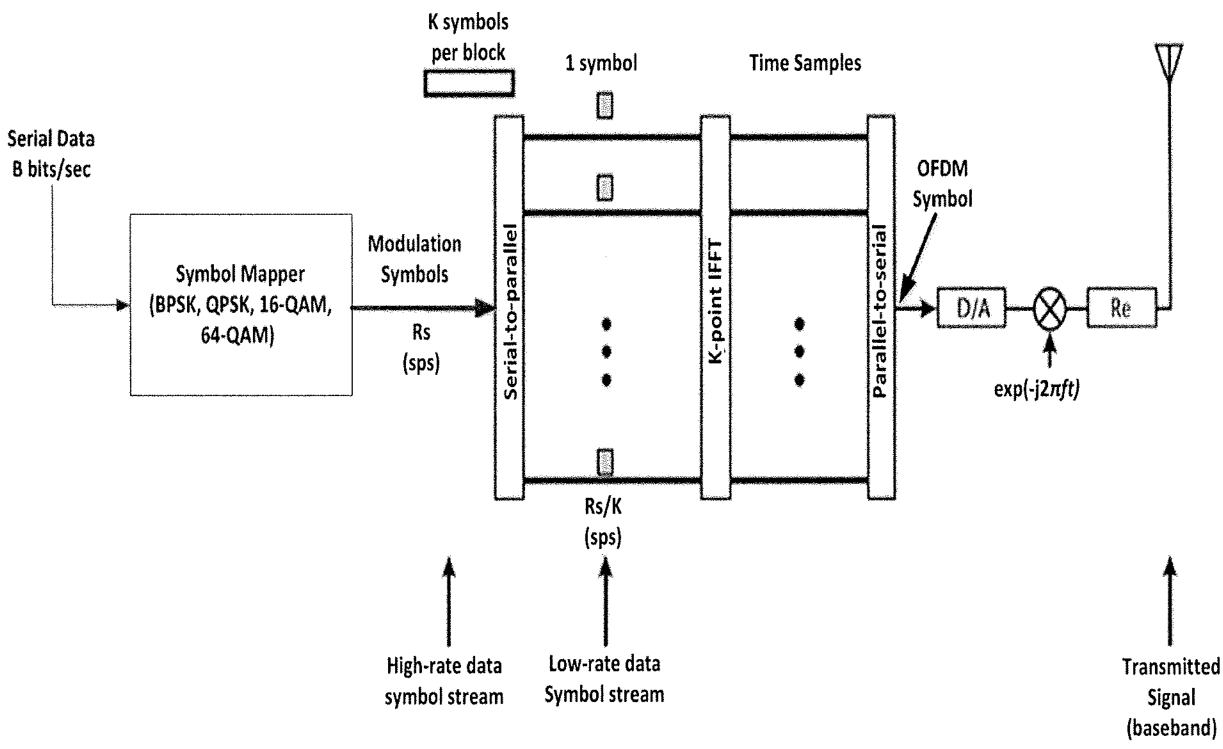

Figure 2.9: OFDM Transmitter (Source: [9])

at a rate of R_s symbols/second (sps). This high-rate stream of modulation symbols is divided into blocks of modulation symbols, and each block contains K modulation symbols. The serial-to-parallel (S/P) converter converts the high-rate stream into K separate low-rate substreams each with a rate of R_s/K symbols per second. The serial-to-parallel converter has as input successive modulation symbols, and outputs K separate substreams. OFDM transmission is thus block based, implying that, during each OFDM symbol interval, K modulation symbols are transmitted in parallel. The modulation symbols can be obtained employing any modulation technique, such as BPSK, QPSK, 16QAM, or 64QAM. The set of K symbols in parallel pass through the IFFT function, which transforms the K symbols into complex time domain symbols.

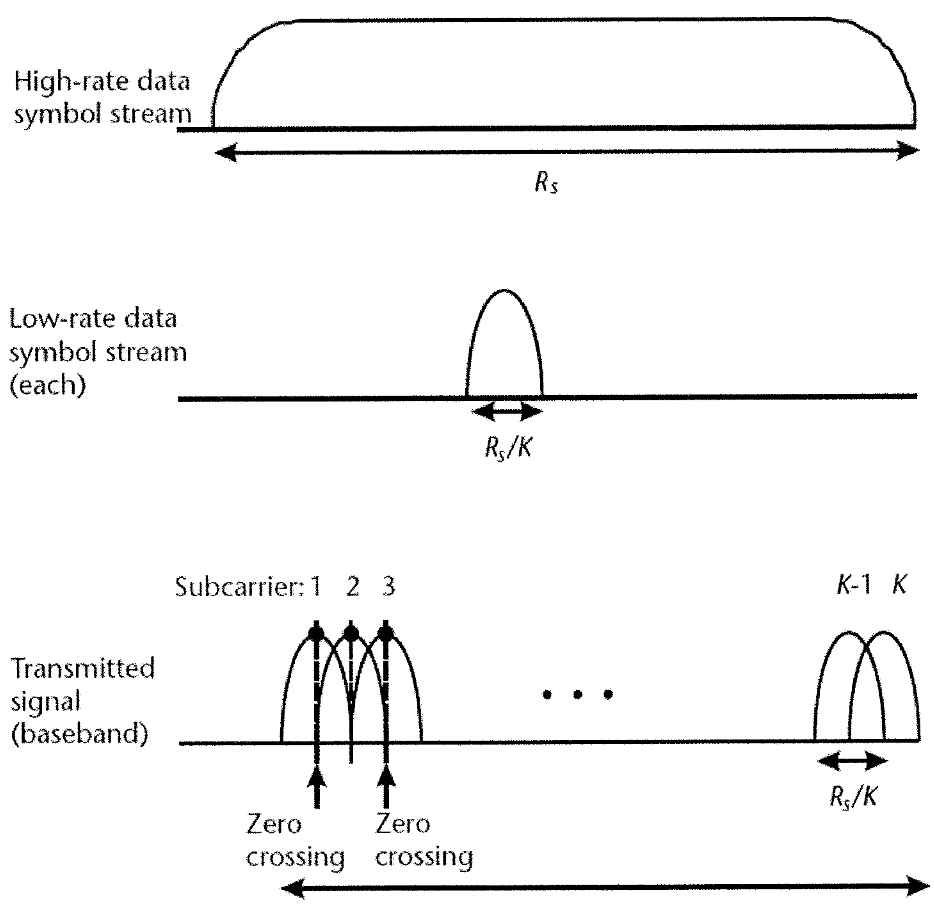

Figure 2.10: The Spectrum of OFDM Signal (Source: [9])

These symbols are then passed through the parallel-to-serial converter in order to be transmitted one after the other as temporal samples. The block of K transformed symbols in series constitute a single block or an OFDM symbol. The output of the D/A converter is then modulated by the sinusoid $exp(-j2\pi ft)$.

Figure 2.10 shows the spectrum of the hight-rate stream, the low-rate substream and the transmitted orthogonal signal. All zero crossings of a subcarrier fall on the peaks of all adjacent subcarriers. Note that even though the figure shows that there are K different subcarriers, all K subcarriers in a block or OFDM symbol are assiged to only one user. What happens if we want more than one user to acces the air interface at one time? The answer is to employ OFDMA which is discussed in the next section.

2.1.2.3 OFDMA Basics

As we saw in the previous section, OFDM's robustness against ISI and multipath fading is a result of employing orthogonal subcarriers. OFDM also achieves better spectral efficiency compared to conventional FDM since the gap between subcarriers is eliminated. However, all subcarriers are assigned to only one user, and only one user can transmitt at a time. OFDMA is a method that assigns different groups of subcarriers in frequency to different users. In this way more than one user can access the air interface at the same time.

Figure 2.11 shows a simplified OFDMA transmitter. The high-rate stream consists of J groups of modulation symbols, with each group containing L symbols. Each of the groups is assigned to a different user. The serial-to-parallel converter separates the high-rate stream into JL separate low-rate substreams. The subcarrier mapper looks at the data symbols and assigns them to their respective subcarriers, each corresponding to a different user. The subcarrier mapper reorders the parallel data symbols according to the particular subcarrier assigned to each user. Each of the parallel data symbols are equivalent to the K data symbols in the OFDM system.

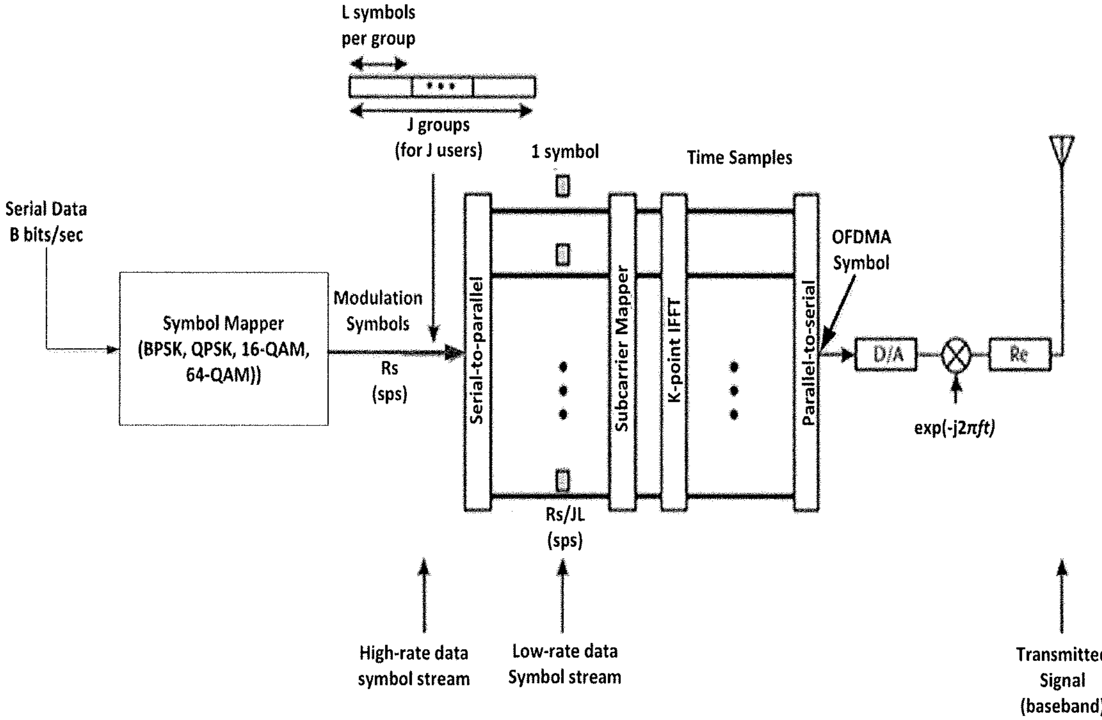

Figure 2.11: OFDMA Transmitter (Source: [9])

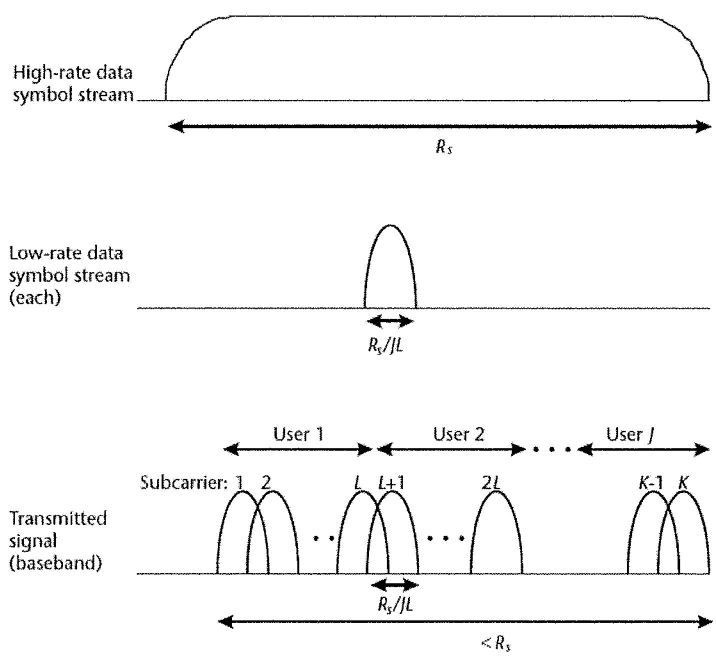

Figure 2.12: The Spectrum of OFDMA Signal (Source: [9])

The parallel symbols are then fed through the IFFT which transforms them to the time domain. The output at the parallel-to-serial converter represents one OFDMA symbol whose spectrum is shown in Figure 2.12. In Figure 2.12, data symbols belonging to a user are carrier by contiguous subcarriers. Like OFDM, OFDMA has good spectral efficiency and it is robust against ISI and multipath fading. OFDMA can also take advantage of multiuser diversity. Different users at different locations likely experience different channel conditions. The system can improve a particular user's link by assigning to that user a set of contiguous subcarriers that experience the best channel condition.

2.1.3 Frequency Domain OFDMA Symbol Consideration

As we have seen in the previous section, the physical layer in WiMAX is based on the use of subcarriers. These subcarriers are generated by n-point IFFT, meaning that we are generating n individual subcarriers. The amount of data that can be carried by each subcarrier depends on the bandwidth of the channel, the modulation scheme used, and the coding rate employed. However, not all subcarriers are used to carry data. There are four subcarrier types identified in WiMAX as shown in Figure 2.13 and explained as follows:

- **Data subcarriers**: used for useful data transmission and their number depend on the size of the channel bandwidth.

- **Pilot subcarriers**: used in uplink and downlink transmission to allow channel estimation, measurements for channel quality indicators such as Signal to Noise Ratio (SNR), frequency offset estimation, etc.

- **Null subcarriers**: have no power and do not transmit any data. Null subcarriers are frequency guard bands to allow the signal to naturally decay.

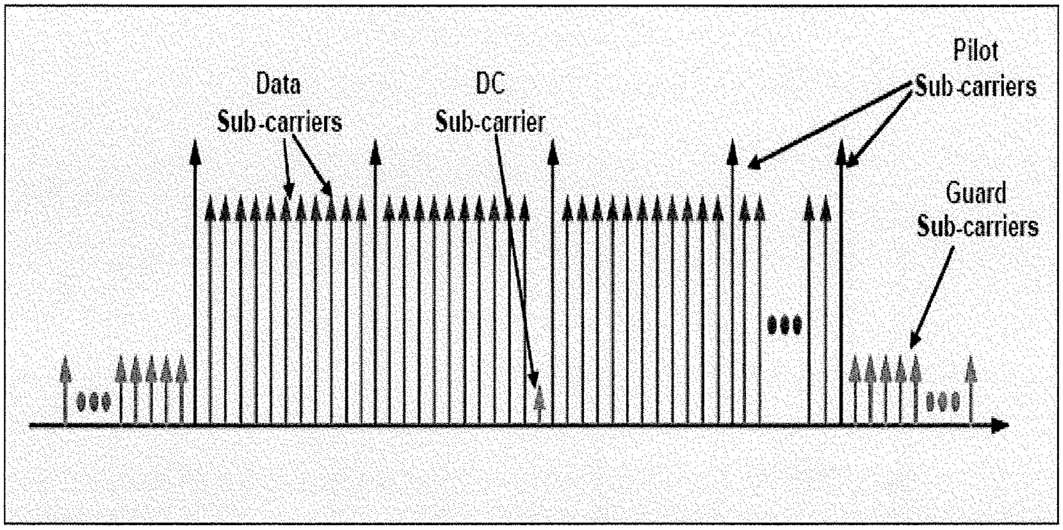

Figure 2.13: Frequency Domain OFDMA Subcarriers (Source: [4])

- **Direct Current (DC) subcarrier**: the DC subcarrier is a subcarrier whose frequency is equal to the RF centre frequency of the transmitting station. In order to simplify Digital-to-Analogue and Analogue-to-Digital Converter operations, the DC subcarrier is null [11].

In the next chapter we will examine how subcarriers are grouped to form subchannels and how data is organized in a hierarchical manner at the physical level.

2.2 WiMAX Medium Access Control Layer

In this section we describe some important IEEE 802.16 Medium Access Control (MAC) functions that are applicable to WiMAX and provide the reader with enough background to understand the concepts presented in the next chapters. Some of the functions include classification and mapping of higher order packets, payload header suppression, fragmentation, packing and scheduling and link adaptation. These functions help WiMAX protocol to achieve two main objectives: to have high bandwidth efficiency and to provide full flexibility for the QoS management. The IEEE 802.16

standard was developed from the beginning for the delivery of broadband services including voice, data, and video. The MAC layer is based on the time-proven Data Over Cable Service Interface Specification (DOCSIS) standard and can support bursty data traffic with high peak rate demand while simultaneously supporting streaming video and latency-sensitive voice traffic over the same channel [13]. The primary task of the WiMAX MAC layer is to provide an interface between the higher network layer and the physical layer. The MAC layer in WiMAX is reservation-based and contention-free. The reservation-based resource allocation allows the WiMAX base station to serve a large number of users as well as the guarantee of QoS for both uplink and downlink traffic.

The WiMAX MAC layer is divided into three sub-layers: service-specific convergence sublayer (CS), MAC common part sublayer (CPS) and security sublayer as shown in Figure 2.14. The CS resides on top of the MAC CPS and its function is to communicate with the network layer. Data received through the CS service access point (SAP) is mapped into MAC service data units (SDUs) received by the MAC common part sublayer through MAC SAP. The WiMAX standard has defined two service specific convergence sub-layers for service mapping to and from the MAC CPS sublayer. These are the Asynchronous Transmission Mode (ATM) and Packet CS [4]. The ATM standard makes use of the ATM CS, while Packet CS is used for IPV4, IPV6 and Ethernet. Some other functions of the CS include suppression of payload header information, rebuilding any suppressed payload header information and enabling QoS and bandwidth allocation based on parameters received from upper layers.

The MAC CPS forms the core of the MAC layer and is designed to support both point-to-multipoint and mesh network architectures. For the purpose of mapping to services on subscriber stations and associating varying levels of QoS, all data communications are in the context of a connection. The MAC defines two kinds of connections: management connections and transport connections. Upon entering a network, a subscriber station is assigned three management connections: basic management connection, primary management connection and secondary management connection. The basic connection is responsible for transferring short and time urgent MAC and Radio Link Control (RLC) management messages. The primary management connection is used by the base station MAC and subscriber station MAC to exchange longer, more delay-tolerant MAC management messages. The secondary management connection is used for the transfer of standard based management messages such as Dynamic Host Configuration Protocol (DHCP), Trivial File Transfer Protocol (TFTP), etc. The transport connections are used to transport data between the users and the base station.

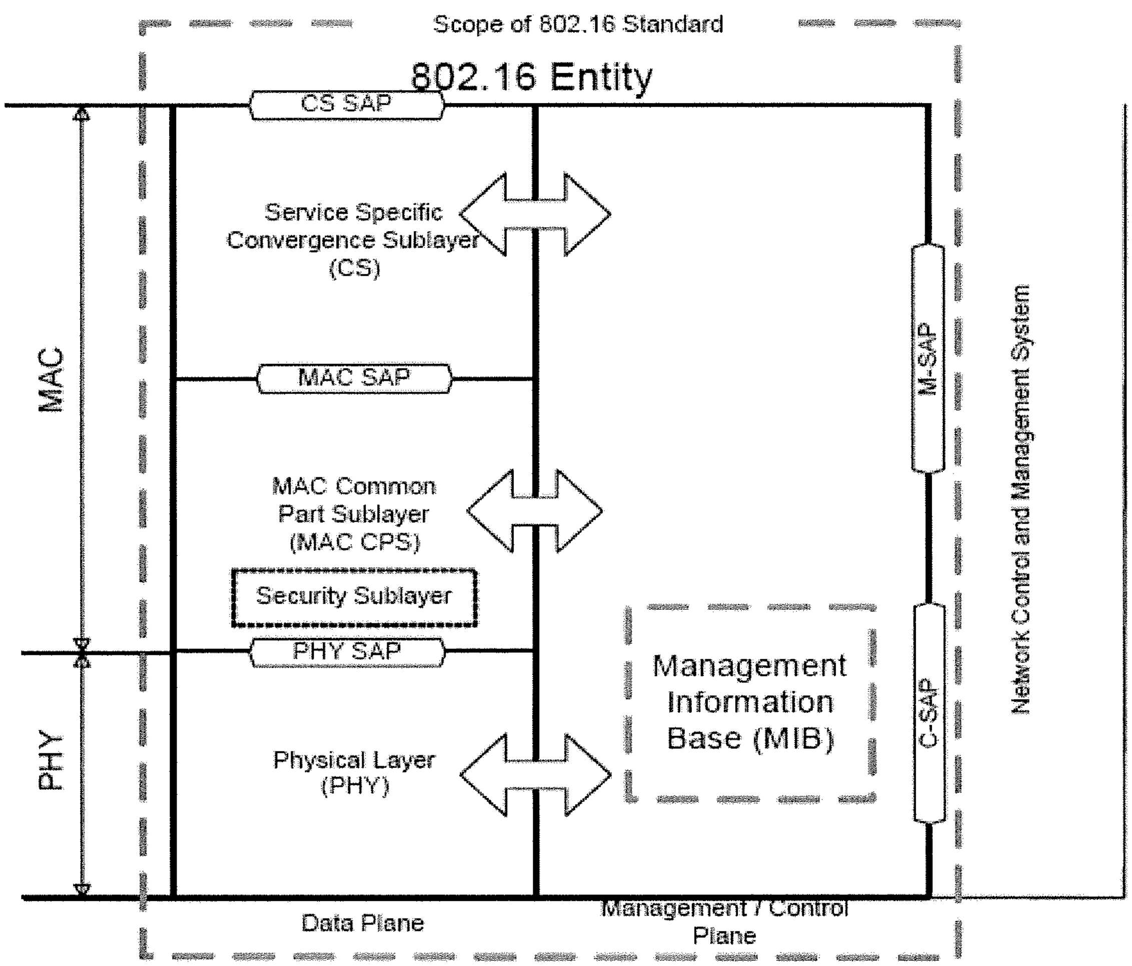

Figure 2.14: IEEE 802.16 Protocol Reference Model (Source: [4])

Some of the functionalities performed by the common part sublayer are:

- **Channel Acquisition**: Whenever a subscriber station enters a network, the MAC protocol includes an initialization procedure that allows the subscriber station to scan its frequency list to find an operating channel. MAC will periodically search and check the uplink and downlink channel descriptor messages, which inform the SS about modulation schemes employed on the communication channel.

- **Ranging and Negotiation**: is the process of acquiring the correct timing offset, frequency offset and power adjustments so that the subscriber station uplink transmissions are aligned with the base station uplink frame. The base station is the entity that commands any timing or power level adjustments.

- **Bandwidth Grant and Request**: A subscriber station uses the bandwidth request (BR) header to request bandwidth from the base station. The MAC layer distinguishes between two classes of requests:

    - **Grant per Connection (GPC)**: For GPC class, bandwidth is granted explicitly to the connection of the subscriber station. It is the class of request used in this thesis.

    - **Grant per subscriber station** (GPSS). For GPSS class, bandwidth is granted to the subscriber station and it is up to the subscriber station to determine how to distribute the bandwidth among its connections.

  WiMAX also allows subscriber stations to request bandwidth via contention or piggyback. In contention mechanisms, the subscriber stations will contend for a slot to send the bandwidth request while in the piggyback mechanism the subscriber stations will attach its bandwidth request onto data packets.

- **Scheduling services:** The common part sublayer maps each connection to a scheduling service, where a scheduling service is associated with pre-determined

QoS parameters. The WiMAX protocol specifies five classes of service which will be discussed in the next subsection.

The third MAC sublayer is the security sublayer which provides authentication, secure key exchange and encryption.

2.2.1 Quality of Service (QOS) Provisioning

Quality of Service capability plays an important role in WiMAX architecture. Without sophisticated QoS, many new wireless services, from legacy data services to complex interactive services, do not work as well as they could. We will present a few parameters employed in describing QoS[14].

- **Bandwidth** is one of the most basic QoS parameters and is defined as the amount of data transmitted by a user per unit time. It is limited by the channel conditions between end user and base station and by the number of end users that are active at any one time since the overall system bandwidth is shared.

- **Latency** is the time it takes a packet of data to get from one designated point to another. Some of the contributors to latency are propagation delay, processing delay and storage delay.

- **Jitter** is the variation of delay over different packets. Jitter can be caused by network congestion, timing drift, or route changes. A jitter buffer can be used to handle jitter.

- **Reliability** is the amount of successfully delivered packets. It is more complicated in wireless networks than fixed networks. This is due to the fact that radio waves propagation is less reliable especially when we are dealing with small antennas and low powers in cluttered areas such as urban center.

Depending on the traffic characteristics of user applications and their associated QoS parameters, WiMAX scheduling services can be classified into five different Classes of Service (CoS) [4] [12][14]:

- **Unsolicited Grant Service (UGS)**: This service is designed to support applications that generate fixed-sized data packets periodically such as T1/E1 and VoIP. This service ensures that the packets are transmitted at periodic intervals with little or no variation so that jitter and latency requirements are satisfied.

- **Real-Time Polling Service (rtPS)**: This service provides support for real time data streams consisting of variable bit-rate (VBR) data that are issued at periodic intervals, such as compressed moving picture express group (MPEG) video or voice telephony. For this service, delay variation or jitter is not as important as in UGS. However, latency requirements need to be satisfied.

- **Extended Real Time Polling (ertPS):** This service provides support for real-time data streams for which the bit rates vary slightly with time. It is implemented for applications such as VoIP with silence suppression. This service is a blend of UGS, which achieves low delay and rtPS which offers more efficient bandwidth utilization.

- **Non Real Time Polling Service (nrtPS):** This service provides support for non-real time applications that require variable size data grants on a regular basis. This service is delay tolerant but requires a minimum reserved traffic rate. It is ideal for transporting delay sensitive traffic such as File Transfer Protocol (FTP).

- **Best-Effort (BE):** This service supports data streams for which no minimum service level is required. It is used to transport delay-insensitive, noncritical background traffic such as E-mail traffic and HTTP.

Even though the IEEE 802.16 standard has provisions to support more than five scheduling services, to the best of our knowledge the latest standard defines only the above mentioned scheduling services. Since the standard does not suggest the scheduling schemes used to manage network resources, a lot of research has been performed to design and implement scheduling algorithms for WiMAX. In the next chapter we will give an overview of the state of research in the area of bandwidth allocation for broadband wireless networks.

2.3 Chapter Summary

In this chapter we provided the reader with a detailed theoretical background of the WiMAX technology with emphasis on the Physical (PHY) and Medium Access Control (MAC) Layers as specified by the IEEE 802.16 standard. BPSK, QPSK, 16-QAM and 64-QAM modulation techniques were discussed and we showed how data bits are modulated onto carriers using these four modulation techniques. We also showed how the modulated carriers are transmitted over the air to maximize throughput and minimize the effect of dispersion and multipath propagation. In particular we identified OFDMA as the most widely used WiMAX access scheme and the scheme used in the thesis. We introduced some of the functions of the Medium Access Control layer and scheduling services support by WiMAX.

Chapter 3 Literature Review

A critical part of the MAC layer specification is packet scheduling, which resolves contention for bandwidth and determines the transmission order of users. Although the IEEE 802-16 standard specifies a service framework and associated bandwidth request/grant mechanisms, it does not specify how the network resources are allocated to users. The implementation of scheduling algorithms is left at the discretions of equipment makers and network operators. Based on the literature survey we classified scheduling algorithms for uplink traffic in WiMAX into three categories: legacy scheduling algorithms, hybrid scheduling algorithms, and Channel Aware Scheduling algorithms.

3.1 Legacy Algorithms

The scheduling algorithms proposed in this category are based on algorithms and methods developed for wired networks, which have been adapted for wireless broadband networks. The simplest algorithm is the Round Robin (RR) scheduler evaluated for performance in WIMAX networks in [15]. The scheduler assigns time slots to each user in equal proportion without priority. Once a user has been serviced it is not visited again until all the other users in the system have been serviced. It is a simple and easy to implement technique, but as the authors found, it cannot guarantee different QoS requirements and it has low bandwidth efficiency.

Reference [16], evaluates the performance of a Weighted Round Robin (WRR) for the uplink traffic. The WRR was first proposed in [17] for ATM networks as an improvement for the round robin algorithm. Packets are first classified into various service classes and then assigned a queue. Each user is assigned a certain weight and the bandwidth is allocated according to the weights in a round robin fashion. WRR ensures that all service classes have access to at least some configured amount of network bandwidth, in order to avoid bandwidth starvation.

Reference [18], evaluates the performance of the Earliest Deadline First (EDF) algorithm, which is found to be suited for systems with guaranteed service requirements. The scheduler assigns deadlines to each packet and it services packets in the order of their deadlines. The packet with the minimum (or earliest) deadline will be serviced first. The EDF algorithm is suitable for UGS, ertPS and rtPS classes since they have strict delay requirements. The remaining two classes, nrtPs and BE do not have delay requirements and are only serviced if there are packets remaining. Hence, this algorithm does not protect against greedy connections. Reference [18], also employs Weighted Fair Queuing (WFQ) for the uplink traffic in WiMAX. WFQ assigns each packet a finish time based on a weight assigned to the user and the size of the packets. Packets are serviced in increasing order of their finish time. The WFQ scheduler gives an advantage when many connections and classes need to be served.

Reference [19], proposes an uplink scheduling algorithm and a token bucket based Call Admission Control (CAC) algorithm. The CAC algorithm assigns thresholds to each class to avoid starvation of lower priority classes. The scheduling algorithm first grants bandwidth to subscriber stations of the UGS class. The algorithm will then allocate bandwidth to subscriber stations of the rtPS class using the Early Deadline First algorithm. The algorithm will allocate minimum required bandwidth to the nrtPs and BE classes, respectively. Each subscriber station is controlled by a token rate and bucket size. A mathematical model is proposed that estimates the appropriate token rate based on the queuing delay and the packet loss requirements.

A more dynamic solution that has been proposed for Passive Optical Networks which can be adapted to WiMAX, called Interleaved Polling with Adaptive Cycle Time (IPACT) [23]. It uses a polling mechanism with Request and Grant messaging between a central office (OLT) and users connected to Optical Network Units (ONUs). A Grant message is used by the OLT to assign a variable-sized transmission slot to an ONU. A Request message is used by the ONU to convey its local buffer occupancy to the OLT. To support different QoS, the authors have added strict priority queuing to the basic

IPACT. Every ONU maintains a separate queue for each CoS in its buffer. The OLT issues "colorless" grants to ONUs, which means that the OLT does not dictate how many bytes from a particular queue an ONU must transmit. Instead, the ONU uses a strict priority policy to determine the order in which the queues are processed. This scheme leads to a phenomenon known as light-load penalty, where the queuing delay for some traffic classes increases when the network load decreases. Another problem with this scheme is that it fails to distribute the upstream bandwidth fairly among the users. As an example, if two identical users access two different ONUs with identical requests, these users can receive very different qualities of service when one of the ONUs is lightly loaded and the other ONU is fully loaded.

In [24], a dynamic bandwidth scheme called Class-of-Service Oriented Packet Scheduling (COPS) is proposed. The COPS scheme employs a credit pooling technique combined with a weighted-share policy to partition the upstream bandwidth among different classes of service and to prevent users from monopolizing the bandwidth. The scheduler also controls the traffic from each ONU in order to place a limit on the number of packets transmitted during each transmission interval. The OLT runs the COPS algorithm once every transmission cycle in order to generate grants within a specific CoS for each ONU. In this sense the grants are “colored”, which means that they dictate how many bytes from a particular CoS the ONU must transmit. COPS algorithm is shown to be superior to other existing dynamic bandwidth allocation schemes for Ethernet Passive Optical Networks, in terms of providing global fairness between different classes of service.

3.2 Hybrid Algorithms

This category of algorithms uses a combination of legacy algorithms presented in the previous section to satisfy the uplink QoS requirements of the multi class traffic as specified by the WiMAX protocol.

Reference [20], proposes a hybrid scheduling algorithm that supports four classes of service flows (UGS, rtPS, nrtPS and BE). The proposed uplink packet scheduling algorithm uses a combination of strict priority service discipline, earliest deadline first (EDF) and weight fair queuing (WFQ). The proposed algorithm consists of three modules: information module, scheduling database module, and service assignment module. The overall allocation of bandwidth is done in a strict priority manner i.e. all the higher priority SSs are allocated bandwidth until they do not have any packets to send. The EDF scheduling algorithm is used for SSs of the rtPS class, WFQ is used for SSs of the nrtPS class and FIFO for SSs of the BE class. One shortfall of this algorithm is monopolization of bandwidth by high priority SSs due to the strict priority bandwidth allocation.

Reference [21], proposes a hybrid scheduling algorithm that uses WRR and RR algorithms with a strict priority mechanism for overall bandwidth allocation. In the initial portion of the algorithm, bandwidth is allocated on a strict priority basis to SSs of the rtPS and nrtPS classes only. After that, the WRR algorithm is used to allocate bandwidth among SSs of rtPS and nrtPS classes until they are satisfied. If any bandwidth remains, it is distributed among the SSs of the BE class using the RR algorithm. The hybrid algorithm is opportunistic in nature as the WRR and RR algorithms select the SSs with the most robust burst profiles. This algorithm will starve lower priority SSs in the presence of a large number of higher priority SSs. The algorithm can also result in low fairness among SSs as it selects SSs with the most robust burst profiles first.

Reference [22], proposes a Multi-class Uplink Fair Scheduling Structure algorithm to satisfy throughput and delay requirements of the multi-class traffic in WiMAX. The model is based on Grant Per Subscriber Station (GPSS) bandwidth grant mode and thus schedulers are implemented at the SSs to distribute the bandwidth granted among their connections. At the SS, Modified WFQ (MWFQ) is used for UGS and rtPS connections, MWRR is used for nrtPS connections and FIFO is used for BE connections.

3.3 Channel Aware Scheduling Algorithms

The algorithms presented in this section consider variations in channel conditions while attempting to satisfy the QoS requirements of the users.

Reference [25], proposes a channel allocation and scheduling algorithm in OFDMA systems for the downlink transmission. The system resources are defined as channel units. Each channel unit is assigned a *representative quality* based on the signal-to-interference ratio. The algorithm is broken down into two stages. First, the scheduler sorts the channel units available within a Transmission Time Interval (TTI) in order of the representative quality. Second, the scheduler analyzes the channel units request from users and proceeds in categorizing the users based on their representative quality. The scheduler starts matching available channel units with user requests. The scheduler starts with the highest quality channel units and allocates them sequentially to the most adequate users. The drawback of this algorithm comes from the fact that users with low representative quality are being serviced last. This results in unfair distribution of resources among users with different representative qualities.

Reference [26], proposes a joint bandwidth allocation and connection admission control algorithm based on the queuing theory. The algorithm employs bandwidth threshold to limit the amount of bandwidth per class of service. Subscriber stations are allocated a utility function and the scheduler allocates bandwidth based on utility starting with the subscriber station with the least utility. Utility function is constructed using average delay and maximum latency for the rtPS class and average and minimum throughput for the nrtPS class. Subscriber stations of the BE class are assigned the highest utility as long as one unit of bandwidth is assigned to them. The unit of bandwidth used in this algorithm is a fixed size packet called a Protocol Data Unit (PDU). In the initial stages of the algorithm, minimum bandwidth is allocated to the subscriber stations in a strict priority order. Afterwards, the algorithm assigns the residual bandwidth in PDU units to the subscriber stations. SSs are sorted in ascending order based on their utility and they are allocated the PDUs based on a water filling mechanism until the bandwidth of each class exceeds the class threshold. The thresholds for the classes are calculated such that QoS requirements of the SSs are met and the system revenue is maximized.

Reference [27], proposes scheduling algorithms for the OFDMA system with a TDD frame structure for both uplink and downlink traffic in WiMAX. An optimization problem is first formulated for the allocation of resources to the subscriber stations. The formulation considers the channel quality in terms of the number of bytes/slot a subscriber station can send on a channel, the number of slots allotted to the subscriber station and the total bandwidth demanded by the subscriber station. A linear programming solution to the optimization problem is devised, in order to maximize the throughput of the subscriber stations. Due to the high complexity of the linear programming solution, a heuristic algorithm is considered that provides close to optimal performance. A definition of fairness is also proposed, and it requires assigning bandwidth to a subscriber station proportional to its throughput requirement. The bandwidth allocation among the scheduling services is done on a priority basis, similar to that of the hybrid algorithm proposed in [20]. More specifically, the base station first attempts to satisfy the requirements of UGS connections followed by the requirements of rtPS and nrtPS connections. Finally, any residual bandwidth is distributed among the BE connections. This mechanism of allocating bandwidth might lead to unfairness among the connections.

In this chapter we presented various proposals of uplink scheduling algorithms for WiMAX. Some of those algorithms were initially developed for wired systems but showed positive results when implemented in WiMAX systems. WiMAX is a fast changing technology as we can see from the addition of the ertPS class in the latest IEEE 802.16 revision. Some of the algorithms reviewed were designed without this additional class of service. Nevertheless, the background information and concepts presented in those algorithms provide a solid foundation for future work.

3.4 Chapter Summary

In this chapter we provided information on the representative uplink scheduling algorithms. Based on the literature survey we classified the scheduling algorithms for uplink traffic in WiMAX in three categories: legacy scheduling algorithms, hybrid scheduling algorithms and Channel Aware Scheduling algorithms.

Chapter 4 System Overview

4.1 Basic System Model

In this thesis we focus on the uplink transmission since collisions may happen when multiple users try to access the common channel in the uplink direction. We consider a WiMAX network consisting of a base station (BS) and N subscriber stations (SS) connected wirelessly in a tree topology as shown in Figure 4.1. Downlink communications occurs from the BS to SS, while uplink communication is defined as communication from SS to the BS. Our scheduling algorithm is employed for uplink traffic.

The scheduling algorithm is employed at the BS. The lack of communicability between subscriber stations means that the only entity that can arbitrate access to the shared channel is the base station. Every SS contains M queues, each serving a class of service (CoS). The CoSs are used for delivering voice, video and data traffic and can be mapped to one of the five standardized classes as defined by the WiMAX protocol. For example, CoS1 can be mapped to Unsolicited Grant Service (UGS), which provides low loss and delay sensitive service for applications such as VoIP; CoS2 can be mapped to Real-Time Polling Service (rtPS), which provides delay sensitive service for applications such as MPEG video; CoS3 can be mapped to Extended Real Time Polling Service (ertPS), which provides low delay; CoS4 can be mapped to Non Real Time Polling Service (nrtPS), which provides minimum reserved traffic rate for FTP applications; and CoS5 can be mapped to Best Effort (BE), which does not require any commitment from the network.

All of the queues in each subscriber station share a common buffer. To ensure that incoming data destined for a specific queue does not take up more than its fair share of the buffer space, we employ a mechanism called Weighted Random Early Detection (WRED) [28].

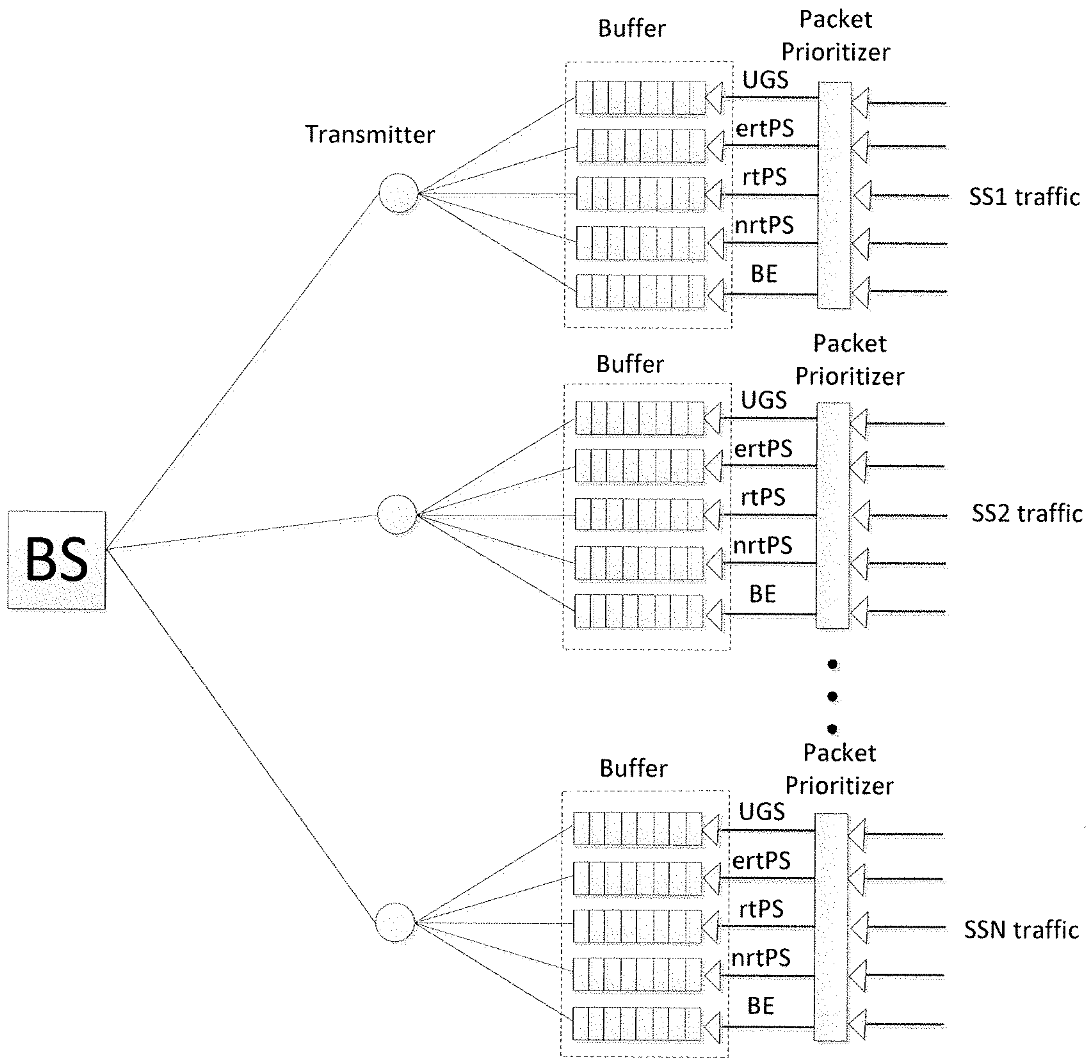

Figure 4.1: Queue Mapping

The WRED operates on the average queue size S_α for every CoS m, where α is from 1 to m. On the arrival of a CoS m packet at the SS buffer, the WRED computes a new value for S_α using an exponential weighted moving average:

$$S_\alpha = \mu \times S_i + (1 - \mu) \times S_\alpha \tag{4.1}$$

where μ is a fixed weight parameter, and S_i is the current queue size for all CoSs 1 to m.

The WRED will then decide to reject or accept a newly arrived packet based on the following outcomes:

- If S_α is less than a predetermined minimum queue threshold (S_{min}), accept packet;
- Else, if S_α exceeds a predetermined maximum queue threshold (S_{max}), reject packet
- Else, the packet will be dropped with the probability P:

$$P = P_d \times \frac{S_\alpha - S_{min}}{S_{max} - S_{min}} \tag{4.2}$$

where P_d is the maximum drop probability. This process is illustrated in Figure 4.2, while a depiction of the five queues employed at each SS serving a specific class of service is shown in Figure 4.1.

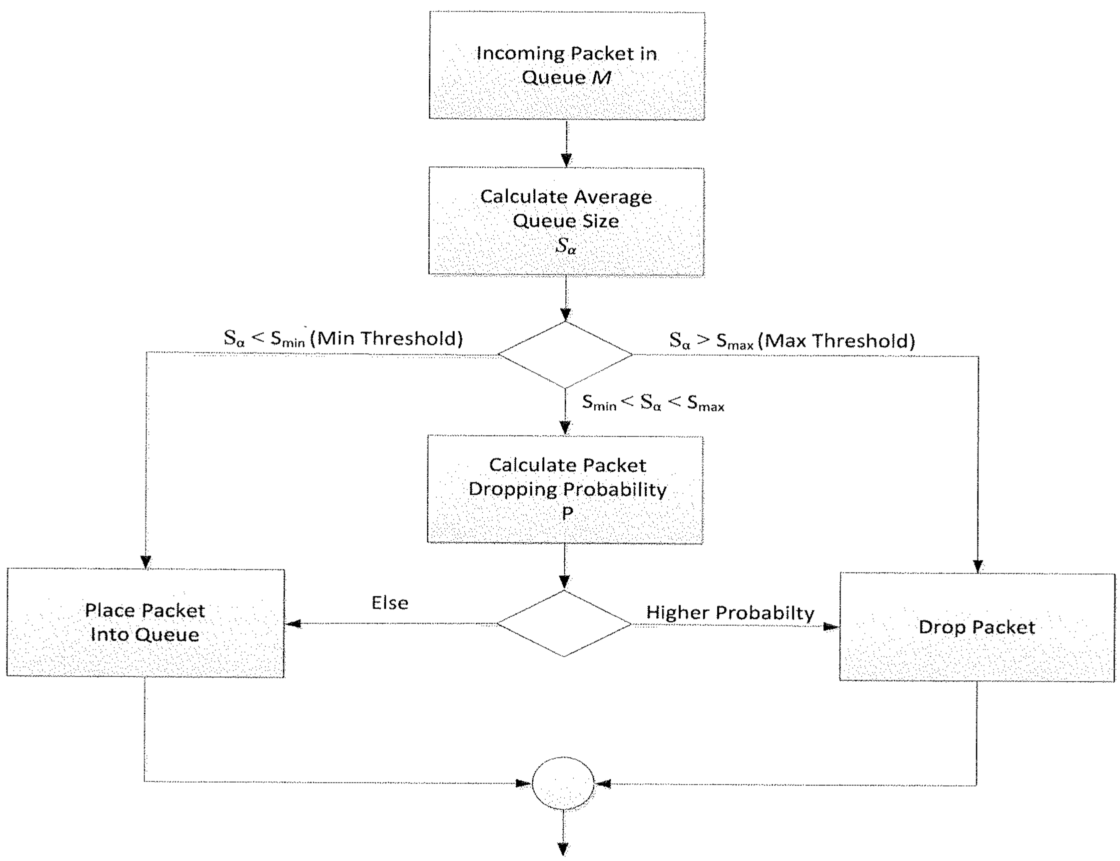

Figure 4.2: WRED Buffer Management Mechanism

4.2 Architecture of the Proposed Scheduling Algorithms

The base station keeps track of available system resources by employing a resource-allocation matrix. The system resources are defined in terms of *slots,* which are the minimum possible data allocation units [4]. In this section we will show how slots are defined and how to construct the slot allocation matrix.

In OFDMA, the radio resources are two-dimensional regions over time and frequency. The time axis is divided into OFDMA symbols and the frequency axis is divided into OFDMA subcarriers. Regulatory bodies throughout the world are allowing

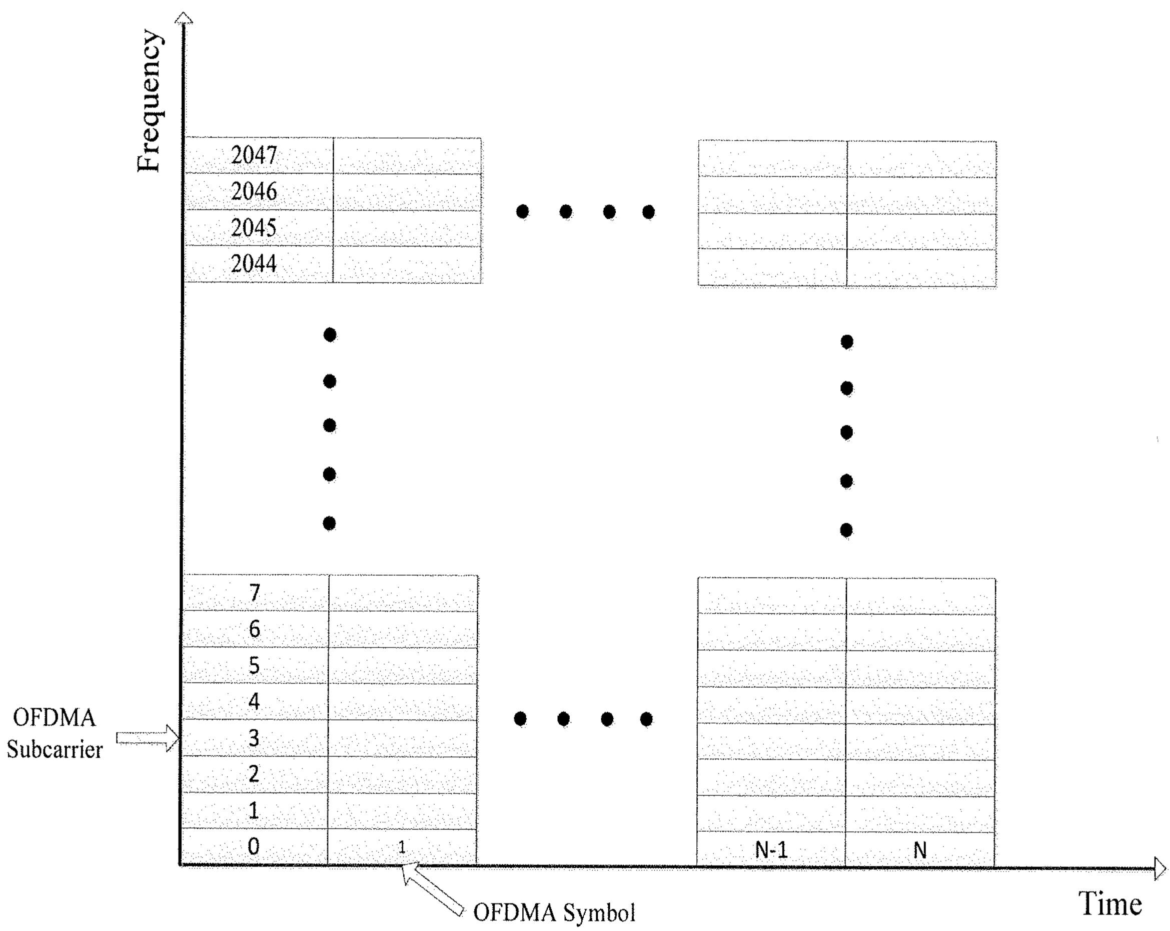

Figure 4.3: OFDMA Resource Matrix

network operators to deploy WiMAX systems in spectrum block that are integer multiples of 5 MHz channel bandwidths [33]. Hence, WiMAX may use 512, 1024 or 2048 FFTs based on whether the channel bandwidth is 5MHz, 10MHz, or 20MHz, respectively. In this thesis we are employing a channel bandwidth of 20 MHz corresponding to a 2048-point FFT as shown in Figure 4.3. To completely describe the allocation matrix we need to find the number of OFDMA symbols that WiMAX can support. The time axis is divided into OFDMA *time symbols* having the structure shown in Figure 4.4, with

$$T_s = T_u + T_g \tag{4.3}$$

where T_u is the useful symbol period, T_g is the guard time also known as cyclic prefix (CP) and T_s is the symbol duration including guard time. In Figure 4.3 each horizontal partition represents a symbol of duration T_s.

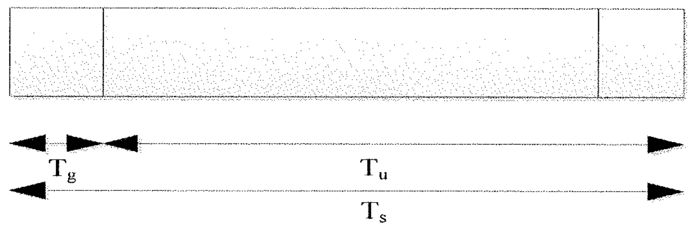

Figure 4.4: Time Domain OFDMA Symbol Structure (Source: [4])

The IEEE 802.16 standard specifies that for transmission, time domain symbols must be grouped together in a frame. The duration of the frame indicates the periodicity of the downlink (DL) frame start preamble in both FDD and TDD cases. Although the IEEE 802.16 standard specifies a number of radio frame sizes ranging from 2 ms to 20 ms, WiMAX systems only support a 5 ms radio frame. The 5 ms frame time gives low-latency jitter to delay-sensitive applications while keeping the framing overhead low [32]. To find the number of symbols per 5 ms radio frame, we need to find the symbol time duration, T_s. T_s can be found by making use of WiMAX primitive parameters such as channel bandwidth, number of subcarriers and sampling factor as shown in Table 4.1

Parameter	Description	Value
BW	Channel Bandwidth	20 MHz
N_{used}	Number of subcarriers	2048
n	Sampling factor	28/25
G	Ratio T_g/T_u	1/8

Table 4.1: OFDMA Symbol Primitive Parameters (Source: [4])

[4]. OFDM analog to digital converters sample the incoming signal using the so called sampling factor n. The sampling factor depends on the channel bandwidth and is set by the WiMAX standard to 28/25 for channel bandwidth of 20 MHz [4]. The sampling factor is used in conjunction with the channel bandwidth (BW) and number of subcarriers, N_{used} to determine the subcarrier spacing and the useful symbol time, T_u. The sampling frequency F_s is calculated using the following formula [4]:

$$F_s = n \times BW = (28/25) \times 20\ MHz = 22.4 MHz \tag{4.4}$$

From the above formula it seems that Nyquist theorem is being violated since the sampling rate is less than twice the highest frequency contained in the signal. Recall that in OFDMA we are dealing with complex signals. Let B denote the double-sided bandwidth of a complex-valued baseband signal (i.e., the spectrum is nonzero only for $-B_{neg} \leq f \leq B_{pos}$, $B = B_{neg} + B_{pos}$). To avoid aliasing, the sampling frequency F_s should be high enough such that the spectral images do not overlap [34] :

$$F_s - B_{neg} \geq B_{pos} \Leftrightarrow F_s \geq B_{pos} + B_{neg} \Leftrightarrow F_s \geq B \tag{4.5}$$

In our case F_s is greater than B times the sampling factor n, which is greater than one. Next, we calculated the subcarrier spacing Δf as follows:

$$\Delta f = F_s/N_{used} = 22.4MHz/2048 = 10.94\ KHz \tag{4.6}$$

Using the subcarrier spacing value, we can calculate the useful symbol time T_u:

$$T_u = 1/\Delta f = 1/10.94\,KHz = 91.4\,\mu s \tag{4.7}$$

The symbol duration T_s is found using Equation 4.3:

$$T_s = T_u + T_g = T_u + G \times T_u = 91.4\mu s + 1/8 \times 91.4\mu s = 102.85\mu s \tag{4.8}$$

Having calculated the symbol duration T_s, we obtain the number of OFDMA symbols contained in one 5 ms frame as following:

$$OFDMA\ symbols = frame\ duration\ /T_s = \frac{5\ ms}{102.85\ \mu s} = 48\ \frac{symbols}{5\ ms\ frame} \tag{4.9}$$

In the frequency domain of OFDMA systems, the encoded and modulated data is mapped into multiple frequency components called *subcarriers*. Recall from Chapter 2, that OFDMA technology defines three types of subcarriers: data, pilot and null subcarriers. While in theory implementation of multiple carriers in OFDMA can be completely orthogonal by using infinite number of iteration in IFFT, in practice there is still a need of guard bands to be provided between two adjacent symbols [10]. Null subcarriers are frequency guard bands employed to allow the signal to naturally decay. WiMAX uses left and right frequency guard bands as follows: 160 subcarriers are allocated to the left guard, 159 subcarriers are allocated to the right guard and one subcarrier allocated as DC subcarrier. The remaining 1728 active subcarriers (data and pilot) are further grouped in multiple subcarriers to form *subchannels*. Subchannelization leads to an efficient utilization of bandwidth by subdivision of an OFDMA frequency symbol into subchannels, thus allowing a given terminal to transmit only a few subcarriers instead of all available subcarriers. WiMAX protocol states that each subchannel contains 18 active subcarriers for a total of 1728/18=96 subchannels [4]. From the above discussion we can conclude that each rectangle in Figure 4.5 has the dimension of one subchannel on the frequency axis and one OFDMA symbol on the time axis.

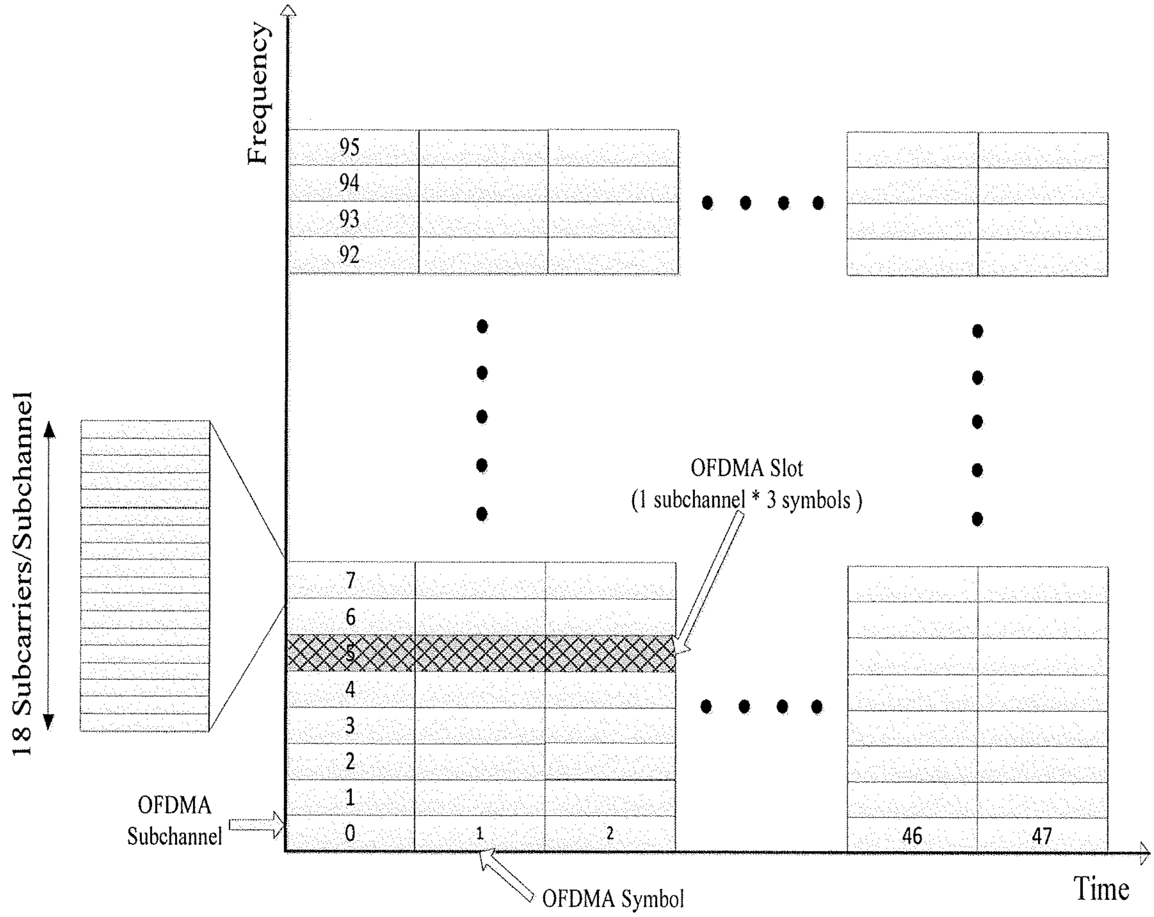

Figure 4.5: OFDMA Sub-channelization Using Active Subcarriers

However, WiMAX protocol states that a *slot* is the minimum data allocation unit in OFDMA. The slot is defined in two dimensions; one in time (OFDMA symbol number) and one in frequency (the subchannel number). The exact slot structure depends on the communication type, which gives us three possible slot definitions as shown in Table 4.2. For downlink Full Usage of the Subchannels (FUSC) and downlink Partial Usage of the Subchannels (PUSC), the pilot subcarriers are allocated first. What remains are data subcarriers, which are divided into subchannels that are used exclusively for data. For uplink PUSC, the set of used subcarriers is first partitioned into subchannels and then the pilot subcarriers are allocated from within each subchannel [11]. Since in this

Communication	Slot Definition
Downlink FUSC	1 subchannel × 1 symbol
Downlink PUSC	1 subchannel × 2 symbols
Uplink PUSC	1 subchannel × 3 symbols

Table 4.2: Slot Definition (Source: [4])

thesis we are designing algorithms for the uplink direction, we will employ the PUSC slot structure, where a given slot has a dimension of one subchannel by three symbols. Hence, each slot contains 3×18=54 subcarriers. Out of the 54 active subcarriers, 48 are data subcarriers and six are pilot subcarriers used for channel estimation as mandated by WiMAX [4].

Having determined the slot dimensions, we can represent the upstream resources as a slot resource matrix containing N_{total} slots as shown in Figure 4.6. Since the time axis contains 48 time domain symbols and a slot has three time domain symbols, we can calculate the number of time slots per 5 ms frame to be 48/3=16 slots. Knowing that the frequency domain axis is divided into 96 equal intervals or subchannels the slot resource matrix shown in Figure 4.6 contains

$$N_{total} = 96\ subchannels \times 16\ slots = 1536\ slots \tag{4.10}$$

This number of slots represents the maximum number of resources that the scheduler will have available to allocate to the contending subscriber stations. The amount of data encoded by one slot depends on the radio link quality, as we will see in the next section.

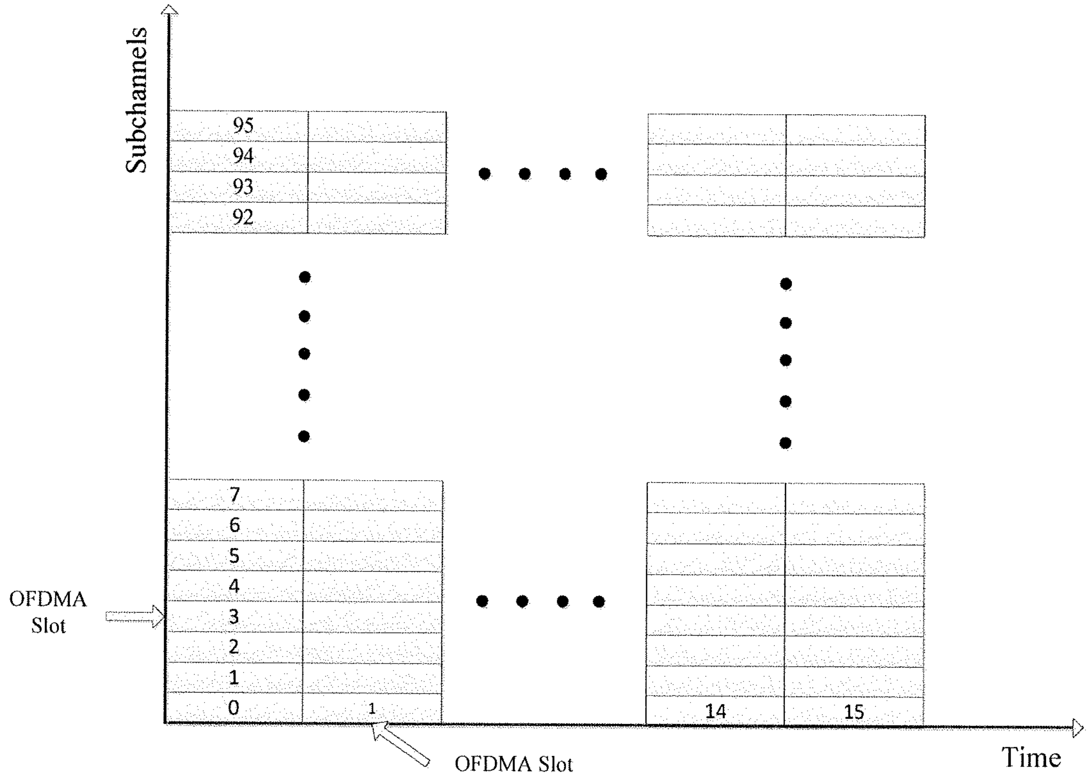

Figure 4.6: Slot Resource Matrix

4.3 Modulation and Coding Parameters

WiMAX system adjusts the signal modulation and coding rates on the basis of the Signal to Noise ratio (SNR) value of the radio link as conveyed by the pilot subcarriers. When the radio link quality is high, the highest modulation scheme and light coding is used giving the system more capacity. During a signal fade or long signal path, the WiMAX system can shift to a lower order modulation scheme with heavier coding to maintain the connection quality and link stability. Table 4.3 shows the SNR threshold values and the recommended modulation and coding rates (CR) required for upstream transmission. We assign a zone number from one to seven to each of the receiver SNR intervals. For example, if a subscriber station experiences a SNR in the interval 5 ~ 8 dB, then that specific subscriber station is said to be in zone two. In each zone i, the maximum possible transmission bit rate per slot is given by R_i, which is calculated as follows. In Chapter 2 we found the number of bits per subcarrier varies with the modulation type used. To obtain the bits per slot values for a specific zone we use the following formula:

$$B_i = N_{data\ subcarriers/slot} \times N_{bits/subcarrier} \times CR \tag{4.11}$$

where B_i is the number of bits per slot, and CR is the coding rate as per Table 4.3. $N_{data\ subcarriers/slot}$ is the number of data subcarriers per slot based on WiMAX standard. $N_{bits/subcarrier}$ is the number of bits modulated per subcarrier and its value depends on the modulation scheme employed as shown in Table 4.3.

To calculate the slot transmission rate for zone *i,* we use the following formula:

$$R_i = \frac{B_i}{Slot\ Time} \tag{4.12}$$

where slot time is 308.571µs.

Zone	Receiver SNR (dB)	Modulation	Coding Rate	Data Sub-Carriers/Slot	Bits/Sub-Carrier	Bits/Slot (Bi)	Slot Time (3*Ts)	Slot Transmission Rate Ri(kbps)
1	5	QPSK	1/2	48	2	48	308.571µs	155.6
2	8		3/4	48	2	72	308.571µs	233.3
3	10.5	16-QAM	1/2	48	4	96	308.571µs	311.1
4	14		3/4	48	4	144	308.571µs	466.6
5	16	64-QAM	1/2	48	6	144	308.571µs	466.6
6	18		2/3	48	6	192	308.571µs	622.2
7	20		3/4	48	6	216	308.571µs	700

Table 4.3: Receiver SNR levels

For example, if a subscriber station is located in zone seven, according to Equation (4.12), the slot transmission rate will be:

$$R_7 = \frac{B_7}{Slot\ Time} = \frac{216\ bits/slot}{308.571\ \mu s} = 700\ kbps \tag{4.13}$$

Similarly, we can obtain the slot transmission rate for all the other zones. The minimum and maximum transmission rates determine what type of services network operators can offer. Also, a WiMAX service provider will need to determine how many service level agreements (SLA) to offer. An SLA is a negotiated agreement between the provider and a customer. The subscription choice typically defines the service level offered (i.e. bit rates) for a given monthly fee. However, some users might try to monopolize the resources above and beyond its agreed share of resources which brings into focus the fairness issue. In this thesis, we will develop an algorithm which guarantees a customer its bit rate based on the SLA and channel conditions, while maintaining fairness among users by placing a cap on the maximum bit rate.

4.4 Channel Model

A significant condition for assessing WiMAX technology is to have an accurate description of the wireless channel. A wireless propagation channel is the medium linking the transmitter and the receiver. In WiMAX, the wireless channel is characterized by Non-Line-of-Sight (NLOS) conditions, meaning that there is no direct or visual line of sight between the transmitting antenna and the receiving antenna. These conditions create a multitude of paths from the transmitter to the receiver, causing the wireless signal to be reflected, diffracted or scattered along the way. The different reflected signal paths arrive at slightly different times, with different amplitudes, and with different phases. These paths superimpose constructively and destructively to cause the received power to go up and down. The variations in loss occur over very small distances (on the order of a wavelength), so it is called fast fading. It was verified, both theoretically and experimentally, that the envelope of a received carrier signal over small distances is Rayleigh distributed [9]. The probability density function of the Rayleigh distribution is given by:

$$f_r(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}, \qquad 0 \leq r \leq \infty \tag{4.14}$$

where σ^2 is the time average power of the received signal. Since the SNR is the received power scaled with the noise power, the probability density function of the SNR according to [8] has an exponential distribution given by:

$$f_\beta(\beta) = \frac{1}{\bar{\beta}} e^{-\frac{\beta}{\bar{\beta}}} \qquad \beta \geq 0 \tag{4.15}$$

where $\bar{\beta}$ is the mean SNR and is a function of distance in our simulation as shown below [35] :

$$\bar{\beta} = \begin{cases} 19 & o\ km < D \leq 2\ km \\ 15 & 2\ km < D \leq 10\ km \\ 7 & 10\ km < D \leq 20\ km \\ 4 & else \end{cases} \tag{4.16}$$

4.5 Transmission Cycle

The base station executes the proposed algorithm to determine which packets the subscriber stations must transmit, and when, from their buffers. A typical transmission cycle begins with the execution of the proposed algorithm in the base station for duration of T_{sch}. After the scheduling is complete, the base station sends Grant messages to the subscriber stations with the downlink data. The subscriber stations receive the Grant messages after a time given by Grant Transmission Time (GTT). Subscriber stations will send their data in parallel during the assigned slot time. Each SS will also send a request message to the BS during its transmission slot.

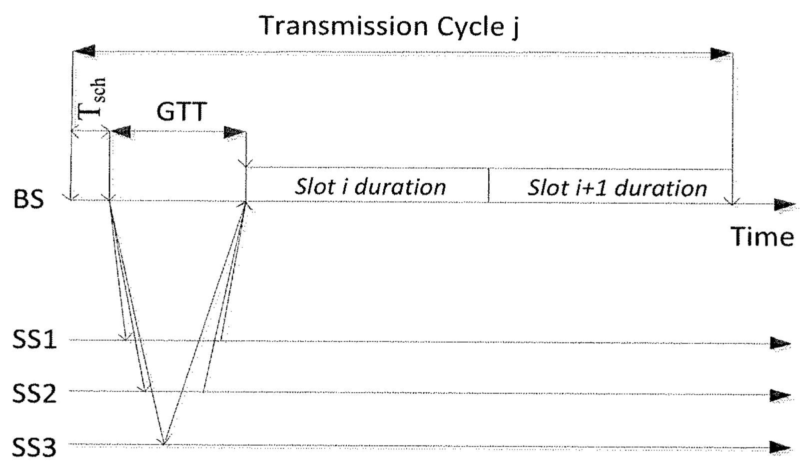

Figure 4.7: Transmission Cycle Example

4.6 Chapter Summary

In this chapter we introduced the basic system model and the architecture of the proposed scheduling algorithms. We saw that in OFDMA, radio resources are two-dimensional regions over time and frequency. This two-dimensional partition gives rise to a slot resource matrix. We completely defined the slot resource matrix in terms of subchannels and OFDMA time symbols. The number of slots in the slot resource matrix represent the maximum number of resources that the scheduler will have available to allocate to contending subscriber stations during each transmission cycle. Based on channel conditions we assigned zones to the WiMAX defined SNR levels and derived the amount of data that each slot can encode. To properly assess the WiMAX technology we have also provided a description of the channel model and transmission cycle.

Chapter 5 Algorithms

In this chapter we introduce two proposed algorithms: Fairness Assured Scheduling Algorithm (FASA) and Fairness Assured Scheduling Algorithm with Quality of Service (FASAQ). In each case, the base station (BS) executes the algorithms to determine how much bandwidth to allocate to contending subscriber stations (SS) during the transmission cycle.

5.1 Fairness Assured Scheduling Algorithm

5.1.1 FASA Scheduler Architecture

Figure 5.1 shows the schematic diagram of FASA implemented in the base station. The base station maintains N credit pools, each used to control the usage of slots by the subscriber stations. The role of the credit pools is to arbitrate the allocation of slots by putting a cap on the amount of data a subscriber station can transmit during each cycle. The credit pool for subscriber station i has two parameters: weight (s_i) and the number of slots (V_i). The value of V_i is the maximum number of slots that a subscriber station can transmit during any particular cycle. At the beginning of every cycle, all the subscriber stations credit pools are initialized with their preconfigured V_i value for the upcoming cycle. A request for slots from a given subscriber station is honoured if there are enough slots in the credit pool. The scheduler allocates grants to subscriber stations using a Round Robin scheme.

5.1.2 Grant Processing

The algorithm executes in two rounds to allocate slots to the subscriber stations. The first round gives rise to the distribution of most of the slots and is mandatory. The second round is designed to redistribute any remaining slots in the credit pool from the first round.

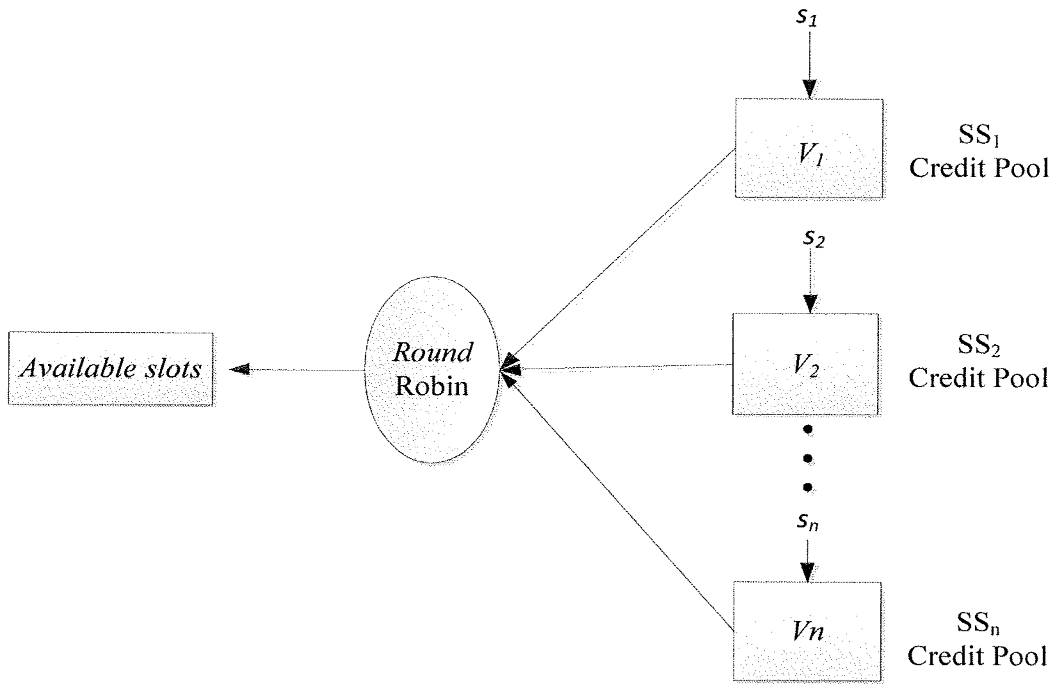

Figure 5.1: Structure of FASA algorithm

Round 1

1. The base station determines the transmission rate of each subscriber station according the following procedure. Based on the received SNR level, the base station chooses the adequate modulation and coding rate schemes required for transmission. The transmission rate is obtained using Equation 4.12 in Section 4.3. Equation 5.1 gives a numerical example for a 64-QAM modulation and 3/4 coding rate corresponding to a SNR of 20 dB or more.

$$R_i = \frac{48 \frac{data\_subcariers}{slot} \times 6 \frac{bits}{data\_subcarrier} \times 3/4}{308.571 \mu s} = 700\ kbps \qquad (5.1)$$

2. The base station converts the bits of data requested by each subscriber station into corresponding slots Q_i according to the following equation

$$Q_i = \left\lceil \frac{q_i}{B_i} \right\rceil \tag{5.2}$$

where q_i is the number of bits awaiting transmission in subscriber station i buffer and B_i is the maximum number of bits per slot that subscriber station i can encode as per Table 4.3.

3. The base station will determine the number of slots V_i that each subscriber station is entitled to receive during a transmission cycle. We want to ensure that the number of slots allocated to each subscriber station translates into equal amount of data transmitted by each subscriber station during the transmission cycle. Recall from Chapter 4 that a zone number is assigned to each subscriber station according to the SNR values experienced. Since zones have different slot transmission rates, any two or more subscriber stations located in different zones, but with the same amount of data to transmit, will require unequal number of slots to transmit the data. For example, a subscriber station lôcated in zone six will have a higher transmission rate than a subscriber station located in zone one. This means that the subscriber station located in zone six will only need a quarter of slots to transmit the same amount of data as a subscriber station located in zone one as per Table 4.3. To account for unequal slot transmission rates, we introduce a normalization parameter α_z, to ensure that all subscriber stations transmit the same amount of data during a transmission cycle:

$$\alpha_z = \frac{R_{low}}{R_z}, \quad R_{low} \leq R_z \tag{5.3}$$

R_{low} is the transmission rate of the lowest zone with an active subscriber station and R_z is the transmission rate of the given zone that we need to normalize. Table 5.1 lists the possible values that the normalization parameter can have.

Zone	Transmission	α_z						
1	155.6	1.000						
2	233.3	0.667	1.000					
3	311.1	0.500	0.750	1.000				
4	466.6	0.333	0.500	0.667	1.000			
5	466.6	0.333	0.500	0.667	1.000	1.000		
6	622.2	0.250	0.375	0.500	0.750	0.750	1.000	
7	700	0.222	0.333	0.444	0.667	0.667	0.889	1.000

Table 5.1: Normalization Parameter Values

The following example illustrates how the table is used. Assume that we have two subscriber stations, SS1 and SS2 transmitting from zones two and six, respectively. Since the lowest zone with an active subscriber station is zone two, it becomes the reference zone and any subscriber station in that zone is assigned a value of one. According to Table 5.1, any subscriber station in zone six will have a value of 0.375. We can now proceed in calculating the number of slots in the credit pool, V_i that the system must make available to each subscriber station at the beginning of each transmission cycle. First, we allocate slots to each subscriber station i according to its weight and normalization parameter α_i using the following formula:

$$V_i = \left(\frac{s_i}{\sum_k s_k} \times N_{total}\right) \times \propto_i \qquad (5.4)$$

$\sum_k s_k$ is the sum of all active subscriber stations' weights, s_i is the weight of an active subscriber station i, α_i is the normalization parameter of subscriber station i according to Table 5.1, and N_{total} is total number of slots. The step is repeated until there are no more N_{total} slots to be allocated.

4. The base station selects the first station SS i stored in the priority database and compares the number of slots demanded Q_i with the number of slots available V_i.
   a. If $Q_i \leq V_i$, then allocate Q_i slots to subscriber station i and subtract them from the total number of available slots, N_{total}. Station is flagged full grant.

$$N_{left} = N_{total} - Q_i \tag{5.5}$$

   b. If $Q_i > V_i$ then allocate V_i slots to subscriber station i and subtract them from the total number of available slots, N_{total}. Station is flagged partial grant and moved to the partial grant set.

$$N_{left} = N_{total} - V_i \tag{5.6}$$

The base station repeats the above steps for each SS in the order stored in the priority database.

Round 2

First, the base station examines the partial grant set. If the partial grant set is empty then finish the allocation. If the partial grant set is not empty, it means there are some SSs that received a partial grant in Round 1. The subscriber station allocates each SS in the partial grant queue a share of the resources proportional to its weight and channel condition. Similar to the first round, the subscriber station takes the first partial grant SS stored in the priority database and compares the number of requested slots to the available slots. This is repeated until there are either no more requests or no more resources available. The base station executes the same algorithm as in Round 1, but this time it only considers the SSs in the partial grant set.

5.2 Fairness Assured Scheduling Algorithm with QoS (FASAQ)

The algorithm introduced in the previous section does not address the issue of Quality of Service (QoS). In this section we build upon FASA to ensure that QoS capabilities are satisfied and introduce Fairness Assured Algorithm with Quality of Service (FASAQ).

5.2.1 FASAQ Scheduler Architecture

The base station maintains N credit pools, each used to control the usage of slots by a subscriber station. The role of the credit pools is to arbitrate the allocation of slots by putting a cap on the amount of data a subscriber station can transmit during each cycle. The credit pool for subscriber station i has two parameters: weight (s_i) and the number of slots (V_i). The value of V_i determines the maximum number of slots that a subscriber station can transmit during any particular cycle. At the beginning of every cycle, all the subscriber station credit pools are initialized with their preconfigured V_i value for the upcoming cycle. The base station also maintains M credit pools, where M is the number of classes of service (CoS) supported by the system. Each CoS credit pool is used to enforce a long-term average rate of CoS traffic transmitted from the subscriber stations. The credit pool has two parameters: weight (w_j) and the number of slots (U_j). Similar to the user credit pools, the CoS credit pools are initialized to their preconfigured values at the beginning of each cycle. A request from a given subscriber station for a CoS j is honoured if there are sufficient slots in both the CoS and subscriber station credit pools.

A fair distribution of CoS j credits between subscriber stations is established through the assignment of weights μ_{ij}. To enforce the weights, we employ an arbiter called Weighted SS Scheduling Order (WSSO) [24]. The inputs to the WSSO arbiter are: the number of CoS j bytes scheduled to date to every SS i (g_{ij}), the total number of CoS j

bytes allocated so far to all SSs (g_j) and the weight μ_{ij}. Ideally, we want the ratio $\frac{g_{ij}}{g_j}$ to be μ_{ij}. Any deviation from this ratio is defined as

$$d_j[i] = \frac{\frac{g_{ij}}{g_j} - \mu_{ij}}{\mu_{ij}} \tag{5.7}$$

The arbiter uses the deviation formula to assign priorities to every subscriber station, such that if $d_j[i] < d_j[k]$, then SS i will be considered for CoS j grants before SS k.

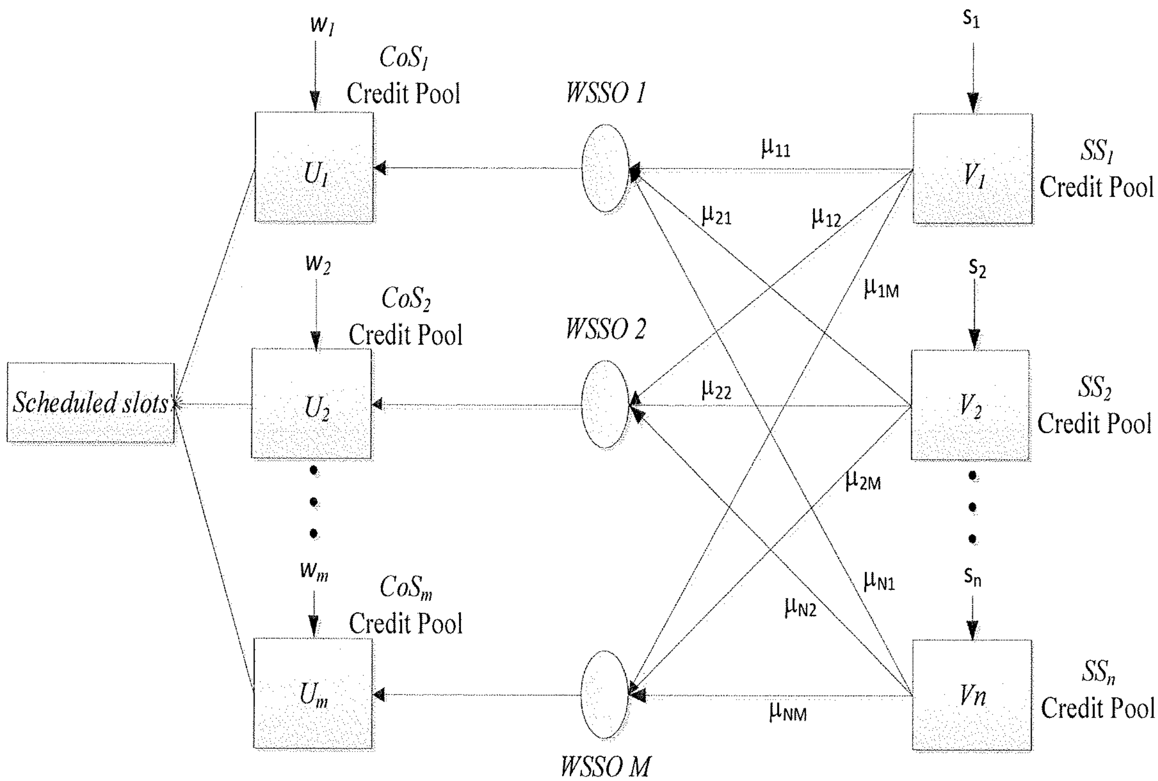

Figure 5.2: Structure of FASAQ Algorithm

5.2.2 Grant Processing

The algorithm executes two rounds to allocate grants to the subscriber stations. The first round gives rise to the distribution of most of the slots and is mandatory. The second round is designed to redistribute any remaining slots in the credit pool from the first round.

Round 1

1. The base station will determine CoS credit pool size U_j and subscriber station credit pool size V_i. The subscriber stations' credit pools V_i are initialized using the procedure outlined for FASA algorithm in previous section. First, we allocate slots to each subscriber station i according to its weight and normalization parameter α_i using the following formula:

$$V_i = \left(\frac{s_i}{\sum_j s_j} \times N_{total}\right) \times \propto_i \tag{5.8}$$

$\sum_j s_j$ is the sum of all active subscriber stations' weights, s_i is the weight of an active subscriber station i, α_i is the normalization parameter of subscriber station i according to Table 5.1, and N_{total} is total number of slots. The step is repeated until there are no more N_{total} slots to be allocated. The CoS credit pools are initialized using the following formula:

$$U_j = \left\lfloor \frac{w_j}{\sum_k w_k} \times N_{total} \right\rfloor \tag{5.9}$$

where N_{total} is the total number of slots available during the transmission cycle, and $\sum_k w_k$ is the sum of CoSs' weight.

2. The base station determines the transmission rate of each subscriber station according the following procedure. Based on the received SNR level, the base

station chooses the adequate modulation and coding rate schemes required for transmission. The transmission rate is obtained using the Equation 4.12 in Section 4.3. Equation 5.12 gives a numerical example for a 64-QAM modulation and 3/4 coding rate corresponding to a SNR of 20 dB or more.

$$\mathrm{R_i} = \frac{48\,\frac{\text{data\_subcariers}}{\text{slot}} \times 6\,\frac{\text{bits}}{\text{data\_subcarrier}} \times 3/4}{308.571\mu\text{s}} = 700 \text{ kbps} \qquad (5.10)$$

3. The base station converts the number of bits of data requested by each subscriber station i for each CoS j, called q_{ij}, into corresponding slots Q_{ij} according to the following equation

$$Q_{ij} = \left\lceil \frac{q_{ij}}{B_i} \right\rceil \qquad (5.11)$$

where q_{ij} is the number of bits requested by station i for CoS j and B_i is the maximum number of bits per slot that station i can encode as per Table 4.3.

4. The base station begins allocating grants as follows. First, grants are processed for CoS 1 for all subscriber stations. Once all CoS 1 grants have been assigned, the base station will process CoS 2 grants. In general, the BS will begin processing CoS j grants after all CoS $(j-1)$ grants have been calculated. The base station takes the first subscriber station stored in the priority queue, SS i, and compares the requested number of slots from this user, Q_{ij} with both V_i and U_j. If there are enough slots available in both the V_i and U_j credit pools, then the SS i will get a full grant and the number of slots allocated are subtracted from each V_i and U_j credit pools. If there are not enough slots in either V_i or U_j credit pools, then the SS i will get the available slots (if any) from the smaller credit pool. The BS will put the SS i into the partial grant set.

Round 2

First, the base station examines the total number of credits in all CoS credit pools. If the total number of credits is zero then finish the allocation. If the total number of credits is not zero, it means that there are some credits still remaining in the CoS credit pools from Round 1. The base station distributes the unused credits between subscriber stations according to their weights w_j:

$$U_j = \begin{cases} \left\lfloor \frac{w_j}{\sum_k w_k} \times N_{unused} \right\rfloor & \mathrm{j} = 1 \\ \left\lfloor \frac{w_j}{\sum_k w_k} \times N_{unused} \right\rfloor + u_{j-1} & \mathrm{j} > 1 \end{cases} \tag{5.12}$$

The first term in the equation represents the fair share of the remaining credits (from Round 1) that will be given to each CoS in the second round. While CoS1 gets its fair share only, a lower priority CoS *j* (*j>1*) gets its fair share plus extra u_{j-1}credits. The u_{j-1} is the total credits that will remain unused in CoS(j-1) credit pool when the processing of grants for that CoS is completed in the second round.

The base station executes the same algorithm as in Round 1, but this time only subscriber stations that have not received full grant will be considered. Like in Round 1, CoS1 grants are processed first, followed by other CoSs in sequence.

5.3 Chapter Summary

In this chapter we have described two proposed algorithms for the uplink traffic: Fairness Assured Scheduling Algorithm (FASA) and Fairness Assured Scheduling Algorithm with Quality of Service (FASAQ). In each case, the base station executes the two algorithms to determine the transmission order of the subscriber stations, while maintaining fairness and QoS capabilities.

Chapter 6 Simulation Results

We have developed an in-house C++ simulation program to test the proposed bandwidth allocation schemes. In this chapter, we will introduce the traffic models for different classes of service and the simulation parameters. We will evaluate and compare the proposed algorithms using metrics such as packet delay, network throughput and packet loss ratio.

6.1 Simulation Setup and Assumptions

We have assumed that the base station is servicing ten subscriber stations, each randomly chosen at distance between 0 and 30 km. During any transmission cycle, the subscriber stations are assumed to be stationary. The network overall capacity is C_p = 66.3 Mbps and user access speed is 10 Mbps. These values are consistent with existing devices capabilities and SLA rates offered by WiMAX operators [38]. The maximum transmission cycle or scheduling interval T_{max} is set to 5 ms and the scheduling time at the BS is T_{sch} = 25 μs.

To obtain an accurate and realistic performance analysis, it is important to simulate the system behaviour with appropriate traffic injected into the network. In the simulation, all subscriber stations (SSs) have identical traffic parameters and offered loads. We have merged the five WiMAX classes of service into three classes: CoS1, CoS2, and CoS3. These new classes of service can be mapped to standardized classes defined by the Differentiated Services architecture [36]. In this thesis, CoS1 has been mapped to Expedited Forwarding (EF), which provides for low loss and delay-sensitive services; CoS2 has been mapped to Assured Forwarding (AF), which provides for low loss and non-delay sensitive service; and CoS3 has been mapped to Best Effort (BE), which does not require any bandwidth guarantee from the network. The three classes share a common buffer space F = 5 Mbytes at each SS.

Traffic for CoS1 is simulated by constant bit rate (CBR) VoIP traffic based on the G.711 standard [37]. The VoIP stream has a data rate r_{CBR} = 66 packets/s and packet length L_{CBR} = 120 bytes for a traffic rate of 63.36 kbps. The amount of CBR traffic is kept constant for all simulations.

The offered load at each subscriber station is varied by changing the rate with which the traffic is generated for CoS2 and CoS3 by using a self-similar traffic model. Studies have shown that network traffic is self-similar and long range dependent [29]. A self-similar process is a process that behaves the same when viewed at different degrees of magnification. In the context of network traffic, self-similarity means that the traffic has similar statistical properties over a range of timescales: milliseconds, seconds, minutes, hours and even days. Statistically merging different data streams, does not result in smoothing of the traffic. For example, bursty data stream that are multiplexed tend to produce bursty aggregate stream. Traffic that exhibits this type of behaviour is said to have long range dependence, which means that the correlation function of the traffic process is a heavy-tailed distribution [23].

One of the simplest heavy-tailed distributions is the Pareto distribution. To generate self-similar traffic, we used the method described in [29], where the resulting traffic is an aggregation of multiple streams, each consisting of alternating Pareto-distributed ON/OFF periods. ON periods correspond to packet trains, packets transmitted back to back or separated only by a relatively small preamble. OFF periods are the periods of silence between packet trains. Pareto distribution is a heavy-tailed distribution with the probability density function (PDF):

$$f_x(x) = \frac{kb^k}{x^{k+1}},\ x \geq b, and\ k, b\ \geq 0 \qquad (6.1)$$

where k is the shape parameter and b is the location parameter. The shape parameter and the location parameter are both positive. Pareto distribution with $1 < k < 2$ has finite mean and infinite variance. In our simulation, the shape parameter for the ON and OFF intervals is set to 1.4 and 1.2, respectively. The choice of the shape parameter was prompted by measurements on network traffic performed in previous studies [30].

We assume that the amount of traffic data generated for CoS3 is twice that of Cos2. The offered load at each subscriber station is varied from 0.1 to 1.0 in different simulations. To achieve the desired offered load at the subscriber station, the average arrival rate of each ON-OFF source is varied in the interval [0.31, 3.31] Mbps for CoS2 and in the interval [0.62, 6.62] Mbps for CoS3. The mean sojourn time of the ON state is set to 50 ms. The mean sojourn time of the OFF state is a function of the subscriber station offered load and the mean sojourn time of the ON state [24]. The average rate of arrival r_i for each CoS is obtained as follows:

$$CoS1 = 0.0636\ Mbps \tag{6.2}$$

$$CoS2 = (\rho_s - 0.0636) \times \frac{1}{3}\ Mbps \tag{6.3}$$

$$CoS3 = (\rho_s - 0.0636) \times \frac{2}{3}\ Mbps \tag{6.4}$$

where ρ_s is the subscriber station load and is varied from 1Mbps to 10 Mbps, which corresponds to 10% and 100% subscriber station loading, respectively.

The offered load of 6.63 Mbps (66.3%) at each subscriber station corresponds to 100% overall load on the WiMAX network. The overall network load, ρ, is found using the following formula

$$\rho = \frac{N \times \sum_{i=1}^{3} r_i}{C_p} \tag{6.5}$$

where ρ is the network load, N is the number of users, r_i is the average arrival rate of CoS i traffic and C_p is the maximum network capacity. For CoS2 and CoS3, the length of packets generated during an ON state follows the tri-modal distribution used in [23]. These three modes correspond to the most frequent packet sizes 64, 594, and 1518 bytes. In our simulation, each of these packets is generated with a frequency of 62%, 10% and 28%, respectively.

We made the following assumption regarding the simulation parameters for CoS credit pools employed by FASAQ. The CBR packets must always conform to the specific profile (w_1, U_1) for CoS1 credit pool. This means that the size of the credit pool U_1 must be sufficiently large to accommodate all requests received for CoS1 grants. We obtain the value of U_1 as follows

$$U_1 = N \times \lceil r_{CBR} \times T_{max} \rceil \times (L_{CBR} + L_{overhead}) \tag{6.6}$$

where N is the number of subscriber stations, $\lceil r_{CBR} \times T_{max} \rceil$ is the number of CBR packets arriving at the each subscriber station during a maximum cycle and the sum in the brackets is the size of CBR packet including overhead. To encapsulate VoIP packets Real-Time Transport Protocol (RTP) is often used and then transmitted over UDP/IP. The RTP and UDP headers consume 12 and 8 bytes, respectively. The IPv4 and IPv6 headers consume 20 and 40 bytes, respectively. In other words the total header overhead is 40 or 60 bytes, depending on whether IPv4 or IPv6 is used. However, using either VoIP aggregation or Robust Header Compression (RHOC) the value of the overhead can be reduced significantly [31]. Based on the two header reduction methods mentioned, we chose $L_{overhead}$ to be 20 bytes in our simulation. The value of U_1 will then be 462 bytes or (462 bytes×8bits/byte)/48 bits/slot = 77 slots. We used the value of 48bits/slot to account for the case when all subscriber stations are experiencing worst channel conditions. Since CoS3 serves best effort data, no bandwidth has been explicitly reserved for it. All the remaining bandwidth is allocated to CoS2 traffic. The best effort traffic will be able to make use of the unused bandwidth in the second round of the algorithm.

We assumed that all the SSs share the WiMAX resources equally. We set the weight of every subscriber station in the WSSO to be the same value $\mu_{ij} = \frac{1}{N}$ in the case of

FASAQ, where N is the number of subscriber stations. In this thesis, packet delay is the time a packet spends in the network from the instant that is generated at the subscriber station until it is delivered to the base station.

The packet loss ratio is given by

$$paket\ loss\ ratio = \frac{number\ of\ CoS\ packets\ dropped}{total\ CoS\ packets\ generated}$$

Network throughput is given by

$$throughput = \frac{total\ amount\ of\ data\ received\ at\ the\ base\ station}{simulation\ time}$$

6.2 Results and Analysis

6.2.1 Performance evaluation of FASA Algorithm

We start our numerical analysis by investigating network throughput of the FASA algorithm. In Figure 6.1 we present the overall network throughput as a function of the subscriber station offered load. Remember that a subscriber station loading of 66.3% corresponds to an overall network loading of 100% as shown by the two horizontal axes. We show two curves on the graph. The bottom curve represents the network throughput with the users randomly distributed around the base station at different distances ranging from 0 to 30 km and experiencing different channel conditions. The top curve represents network throughput where all the users are placed near the base station and they all experience optimal channel conditions and maximum transmission rates.

As expected the network throughput increases as the network loading increases until the network becomes fully loaded. Once the network is fully loaded maximum capacity is reached and the graph starts levelling off. Although the overall throughput of the bottom curve seems low, this is justifiable as follows. At any one time, the users are randomly distributed around the base station, which results in different transmission rates. Some users might experience optimal channel conditions, allowing them to transmit at the maximum allowable transmission rates. At the same time, some users might experience poor channel conditions which result in low transmission rates. However, collectively users experience conditions somewhere in between the two extremes. In our case, the bottom curve shows the overall throughput below 50%, which means that, on average; users are located somewhere between zones three and four.

As mentioned, the top curve represents the case in which all users experience optimal channel conditions. We intentionally placed all the users in zone seven (ie best channel conditions) to test the validity of our algorithm. In theory, the overall network throughput should be 66.3 Mbps. Our experimental results, as represented by the top curve, show that we are achieving a throughput close to but slightly below the maximum network throughput. The discrepancy can be explained as follows. Remember that the minimum resource allocation unit is a slot, which can contain anywhere between 48 and

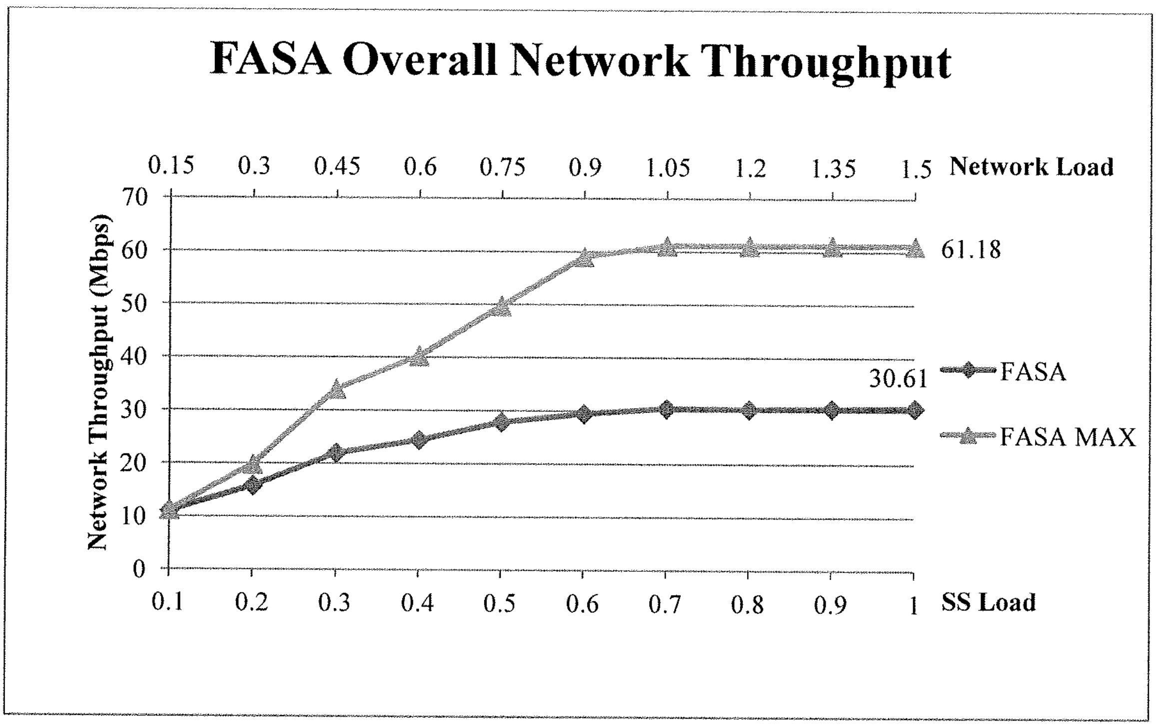

Figure 6.1: FASA Network Throughput

216 bits, depending on the channel conditions. If the subscriber stations want to transmit bits that are not exact multiples of the above mentioned numbers, then some transmitted slots may not be full. This results in what is known as "slot packing problem", which explains the slightly lower network throughput achieved.

Let us turn our attention to the average delay experienced by a packet in the network with FASA scheme, as shown in Figure 6.2. In FASA, the average delay approaches 120ms which can be explained as follows. Recall that during the source ON interval, the traffic is being generated at a rate of 10 Mbps for an average duration of 50 ms, which results in the total of 62500 bytes. The number of transmission cycles needed to transmit this data to the base station will be

$$\frac{62500\, bytes}{(B_{median} \times 1/8 \times N_{total})/N} = \frac{62500\, bytes}{2765\, bytes} = 23 \text{ cycles} \tag{6.7}$$

where B_{median} is the number of bits per slot, N is the number of subscriber stations and N_{total} is the total number of slots. The denominator indicates the maximum number of slots that can be given to a subscriber station during a cycle. We used a value of 144 bits for B_{median} since, as we already discussed, the subscriber stations will be located somewhere in the middle zone. When the network load becomes fully loaded, the transmission time to transmit the above data will be 5 ms times 23 cycles for a total of 115 ms. This value is close to our simulated value of 121 ms. As the network becomes fully loaded the curve starts flattening, meaning that any incoming packets will be dropped if the user buffer is full and will not contribute to the average delay statistics.

Next we analyze the per user throughput as shown in Figure 6.3 for FASA scheme. Each bar represents the throughput of one subscriber station. Remember that due to location and channel conditions, an equal distribution of slots among users does not translate into equal user throughput.

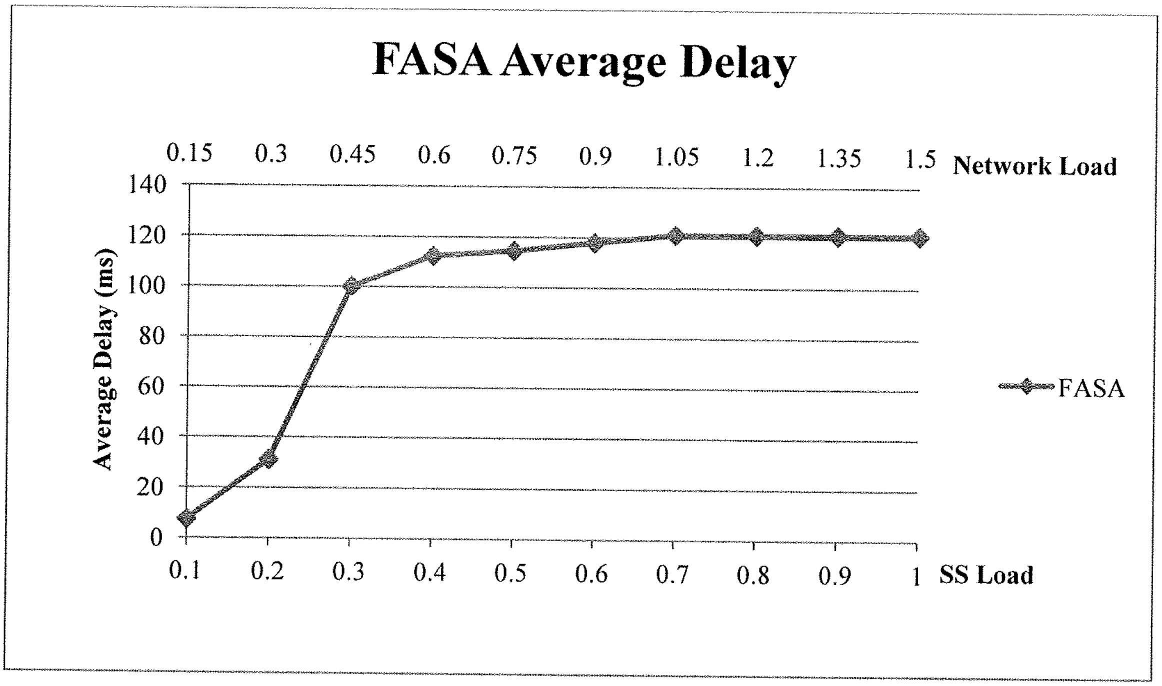

Figure 6.2: FASA Average Delay

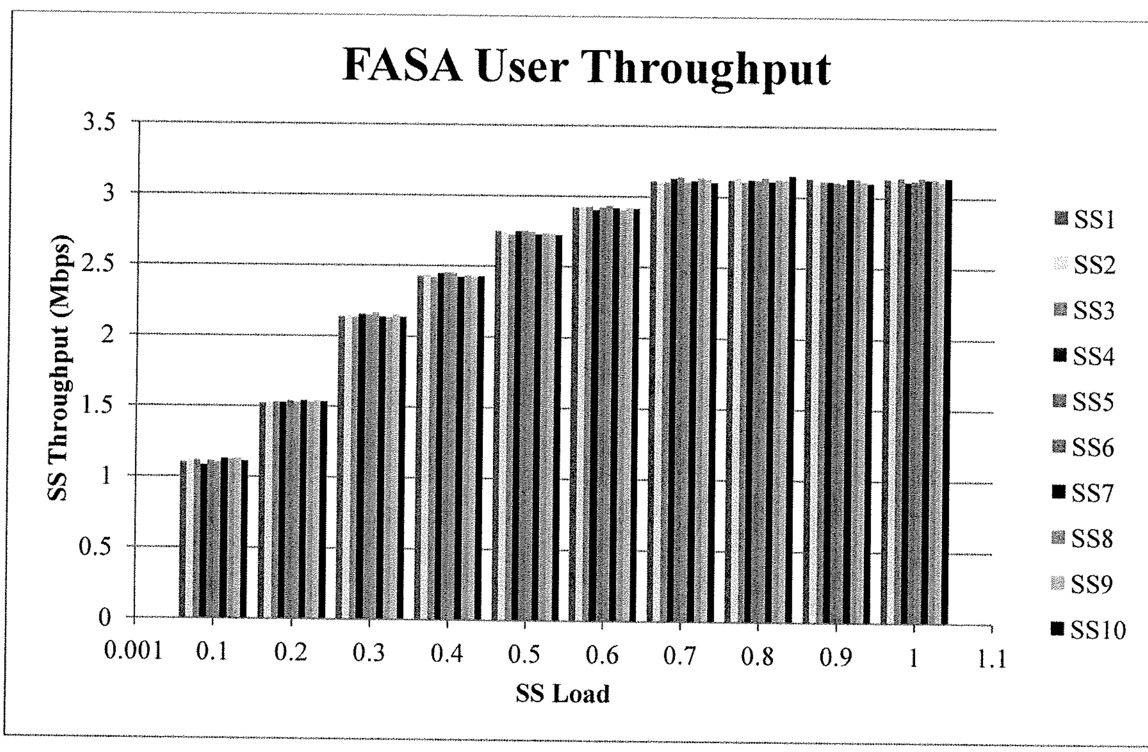

Figure 6.3: FASA User Throughput

FASA is able to ensure equitable user throughput by employing the normalization parameter, which results in different amount of slot grants to subscriber stations based on channel conditions. As we can see in Figure 6.3, subscriber stations are experiencing similar throughput which is what we expected.

6.2.2 Performance Evaluation of FASAQ Algorithm

We now analyze the performance of the FASAQ algorithm which adds Quality of Service (QoS) to the previous algorithm. We are comparing the performance of the FASAQ algorithm with the same algorithm but when the zoning concept is not used (i.e. turned off). We call this algorithm Quality of Service Assured Algorithm (QASA). Recall that in the case of FASA and FASAQ, user credit pools are filled with slots and the size of the credit pool varies according to channel conditions. In the case of QASA, the total amount of available slots, N_{total} are distributed equally among subscriber stations without regard to channel conditions. As we already mentioned, users experience different channel conditions, and hence different transmission rates. As a result, equal amount of slots does not translate into equal subscriber station throughput. Thus, we expect that with QASA, subscriber stations experiencing poor channel conditions will have a lower throughput than subscriber stations experiencing good channel conditions. Figure 6.4 shows the overall network throughput as a function of network loading for both FASAQ and QASA. Comparing the network throughput of FASAQ with the network throughput of FASA of Figure 6.1, we see that adding QoS does not change the overall network throughput. This is understandable since QoS is only used to guarantee a certain level of performance to a data flow. Both FASAQ and QASA achieve their maximum throughput once the network becomes fully loaded. Since users are randomly distributed around the base station, they will not be able to transmit at maximum theoretical rates. This explains why our simulated throughput is less than the maximum theoretical throughput. In Figure 6.4, we see that the FASAQ network throughput is higher than the QASA network throughput. This reason is that QASA is blind to channel conditions when it allocates slots.

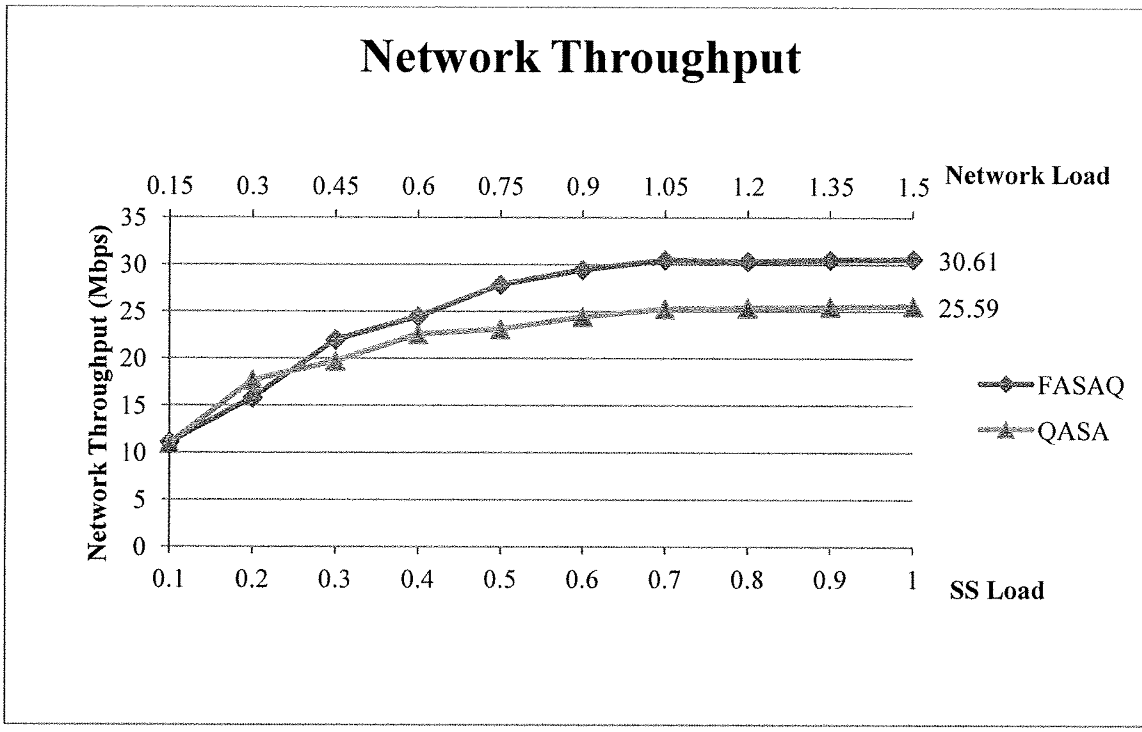

Figure 6.4: FASAQ and QASA Network Throughput

We now want to see if the requirements of different classes of service are being satisfied by the FASAQ algorithm. We begin by looking at average delay as shown in Figure 6.5. We start with the average delay for CoS1 data. For both FASAQ and QASA CoS1 delay increases very little as the offered load is increased. When a CoS1 packet arrives in the user buffer, it has to wait on average half a cycle to be polled and one cycle to receive a grant. Since the transmission cycle time is 5 ms, the average delay experience by CoS1 packets is 7.5 ms.

The 143 ms average delay for CoS2 at heavy loads can be explained as follows. The average ON time for CoS2 source is 50 ms, and the data rate at each subscriber station is 10 Mbps. Thus, the total number of bytes generated is equal to 62500 bytes. We use the following equation to calculate the number of cycles needed to transmit this data to the base station:

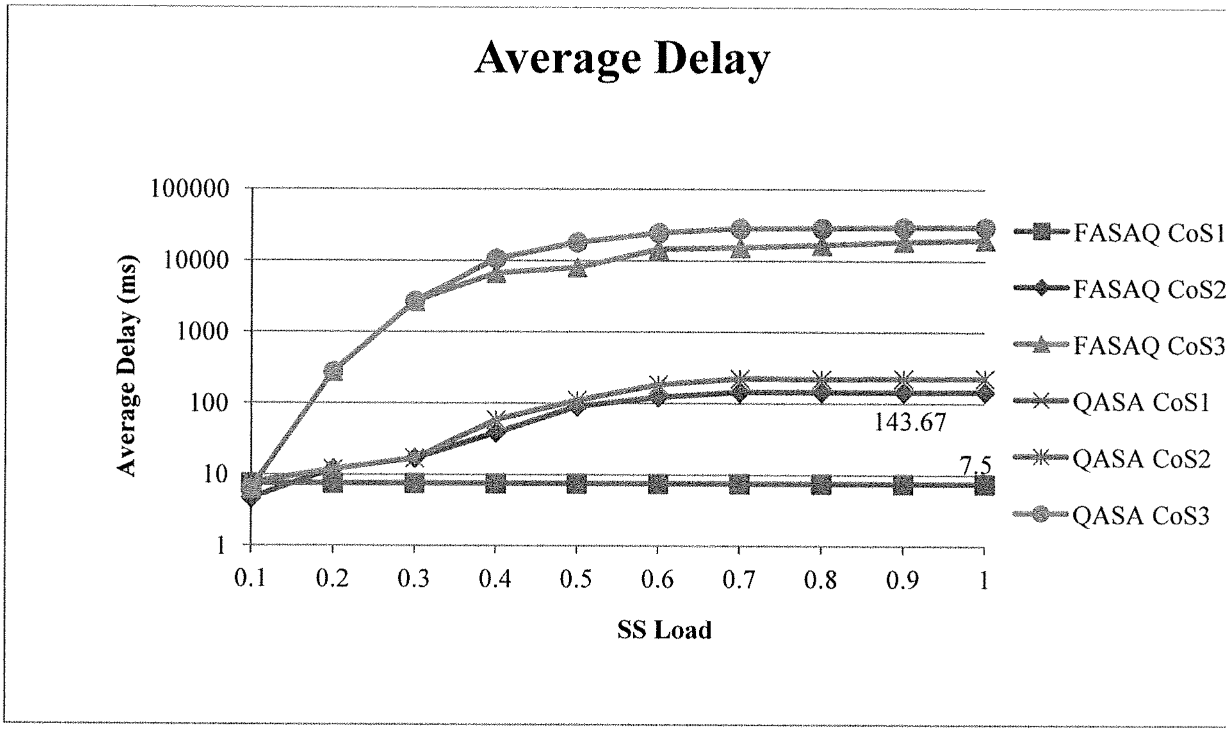

Figure 6.5: FASAQ and QASA Average Delay

$$\frac{62500\, bytes}{(B_{median}\times 1/8\times N_{total})/N - U_1/N} = \frac{62500\, bytes}{2719\, bytes} = 23\ cycles \qquad (6.8)$$

where N is the number of subscriber stations, U_1 is the size of CoS1 credit pool in bytes, B_{median} is the number of bits per slot, and N_{total} is the total number of slots. The denominator of the fraction indicates the maximum number of grants that can be given to CoS2 traffic during a transmission cycle. We subtract U_1/N bytes from the total available grants for CoS2 since we need to take into account the grants allocated to CoS1. The total time needed to transmit the data will be 5 ms times 23 cycles for a total of 115 ms, which is close to the simulated values. QASA experiences slightly higher average delay performance. This is understandable since users that are in poor channel conditions will get the same amount of slots as users in good channel conditions. This means that the users in poor channel conditions will have to wait longer to transmit their packets which contributes to a higher delay. In both cases the CoS3 delay is higher than

CoS2 delay, since a packet of CoS3 type has lower priority compared to a CoS2 meaning it has to wait longer time in the user buffer before being transmitted.

The next metric to be analyzed is the maximum packet delay as shown in Figure 6.6. In both FASAQ and QASA, CoS1 maximum delay settles at about 10 ms. This corresponds to twice the cycle time since a packet that arrives in the user buffer immediately after the request message has been sent has to wait an entire cycle to be pooled and another cycle to receive a grant. Upon arriving at the subscriber station's buffer, CoS2 packets, in the worst case scenario, will encounter a full buffer. The number of cycles a CoS2 packet has to wait in the buffer is given by:

$$\frac{5\ Mbytes}{(B_{median}\times 1/8\times N_{total})/N-U_1/N}=\frac{5000000\ bytes}{1873\ bytes}=1839\ cycles \qquad (6.9)$$

where N is the number of subscriber stations, U_1 is the size of CoS1 credit pool in bytes, B_{median} is the number of bits per slot, and N_{total} is the total number of slots. This is equivalent to 1839 cycles times 5 ms for a total of 9195 ms. Our simulated results of 11463 ms are comparable in value with the calculated results.

We proceed in analysing the average CoS throughput as shown in Figure 6.7. Since CoS1 traffic remains unchanged, the output remains constant as the network load is varied. For CoS1 traffic, we had to take into account CBR packet overhead, which results in a CoS1 traffic rate greater than 63.36 kbps as reflected in Figure 6.7. Since Cos2 traffic has higher priority compared to CoS3 traffic, at any given time more CoS2 traffic should get transmitted compared to CoS3 traffic. As the network load increases CoS3 traffic has less chance of getting transmitted. This is evident from the graphs of CoS2 and CoS3, which shows that at a network load of 0.5 the buffer start filling up. Once the network is fully loaded CoS2 queues in all subscriber stations are constantly full which means they will have priority over CoS3 traffic resulting in deterioration of CoS3 traffic rates as seen in the graph. Also, average CoS throughput is lower in the case of QASA. Recall that QASA algorithm does not use the normalization constant to take into account channel conditions. Subscriber stations are getting equal number of

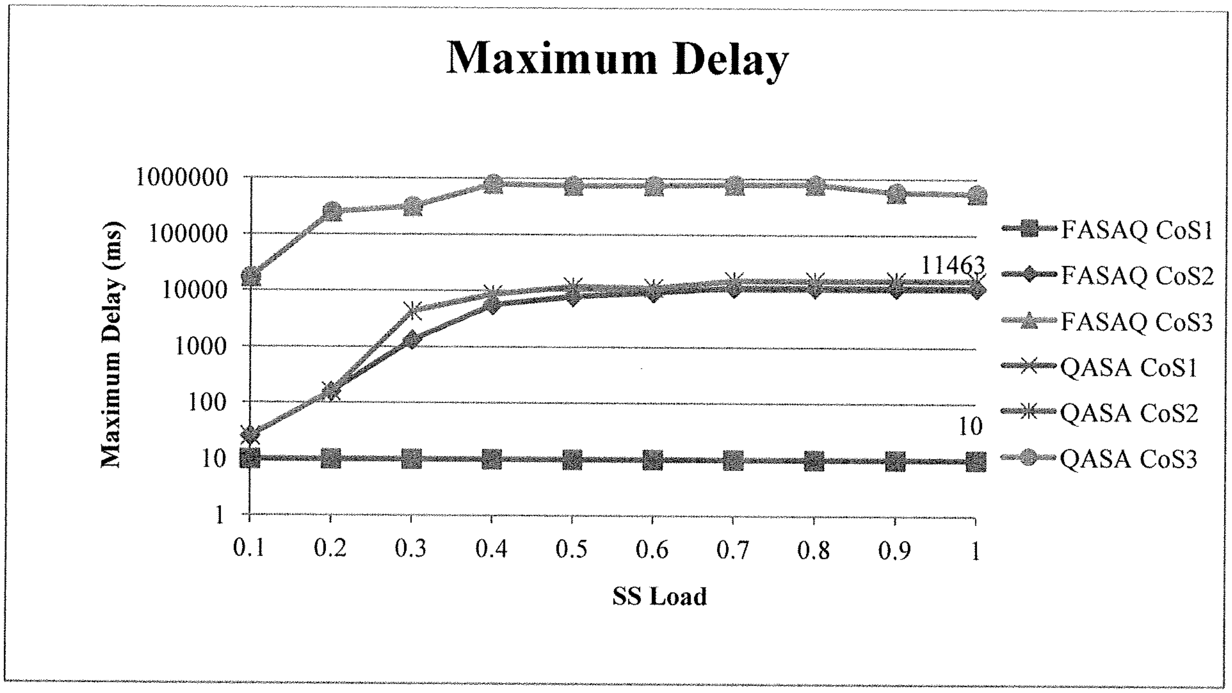

Figure 6.6: FASAQ and QASA Maximum Delay

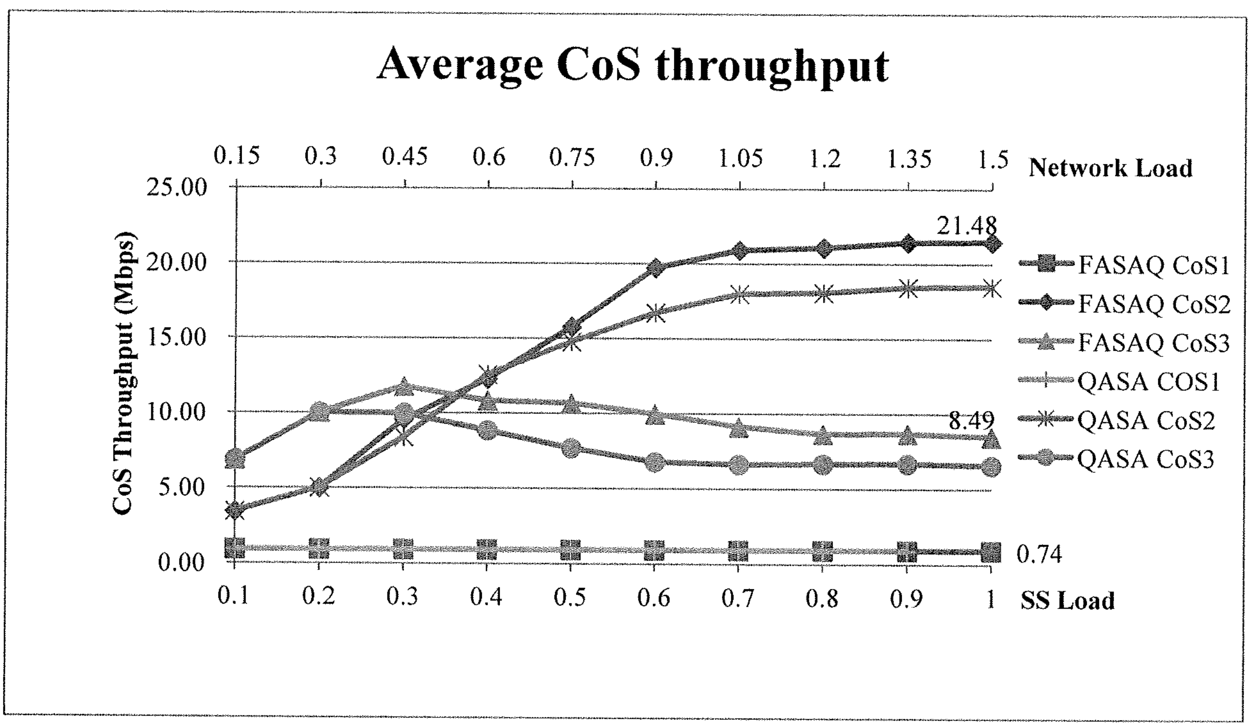

Figure 6.7: FASAQ and QASA Average CoS Throughput

slots regardless of channel conditions. This means that the amount of data sent will be lower than in the case of FASAQ, in which case subscriber stations with poor channel conditions will get more slots resulting in higher throughput. Figure 6.8 shows the average packet loss ratio for the three CoSs as a function of network load. The CoS1 packet loss ratio remains zero throughout the simulation, which means that VoIP traffic will not be degraded by the presence of other CoS traffic. For both FASAQ and QASA the packet loss ratio tends to level off once the network is fully loaded. At high load, the number of packets generated is greater than the number of packets transmitted, thus the buffer gets full on a regular basis. This causes the incoming packets to be dropped. In the case of QASA, users with poor channel conditions are not getting slots according to the channel conditions. Since these users experience lower transmission rates, their queues tend to fill up faster than in the FASAQ case causing more packets to be dropped.

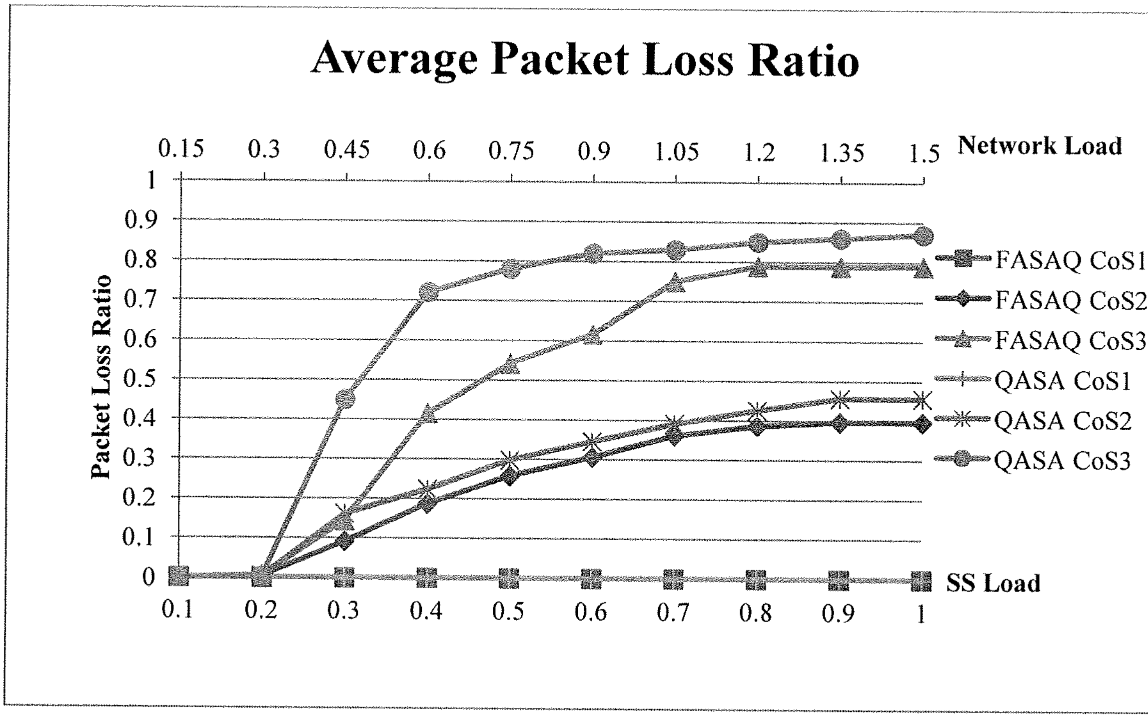

Figure 6.8: FASAQ and QASA Average Packet Loss Ratio

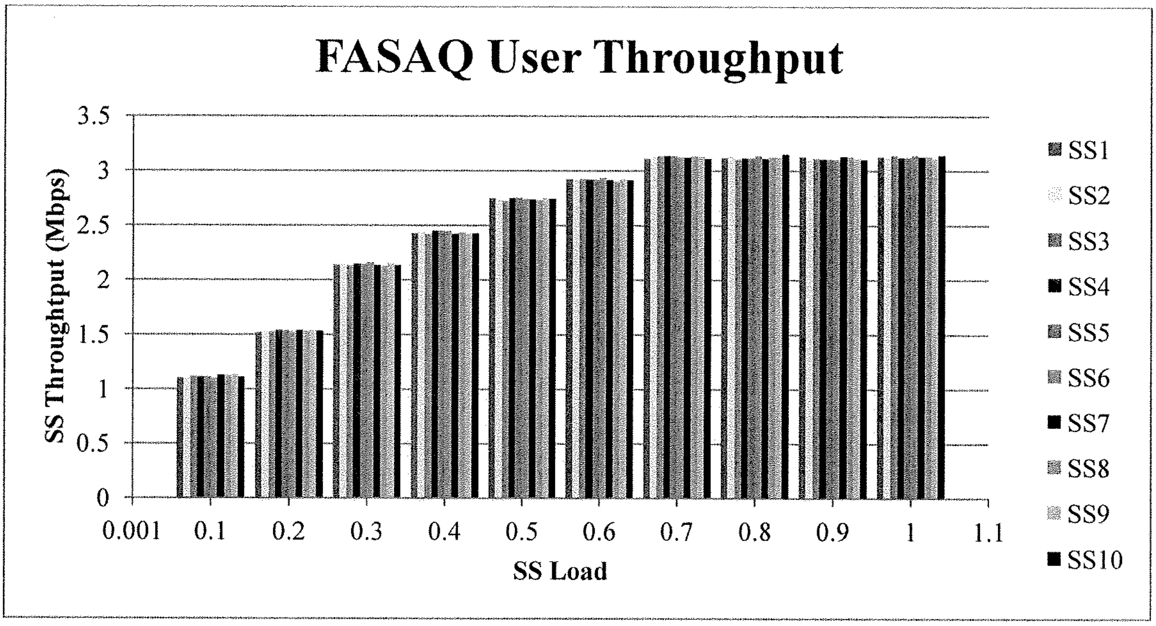

Figure 6.9: FASAQ User Throughput

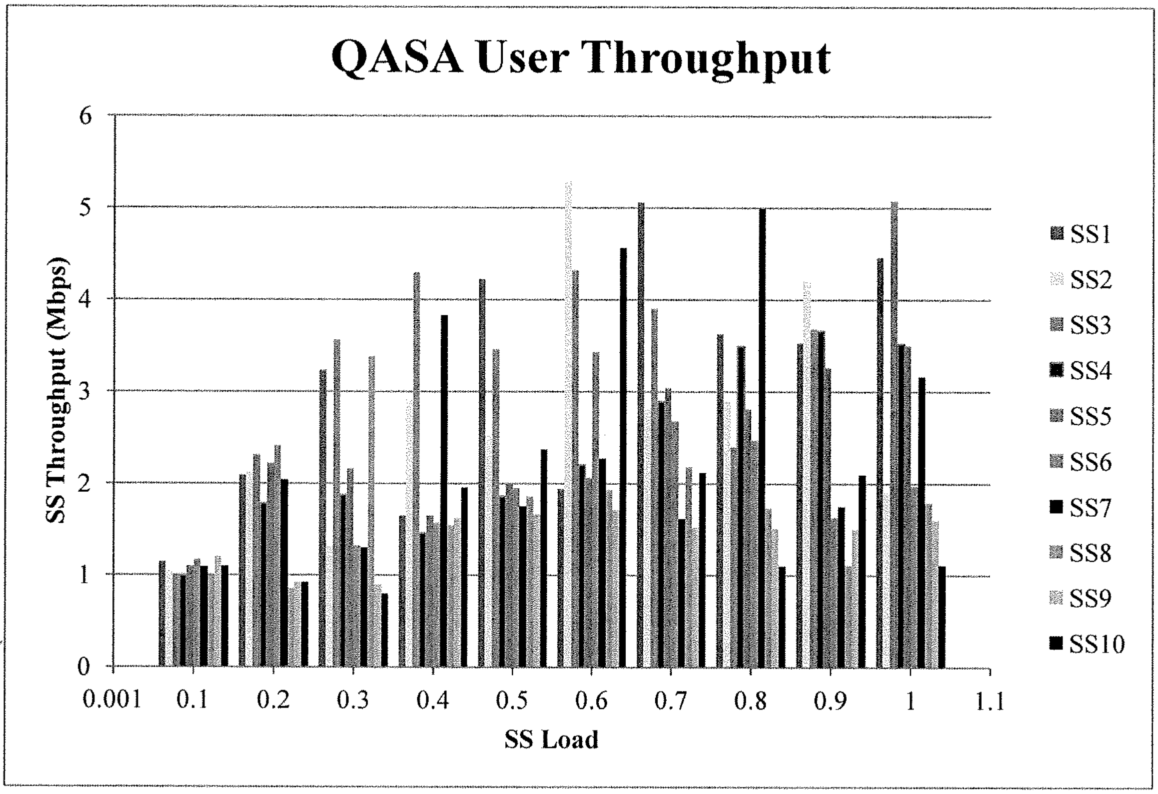

Figure 6.10: QASA User Throughput

Figure 6.9 and Figure 6.10 show the individual user throughput. Recall that QASA assigns equal number of slots to users. This means that users with poor channel conditions will not get the same transmission rates as users with good channel conditions. As we can see in Figure 6.10, users do not get a fair share of the transmission bandwidth in the case of QASA. In the case of FASAQ users are assigned slots depending on the channel conditions. This means users get assigned different number of slots which ensures a fair distribution of bandwidth among users, as evident in Figure 6.9.

6.3 Chapter Summary

In this chapter we evaluated the performance of the proposed algorithms. We presented the simulation environment and the values for the simulation parameters, including any assumption made. We have used metrics such as packet delay, network throughput and packet loss ratio to evaluate and compare the proposed algorithms.

Chapter 7 Conclusion and Future Work

In recent years, ever increasing user demands for high data rates, mobility, and connectivity has led to the development of new wireless broadband technologies. In this thesis, we have introduced WiMAX technology, which is emerging as the leading wireless access technology that provides cost-competitive, ubiquitous broadband wireless access and Quality of Service capabilities. We introduced the Physical and Medium Access Control layer protocols as specified by the WiMAX standard with emphasis on concepts applicable to this thesis. Theoretical background on the modulation and multiplexing schemes applicable to WiMAX was provided. The dominant access technology used in WiMAX is OFDMA, which is a powerful multi-user access technique. WiMAX defines a *slot* as the smallest unit of Physical layer resource and is represented in two dimensions: on in time (time symbols) and one in frequency (subcarriers). We have constructed a slot resource matrix to represent the total available resources to be distributed to subscriber stations.

Due to the multipoint-to-point architecture of WiMAX in the uplink direction, subscriber stations share the wireless resources and capacity. This may result in collisions of data signals from different subscriber stations as they try to access the shared wireless channel. Since there is no direct communications between the subscriber stations, a Medium Access Control mechanism is required to mediate access between contending subscriber stations. Also, the Medium Access Control mechanism has to ensure that users' Quality of Service requirements are satisfied. WiMAX specifies five classes of service, but the scheduling mechanism for these classes is left open to vendor and network provider implementation.

The first algorithm, called Fairness Assured Scheduling Algorithm (FASA), employs a credit pooling technique to allocate resource blocks among users in a fair manner based on the channel conditions. Channel conditions are conveyed by the use of subcarriers called *pilots* and determine the modulation and coding rates employed during the transmission. Based on the modulation and coding rates employed, subscriber stations transmit data at different slot transmission rates. We have defined seven distinct zones characterized by different slot transmission rates. The base station assigns slots to subscriber stations based on their zone number and normalization parameter in such a way to ensure equal data transmission rates among subscriber stations. In this way, FASA guarantees fair distribution of WiMAX resources between subscriber stations based on their service level agreement, regardless of their position or channel conditions. Simulation results showed that FASA was able to ensure almost identical throughput for each subscriber station. FASA overall throughput is less that 50% of the WiMAX overall maximum throughput once the network becomes fully loaded. Due to dissimilar channel conditions, subscriber stations experience different transmission rates, which are lower than the maximum transmission rate. However, once the users are transmitting at maximum rates under the optimal channel conditions, FASA reaches about 93% of the WiMAX maximum theoretical throughput.

The second algorithm, called Fairness Assured with Quality of Service (FASAQ), adds QoS capabilities to FASA. Various classes of service are transmitted based on their QoS requirements, transmission history and their channel conditions. A weighted round robin arbitration mechanism was employed in FASAQ to determine the priority of users by their long term transmission records. Simulation results showed that FASAQ was able to satisfy QoS requirements and guarantee fair distribution of WiMAX resources between subscriber stations regardless of channel conditions. FASAQ also assigns zones to different subscriber stations. Subscriber stations receive slots proportional to their zone number and a normalization parameter to ensure equal data transmission rates among subscriber stations. We compared FASAQ with an algorithm called QASA, which does not take into account channel conditions and hence does not employ the normalization parameter. In the case of QASA, users get an equal amount of slots regardless of channel conditions. Since slots have different transmission rates, this results in different subscriber throughput. Simulation results showed that when zones are not employed, subscriber stations will not get their fair share of the bandwidth as promised.

With respect to our research, in the future we will try to address the slot packing problem since, as seen from the simulation results, it has a negative effect on the overall performance of the system. The slot packing problem arises due to the fact that slots are fixed bit size and data requested from subscriber station has variable bit size. It may happen that data requested from subscriber stations does not completely fill up the slots. In this case, some slots are transmitted at less than full capacity resulting in wasted capacity. The goal of future work is to minimize the number of slots that are not completely filled up. In this thesis the subscriber stations were stationary during the transmission cycle, meaning that their channel conditions were not changing. WiMAX is developing a mobile version of the standard and we will like to evaluate the performance of the algorithms with the subscriber stations moving at high speeds.

Bibliography

[1] International Telecommunication Union, "The world in 2010," 2011, [Online]. Available: http://library.rit.edu/userservices/pubschol/IEEEGuidelines.pdf

[2] Bluetooth Wireless Technology, 2011, [Online]. Available: http://www.bluetooth.org/apps/content/

[3] Matthew Gast, "802.11 Wireless Networks: The Definitive Guide", second ed. New York: O'Reilly Media Inc., 2005.

[4] IEEE Std 802.16-2009, "IEEE Standard for Local and Metropolitan Area Network-Part 15: Air Interface for Broadband Wireless Access Systems", May 2009, [Online]. Available: http://standards.ieee.org/getieee802/download/802.16-2009.pdf

[5] Andrew S. Tanenbaum, "Computer Networks", fourth ed. New York: Prentice Hall, 2002, Ch.1, Sec. 1.2, pp. 19-22.

[6] WiMAX Forum Network Working Group, "WiMAX Forum Network Architecture Stage 2: Architecture Tenets, Reference Model and Reference Points—Release 1, Version 1.2", WiMAX Forum, January 2008. [Online]. Available: http://www.wimaxforum.org/about

[7] WiMAX Forum Network Working Group, "Mobile System Profile, Release 1.0 approved Specification, Revision 1.7.1", July 2008, [Online]. Available: http://www.wimaxforum.org/ .

[8] A.F. Molisch, "Wireless communications", 2^{nd} ed. New York: Wiley, 2011, Ch. 17, Sec. 5, pp. 378-379.

[9] S.C. Yang, "OFDMA System Analysis and Design", Boston, MA: Artech House, 2010, Ch. 3, Sec. 4, pp. 51-52.

[10] A. Kumar, "Mobile Broadcasting with WiMAX: Principles, Technology and Application", Ney York, NY: Focal Press, 2008, Ch. 4, pp. 73-83.

[11] L. Nuaymim, "WiMAX: Technology for Broadband Wireless Access", Hobokne, NJ: Wiley & Sons, 2007, Ch. 5, Sec. 2.

[12] WiMAX Forum, "Mobile WiMAX Part1: Overview and performance Evaluation", August 2006, [Online]. Available: www.wimaxforum.org/.../Mobile WiMAX Part1 Overview and Performance.pdf

[13] G. Nair, J. Chou, T. Madejski, K. Perycz, P. Putzolu and J. Sydir "IEEE 802.16 Medium Access Control and Service Provisioning", *Intel Technology Journal*, vol 08, August 2004, pp 213-228.

[14] Y. Xiao, "WiMAX/MobileFi: advanced research and technology", Boca Raton, FL: Auerbach Publications, 2008.

[15] A. H. Rashwan, H. M. ElBadawy, H. Ali, "Comparative Assessments for Different WiMAX Scheduling Algorithms", *World Congress on Engineering and Computer Science*, vol.1, San Francisco, October 20-22, 2009.

[16] C.Cicconetti, A.Erta, L.Lenzini and E.Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support", *IEEE Transactions on Mobile Computing*, vol.6, no.1, pp.26-38, January 2007.

[17] M.Katevenis, S.Sidiropolous and C.Courcoubetis, "Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip", *IEEE Journal on Selected Areas in Communications*, vol.9, pp.1265-1279, October 1991.

[18] N.Ruangchaijatupon, L.Wang and Y.Ji, "A Study on the Performance of Scheduling Schemes for Broadband Wireless Access Networks", *Proceedings of International Symposium on Communications and Information Technology*, pp. 1008-1012, Bangkok, Thailand, October 2006.

[19] T.Tsai, C.Jiang and C.Wang, "CAC and Packet Scheduling Using Token Bucket for IEEE 802.16 Networks", *Journal of Communications*, vol.1, no. 2, pp.30-37, May 2006.

[20] K. Wongthavarawat, and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems", *International Journal of Communication Systems*, vol. 16, issue 1, pp. 81-96, February 2003.

[21] M.Settembre, M.Puleri, S.Garritano, P.Testa, R.Albanese, M.Mancini and V.Lo Curto, "Performance analysis of an efficient packet-based IEEE 802.16 MAC supporting adaptive modulation and coding", *Proceedings of International Symposium on Computer Networks*, pp.11-16, Istanbul, Turkey, June 2006.

[22] J.Lin and H.Sirisena, "Quality of Service Scheduling in IEEE 802.16 Broadband Wireless Networks", *Proceedings of First International Conference on Industrial and Information Systems*, pp.396-401, Peradeniya , Sri Lanka, August 2006.

[23] G. Kramer, B. Mukherjee, and G. Pesavento, "Interleaved polling with adaptive cycle time (IPACT): A dynamic bandwidth distribution scheme in an optical access network," *Photon Network Communications*, vol. 4, no. 1, pp. 89-107, Jan. 2002.

[24] H. Naser and H. T. Mouftah, "A joint-ONU interval-based dynamic scheduling algorithm for Ethernet passive optical networks," *IEEE/ACM Transactions on Networking,* vol.14, no. 4, pp. 889-899, August 2006.

[25] S. Hwang and H. Cho, "A Novel Channel Allocation and Scheduling Algorithm in OFDMA System", *IEEE 64th Vehicular Technology Conference,* pp. 1-4, Montreal, Canada, 25-28 September, 2006.

[26] D.Niyato and E.Hossain, "A Queuing-Theoretic Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks", *IEEE Transactions on Computers*, vol.55, no.11, pp. 1473-1488, November 2006.

[27] V.Singh and V.Sharma, "Efficient and fair scheduling of uplink and downlink in IEEE 802.16 OFDMA networks", *Proceedings of IEEE Wireless Comms. and Networking Conference*, pp.984-990, Las Vegas, USA, September 2006.

[28] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance", *IEEE/ACM Trans. Networks*, vol. 1, no. 4, pp. 397-413, Aug. 1993.

[29] K. Park and W. Willinger, *Self-similar network traffic: An overview*, In K. Park and W.Willinger, editors, Self-Similar Network Traffic and Performance Evaluation. WileyInterscience, 2000.

[30] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "*On the Self-Similar Nature ofEthernet Traffic (Extended Version)*," IEEE/ACM Transactions on Networking, vol. 2, no. 1, pp. 1-15, February 1994.

[31] M. Katz and F. Fitzek, "WiMAX evolution: emerging technologies and applications", Chichester, West Sussex: John Wiley &Sons Ltd, 2009, ch2, sec 2.3, pp25-31.

[32] Raj Jain, Chakchai So-In and Abdel-Karim Al Tamimi, "System-Level Modeling of IEEE 802.16e Mobile WiMAX Networks: Key Issues". *IEEE Wireless Communications*, vol. 5, no. 5, pp 73-79, 17 October 2008.

[33] S. Ahmadi, "Mobile WiMAX: A Systems Approach to Understanding IEEE 802.16m Radio Access Technology", Burlington, MA: Academic Press, 2011, ch.9 , sec. 4.

[34] M. Valkama and M. Renfors, "Advanced I/Q signal processing for communication systems," tutorial paper in *2003 Software Defined Radio Conf.* (SDR'03), Orlando, FL, USA, Nov. 2003.

[35] Peng Sun and Hassan Naser, "Dynamic Resource Allocation Algorithms for Long Term Evolution (LTE) Wireless Broadband Networks", A thesis submitted to the Faculty of Graduate Studies, Lakehead University, September 2010.

[36] Internet Engineering Task Force (IETF), Differentiated Services, [Online]. Available: http://www.ietf.org/html.charters/OLD/diffserv-charter.html

[37] International Telecommunication Union (ITU), G.711 standard, [Online]. Available: http://www.itu.int/ITU-T/recommendations/related_ps.aspx?id_prod=911

[38] WiMAX Global Providers, [Online]. Available: http://www.intel.com/content/www/us/en/wireless-network/wimax-providers.html

[39] Mirentech Ltd, [Online]. Available: http://mirentech.co.uk/cms/app/webroot/files/upload/image/voip_clip_image001.gif