Resource Allocation in IEEE 802.16 Mobile WiMAX

Chakchai So-In, Raj Jain, and Abdel Karim Al-Tamimi
Department of Computer Science and Engineering
Washington University in St. Louis
St. Louis, MO 63130 USA
cs5, Jain, and aa7 @cse.wustl.edu

Published in "*Orthogonal Frequency Division Multiple Access (OFDMA)*," Edited by Tao Jiang, Lingyang Song, Yan Zhang, Auerbach Publications, CRC Press, ISBN: 1420088246, April 2010.

Chapter Summary

In this chapter, we focus on the management of resource allocation and scheduling in IEEE 802.16e based mobile WiMAX networks. Since mobile WiMAX uses orthogonal frequency division multiple access (OFDMA), the scheduling issues can apply for other OFDMA-based networks as well. Unlike wireless LANs, WiMAX networks incorporate several Quality of Service (QoS) mechanisms at the media access control (MAC) layer for guaranteed services for data, voice, and video. The problem of assuring QoS is basically that of how to allocate available resources among users in order to meet the QoS criteria such as delay, delay jitter, and throughput requirements. IEEE standard does not include a standard scheduling mechanism and leaves it for vendor differentiation. Scheduling is, therefore, of special interest to all WiMAX equipment makers and service providers. This chapter mainly discusses the key issues and design factors to be considered for scheduler design. Before discussing the scheduling disciplines in detail, the three main services - voice, video, and data are discussed. Understanding the nature of traffic classes gives an insight on how to design the scheduler to guarantee the service. Moreover, the capacity limitation of each service is analyzed not only for capacity planning but also for simulation validation purposes. Finally, we present a brief survey of recent scheduling research. We classify the proposed mechanisms based on the use of channel conditions. The goals of scheduling are to achieve the optimal usage of resources, to assure the QoS guarantees, to maximize goodput, and to minimize power consumption while ensuring feasible algorithm complexity and system scalability.

The organization of this chapter, shown in Table 1, is as follow: In Section 1 we give an overview to WiMAX in order to give readers better understanding of WiMAX characteristics such as physical (PHY) and MAC. In Section 2 the capacity limitation of each service is analyzed not only for capacity planning but also for simulation validation purposes. Then, we emphasize key issues such as two-dimensional downlink mapping (Section 3) and bandwidth request mechanism (Section 4) in order to design the scheduling disciplines, and then we discuss the key factors that the scheduling designer needs to consider as well as a brief survey on WiMAX scheduling in Section 5. We revisit open issues and potential research areas in Section 6.

Table 1: Structure of Chapter 1

- Introduction: PHY and MAC, registration process
- Capacity estimation
- Downlink mapping
- Bandwidth request mechanisms
- QoS schedulers: channel-aware and channel-unaware
- Conclusions and open research issues

Table 2: Constraints and assumptions

- Partially Used Subchannelization (PUSC) is assumed. This is the most commonly used subchannelization.
- Mobile WiMAX: We focus on mobile WiMAX. Since it uses OFDMA, the scheduling issues can apply to other OFDMA-based networks.
- Some overview sections may be revisited. For example, WiMAX frame structure is briefly described again in the capacity estimation section to allow understanding effect of various framing constraints on the goodput calculation.

1 Introduction to Resource Allocation in WiMAX [1, 2]

In this section, general concepts of resource allocation in OFDMA are explained and compared to Single Carrier (SC) and Orthogonal Frequency Division Multiplexing (OFDM). Frame structure and subscriber initialization are introduced so that the reader can understand the constraints in designing resource allocation schemes.

IEEE 802.16e [5, 6] is one of a set of telecommunications technology standards aimed at providing wireless access over long distances in a variety of ways - from point-to-point links to full mobile cellular type access as shown in Figure 1. It covers a metropolitan area of several kilometers and is also called WirelessMAN. Theoretically, a WiMAX base station can provide broadband wireless access up to 30 miles (50 kms) for fixed stations and 3 to 10 miles (5 to 15 kms) for mobile stations with a maximum data rate of up to 70 Mbps compared to 802.11a with 54 Mbps up to several hundred meters, Enhanced Data Rates for Global Evolution (EDGE) with 384 kbps to a few kms, and Code-Division Multiple Access 2000 (CDMA2000) with 2 Mbps for few kms.

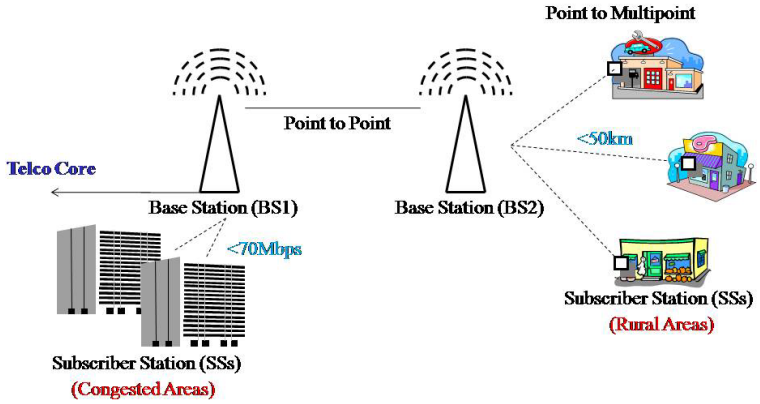

Figure 1: WiMAX Deployment Scenarios

IEEE 802.16 standards group has been developing the standards for broadband (high-speed) wireless access (BWA) in a metropolitan area. Since 2001, a number of variants of these standards have been issued and are still being developed. Like any other standards, these specifications are also a compromise of several competing proposals and contain numerous optional features and mechanisms. The **W**orldwide **I**nteroperability for **M**icrowave **Acc**ess Forum or WiMAX Forum is a group of 400+ networking equipment vendors, service providers, component manufacturers and users that decide which of the numerous options allowed in the IEEE 802.16 standards should be implemented so that equipment from different vendors will inter-operate. Several features such as unlicensed band operation, 60 GHz operation, while specified in the IEEE 802.16 are not a part of WiMAX networks since it is not currently in the profiles agreed at the WiMAX Forum. For an equipment to be certified as WiMAX compliant, the equipment has to pass the inter-operability tests specified by the WiMAX Forum. For the rest of this chapter, the terms WiMAX and the IEEE 802.16 are used interchangeably.

1.1 Key Features of WiMAX Networks

The eight key features of WiMAX networks that differentiate it from other metropolitan area wireless access technologies are:

1) Its use of Orthogonal Frequency Division Multiple Access (OFDMA)
2) Scalable use of any spectrum width (varying from 1.25 MHz to 28 MHz)
3) Time and Frequency Division Duplexing (TDD and FDD)
4) Advanced antenna techniques such as beam forming, multiple input multiple output (MIMO)
5) Per subscriber adaptive modulation

6) Advanced coding techniques such as space-time coding and turbo coding
7) Strong security
8) Multiple QoS classes suitable not only for voice but designed specifically for a combination of data, voice and video services.

Guaranteeing quality of service for a combination of data, voice and video services (or triple play) is important. Unlike voice services, which make symmetric use of uplink (subscriber to base station) and downlink (base station to subscriber), data and video services make a very asymmetric use of link capacities and are, therefore, better served by time division duplexing (TDD) than frequency division duplexing (FDD). This is because TDD allows the service provider to decide the ratio of uplink and downlink transmission times and match it to the expected usage. Most importantly, paired spectrum is not required. Thus, TDD will be the main focus of this chapter. However, the techniques mentioned here can be used for WiMAX networks using FDD as well.

In terms of guaranteed services, WiMAX includes several quality of service (QoS) mechanisms at the MAC layer. Typically, the QoS support in wireless networks is much more challenging than that in wired networks because the characteristics of the wireless link are highly variable and unpredictable both on a time-dependent basis and a location dependent basis. With longer distances, multipath and fading effects also need to be considered. The Request/Grant mechanism is used for mobile stations (MSs) to access the media with a centralized control at the base stations (BSs). WiMAX is a connection-oriented technology (with 16 bits connection id or CID shared for downlink and uplink). Therefore, MSs are not allowed to access the wireless media unless they register and request the bandwidth allocations from the BS first except during certain time slots reserved specifically for contention-based access.

To meet QoS requirements especially for voice and video transmissions with the delay and delay jitter constraints, the key issues are how to allocate resources among the users not only to achieve those constraints but also to maximize goodput (throughput after overheads such as preamble, management messages, level headers, and so on) and to minimize power consumption while keeping feasible algorithm complexity and ensuring system scalability. IEEE 802.16 standard does not specify any resource allocation mechanisms or admission control mechanisms. Although, a number of scheduling algorithms have been proposed in the literature such as Fair Scheduling [26], Distributed Fair Scheduling [27], MaxMin Fair Scheduling [28], Channel State Dependent Round Robin (CSD-RR) [29], Feasible Earliest Due Date (FEDD) [30], and Energy Efficient Scheduling [31]. These algorithms cannot be directly used for WiMAX due to the specific features of the technology. Examples of these specific features are: the Request/Grant mechanism, Orthogonal Frequency Division Multiple Access (OFDMA) vs. Carrier Sense Multiple Access/ Collision Avoidance (CSMA-CA) for wireless LANs, the allocation unit being a slot with specific subchannel and time duration, the definition of fixed frame length and the guaranteed QoS.

The purpose of this section is to provide a brief overview about WiMAX characteristics that need to be considered in developing a scheduler. Therefore, in Section 1.2, we provide a brief introduction to various WiMAX physical layers (PHYs) while we focus on the OFDMA based PHY in the rest of the chapter. Section 1.3 gives an overview of WiMAX frame structure, downlink map (DL-MAP) and uplink map (UL-MAP) for OFDMA. Then, subscriber initialization process is described briefly in Section 1.4. Finally, WiMAX QoS classes are discussed.

1.2 IEEE 802.16 PHYs - Single Carrier (SC), OFDM and OFDMA

IEEE 802.16 supports a variety of physical layers. Each of these has its own distinct characteristics. First, WirelessMAN-SC (Single Carrier) PHY is designed for 10 to 60 GHz spectrum. While IEEE has

standardized this PHY, there are not many products implementing it because this PHY requires line of sight (LOS) communication. Rain attenuation and multipath also affect reliability of the network at these frequencies. To allow non-line of sight (NLOS) communication, IEEE 802.16 designed the Orthogonal Frequency Division Multiplexing (OFDM) PHY using spectrum below 11 GHz. This PHY, popularly known as IEEE 802.16d, is designed for fixed subscriber stations. WiMAX Forum has approved several profiles using this PHY. Most of the current WiMAX products implement this PHY. In this PHY, multiple subscribers use a time division multiple access (TDMA) to share the media. OFDM is a multi-carrier transmission in which thousands of subcarriers are transmitted and each user is given complete control of all subcarriers. The scheduling decision is simply to decide what time slots should be allocated to each subscriber. For mobile users, it is better to reduce the number of subcarriers and to have higher signal power per subscriber. Therefore, multiple users are allowed to transmit using different subcarriers in the same time slot. The scheduling decision then is to decide which subcarriers and what time slots should be allocated to which user. This combination of time division and frequency division multiple access in conjunction with OFDM is called Orthogonal Frequency Division Multiple Access (OFDMA). Figure 2 illustrates a schematic view of the three 802.16 PHYs discussed above. The details of these interfaces can be found in [5, 6].

The scheduler for WirelessMAN-SC can be fairly simple because only time domain is considered. The entire frequency channel is given to the MS. For OFDM, it is more complex since each subchannel can be modulated differently, but it is still only in time domain. On the contrary, both time and frequency domains need to be considered for OFDMA. The OFDMA scheduler is the most complex one because each MS can receive some portions of the allocation for the combination of time and frequency so that the channel capacity is efficiently utilized. It can be shown that the OFDMA outperforms the OFDM [32]. The current direction of WiMAX forum, as well as most WiMAX equipment manufacturers, is to concentrate on Mobile WiMAX, which requires OFDMA PHY. The authors of this chapter have been actively participating in the WiMAX Forum activities. The Application Working Group (AWG) considers scheduling crucial for ensuring optimal performance for Mobile WiMAX applications. Thus, the OFDMA will be our focus for the rest of this chapter.

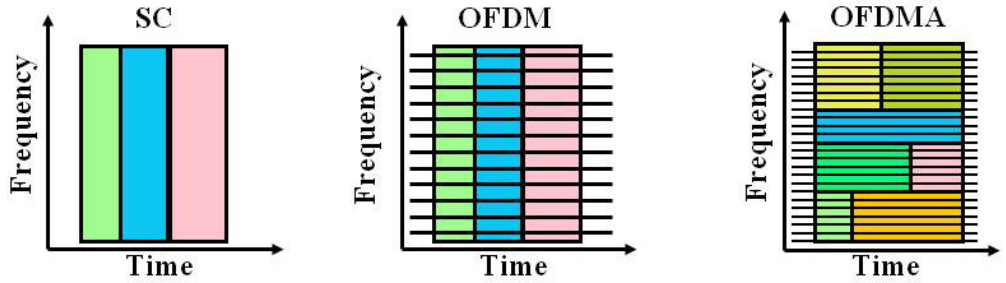

Figure 2: IEEE 802.16 PHYs: SC, OFDM and OFDMA

1.3 WiMAX Frame Structure

IEEE 802.16 standard defines a frame structure as depicted logically in Figures 3 and 4. Each frame consists of downlink (DL) and uplink (UL) subframes. A preamble is used for time synchronization. The downlink map (DL-MAP) and uplink map (UL-MAP) define the burst-start time and burst-end time, modulation types and forward error control (FEC) for each MS. Frame Control Header (FCH) defines these MAP's lengths and usable subcarriers. The MS allocation is in terms of bursts. In the figure, we show one burst per MS; however, WiMAX supports multiple MSs in a single burst in order to reduce the burst overhead. Each burst can contain multiple MAC protocol data units (MPDUs) - the smallest unit from MAC to physical layer. Basically each MPDU is a MAC frame with MAC header (6 bytes), other subheaders such as fragmentation, packing, and grant management (GM) subheaders (2 bytes each) if needed and finally a variable length of payload.

Due to the nature of wireless media, the channel state condition keeps changing over time. Therefore, WiMAX supports adaptive modulation and coding, i.e., the modulation and coding can be changed adaptively depending on the channel condition. Either MS or BS can do the estimation and then BS decides the most efficient modulation and coding scheme. Channel Quality Indicator (CQI) is used to pass the channel state condition information. Figure 4 also shows TTG and RTG gaps. Transmit-receive Transition Gap (TTG) is when the BS switches from transmit to receive mode and Receive-transmit Transition Gap (RTG) occurs when BS switches from receive to transmit mode. The MSs also use these gaps in the opposite way.

To design a WiMAX scheduler, some parameters and attributes need to be considered. For example, number of burst per frame - more bursts result in a larger burst overhead in the form of DL-MAP and UL-MAP information elements (IEs). For uplink, usually there is one burst per subscriber. Note that "burst" usually is defined when there is a different physical mode such as one MS uses QPSK1/4 and another may use 64-QAM3/4. Moreover, all UL data bursts are allocated as horizontal stripes, that is, the transmission starts at a particular slot and continues until the end of UL subframe. Then it continues on the next subchannel. This minimizes the number of subcarriers used by the MS and thus maximizes the power per subcarrier and hence the signal to noise ratio.

For downlink, although the standard allows more than one burst per subscriber, doing so increases DL-MAP overhead. The standard also allows more than one connection packed into one burst with the increased DL-MAP IE size. Moreover, it is possible to pack multiple subscribers into one burst particularly if they are parts of the same physical node. In this scenario, the unique connection identifier (CID) helps separate the subscribers. Packing multiple subscribers in one burst reduces DL-MAP overhead. However, with increase of burst size, there is a decoding delay at the receiving end that needs to be concerned. The DL and UL MAPs are modulated with reliable modulation and coding such as BPSK or QPSK. Also these regions usually require 2 or 4 repetitions depending on the channel condition.

Secondly, in the downlink direction, IEEE 802.16e standard requires that all DL data bursts be rectangular. In fact, the two-dimensional rectangular mapping problem is a variation of bin packing problem, in which one is given bins to be filled with objects. The bins can be in two or more dimensions. If we restrict the bins to two dimensions, we have a "tiling" problem where the objective is to fill a given shape bin with tiles of a given shape. For example, tiling circles in a circle, triangles in a circle, squares in a rectangle etc. We discuss this issue in Section 3.

Thirdly, the number of MPDUs in a burst and their sizes is important. Each MPDU has a MAC header overhead 6 bytes. One can have large MPDUs but then the MPDU loss probability due to bit errors is higher. On the other hand, the MPDU header overhead is significant if there are many small MPDUs. WiMAX provides a packing feature so that instead of 6 bytes, only 2 bytes of overhead are added.

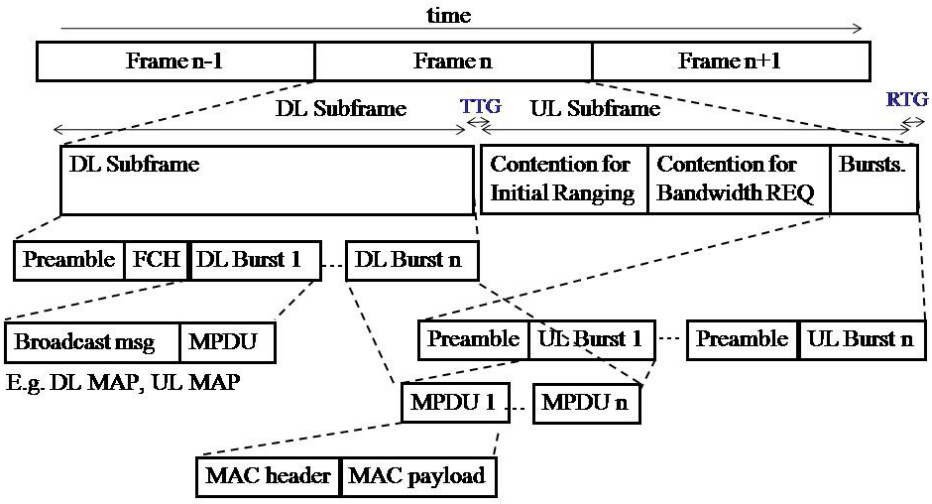

Figure 3: IEEE 802.16 Frame Structure in TDD mode

1.4 Subscriber Initialization Overview

This section gives an overview of MS initialization process. The details can be found in [5, 6]. Since WiMAX is a connection-oriented technology, each MS must register and do the setup process such as the agreement on modulation and coding schemes and QoS requirements. Basically, when the MS joins the network, it first scans for the downlink channel and obtains link parameters. Then, the MS goes through the ranging process that includes basic capability negotiation such as how much power needs to reach the BS. The MS uses a backoff mechanism if there is a contention during ranging. After the basic capabilities have been negotiated and QoS service class has been set up, the MS goes through the authorization and key exchange processes. Once these processes are complete, the MS registers with the BS and receives an IP address and is ready to transfer data. To transmit or receive data, the MS must request the bandwidth either explicitly or implicitly (as described in Section I.G. The BS makes allocation decisions to grant the bandwidth for the MSs via DL-MAP entries for downlink (the MS receives the data) and via UL-MAP entries for uplink (the MS transmits the data).

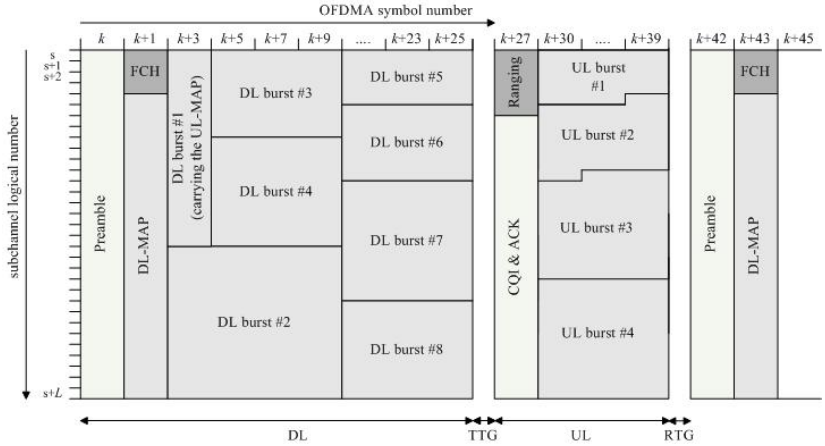

Figure 4: A sample OFDMA TDD frame structure [5]

1.5 WiMAX QoS Service Classes

IEEE 802.16 defines five QoS service classes: Unsolicited Grant Scheme (UGS), Extended Real Time Polling Service (ertPS), Real Time Polling Service (rtPS), Non Real Time Polling Service (nrtPS) and Best Effort Service (BE). Each of these has its own QoS parameters such as the way to request

bandwidth, minimum throughput requirement and delay/jitter constraints. Table 3 summarizes the QoS service classes and Table 4 presents a comparison of these classes.

UGS: This service class provides a fixed periodic bandwidth allocation. Once the connection is setup, there is no need to send any other requests. This service is designed for constant bit rate (CBR) real-time traffic such as E1/T1 circuit emulation. The main QoS parameters are maximum sustained rate (MST), maximum latency and tolerated jitter (the maximum delay variation).

ertPS: This service is designed to support VoIP with silence suppression. No traffic is sent during silent periods. ertPS service is similar to UGS in that the BS allocates the maximum sustained rate in active mode, but no bandwidth is allocated during the silent period. There is a need to have the BS poll the MS during the silent period to determine if the silent period has ended. The QoS parameters are the same as those in UGS.

rtPS: This service class is for variable bit rate (VBR) real-time traffic such as MPEG compressed video. Unlike UGS, rtPS bandwidth requirements vary and so the BS needs to regularly poll each MS to determine what allocations need to be made. The QoS parameters are similar to the UGS but minimum reserved traffic rate and maximum sustained traffic rate need to be specified separately. For UGS and ertPS services, these two parameters are the same, if present.

nrtPS: This service class is for non-real-time VBR traffic with no delay guarantee. Only minimum rate is guaranteed. File Transfer Protocol (FTP) traffic is an example of applications using this service class.

BE: Most of data traffic falls into this category. This service class guarantees neither delay nor throughput. The bandwidth will be granted to the MS if and only if there is a left-over bandwidth from other classes. In practice most implementations allow specifying minimum reserved traffic rate and maximum sustained traffic rate even for this class

Table 3: QoS Service Class Summary

QoS	Applications	Scheduling	Bandwidth Requests	Parameters
UGS	CBR real-time periodic traffic, e.g., a T1 connection	Static allocation; Grant = MST	Reserves BW during setup. Poll-me (PM) bit for unicast polling. No piggyback requests. No bandwidth stealing. No other kinds of polling. GM can be used to for bandwidth request in case of rate mismatch compensation (BS may grant up to 1% additional bandwidth)	Maximum Sustained Traffic Rate = Minimum Reserved Traffic Rate, Maximum Latency, Tolerated Jitter, Uplink Grant Scheduling Type and Unsolicited Grant Interval
ertPS	VoIP with silence suppression/ Video conference (real-time variable-size periodic data)	Dynamic allocation; Grant = MST if active, or 0 if inactive.	Reserves BW during setup. Allows piggyback requests. Allows bandwidth stealing. Allows all kinds of polling (unicast, multicast, broadcast) or Codeword over CQICH	Maximum Sustained Traffic Rate (MST) = Minimum Reserved Traffic Rate, Maximum Latency and Unsolicited Grant Interval

rtPS	Real-time Video (real-time variable-size data on periodic basis)	Dynamic allocation	Allows piggyback Allows bandwidth stealing Allows unicast polling	Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Maximum Latency and Uplink Grant Scheduling Type,
nrtPS	FTP (variable size data)	Dynamic allocation	Allows piggyback Allows bandwidth stealing Allows all kinds of polling (unicast, multicast, broadcast) Unicast polling interval: one second or less	Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Traffic Priority and Uplink Grant Scheduling Type
BE	Web traffic	Dynamic allocation	Allows piggyback Allows bandwidth stealing Allows all kinds of polling (unicast, multicast, broadcast)	Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Traffic Priority and Uplink Grant Scheduling Type

Table 4: Comparison of WiMAX QoS Service Classes

QoS	Pros	Cons
UGS	No overhead. Meet guaranteed latency of MS requests for real-time service	Bandwidth may not be utilized fully since allocations are granted regardless of current need.
ertPS	Optimal latency and data overhead efficiency	Need to use the polling mechanism (to meet the delay guarantee) and a mechanism to let the BS know when the traffic starts during the silent period.
rtPS	Optimal data transport efficiency	Require the overhead of bandwidth request and the polling latency (to meet the delay guarantee)
nrtPS	Provide efficient service for non-real-time traffic with minimum reserved rate	N/A
BE	Provide efficient service for BE traffic	No service guarantee; some connections may starve for long period of time.

1.6 Section Summary

In this section an overview of WiMAX was given to provide a better understanding of resource allocation management to be discussed in the rest of this chapter. Both PHY and MAC descriptions and characteristics were explained along with the subscriber initialization process. A brief overview of WiMAX QoS was also discussed.

2 Triple Play Capacity Estimations [3]

This section describes the capacity estimation for triple play services: voice, video, and data. The purpose is to explain a simple model to estimate the number of supported users. The results can be used not only for capacity planning but also for simulation validation. Note that this model is simple enough to be programmed in a spread sheet program such as Microsoft Excel [3].

In order to estimate the capacity, users need better understanding not only of WiMAX system and configuration parameters but also of the traffic models. As a result, we organize this section as follow:

an overview of WiMAX physical layer framing is provided in Section 2.1 so as to allow better understanding of parameters and configurations used later in the analysis. In Section 2.2, WiMAX system and configuration parameters are discussed. We present three sample workloads consisting of Mobile TV, VoIP, and data applications (Web traffic) in Section 2.3. Section 2.4 explains the analysis of overheads; namely, upper layer overheads and MAC and PHY overheads and also ways to reduce these overheads. Section 2.5 presents parameters of a sample WiMAX system that we used to illustrate the capacity estimation procedure. Moreover, the number of users supported for the three workloads are finally presented. Finally conclusions are drawn in Section 2.5.

2.1 WiMAX OFDMA Modulation and Coding Schemes

One of the key developments of the last decade in the field of wireless broadband is the practical adoption and cost effective implementation of orthogonal frequency division multiple access (OFDMA). Today, almost all upcoming broadband access technologies including WiMAX and its competitors use OFDMA. For performance modeling of WiMAX, it is important to understand OFDMA and hence we provide a very brief explanation that helps us introduce the terms that are used later in our analysis. For further details, we refer the reader to one of several good books on WiMAX [11, 12, 13].

Unlike WiFi and many cellular technologies which use fixed width channels, WiMAX allows almost any available spectrum width to be used. Allowed channel bandwidths vary from 1.25 MHz to 28 MHz. The channel is divided into many equally spaced subcarriers. For example, a 10 MHz channel is divided into 1024 subcarriers some of which are used for data transmission while others are reserved for monitoring the quality of the channel (pilot subcarriers), for providing safety zone (guard subcarriers) between the channels, or for use as a reference frequency (DC subcarrier).

The data and pilot subcarriers are modulated using one of several available MCS (Modulation and Coding Schemes). Quadrature Phase Shift Keying (QPSK) and Quadrature Amplitude Modulation (QAM) are examples of modulation methods. Coding refers to the forward error correction (FEC) bits. Thus, QAM-64 1/3 indicates an MCS with 8-bit (64 combinations) QAM modulated symbols and the error corrections bits take up ⅔ of the bits leaving only 1/3 for data.

In traditional cellular networks, the downlink - Base station (BS) to Mobile Station (MS) - and uplink (MS to BS) use different frequencies. This is called frequency division duplexing (FDD). WiMAX allows FDD but also allows time division duplexing (TDD) in which the downlink (DL) and uplink (UL) share the same frequency but alternate in time. The transmission consists of frames as shown in Figure 5. The DL subframe and UL subframe are separated by a TTG (transmit to transmit gap) and RTG (receive to transmit gap). The frames are shown in two dimensions with frequency along the vertical axis and time along the horizontal axis.

In OFDMA, each MS is allocated only a subset of the subcarriers. The available subcarriers are grouped in to a few subchannels and the MS is allocated one or more subchannels for a specified number of symbols. There are a number of ways to group subcarriers in subchannels of these Partially Used Subchannelization (PUSC) is the most common. In PUSC, subcarriers forming a subchannel are selected randomly from all available subcarriers. Thus, the subcarriers forming a subchannel may not be adjacent in frequency.

Users are allocated variable number of “slots” in the downlink and uplink. The exact definition of slots depends upon the subchannelization method and on the direction of transmission (DL or UL). Figure 5

shows slot formation for PUSC. In uplink (Figure 5a), a slot consists of 6 "tiles" where each tile consists of 4 subcarriers over 3 symbol times. Of the 12 subcarrier-symbol combinations in a tile, 4 are used for pilot and 8 are used for data. The slot, therefore, consists of 24 subcarriers over 3 symbol times. The 24 subcarriers form a subchannel and thus at 10 MHz, 1024 subcarriers form 35 UL subchannels. The slot formation in downlink is different and is shown in Fig 2b. In the downlink, a slot consists of 2 clusters where each cluster consists of 14 subcarriers over 2 symbol times. Thus, a slot consists of 28 subcarriers over two symbol times. The group of 28 subcarriers is called a subchannel resulting in 30 DL subchannels from 1024 subcarriers at 10 MHz.

The WiMAX DL subframe, as shown in Figure 4, starts with one symbol-column of preamble. Other than preamble, all other transmissions use slots as discussed above. The first field in DL subframe after the preamble is a 24-bit Frame Control Header (FCH). For high reliability, FCH is transmitted with the most robust MCS (QPSK ½) and is repeated 4 times. Next field is DL-MAP which specifies the burst profile of all user bursts in the DL subframe. DL-MAP has a fixed part which is always transmitted and a variable part which depends upon the number of bursts in DL subframe. This is followed by UL-MAP which specifies the burst profile for all bursts in the UL subframe. It also consists of a fixed part and a variable part. Both DL and UL MAPs are transmitted using QPSK ½ MCS.

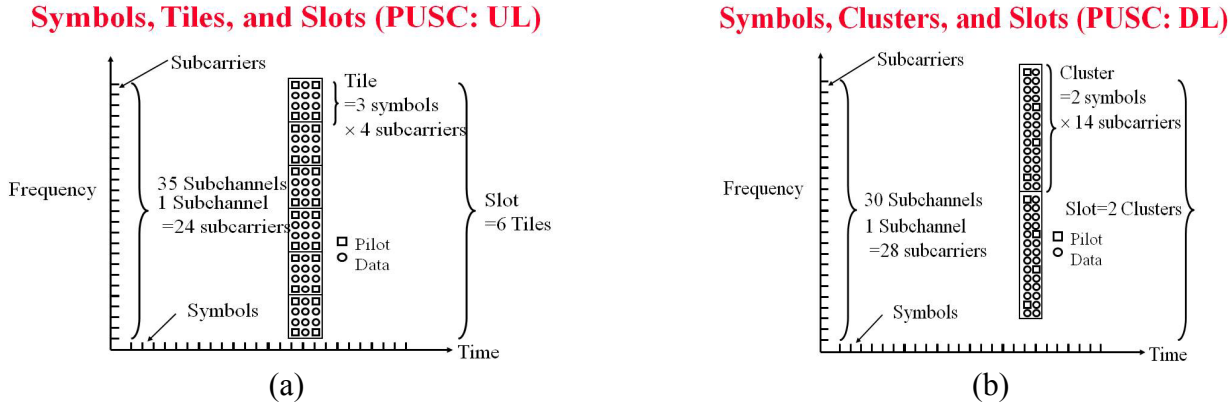

Figure 5: Symbols, Tiles, Clusters, and Slots

2.2 WiMAX configuration parameters and characteristics

The key parameters of WiMAX PHY are summarized in Table 5 through 7.

Table 5: OFDMA Parameters for WiMAX [7]

Parameters	**Values**						
System bandwidth (MHz)	1.25	5	10	20	3.5	7	8.75
Sampling factor	28/25				8/7		
Sampling frequency ($F_{s,MHz}$)	1.4	5.6	11.2	22.4	4	8	10
Sample time ($1/F_s$,nsec)	714.3	178.6	89.3	44.6	250	125	100
FFT size (N_{FFT})	128	512	1024	2048	512	1024	1024
Subcarrier spacing (Δf, kHz)	10.93				7.81		9.76
Useful symbol time ($T_b=1/\Delta f$, µs)	91.4				128		102.4
Guard time ($T_g = T_b/8$,	11.4				16		12.8

μs)			
OFDMA symbol time ($T_s=T_b+T_g$, μs)	102.8	144	115.2

Table 5 lists the OFDMA parameters for various channel widths. Note that the product of subcarrier spacing and FFT size is equal to the product of channel bandwidth and sampling factor. For example, for 10 MHz channel, 10.93kHz×1024 = 10MHz×(28/25). This table shows that at 10 MHz the OFDMA symbol time is 102.8 μs and so there are 48.6 symbols in a 5 ms frame. Of these, 1.6 symbols are used for TTG and RTG leaving 47 symbols. If *n* of these are used for DL then *47-n* are available for uplink. Since DL slots occupy 2 symbols and UL slots occupy 3 symbols, it is best to divide these 47 symbols such that *47-n* is a multiple of 3 and *n* is of the form *2k+1*. For a DL:UL ratio of 2:1, these considerations would result in a DL subframe of 29 symbols and UL subframe of 18 symbols. In this case, the DL subframe will consists of a total of 14×30 or 420 slots. The UL subframe will consist of 6×35 or 210 slots.

Table 6 lists the number data, pilot, and guard subcarriers for various channel widths. A PUSC subchannelization is assumed, which is the most common subchannelization.

Table 6: Number of Subcarriers in PUSC [14]

Parameters	**Values**				
(a) DL					
System bandwidth (MHz)	1.25	2.5	5.	10	20
FFT size	128	N/A	512	1024	2084
# of guard subcarriers	43	N/A	91	183	367
# of used subcarriers	85	N/A	421	841	1681
# of pilot subcarriers	12	N/A	60	120	240
# of data subcarriers	72	N/A	360	720	140
(b) UL					
System bandwidth (MHz)	1.25	2.5	5.	10	20
FFT size	128	N/A	512	1024	2084
# of guard subcarriers	31	N/A	103	183	367
# of used subcarriers	97	N/A	409	841	1681

Table 7: Slot Capacity for various MCSs

MCS	**Bits per symbol**	**Coding Rate**	**DL Bytes per slot**	**UL bytes per slot**
QPSK ⅛	2	0.125	1.5	1.5
QPSK ¼	2	0.25	3	3
QPSK ½	2	0.5	6	6
QPSK ¾	2	0.75	9	9
QAM-16 ½	4	0.5	12	12
QAM-16 ⅔	4	0.67	16	16
QAM-16 ¾	4	0.75	18	16
QAM-64 ½	6	0.6	18	16
QAM-64 ⅔	6	0.67	24	16
QAM-64 ¾	6	0.75	27	N/A

QAM-64 5/6	6	0.83	30	N/A

Table 7 lists the number of bytes per slot for various MCS values. For each MCS, the number of bytes is equal to (#bits per symbols × Coding Rate × 48 data subcarriers and symbols per slot / 8 bits). Note that for UL, the maximum MCS level is QAM-16 ⅔ [7].

This analysis method can be used for any allowed channel width, any frame duration, or any subchannelization. We assume a 10 MHz WiMAX TDD system with 5 ms frame duration, PUSC subchannelization mode, and a DL:UL ratio of 2:1. These are the default values recommended by WiMAX forum system evaluation methodology and are also common values used in practice.

The number of DL and UL slots for this configuration can be computed as shown in Table 8.

Table 8: WiMAX System Configuration

Configurations	**Downlink**	**Uplink**
DL/UL Symbols excluding preamble	28	18
Ranging, CQI and ACK (column symbols)	N/A	3
# of symbol columns per Cluster[1]/ Tile[2]	2	3
# of subcarriers per Cluster[1]/ Tile[2]	14	4
Symbols × Subcarriers per Cluster[1]/ Tile[2]	28	12
Symbols × Data Subcarriers per Cluster[1]/ Tile[2]	24	8
# of pilots per Cluster[1]/ Tile[2]	4	4
# of clusters[1]/ #Tiles[2] per Slot	2	6
Subcarriers × Symbols per Slot	56	72
Data Subcarriers × Symbols per Slot	48	48
Data Subcarriers × Symbols per DL/UL Subframe	23,520	12,600
Number of Slots	**420**	**175**

[1]Cluster for DL and [2]Tile for UL

2.3 Traffic Models and Workload Characteristics

In this capacity modeling study, three sample workloads consisting of VoIP, Mobile TV, and Web traffic are used. Note that to simplify the capacity estimation model, only average packet size is used in the model. Second order statistics (e.g., standard deviation) are not modeled.

First, the VoIP workload is symmetric in that DL data rate is equal to the UL data rate. It consists of very small packets that are generated periodically. The packet size and the period depend upon the Vocoder used. G723.1 is used in our analysis and results in a data rate of 5.3 kbps, 20 bytes voice packet every 30 ms.

Second, the Mobile TV workload depends upon the quality and size of the display. In our analysis, a sample measurement on a small screen Mobile TV device produced an average packet size of 984 bytes every 30 ms resulting in an average data rate of 350.4 kbps [15, 16]. Note that Mobile TV workload is highly asymmetric with almost all of the traffic going downlink.

Finally, for data workload, we selected the Hypertext Transfer Protocol (HTTP) workload recommended by the 3rd Generation Partnership Project (3GPP) [8].

The characteristic summary of the three workloads are presented in Table 9.

Table 9: Workload Characteristics

Parameters	Mobile TV	VoIP	Data
Type of transport layer	RTP	RTP	TCP
Average packet Size (bytes)	983.5	20.0	1200.2
Average data rate (kbps) w/o headers	350.0	5.3	14.5
UL:DL traffic ratio	0	1	0.006
Silence suppression (VOIP only)	N/A	Yes	N/A
Fraction of time user is active		0.5	
ROHC packet type	1	1	TCP
Overhead with ROHC (bytes)	1	1	8
Payload Header Suppression (PHS)	No	No	No
MAC SDU size with header	984.5	21.0	1208.2
Data rate (kbps) after headers	350.4	5.6	14.6
Bytes/Frame per user (DL)	219.0	3.5	9.1
Bytes/Frame per user (UL)	0.0	3.5	0.1

2.4 Overhead Analysis

In this section, we consider both upper (Network, Transport, and so on) and lower (MAC and PHY) layer overheads. We consider only Real Time Transport (RTP) or TCP and IP for upper layer, and these overheads can apply for both downlink and uplink. Next, MAC overhead basically consists of MAC header and other subheaders. Finally, the PHY overhead can be divided into DL overhead and UL overhead. Each of these overheads is discussed next.

2.4.1 Upper Layer Overhead

Table 9, which lists the characteristics of our Mobile TV, VoIP, and data workloads, includes the type of transport layer used: Real Time Transport (RTP) or TCP. This affects the upper layer protocol overhead. RTP over UDP over IP (12+8+20) or TCP over IP (20+20), both can results in a per packet header overhead of 40 bytes. This is significant and can severely reduce the capacity of any wireless system.

There are two ways to reduce upper layer overheads and to improve the number of supported users. These are Payload Header Suppression (PHS) and Robust Header Compression (ROHC). PHS is a WiMAX feature. It allows the sender to not send fixed portions of the headers and can reduce the 40 byte header overhead down to 3 bytes. ROHC, specified by the Internet Engineering Task Force (IETF), is another higher layer compression scheme. It can reduce the higher layer overhead to 1 to 3 bytes. In our analysis, we use ROHC-RTP packet type 0 with R-0 mode. In this mode, all RTP sequence numbers functions are known to the decompressor. This results in a net higher layer overhead of just 1 byte [9, 10].

For small packet size workloads, such as VoIP, header suppression and compression can make a significant impact on the capacity. We have seen several published studies that use uncompressed headers resulting in significantly reduced performance which would not be the case in practice.

PHS or ROHC can significantly improve the capacity and should be used in any capacity planning or estimation.

One option with VoIP traffic is that of silence suppression which if implemented can increase the VoIP capacity by the inverse of fraction of time the user is active (not silent). As a result in this analysis, given silence suppression option, a number of supported users are twice of that without this option.

2.4.2 Lower Layer Overhead

In this section, we analyze the overheads at MAC and PHY layer. Basically, there is 6-byte MAC header and several 2-byte subheaders. For PHY overheads, both downlink and uplink overhead are discussed in details.

2.4.2.1 MAC Overhead

At MAC layer, the smallest unit is MAC protocol data unit (MPDU). As shown in Figure 6, each MPDU has at least 6 bytes of MAC header and a variable length payload consisting of a number of optional subheaders, data, and an optional 4-byte CRC. The optional subheaders include fragmentation, packing, mesh and general subheaders. Each of these is 2 bytes long.

In addition to generic MAC PDUs, there are bandwidth request PDUs. These are 6 bytes in length. Bandwidth requests can also be piggybacked on data PDUs as a 2-byte subheader.

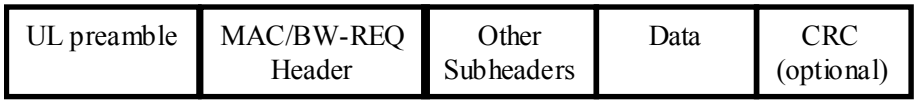

Figure 6: UL burst preamble and MAC PDU (MPDU)

Consider fragmentation and packing subheaders. As shown in Table 9, the user bytes per frame in downlink are 219, 3.5, and 9.1 bytes for Mobile TV, VoIP, and Web, respectively. In each frame, a 2-bytes fragmentation subheader is needed for all types of traffic. Packing is not used for the simple scheduler used here.

However, in enhanced scheduler, given a variation of deadline, packing multiple SDU is possible. Table 10 shows an example when deadline is put into consideration. In this analysis, the deadline of Mobile TV, VoIP, and Web traffic are set to 10, 60 and 250 ms. As a result, 437.9, 42, and 454.9 bytes are allocated per user. These configuration results in one 2-byte fragmentation overhead for mobile TV and web traffic but two 2-byte packing overheads with no fragmentation for VoIP. Table 10 also shows the detailed explanation of fragmentation and packing overheads in downlink. Note that the calculation for uplink is very similar.

Table 10: Downlink Fragmentation and Packing Subheaders

Parameters	Mobile TV	VoIP	Data
Average packet Size with higher level header (bytes)	984.5	21.0	1208.2
Simple Scheduler (every frame scheduling)			
Bytes/5 ms frame per user	219.0	3.5	9.1
Number of fragmentation subheaders	1	1	1
Number of packing subheaders	0	0	0
Enhanced Scheduler (scheduling within deadline)			
Deadline (ms)	10	60	250

Bytes/5 ms frame per user	437.9	42.0	454.9
Number of fragmentation subheaders	1	0	1
Number of packing subheaders	0	2	0

2.4.2.2 Downlink Overhead

In DL subframe, overhead consist of preamble, FCH, DL-MAP and UL-MAP. The MAP entries can result in a significant amount of overhead since they are repeated 4 times. WiMAX Forum recommends using compressed MAP [7], which reduces the DL-MAP entry overhead to 11 bytes including 4 bytes for Cyclic Redundancy Check (CRC) [5, 6]. The fixed UL-MAP is 6 bytes long with an optional 4-byte CRC. With a repetition code of 4 and QPSK½, both fixed DL-MAP and UL-MAP take up 16 slots.

The variable part of DL-MAP consists of one entry per bursts and requires 60 bits per entry. Similarly, the variable part of UL-MAP consists of one entry per bursts and requires 52 bits per entry. These are all repeated 4 times and use only QPSK ½ MCS. It should be pointed out that repetition consists of repeating slots (and not bytes). Thus, both DL and UL MAPs entries also take up 16 slots each per burst.

2.4.2.3 Uplink Overhead

The UL subframe also has fixed and variable parts (See Figure 4). Ranging and contention are in the fixed portion. Their size is defined by the network administrator. These regions are allocated not in units of slots but in units of "transmission opportunities". For example, in CDMA initial ranging, one opportunity is 6 subchannels and 2 symbol times.

The other fixed portion is channel quality indication (CQI) and acknowledgements (ACK). These regions are also defined by the network administrator. Obviously, more fixed portions are allocated; less number of slots is available for the user workloads. In our analysis, we allocated three OFDM symbol columns for all fixed regions.

Each UL burst begins with a UL preamble. One OFDM symbol is used for short preamble and two for long preamble. In this analysis, we do not consider one short symbol (a fraction of one slot); however, users can add an appropriate size of this symbol to the analysis.

2.5 Pitfalls

Many WiMAX analyses ignore the overheads described in Section 2.3, namely, UL-MAP, DL-MAP, and MAC overheads. In this section, we show that these overheads have a significant impact on the number of users supported. Since some of these overheads depend upon the number of users, the scheduler needs to be aware of this additional need while admitting and scheduling the users.

Given the user workload characteristics and the overheads discussed so far, it is straightforward to compute the system capacity for any given workload. Using the slot capacity indicated in Table 7, for various MCS, we can compute the number of users supported.

One way to compute the number of users is simply to divide the channel capacity by the bytes required by the user payload and overhead [1]. This is shown in Table 11. The table assumes QPSK ½ MCS for all users. This can be repeated for other MCS. The final results are as shown in Figure 7. The number of users supported varies from 2 to 46 depending upon the workload and the MCS.

Table 11: Capacity Estimation using a Simple Scheduler

Parameters	Mobile TV	VoIP	Data
MAC SDU size with header (bytes)	984.5	21.0	1208.2
Data rate (kbps) with upper layer headers	350.4	5.6	14.6
(a) DL			
Bytes/5 ms frame per user (DL)	219.0	3.5	9.1
Number of fragmentation subheaders	1	1	1
Number of packing subheaders	0	0	0
DL data slots per user with MAC header + packing and fragmentation subheaders	38	2	3
Total slots per user (Data + DL-MAP IE + UL-MAP IE)	46	18	19
Number of users (DL) slot-based	8	22	21
Number of users (DL) upper bound (w rounding error)	9	35	33
(b) UL			
Bytes/5ms Frame per user (UL)	0.0	3.5	0.1
# of fragmentation subheaders	0	1	1
# of packing subheaders	0	0	0
UL data slots per user with MAC header + packing and fragmentation subheaders	0	2	2
Number of users (UL)	∞	87	87
Number of users (min of UL and DL)	9	35	33
Number of users with silence suppression	**9**	**70**	**33**

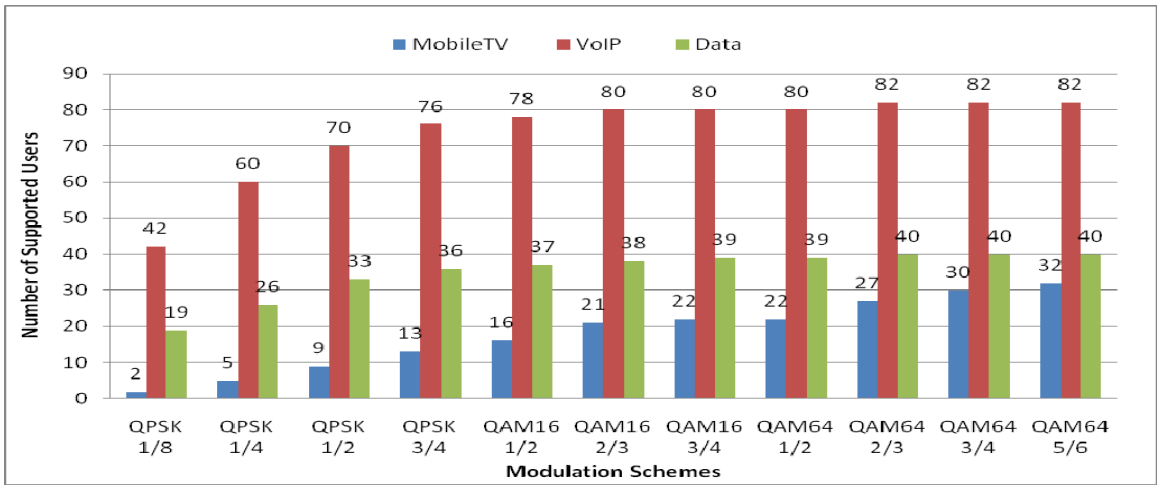

Figure 7: Number of users supported in lossless channel (Simple scheduler)

The main problem with the analysis presented above is that it assumes that every user is scheduled in every frame. Since there is a significant per burst overhead, this type of allocation will result in too much overhead and too little capacity. Also, since every packet (SDU) is fragmented, a 2-byte fragmentation subheader is added to each MAC PDU.

What we discussed above is a common pitfall. The analysis assumes a dumb scheduler. A smarter scheduler will try to aggregate payloads for each user and thus minimizing the number of bursts. We call this enhanced scheduler. It works as follows. Given *n* users with any particular workload, we divide the users in *k* groups of *n/k* users each. The first group is scheduled in the first frame; the second group is

scheduled in the second frame, and so on. The cycle is repeated every *k* frames. Of course, *k* should be selected to match the delay requirements of the workload. For example, with VoIP users, a VoIP packet is generated every 30 ms but assuming 60 ms is an acceptable delay, we can schedule a VoIP user every 12th WiMAX frame (recall that each WiMAX frame is 5 ms) and send two VoIP packets in one frame as compared to the previous scheduler which would send 1/6th of the VoIP packet in every frame and thereby aggravating the problem of small payloads. A 2-byte×2 packing overhead has to be added in the MAC payload along with the two SDUs.

Table 12 shows the capacity analysis for the three workloads with QPSK ½ MCS and the enhanced scheduler. The results for other MCS can be similarly computed. These results are plotted in Figure 8. Note that the number of users supported has gone up 2 to 600. Compared to Figure 7, there is a capacity improvement by a factor of 1 to 25 depending upon the workload and MCS.

Proper scheduling can change the capacity by an order of magnitude. Making less frequent but bigger allocations can reduce the overhead significantly.

Table 12: Capacity Estimation using an Enhanced Scheduler

Parameters	**Mobile TV**	**VoIP**	**Data**
MAC SDU size with header (bytes)	984.5	21.0	1208.2
Data rate (kbps) with upper layer headers	350.4	2.8	14.6
Deadline (ms)	**10**	**60**	**250**
(a) DL			
Bytes/5 ms frame per user (DL)	437.9	42.0	454.9
Number of fragmentation subheaders	1	0	1
Number of packing subheaders	0	2	0
DL data slots per user with MAC header + packing and fragmentation subheaders	75	9	78
Total slots per user (Data + DL-MAP IE + UL-MAP IE)	83	25	94
Number of users (DL) slot-based	8	192	200
Number of users (DL) upper bound (w rounding error)	10	269	233
(b) UL			
Bytes/5 ms frame per user (UL)	0.0	42.0	2.9
Number of fragmentation subheaders	1	0	1
Number of packing subheaders	0	2	0
UL data slots per user with MAC header + packing and fragmentation subheaders	0	9	2
Number of users (UL)	∞	228	4350
Net number of users (min of UL and DL)	10	228	233
Number of users with silence suppression	**10**	**456**	**233**

Note that the per user overheads impact the downlink capacity more than the uplink capacity. The downlink subframe has DL-MAP and UL-MAP entries for all DL and UL bursts, and these entries can take up a significant part of the capacity and so minimizing the number of bursts increases the capacity.

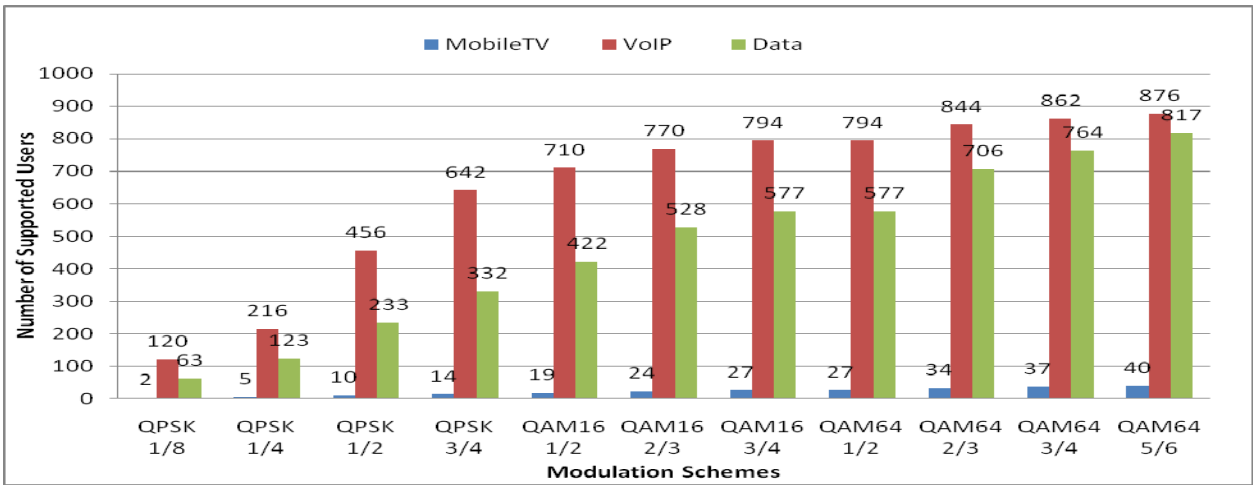

Figure 8: Number of users supported in lossless channel (Enhanced Scheduler)

Note that there is a limit to aggregation of payloads and minimization of bursts. First, the delay requirements for the payload should be met, and so a burst may have to be scheduled even if the payload size is small. In these cases, multi-user bursts in which the payload for multiple users is aggregated in one DL burst can help reduce the number of bursts. This is allowed by the IEEE 802.16e standards and applies only to the downlink bursts.

The second consideration is that the payload cannot be aggregated beyond the frame size. For example, with QPSK ½, a Mobile TV application will generate enough load to fill the entire DL subframe every 10 ms or every 2 frames. This is much smaller than the required delay of 30 ms between the frames.

2.6 Section Summary

In this section, we explained how to compute the capacity of a WiMAX system and account for various overheads. We illustrated the methodology using three sample workloads consisting of Mobile TV, VoIP, and data users.

Analysis such as the one presented in this section can be easily programmed in a simple program or a spread sheet and effects of various parameters can be analyzed instantaneously. This can be used to study the sensitivity to various parameters so that parameters that have significant impact can be analyzed in detail by simulation. This analysis can also be used to validate simulations.

However, there are a few assumptions in the analysis such as the effect of bandwidth request mechanism, two-dimensional downlink mapping, and the imprecise calculation of slot-based vs. bytes-based. Moreover, we do not consider (H)ARQ and the error-free channel is given as one of our assumption (the error-prone channel analysis was shown in [3]). In addition, the number of supported users is calculated with the assumption that there is only one traffic type. Finally, fixed UL-MAP is always in the DL subframe though there is no UL traffic such as Mobile TV.

We show that proper accounting of overheads is important in capacity estimation. A number of methods are available to reduce these overheads and these should be used in all deployments. In particular, robust header compression or payload header suppression, compressed MAPs are examples of methods for reducing the overhead.

Proper scheduling of user payloads can change the capacity by an order of magnitude. The users should be scheduled so that their number of bursts is minimized while still meeting their delay constraint. This reduces the overhead significantly particularly for small packet traffic such as VoIP.

3 Downlink Burst Mapping Algorithms [4]

In this section, we focus on a two-dimensional bust mapping issue in WiMAX networks, which many of the scheduling proposals ignore. Unlike uplink resource allocation, horizontal strip-based mapping, the downlink burst mapping requires a rectangular shape so perhaps because of this mapping constraint, QoS requirements may not be met although the scheduler allocates enough slots for a subscriber.

Again we revisit the idea of using Orthogonal Frequency Division Multiple Access (OFDMA) technique in order to achieve higher data rate, longer distance, and mobility for IEEE 802.16e Mobile WiMAX. Basically, the entire channel is divided into multiple subcarriers. The number of subcarriers is proportional to the channel spectral width. These subcarriers are grouped into a number of subchannels. Then, each mobile station (MS) is assigned a group of subchannels for some OFDMA symbol times as shown by the two dimensional diagram in Figure 9.

In this figure the vertical axis is frequency or subcarriers or logical subchannels, and the horizontal axis is time or OFDMA symbols. Mobile WiMAX uses a fixed frame-based allocation. Basically, each frame is of 5 ms duration [6]. Bi-directional communication can be achieved by frequency division duplexing (FDD) in which uplink and downlink use different frequency bands or time division duplexing (TDD) in which the downlink (DL) traffic follows the uplink (UL) traffic in time domain.

Figure 8 shows both a downlink (DL) subframe and uplink (UL) subframe of a TDD WiMAX system. In FDD, the two subframes are parallel in time. For data traffic, TDD provides a flexible partitioning of the frame into DL and UL subframes. We use a TDD system for the rest of the chapter; however, the mapping algorithm we have introduced can be used for both systems.

Mobile WiMAX frame starts with a downlink preamble and a frame control header (FCH) followed by the downlink map (DL-MAP) and uplink map (UL-MAP). These maps contain the information elements that specify the burst profile for each burst. The profile consists of burst-start time, burst-end time, modulation type, forward error control (FEC) used or to be used in the burst. Although we limit the discussion to one subscriber per burst, our algorithm can be easily applied to the case of multiple subscribers per burst, which is allowed by the standard.

In Mobile WiMAX systems, the base station (BS) has full control over resource allocations to various MSs in both DL and UL. In DL, BS decides the burst size based on the packets waiting in the queue to be sent to various subscribers. In UL, MSs send bandwidth requests for each connection that they have set up. Each connection has an agreed quality of service (QoS) requirement that is negotiated between the BS and MS at the time of connection setup. BS grants transmit opportunities to various MSs based on their bandwidth requests and QoS. Note that while the bandwidth requests are made separately for each connection, the grants (allocations) are made per MS in the sense that the MS can use its uplink allocation to transmit packets of any of its connections.

Figure 8 also shows a ranging region in the uplink subframe. Ranging is used to determine the distances between the BS and various MSs so that the transmission start times at these stations can be properly synchronized. Ranging also helps to set the right transmit power level for each MS. Some slots are also reserved for sending bandwidth requests, channel quality indicator (CQI), and acknowledgements (ACK).

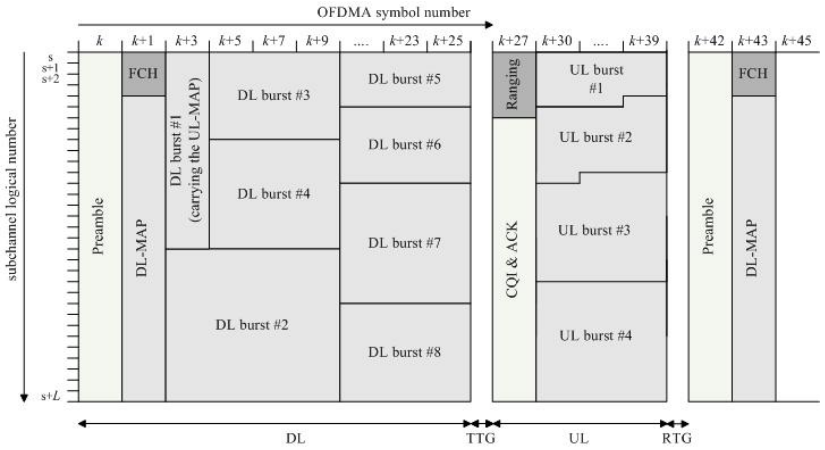

Figure 8: A sample OFDMA TDD frame structure [5]

Note that each UL data burst is allocated as a horizontal strip. The transmission starts at a particular slot and continues until the end of UL subframe. After that, the allocation continues on the next subchannel. The horizontal allocation is used to minimize the number of subcarriers for each MS. This maximizes the power per subcarrier and hence the signal to noise ratio (SNR).

IEEE 802.16e Mobile WiMAX standard requires that all DL data bursts be rectangular in shape. Although the standard allows more than one burst per subscriber, it increases DL-MAP overhead. This particular case may be used when the subscriber really needs a different reliable channel say different MCSs for different connections. The standard also allows more than one connections packing into one burst with the increase of DL-MAP IE size [5, 6]; however, the problem of rectangular mapping still remains. Our algorithm primarily maps the resource for each subscriber into a downlink burst in a rectangular fashion. We do not consider more than one burst per subscriber. However, it is possible to pack multiple subscribers into one burst particularly if they are parts of the same physical mode. In this scenario, the unique connection identifier (CID) helps separate the subscribers. Packing multiple subscribers in one burst reduces DL-MAP overhead and the proposed algorithm can be applied directly to this combined resource allocation. This rectangular criterion requires an efficient two-dimensional mapping algorithm. This is the main focus of this section.

To assure QoS requirements, downlink resource scheduling can be done in three steps as shown in Figure 9. In the first step, before accepting new connections, the admission control module consults the scheduler to ensure that the required QoS of the new connection can be met. In this step, basically the scheduler computes the resource allocation (number of slots to be allocated) for that new connection and makes sure that allowing it will not violate the QoS for existing connections.

Then for each frame, the resource allocation can be done first without any shape constraints and based solely on demand (the number of packets to be sent to/from a station), capacity (total available slots), and quality of service (QoS). Finally, in downlink subframe this resource allocation is mapped in to the Mobile WiMAX frame in rectangular regions. At this step, the mapping module informs the scheduler and admission control modules if the resource allocation can be met without any QoS violations.

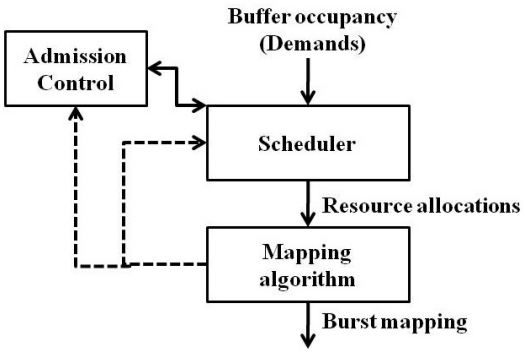

Figure 9: Three steps of downlink allocation

The two-dimensional rectangle mapping problem is a variation of bin or strip packing problem, in which one is given bins to be filled with objects. The bin packing problems are known to be NP hard [24]. The complexity of the solution grows exponentially with the number of objects or bins. There have been many attempts to overcome these problems as stated in [21, 22, 23]. However, there is no easy way to achieve the optimality with simple computation. Examples of simple approaches are to apply first-fit, next-fit, best-fit, or bottom-left allocation [21]. Moreover, many heuristic approaches have been introduced to alleviate the computational complexity such as level approach [21] is used to pack the fixed dimensional bins with non-increasing height from bottom to top and left to right, and then move to the next level when it reaches maximum allowable width W. Reverse-fit [25] approach is another heuristic example; the algorithm applies two level approaches, allocating the bins with decreasing height from bottom up and left to right, and then allocating the bins from top to bottom and right to left when it reaches the width W.

As a result, we propose a simple heuristic algorithm for two-dimensional rectangular mapping for downlink bursts in IEEE 802.16e Mobile WiMAX, that is, all resource allocations need to be mapped in to Mobile WiMAX frame. In fact, we basically apply the concept of largest area first and level mapping approach together. The section is organized as follows: two-dimensional rectangle mapping problem is revisited in Section 3.1. Section 3.2 briefly describes some of the related work. Our heuristic algorithm for two-dimensional downlink burst mapping is described in Section 3.3. Then, performance evaluation is presented in Section 3.4, and finally the conclusions are drawn.

3.1 Downlink Burst Mapping Problem Statement

In IEEE 802.16e Mobile WiMAX, the two-dimensional downlink burst mapping can be stated as follows:

1) We are given a fixed rectangular bin B of width W and height H. The bin B has an area A equal to $W \times H$.
2) We are also given a set of n items $\{b_1, b_2, ..., b_n\}$. The ith item b_i has an area A_i
3) We need to *determin*e a rectangular shape for the ith item with width W_i and height H_i such that $\boldsymbol{A_i \leq W_i \times H_i}$.
4) Width $W_i \leq W$ for all i. Similarly, height $H_i \leq H$ for all i.
5) W_i , H_i, W, and H are all integers.
6) Since the mapped region is more than the desired allocation A_i, the extra resource is wasted and so, $W_i \times H_i - A_i$, should be minimized.
7) Due to the rectangular considerations, all n item bins may not fit the big bin B, the goal is to minimize the additional resource width W that is required to fit all n item bins.

The unit of allocation in IEEE 802.16e Mobile WiMAX is “slot.” The definition of slot depends upon

the subchannelization mode and link direction (DL or UL) [5, 6]. In this section, we assume the Partially Used Sub-Channelization (PUSC) mode, which is the most commonly used mode [7]. The analysis is applicable to other modes as well. Consider downlink PUSC with 10 MHz channel. With WiMAX forum specified parameters, 10 MHz channel requires 1,024 subcarriers. In the downlink, these 1,024 subcarriers are grouped in to 30 subchannels with each subchannel consisting of fixed 28 subcarriers. A 5 ms frame and a 2:1 DL:UL ratio result in 14 slot columns in DL [3]. Thus, for this parameter set, the DL subframe consists of 14 slot columns and 30 rows resulting in a total of 420 slots. Of these we allow 12 slot columns for QoS sensitive traffic. Rectangle mapping may require *one more slot column*. The remaining space is for maps and FCH.

3.2 Design Factors

As indicated earlier, the burst mapping problem is NP hard [24]. A heuristic algorithm is proposed in this section. There were four considerations in designing this algorithm as discussed below.

1) The resource mapping should maximize the throughput and minimize the over allocation and unused spaces.
2) The mapping algorithm should be simple and fast so that large number of users and bursts can be handled efficiently
3) The algorithm should be aware of variable components of DL-MAP and UL-MAP. This variable portion consists of burst profile of each burst and therefore depends upon the number of bursts [3]
4) The resource mapping should be such that the energy consumption of the MS is minimized. The MS needs to be active during the burst and so this duration width should be minimized.

3.3 Related Work

Two-dimensional mapping for Orthogonal Frequency Division Multiple Access (OFDMA) was introduced by Yehuda Ben-Shimol and his colleagues [20]. The algorithm simply assigns the resource allocation row by row with largest resource allocation first. There is no detail explanation how to map the resources to unused space in a frame when their sizes span over multiple rows.

Another rectangular mapping algorithm was introduced by Takeo Ohseki and his colleagues [18]. They basically allocate in time domain first and then the frequency domain (left to right and top to bottom). The algorithm is similar to the algorithm in [20] but allows a burst compaction if there are more than one burst that belongs to the same physical mode. The algorithm does not consider an unused space. Without this consideration, the algorithm results in reduced throughput.

Claude Desset and his colleges [17] showed that only 8 users at maximum can be supported with binary-tree full search algorithm for burst mapping in rectangular shape. However this work does not consider the variation of possible mapping pairs for each particular burst. An optimization was introduced but the purpose was to reduce only the number of allocated OFDM symbols.

Bacioccola et al [19] presented an algorithm that basically allocates from right to left and bottom to top. They map a single allocation in to multiple rectangular areas that may result in increased DL MAP elements overhead. In our assumption, we have only one bust per subscriber.

Table 13: Two-dimensional rectangular mapping for downlink on WiMAX networks

	Algorithm Descriptions	**Pros**	**Cons**

Yehuda Ben-Shimol et al [20]	Assign the resource allocation row by row with largest resource allocation first	Simple	There is no detailed explanation of how to map the resources to unused space in a frame when their sizes span over multiple rows
Takeo Ohseki et al [18].	Allocate in time domain first and then the frequency domain (left to right and top to bottom).	Allows burst compaction if there are more than one bursts that belongs to the same physical node	The algorithm does not consider the unused space.
Claude Desset et al [17]	Binary-tree full search algorithm	Optimize frame utilization	Only 8 users at maximum can be supported
Bacioccola et al [19]	Allocate from right to left and bottom to top	Optimize frame utilization	They map a single allocation in to multiple rectangular areas that may result in increased DL MAP elements overhead

Table 13 also shows the mapping solution on WiMAX networks comparatively. In this section we propose a two-dimensional downlink burst mapping algorithm called Efficient One Column Striping with non-increasing Area first (eOCSA). The four considerations in its design have already been mentioned in Section 3.2. The algorithm is simple to implement and provides very good throughput efficiency. In performance evaluation section, we show that since eOCSA considers only one best mapping-pair either the least width or height, eOCSA lowers down the complexity to $O(n^2)$. The efficiency of the algorithm is approximately 93% with one additional column in average.

3.4 eOCSA Algorithm

In this section, we describe our two-dimensional rectangular burst mapping algorithm. We divide the resource scheduling problem in two steps. In the first step, the scheduler computes the allocation of each user based solely on its demand, quality of service (throughput and delay) guarantee, and available capacity. The rectangular mapping constraint is not considered in this first step and is the main task in the second part.

To maximize the throughput, the resource allocations are sorted in a decreasing order (largest first). To allow space for variable parts of downlink and uplink maps, the resource allocations are mapped from right to left and from bottom to top of the DL subframe. Given a burst area, there are many possible height and width combinations that can contain that area; we choose the pair that is smallest in width. This allows the receiving MS to shut down its electronic circuit for most of the remainder of the DL subframe, thereby, saving energy. Note that in the worst case, *eOCSA can require maximum number of additional columns equal to the number of resource allocations in the frame. In practice, the required additional columns are much less and generally close to one*.

3.4.1 Algorithm Description

eOCSA consists of four steps as follows:

First, given a set of resource allocations $\{A_i\}$, we sort the set in a decreasing order and select the largest element to map.

Second step, *vertical mapping*, consists of mapping this resource allocation to the DL subframe. Given an area A_i, the algorithm maps the width-height pair (W_i, H_i) for the burst as follows:

$$W_i = \lceil A_i/H \rceil$$
$$H_i = \lceil A_i/W_i \rceil$$

Here, $\lceil \rceil$ denotes the ceiling function, and H is the maximum available height (DL subframe). With our 10 MHz Mobile WiMAX, H is 30 subchannels. Note that this ensures that the mapped region is bigger than the required allocation ($W_i \times H_i \geq A_i$) and that the rectangle has the minimum possible width (minimizes MS active time and energy).

After a resource allocation is mapped to DL subframe, some space may remain unallocated above the just mapped burst. In the third step, *horizontal mapping*, eOCSA algorithm tries to assign this space (which we call a **strip**) to the next largest element, say, jth allocation, that can be fitted in. In this step the region width is fixed, and it is used to determine the required height for the next largest element that can be fitted within this available region:

$$\text{Find largest } A_j, \text{ such that } A_j < W_i \times H_0$$
$$H_j = \lceil A_j/W_i \rceil$$
$$W_j = \lceil A_j/H_j \rceil$$

Here, $H_0 = H - H_i$ is the maximum available height in the strip. This step is repeated till either no space is left vertically, or there is no allocation that can be fitted in the available space. If no allocations can be found to fit, we move back to step 2 and select the next largest allocation to map in to DL subframe. The process of moving vertically and then horizontally from right to left and bottom to top is shown in Figure 10.

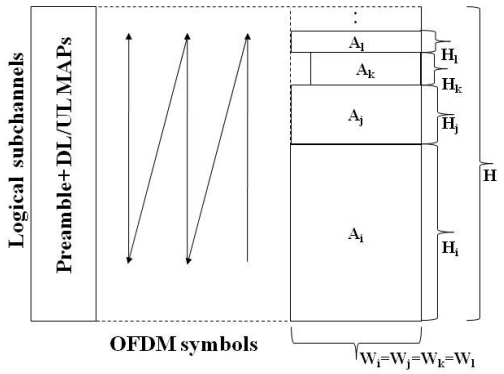

Figure 10: An example of mapping downlink burst using eOCSA

Figure 11 shows a flow chart of eOCSA algorithm, and a pseudo code showing nesting of various steps is presented in Figure 12. Notice that the computational complexity in worst case of eOCSA is in the order of O(n^2), where n is the number of resource allocations within a frame

$$Complexity = O(sorting) + O(allocation) = O(n \log n) + O(n^2)$$

Moreover, to achieve higher frame utilization, either vertical or horizontal mapping step can be added to eOCSA with one more level of complexity.

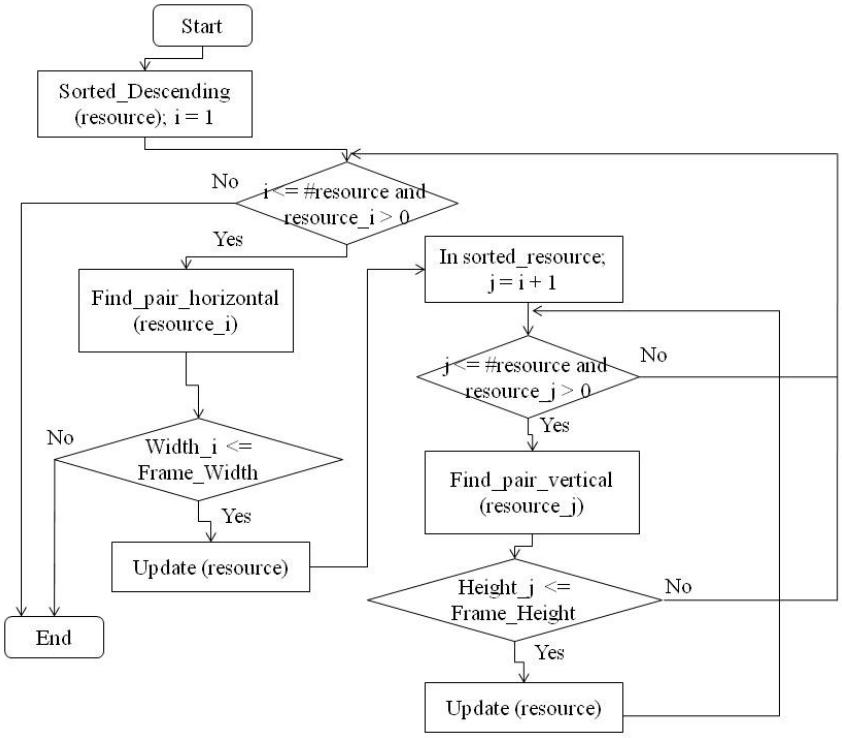

Figure 11: Flow chart of eOCSA burst mapping algorithm

```
sorted_allocations = Sort (resource_allocations)                    //1st step
FOR each unmapped element in sorted_allocations                     //2nd step
    Vertical_Mapping (&start_strip_i, &end_strip_i, &height_i)
    FOR each unmapped element in sorted_allocations                 //3rd step
        Horizontal_Mapping (start_strip_i, end_strip_i, height_i,
        &sub_height_j)
    END FOR
END FOR
```

Figure 12: Steps in eOCSA Algorithm

Note that without additional columns' consideration; actually eOCSA can also roll the additional columns needed for the current frame to the next frame before beginning the next frame mapping. However, this may cause an extra delay. Moreover, without the extra columns a priority mechanism needs to be applied. For example, the resource allocation with the highest priority is moved to the beginning of the mapping queue and so being mapped regardless of the largest size consideration. However, this may lead to more unused space.

3.4.2 eOCSA Example

In this section, we provide an example that helps explain our algorithm. The main idea is to strictly map all resource allocations in to Mobile WiMAX frame to meet the QoS requirements.

In this particular example, the scheduler makes an allocation decision for ten MSs in a Mobile WiMAX DL subframe. Table 14 shows a simple example for ten MSs randomly chosen. These MSs have been allocated A_1 through A_{10} by the scheduler as shown in the first row of the table. Basically, the sum of all resource allocations is 360 or 12×30.

Table 14: Example I: Ten random resource allocations

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
Allocations (slots)	71	127	27	99	15	3	5	9	2	2
Mapping Width,Height	3,24	5,26	1,27	4,25	3,3	4,1	5,1	4,3	4,1	3,1
Over allocation	1	3	0	1	0	1	0	3	2	1

First, the algorithm sorts all resource allocations in decreasing order of area (Step 1). That results in A_2, A_4, A_1, A_3, A_5, A_8, A_7, A_6, A_9, and A_{10}, respectively. The DL subframe area mapping is done from right to left and bottom to top. The largest resource allocation $A_2=127$ is chosen first. Applying step 2 we get a width of $\lceil 127/30 \rceil = 5$ columns and a height of $\lceil 127/5 \rceil = 26$. The rectangle 5×26 results in an over allocation of just 3 slots.

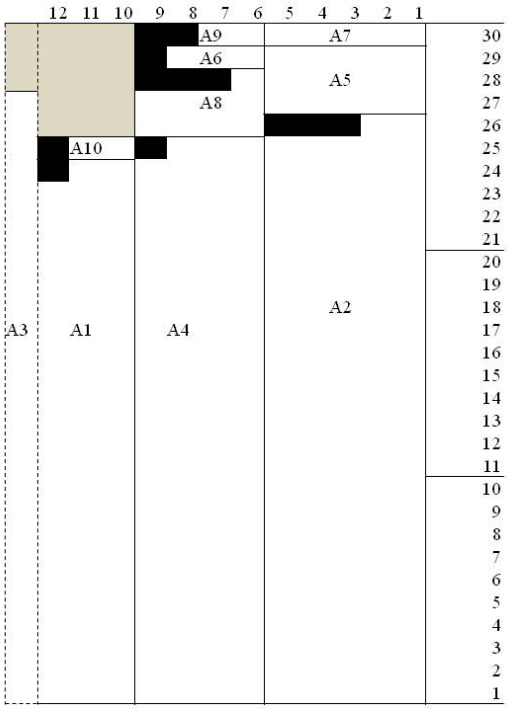

Figure 13: Example I: Two-dimensional downlink burst mapping

Mapping of A_2 leaves a strip of 5×4. In step 3, the algorithm chooses the next largest resource allocation that can fit in to this space. It is $A_5 = 15$. This is mapped as $5\times\lceil 15/5 \rceil$ or 5×3, resulting in no over allocation slot and leaves a space of 5×1 on top. A_7 can perfectly fit within this space with 5×1. Since there is no left-over space within this strip, we repeat step 2 by moving horizontally to the left. $A_4 = 99$, the next largest resource allocation, is mapped in to DL subframe in to a rectangle of width $\lceil 99/30 \rceil = 4$ and height $\lceil 99/4 \rceil = 25$. The rectangular mapping of 4×25 results in an over allocation of 1 and a left-over strip of 4×5 on the top of the mapping.

We move to step 3 to fill this 4×5 strip. At this time, $A_8 = 9$ being mapped to a rectangle of $\lceil 9/4 \rceil \times 4$ results in an over allocation of 3 and a left-over space of 4×2 on the top. Before we move to step 2, A_6 and A_9 are mapped and result in one and two over allocation slots respectively,

The next largest resource allocation, A_1, is mapped to 3×24 with one over allocation slot and a 3×5 left-over space. The only resource here, A_{10}, can fit within this space and results in one over allocation slot.

At this time although there is still 3×4 left-over space, $A_3 = 27$, the only one unmapped resource allocation, can't fit in this space, and unfortunately the algorithm reaches the maximum frame width. As a result, the algorithm needs to use one additional column and then map A_3 as 1×27, and finally the algorithm terminates.

In this particular example, the total of over allocation slots is 1 + 3 + 1 + 1 + 3 + 2 + 1= 12, and the total of unused slots is $1 \times 3 + 3 \times 5 = 18$ as shown by the dark and light shaded areas in Figure 13 respectively. The efficiency of the algorithm (percentage of space used) is 92.30% with over allocation slots and unused slots being counted as wasted.

3.5 Performance evaluation

In this section, we present numerical results comparing eOCSA with the ideal (full-search) algorithm.

To assure the QoS requirements, we assume that the scheduler strictly allocates the resource allocation in each frame, and the total resource allocation slots are 360 slots. We also assume each MS needs one burst. The number of MSs is randomly chosen from 1 to 49. The resource allocation for each MS is also randomly generated in the range from 1 to 360 slots. The over allocations and unused slots are averaged and normalized over 100 trials.

The results are shown in Figure 14 in terms of the normalized over allocations and unused slots versus the number of MSs. On average, the normalized over allocation and unused slots are 0.0088 and 0.0614 compared to the ideal mapping. These normalized numbers includes the additional columns required to guarantee mapping of all resource allocations.

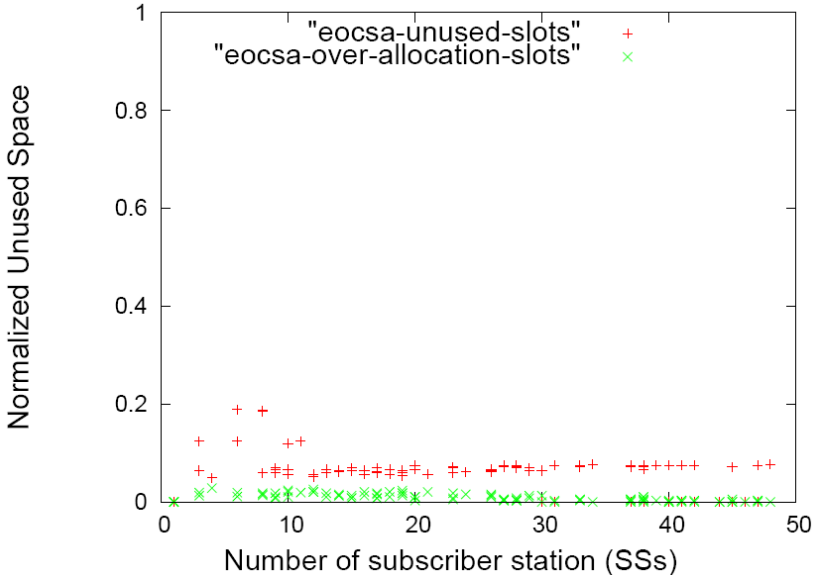

Figure 14: Normalized unused space vs. number of MSs (eOCSA)

In addition to comparing eOCSA with the ideal mapping, we also choose to compare eOCSA with the mapping algorithm by Takeo Ohseki et al. [18]. Each resource allocation is treated as a single burst. We could not compare eOCSA with other published algorithms for various reasons. For example, Yehuda Ben-Shimol et al. [20] provide no details of how to map the resources to unused spaces if their sizes are over multiple rows. Bacioccola et al [19], assume that it is possible to have more than one burst per subscriber. This violates our goal of minimizing burst overhead. Our analysis shows eOCSA can support more than 30 subscribers for the case where binary-tree full search supports only 8 subscribers [17].

With the same configuration, Figure 15 shows the results of the algorithm by Takeo Ohseki et's algorithm, again compared to the ideal mapping. On the average, the normalized unused slots and over allocation slots are 0.5198 and 0.0029, respectively. The average number of additional columns is 16.93 columns compared with only 0.93 (or 1) additional column for eOCSA. This behavior is because the unused slots are not considered. Note that if the scheduler debits the over allocations and unused slots from future allocations, the number of additional columns in some frames can be reduced.

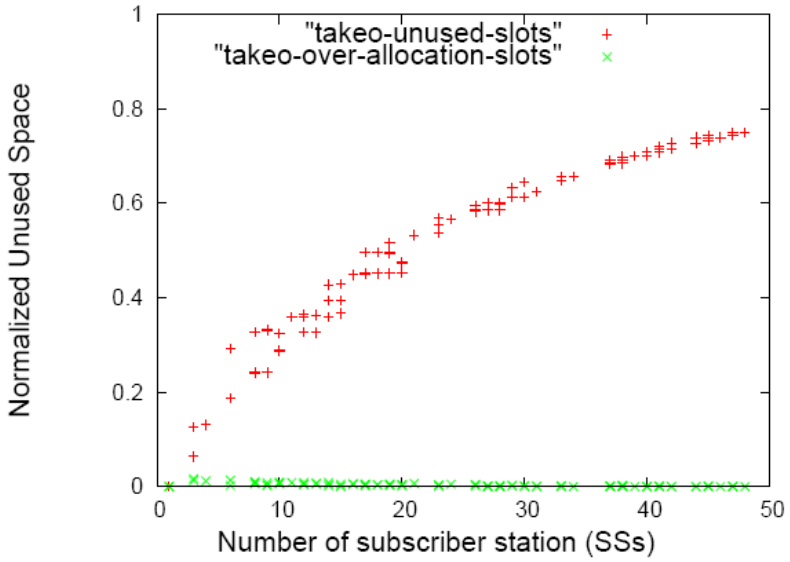

Figure 15: Normalized unused space vs. number of MSs (Takeo Ohseki et al.)

3.6 Section Summary

In this section, we introduced a heuristic algorithm called eOCSA for two-dimensional downlink burst mapping for IEEE 802.16e Mobile WiMAX networks. The algorithm meets the rectangular allocation constraint and achieves high throughput by minimizing left-over space, and optimizes the energy consumption at MS by minimizing the receive time for the MS.

To maximize the throughput, the algorithm considers the mapping in a decreasing order of the size of resource allocations. The mapping is done from right to left and bottom to top of the DL subframe. This allows space for variable portions of the DL-MAP and UL-MAP to be adjusted accordingly in the left part of the subframe.

Since eOCSA is a heuristic algorithm, there is a tradeoff between the throughput optimality and the computation complexity. The throughput may be improved with more complex algorithms such as by making a recursion for both vertical and horizontal mapping.

We also compared the performance of eOCSA with ideal full-search algorithm and found that eOCSA provides 93% throughput compared to a full search algorithm. On average, the number of columns required is one column more than that required to accommodate the sum of all allocations.

4 Bandwidth Request Mechanisms

Consider the BS scheduler. This scheduler has to decide slot allocation for traffic going to various MSs. It also has to grant slots to various MSs to be able to send the traffic upward. For downlink, the BS has complete knowledge of the traffic such as queue length and packet size to help make the scheduling decision.

For uplink traffic, the MSs need to send the Bandwidth Request (BWR) packets to the BS, which then decides how many slots are granted to each MS in the subsequent uplink subframes. Although originally the standard allowed BS to allocate the bandwidth per connection - Grant Per Connection (GPC) or per station - Grant Per Subscriber Station (GPSS), the latest version of the standard recommends only GPSS and leaves the allocation for each connection to the MS scheduler.

Basically, there are two types of BWRs: incremental or aggregate. There are a number of ways to request bandwidth. These methods can be categorized as implicit or explicit based on the need for polling as shown in Tables 15 and 16. As indicated in these two tables, the BWR mechanisms are: unsolicited request, poll-me bit, piggybacking, bandwidth stealing, codeword over Channel Quality Indicator Channel (CQICH), CDMA code-based BWR, unicast polling, multicast polling, broadcast polling and group polling. Table 17 provides a comparison of these mechanisms. The optimal way to request the bandwidth for a given QoS requirement is still in open research area [33, 34, 35, 36, 37, 38, 39, 40, 41, 42].

Mobile WiMAX offers many types of bandwidth request mechanisms and so obviously there is a trade-off between them between the flexibility of resource utilization and QoS requirements. For example, unicast polling can guarantee the delay; however, each polling can be wasted if there are no enqueued packets at the MS. On the other hand, multicast or broadcast polling may utilize the resource but the delay can not be guaranteed.

In order to guarantee the delay, polling in every frame is the best way to ensure the delay bound; however, this results in a significant polling overhead as mentioned earlier. Some research papers recommend polling in every video frame such as one every 20 ms [92] because video frame is generated every 20 ms; however, the polling optimization is still in an open research.

For ertPS, VoIP traffic, based on the Voice Modeling such as Adaptive Multi-Rate (AMR), only 33 bytes are sent every 20 ms during the active period and 7 bytes for inactive period. Also, for Enhanced Variable Rate Codec (EVRC), speech codec in CDMA networks, the silent period can be up to 60% [43, 44, 45]. Schedulers for voice users need to be aware of silent periods. Bandwidth is wasted if an allocation is made when there are no packets (which happens with UGS). With rtPS or ertPS in uplink direction, although the throughput can be optimized, the deadline is the main factor to be considered. The key issue is how to let the BS know whether there is a packet to transmit or not. The polling mechanism should be smart enough so that once there is traffic, the BS allocates a grant for the MS in order to send the bandwidth request and then transmit the packet within the maximum allowable delay. Moreover, BS does not need to allocate the bandwidth during the silent period. To indicate the end of a silent period, a MS can piggyback a zero bandwidth request, make use of a reserved bit in the MAC header to indicate their on/off states [45], or send a management message directly to the BS.

There is also a provision for a contention region and a CDMA bandwidth requests. The number of contention slots should be close to the number of connection enqueued so there is no extra delay in contention resolution. Obviously this region should be adaptively changed over time. Therefore, BS needs to make a prediction on how many MSs and/or connections are going to send the bandwidth request.

During the active period, the MS can use piggybacking or bandwidth stealing mechanisms in order to reduce the overhead and delay, and use contention region (WiMAX) or CDMA bandwidth request (Mobile WiMAX) when MS starts sending the packets. The scheduler should be aware of this and should make predictions accordingly.

In addition, there has been some research on how to optimize the backoff algorithm including backoff start and stop timers. In fact, the efficiency is just 33% with the random binary exponential backoff [33].

For UGS, the scheduler needs to be aware of the resource requirements and should be able to schedule the flows so that the resources can be optimized. For example, given ten UGS flows, each flow requiring 500 bytes every 5 frames, if only 2,500 bytes are allowed in one frame, all 10 flows can not start in the same frame. The scheduler needs to rearrange these flows in order to meet the customer satisfaction, especially delay-jitter. The problem gets more difficult when the UGS flows dynamically join and leave.

Video applications also have their own characteristics such as the size and the duration of Intra Coded Pictures (I-frame), Bi-directionally predicted pictures (B-frame) and Predicted Pictures (P-frame) frames for MPEG video. Basically I-frames are very large and occur periodically. Therefore, the scheduler can use this information to avoid overlapping among connections. The BS can delay accepting new connections so that the new connection's I-frames do not overlap with the exiting connections' I-frames [46].

For rtPS, there is also a strict or loose requirement of delay. If any packets are over the deadline, those packets will be dropped.

For non real-time traffic, nrtPS and best effort, fairness is an issue. The problem is whether the scheduler should be fair in a short-term or a long-term. For example, over one second, a flow can transmit 1 byte every 5 ms or 200 bytes every 1 second.

As a result, with the combination of all types of traffic and many types of bandwidth request mechanisms, WiMAX scheduler design is complicated.

Table 15: Implicit request of bandwidth request mechanism

Types	Mechanisms	Overhead	QoS classes
Unsolicited request	Periodically allocates bandwidth at setup stage	N/A	UGS and ertPS
Poll-me bit (PM)	Asks BS to poll for non-UGS connection	N/A (implicitly in MAC header)	UGS
Piggybacking	Piggyback BWR over any other MAC packets being sent to the BS.	Grant management (GM) subheader (2 bytes)	ertPS, rtPS, nrtPS and BE
Bandwidth stealing	Sends BWR instead of general MAC packet	BWR (6 bytes = MAC header)	nrtPS and BE
Contention region (WiMAX)	MSs use contention regions to send BWR.	Adjustable	ertPS, nrtPS and BE
Codeword over CQICH	Specifies codeword over CQICH to indicate the request to change the grant size	N/A	ertPS
CDMA code-based BWR (Mobile WiMAX)	MS chooses one of the CDMA request codes (256 codewords) from those set aside for bandwidth requests.	N/A	nrtPS and BE

Table 16: Explicit request of bandwidth request mechanism

Types	Mechanisms	Overhead	QoS classes
Unicast Polling	BS polls each MS individually and periodically.	BWR (6 bytes) per user	ertPS, rtPS, nrtPS and BE
Multicast Polling	BS polls a multicast group of MSs (using multicast CID)	BWR (6 bytes) per multicast	ertPS, nrtPS and BE
Broadcast Polling	BS polls all MSs.	Adjustable	ertPS, nrtPS and BE
Group Polling	BS polls a group of MSs periodically (not using multicast CID).	BWR (6 bytes) per group	ertPS, rtPS, nrtPS and BE

Table 17: Comparisons of bandwidth request mechanism

Types	Pros	Cons
Unsolicited request	No overhead and meet guaranteed latency of MS for real-time service	Wasted bandwidth if bandwidth is granted and the flow has no packets to send.
Poll me bit	No overhead	Still needs the unicast polling
Piggybacking	Do not need to wait for poll Less overhead; 2 bytes vs. 6 bytes	N/A
Bandwidth stealing	Do not need to wait for poll	6 bytes overhead
Contention Region	Reduced polling overhead	Need the backoff mechanism
Codeword over CQICH	Makes use of CQI channel	Limit number of bandwidth on CQICH
CDMA code-based BWR	Reduced polling overhead compared to contention region	Results in one more frame delay compared to contention region
Unicast Polling	Guarantees that MS has a chance to ask for bandwidth	More overhead (6 bytes per MS) periodically
Multicast Polling	Reduced polling overhead	Some MSs may not get a chance to request bandwidth; need contention resolution technique.
Broadcast Polling	Reduced polling overhead	Some MSs may not get a chance to request bandwidth; need contention resolution technique.
Group Polling	Reduced polling overhead	Some MSs may not get a chance to request bandwidth; need contention resolution technique.

5 WiMAX QoS and Scheduler [1]

In this section, we focus on Mobile WiMAX scheduling architecture: downlink/uplink at the base station and uplink at the subscriber. Several factors for scheduler design are discussed such as QoS assurance, throughput optimization, fairness, energy consumption, and implementation complexity. Finally, a brief survey of proposed WiMAX scheduling algorithms is presented. The algorithms are classified according to their channel awareness/unawareness.

Note that connection admission control (CAC) plays an important role in assuring the QoS requirements, and it needs to be designed along with the scheduler. Before joining the network, the subscribers need to have a permission from the BS to transmit data with a QoS agreement. The CAC basically maintains the current system load and QoS parameters for each existing connection. Then, it can make a decision if a new connection should be admitted and if admitted, what QoS the BS can provide. It should be obvious that if the CAC cannot support at least the minimum reserved rate for a new flow, that connection should be rejected. Otherwise, the QoS requirements of the existing flows can be broken. For example, instead of admitting another UGS flow, a BE flow is accepted if there is no way to guarantee the maximum allowable delay. We do not include the survey on CAC in this chapter; however, further information can be found in [81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91].

Scheduling is the main component of the MAC layer that helps assure QoS to various service classes. The scheduler works as a distributor to allocate the resources among MSs. The allocated resource can be defined as a number of slots and then these slots are mapped into a number of subchannels (each subchannel is a group of multiple physical subcarriers) and time duration (OFDM symbols). In OFDMA, the smallest logical unit for bandwidth allocation is a slot. The definition of slot depends upon the direction of traffic (downlink/uplink) and subchannelization modes. For example, in Partially Used Subchannelization (PUSC) mode in downlink, one slot is equal to twenty four subcarriers (one subchannel) for three OFDM symbols duration. In the same mode for uplink, one slot is fourteen subcarriers (one uplink subchannel) for two OFDM symbols duration.

The mapping process from logical subchannel to multiple physical subcarriers is called a permutation. PUSC, discussed above is one of the permutation modes. Others include Fully Used Subchannelization (FUSC) and Adaptive Modulation and Coding (band-AMC). The term band-AMC distinguishes the permutation from adaptive modulation and coding (AMC) MCS selection procedure. Basically there are two types of permutations: distributed and adjacent. The distributed subcarrier permutation is suitable for mobile users while adjacent permutation is for fixed (stationary) users. The detailed information again can be found in [5, 6].

After the scheduler logically assigns the resource in terms of number of slots, it may also have to consider the physical allocation, e.g., the subcarrier allocation. In systems with Single Carrier PHY, the scheduler assigns the entire frequency channel to a MS. Therefore, the main task is to decide how to allocate the number of slots in a frame for each user. In systems with OFDM PHY, the scheduler considers the modulation schemes for various subcarriers and decides the number of slots allocated. In systems with OFDMA PHY, the scheduler needs to take into consideration the fact that a subset of subcarriers is assigned to each user.

Scheduler designers need to consider the allocations logically and physically. Logically, the scheduler should calculate the number of slots based on QoS service classes. Physically, the scheduler needs to select which subchannels and time intervals are suitable for each user. The goal is to minimize power consumption, to minimize bit error rate and to maximize the total throughput.

There are three distinct scheduling processes: two at the BS - one for downlink and the other for uplink and one at the MS for uplink as shown in Figure 16. At the BS, packets from the upper layer are put into different queues, which ideally is per-CID queue in order to prevent head of line (HOL) blocking. However, the optimization of queue can be done and the number of required queues can be reduced. Then, based on the QoS parameters and some extra information such as the channel state condition, the DL-BS scheduler decides which queue to service and how many service data units (SDUs) should be transmitted to the MSs.

Since the BS controls the access to the medium, the UL-BS scheduler makes the allocation decision based on the bandwidth requests from the MSs and the associated QoS parameters. Several ways to send bandwidth requests were described earlier in Section I.G. Finally, the third scheduler is at the MS. Once the UL-BS grants the bandwidth for the MS, the MS scheduler decides which queue should use what part of that allocation. Recall that while the requests are per connections, the grants are per subscriber and the subscriber is free to choose the appropriate queue to service. The MS scheduler needs a mechanism to allocate the bandwidth in an efficient way.

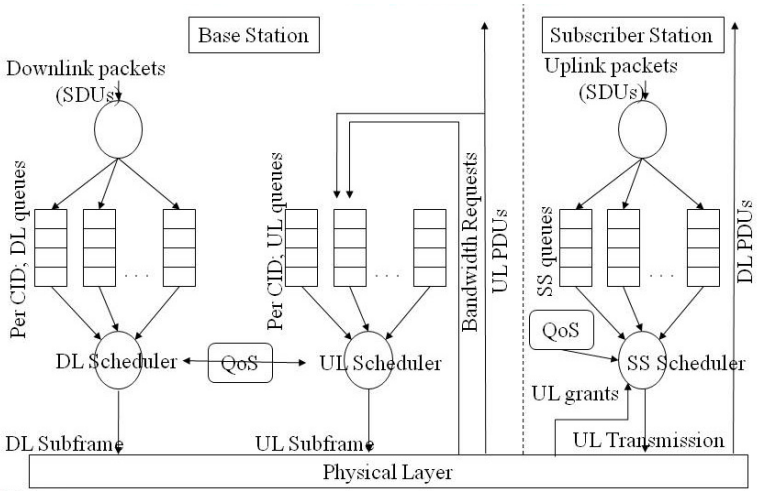

Figure 16: Component Schedulers at BS and MSs

5.1 Design Factors

To decide which queue to service and how much data to transmit, one can use a very simple scheduling technique such as First In First Out (FIFO). This technique is very simple but unfair. A little more complicated scheduling technique is Round Robin (RR). This technique provides the fairness among the users but it may not meet the QoS requirements. Also, the definition of fairness is questionable if the packet size is variable. In this section, we describe the factors that the scheduler designers need to consider. Then, we present a survey of recent scheduling proposals in Section 5.2.

QoS Parameters: The first factor is whether the scheduler can assure the QoS requirements for various service classes. The main parameters are the minimum reserved traffic, the maximum allowable delay and the tolerated jitters. For example, the scheduler may need to reschedule or interleave packets in order to meet the delay and throughput requirements. Earliest Deadline First (EDF) [47] is an example of a technique used to guarantee the delay requirement. Similarly, Largest Weighted Delay First (LWDF) has been used to guarantee the minimum throughput [48].

Throughput Optimization: Since the resources in wireless networks are limited, another important consideration is how to maximize the total system throughput. The metrics here could be the maximum number of supported MSs or whether the link is fully utilized. One of the best ways to represent throughput is using the goodput, which is the actual transmitted data not including the overhead and lost packets. The overheads include MAC overhead, fragmentation and packing overheads and burst overhead. This leads to the discussion of how to optimize the number of bursts per frame and how to pack or fragment the SDUs into MPDUs.

The bandwidth request is indicated in number of bytes. This does not translate straight forwardly to number of slots since one slot can contain different number of bytes depending upon the modulation

technique used. For example, with Quadrature Phase-Shift Keying 1/2 (QPSK1/2), the number of bits per symbol is 1. Together with PUSC at 10 MHz system bandwidth and 1024 Fast Fourier transform (FFT), that leads to 6 bytes per slot. If the MS asks for 7 bytes, the BS needs to give 2 slots thereby consuming 12 bytes. Moreover, the percentage of packet lost is also important. The scheduler needs to use the channel state condition information and the resulting bit error rate in deciding the modulation and coding scheme for each user.

Fairness: Aside from assuring the QoS requirements, the left-over resources should be allocated fairly. The time to converge to fairness is important since the fairness can be defined as short term or long term. The short-term fairness implies long term fairness but not vice versa [49].

Energy Consumption and Power Control: The scheduler needs to consider the maximum power allowable. Given the Bit Error Rate (BER) and Signal to Noise Ratio (SNR) that the BS can accept for transmitted data; the scheduler can calculate the suitable power to use for each MS depending upon their location. For mobile users, the power is very limited. Therefore, MS scheduler also needs to optimize the transmission power.

Implementation Complexity: Since the BS has to handle many simultaneous connections and decisions have to be made within 5 ms WiMAX frame duration [7], the scheduling algorithms have to be simple, fast and use minimum resources such as memory. The same applies to the scheduler at the MS.

Scalability: The algorithm should efficiently operate as the number of connections increases.

5.2 Classification of Schedulers

In this section, we present a survey of recent scheduler proposals for WiMAX. Most of these proposals focus on the scheduler at BS, especially DL-BS scheduler. For this scheduler, the queue length and packet size information are easily available. To guarantee the QoS for MS at UL-BS scheduler, the polling mechanism will be involved. Once the QoS can be assured, how to split the allocated bandwidth among the connections depends on the MS scheduler.

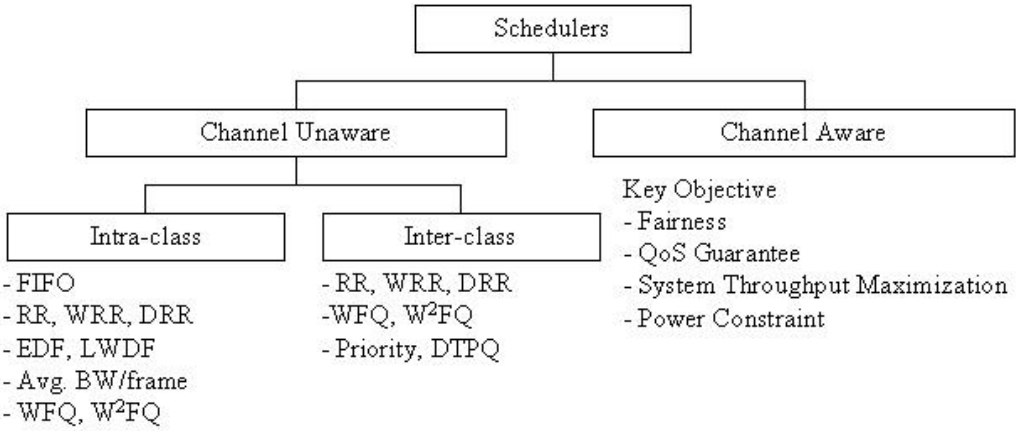

Figure 17: Classifications of WiMAX schedulers

Recently published scheduling techniques for WiMAX can be classified into two main categories: channel-unaware schedulers and channel-aware schedulers as shown in Figure 17. Basically, the channel-unaware schedulers use no information of the channel state condition in making the scheduling decision. In the discussion that follows, we apply the metrics discussed earlier in Section 5.1 to schedulers in each of these two categories.

Channel-unaware schedulers generally assume error-free channel since it makes it easier to prove assurance of QoS. However, in wireless environment where there is a high variability of radio link such as signal attenuation, fading, interference and noise, the channel-awareness is important. Ideally, scheduler designers should take into account the channel condition in order to optimally and efficiently make the allocation decision.

5.2.1 Channel-Unaware Schedulers

This type of schedulers makes no use of channel state condition such as the power level and channel error and loss rates. These basically assure the QoS requirements among five classes - mainly the delay and throughput constraints. Although, jitter is also one of the QoS parameters, so far none of the published algorithms can guarantee jitter. A comparison of the scheduling disciplines is presented in Table 18 and also the mappings between the scheduling algorithms and the QoS classes are shown in Table 19

Round Robin (RR) algorithm: Aside from FIFO, round-robin allocation can be considered the very first simple scheduling algorithm. RR fairly assigns the allocation one by one to all connections. The fairness considerations need to include whether allocation is for a given number of packets or a given number of bytes. With packet based allocation, stations with larger packets have an unfair advantage.

Moreover, RR may be non-work conserving in the sense that the allocation is still made for connections that may have nothing to transmit. Therefore, some modifications need to be made to skip the idle connections and allocate only to active connections. However, now the issues become how to calculate average data rate or minimum reserved traffic at given time and how to allow for the possibility that an idle connection later has more traffic than average? Another issue is what should be the duration of fairness? For example, to achieve the same average data rate, the scheduler can allocate 100 bytes every frame for 10 frames or 1000 bytes every 10^{th} frame.

Since RR cannot assure QoS for different service classes, RR with weight, Weighted Round Robin (WRR), has been applied for WiMAX scheduling [50, 51, 52]. The weight can be used to adjust for the throughput and delay requirements and can also used for inter-class priority. Basically the weights are in terms of queue length and packet delay or the number of slots. The weight is dynamically changed over time. In order to avoid the issue of missed opportunities, variants of RR such as Deficit Round Robin (DRR) or Deficit Weighted Round Robin (DWRR) can be used for the variable size packets [50]. The main advantage of these variations of RR is their simplicity. The complexity is *O(1)* compared to *O(log(N))* and *O(N)* for other fair queuing algorithms. Here, *N* is the number of queues.

Priority-based algorithm (PR): In order to guarantee the QoS to different classes of service, the priority-based scheme can be used in a WiMAX scheduler [53]. For example, the priority order can be: UGS, ertPS, rtPS, nrtPS and BE respectively. Or packets with the largest delay can be considered at the highest priority. Queue length can be also used to set priority level, e.g., more bandwidth is allocated to connections with longer queues [54].

The direct negative effect of this scheme is that it may starve some connections of lower priority service class. The throughput can be lower due to increased number of missed deadlines for the lower service classes' traffic. To mitigate this problem, Deficit Fair Priority Queuing (DFPQ) with a counter was introduced to maintain the maximum allowable bandwidth for each service class [55]. The counter decreases according to the size of the packets. The scheduler moves to another class once the counter falls to zero. DFPQ has also been used for inter-class scheduling [56].

Weighted Fair Queuing algorithm (WFQ): WFQ is an approximation of General Processor Sharing (GPS). WFQ does not make the assumption of infinitesimal packet size. Basically, each connection has its own FIFO queue and the weight can be dynamically assigned for each queue. The resources are shared in proportion of the weight. For data packets in wired networks with leaky bucket, an end-to-end delay bound can be provably guaranteed. With the dynamic change of weight, WFQ can be also used to guarantee the data rate. The main disadvantage of WFQ is the complexity, which could be *O(N)*.

To keep the delay bound and to achieve worst-case fairness property, a slight modification of the WFQ, Worst-case fair Weighted Fair Queueing (WF^2Q) was introduced. Similar to WFQ, WF^2Q uses a virtual time concept. The virtual finish time is the time GPS would have finished sending the packet. WF^2Q looks for the packet with the smallest virtual finishing time and whose virtual start time has already occurred instead of searching for the smallest virtual finishing time of all packets in the queue. The virtual start time is the time GPS starts to send the packet [57]. Note that in [57], the authors also introduced the concept of flow compensation with leading and lagging flow.

Similar to WRR, in achieving the QoS assurance (throughput, delay and jitter requirements), procedure to calculate the weight plays the important role. The weight can be based on several parameters. Aside from queue length and packet delay we mentioned above, the size of bandwidth request can be used to determine the weight of queue (the larger the size, the more the bandwidth) [58]. The ratio of a connection's average data rate to the total average data rate can be used to determine the weight of the connection [59]. The minimum reserved rate can be used as the weight [33]. The pricing can be also used as a weight [60]. Here, the goal is to maximize service provider revenue.

Delay-based algorithm: This set of schemes is specifically designed for real-time traffic such as UGS, ertPS and rtPS service classes, for which the delay bound is the primary QoS parameter and basically the packets with unacceptable delays are discarded. Earliest Deadline First (EDF) is the basic algorithm for scheduler to serve the connection based on the deadline. Largest Weighted Delay First (LWDF) [48] chooses the packet with the largest delay to avoid the missing the deadline.

Delay Threshold Priority Queuing (DTPQ) [61] was proposed for use when both real-time and non real-time traffic are present. A simple solution would be to assign higher priority to real-time traffic but that could harm the non real-time traffic. Therefore, urgency of the real-time traffic is taken into account only when the head-of-line (HOL) packet delay exceeds a given delay threshold. This scheme is based on the tradeoff of the packet loss rate performance of rtPS with average data throughput of nrtPS with a fixed data rate. Rather than fixing the delay, the author also introduced an adaptive delay threshold-based priority queuing scheme which takes both the urgency and channel state condition for real-time users adaptively into consideration [62].

Table 18: Comparison of Channel-Unaware Schedulers

Scheduling	Pros	Cons
RR	Very simple	Unfairness (variable packet size), cannot meet QoS requirements
WRR	Simple; meets the throughput guarantee	Unfairness (variable packet size)
DRR/DFRR	Simple, supports variable packet size	Not fair on a short time scale
Priority	Simple; meets the delay guarantee	Some flows may starve, lower throughput

DTPQ	Trades-off the packet loss rate of rtPS and average data throughput of nrtPS	Lower throughput
EDF	Meets the delay guarantee	Non-work conservative
LWDF	Guarantees the minimum throughput	N/A
WFQ	With proper and dynamic weight, guarantees throughput and delay, Fairness	Complex
WF^2Q	WFQ with worst-case fairness property	Complex

QoS Service Classes: Since the primary goal of a WiMAX scheduler is to assure the QoS requirements, the scheduler needs to support at least the five basic classes of services with QoS assurance. To ensure this, some proposed algorithms have indirectly applied or modified existing scheduling disciplines for each WiMAX QoS class of services because each class has its own distinct characteristics such as the hard-bound delay for rtPS and ertPS. Also, the schedulers have to consider both how to schedule within the class and between the classes. To schedule within the class, RR and WFQ are the common approaches for nrtPS and BE and EDF for UGS and rtPS [63]. The priority-based algorithm is commonly used for the scheduling between the classes. For example, the UGS and rtPS are given the same priority which is also the highest priority [58]. Table 19 shows recently proposed algorithms for each service class.

Table 19: Class based Schedulers

QoS classes	Scheduling Disciplines
UGS	Average Bandwidth in every frame, EDF
ertPS	Average Bandwidth in every frame, EDF, LWDF
rtPS	EDF, DTPQ, LWDF, W^2FQ
nrtPS	WFQ, W^2FQ
BE	RR, Equally distribute, WFQ, W^2FQ

To meet the QoS requirements, “two-step scheduler [64]” is a generic name for schedulers that try first to allocate the bandwidth to meet the minimum QoS requirements - basically the throughput in terms of the number of slots or subcarrier and time duration and delay constraints. Then, in the second step, they consider how to allocate the slots for each connection. This second step of allocating slots and subcarriers is still an open research area. The goal should to optimize the total goodput, to minimize the power and to optimize delay and jitter.

5.2.2 Channel-Aware Schedulers

The scheduling disciplines we discussed so far make no use of the channel state condition. In other words, they assume perfect channel condition, no loss and unlimited power source. However, due to the nature of wireless medium and the user mobility, these assumptions are not valid. For example, a MS may receive allocation but may not be able to transmit successfully due to high loss rate. In this section, we discuss the use of channel state conditions and the power constraints in scheduling decisions.

Basically, the BS downlink scheduler can use the Carrier to Interference and Noise Ratio (CINR) which is reported back from the MS via the CQI channel. For UL scheduling, the CINR is measured directly on previous transmissions from the same MS. Most of the purposed algorithms have the common assumption that the channel condition does not change within the frame period due to the way the MS

can send the feedback information. Also, it is assumed that the channel information is known at transmitter and receiver.

In general, schedulers favor the users with better channel quality since to exploit the multiuser diversity and channel fading, the optimal resource allocation is to schedule the user with the best channel. However, the schedulers also need to concern about other users to meet their QoS requirements such as the minimum reserved rate and so may introduce some compensation mechanisms. The schedulers basically use the property of multi-user diversity in order to increase the system throughput and to support more users.

The channel state condition can be directly used to help the scheduler make a better decision. A simple approach would be to give a priority based on the channel error, or perhaps the scheduler does not allocate any resources for the MS with high error rate because the packets would be dropped anyway. However, the minimum reserved rate needs to be maintained for these MSs. The concept of lagging flow has been used to represent the flow or connection that is behind its minimum reserved rate [66]. If and how the compensation mechanism should be put into consideration are still open questions.

The channel aware schemes can be classified into four classes based on the primary objective: fairness, QoS guarantee, system throughput maximization, or power optimization. Discussion on schemes with these objectives as follows.

Fairness: Unlike in Wireless LAN networks, WiMAX users pay for their QoS assurance. Thus, in [67] the argument of what is the level of QoS was brought on due to the question whether the service provider should provide a fixed number of slots. If the user happens to choose a bad location (such as the basement of a building on the edge of the cell), the provider will have to allocate a significant number of slots to provide the same quality of service as a user who is outside and near the base station. Since the providers have no control over the locations of users, they can argue that they will provide the same resources to all users and the throughput observed by the user will depend upon their location. A generalized weighted fairness concept, which equalizes a weighted sum of the slots and the bytes, was introduced in [65]. WiMAX equipment manufacturers can implement generalized fairness. The service providers can then set a weight parameter to any desired value and achieve either slot fairness or throughput fairness or some combination of the two.

Consider leading/lagging mechanisms. Intuitively if the lagging MSs should be compensated, the allocation should be from the bandwidth left-over either due to a low channel error rate or due to a flow not needing its allocation. It should not take the bandwidth from other well-behaved flows. In case, there is still left-over bandwidth, the leading flow can also gain the advantage of that left-over. However, another approach can be by taking some portion of the bandwidth from the leading flows to the lagging flows. Due to the unpredictable channel state condition, once there is an error, the credit history can be built based on the lagging flows and the scheduler allocates the bandwidth based on the ratio of their credits to theirs minimum reserved rates when the error rate is acceptable [68].

The design objective of Proportional Fairness Scheme (PFS) [68] is to maximize the long-term fairness. PFS uses the ratio of channel capacity (denoted as $W_i(t)$) to the long-term throughput (denoted as $R_i(t)$) in a given time window T_i of queue i as the preference metric instead of the current achievable data rate. $R_i(t)$ can be calculated by exponentially averaging the i^{th} queue's throughput in terms of T_i. Then, the user with the highest ratio of $W_i(t)/R_i(t)$ receives the transmission from the BS. The PFS was originally designed for wireless communication systems with saturated queues in which it is difficult to meet the QoS requirements especially the delay and minimum throughput guarantees [72, 70, 71].

QoS Guarantee: Modified Largest Weighted Delay First (M-LWDF) [66] can provide the QoS guarantee by ensuring minimum throughput guarantee for each user. And, it is provable that the throughput is optimal for LWDF [48]. The algorithm can achieve the optimal whenever there is a feasible set of minimal rates area. The algorithm explicitly uses both current channel condition and the state of the queue into account. The scheme serves the queue j for which "$\rho_i W_j(t) r_j(t)$" is maximal, where ρ_i is a constant which could be different for different service classes (the difficulty is how to find the optimal value of ρ_i). $W_i(t)$ can be either the delay of the head of line packet or the queue length. $r_i(t)$ is the channel capacity with respect to the flow j for each traffic class.

The channel state information is indirectly derived from the normalized channel gain in [67]. The channel gain is the ratio of the square of noise at the receiver and the variance of Additive White Gaussian Noise (AWGN). Then, the channel gain and the buffer state information are both used to decide which subcarriers should be assigned to each user. The buffers state information consists of head of line delay (HOL_delay), mean windowed arrival *(a)* and mean windowed throughput *(d)*. *"a"* and *"d"* are averaged over a sliding-window. This algorithm is the extension of M-LWDF, but schedules the users on each subcarrier during every time slot. For each subcarrier k, the user selection for the subcarrier is expressed by

$$i = \max\left[channel\_gain(i,k) \times HOL\_delay(i) \times \{a(i)/d(i)\}\right]$$

The efficiency of radio resource usage and the urgency (time-utility as a function of the delay) are the two factors for making the scheduling decision in Urgency and Efficiency based Packet Scheduling (UEPS) for rtPS with delay bound and nrtPS with throughput requirements [69]. UEPS outperforms both PFS and M-LWDF in terms of better throughput with QoS assurance. The scheduler first calculates the priority value for each user based on the urgency factor expressed by the time-utility function (denoted as $U'_i(t)$) × the ratio of the current channel state to the average (denoted as $R_i(t)/R'_i(t)$). After that, the subchannel is allocated to each selected user i where:

$$i = \max\left|U'_i(t)\right| \times \left[R_i(t)/R'_i(t)\right]$$

A modification of M-LWDF has been proposed to support multiple traffic classes [73]. The UEPS is not always efficient when the scheduler provides higher priority to nrtPS and BE traffic than rtPS, which may be near their deadlines. The modification of M-LWDF handles QoS traffic and BE traffic separately. The HOL packet's waiting time is used for QoS traffic metric and the queue length for BE traffic.

Packet Loss Fair Scheduling (PLFS) was introduced in [75] in order to employ both AMC and packet loss information for real-time traffic. The algorithm selects the highest priority user among all users and then the next highest priority if any subchannels are left. The selection rule is:

$$j = \max\left\{\left[A_k(t)/A_k\right] \times \left[PLR_i(t)/(PLR_{req,i} \times D_{\max,i})\right]\right\}$$

$$PLR_i(t) = PER_i(t) + PDR_i(t) \le PLR_{req,i}$$

Where, $A_k(t)$ is the state of channel in terms of MCS (modulation and coding scheme) level of user k at time t. PLR is the current packet loss rate; basically distributed proportionally among for all users. The PLR is defined in terms of the sum of the packet error rate (PER) derived from the channel impairments and packet dropping rate (PDR) derived from the packets exceeding the required maximum delay D_{max}. The $PLR_i(t)$ should be less than some threshold $PLR_{req,i}$ for user i in the equation. PLFS results in both

short-term and long-term fairness for diverse real-time traffic compared to just long-term fairness for M-LWDF.

System Throughput Maximization: A few schemes, e.g., [77], focus on maximizing the total system throughput. They use a heuristic approach of allocating a subchannel to the MS that can transmit the maximum amount of data on the subchannel. Suppose a BS has n users and m subchannels, let λ_i be the total uplink demand (bytes in a given frame) for its UGS connections, R_{ij} be the rate for MS_i on channel j (bytes/slot in the frame), N_{ij} be the number of slots allocated to MS_i on channel j, the goal of scheduling is to minimize the unsatisfied demand, that is,

$$\text{Minimize} \sum_{1\le i\le n}\left[\lambda_i - (\sum_{1\le j\le m} R_{ij}N_{ij})\right]$$

subject to the following constraints:

$$\sum_{1\le i\le n} N_{ij} < N'_j \text{ and } \sum_{1\le j\le m} R_{ij}N_{ij} \le \lambda_i$$

Here, N'_j is the total number of slots available for data transmission in the jth subchannel. A linear programming approach was introduced to solve this problem, but the main issue is the complexity, which is $O(n^3m^3N)$. Therefore, a heuristic approach with a complexity of only $O(nmN)$, was also introduced by assigning channels to the MSs that can transmit maximum amount of data.

The modulation scheme can be also used as the scheduling decision factor because this is an indirect indication of the channel state condition. For example, Binary Phase Shift Keying (BPSK) is used for the worst channel condition and Quadrature Amplitude Modulation 3/4 (QAM3/4) is for better channel conditions. Cross-Layer schedulers in [72, 73, 74] employ adaptive modulation and coding for all connections. Then, the priority is assigned dynamically for each connection based on the channel and service quality (basically the delay requirement).

It is possible to allocate the minimum number of slots derived from the minimum modulation scheme to each connection and then adjust the weight according to the exponent (p) of the instant modulation scheme over the minimum modulation scheme [78]. This scheme obviously favors the connections with better modulation scheme (higher p). Users with better channel conditions receive exponentially higher bandwidth. Two issues with this scheme are that additional mechanisms are required if the total slots are less than the total minimum required slots. And, under perfect channel conditions, connections with zero minimum bandwidth can gain higher bandwidth than those with non-zero minimum bandwidth.

Table 20: Comparison of Channel-Aware Schedulers

Scheduling	Pros	Cons	Traffic Classes
Proportional Fairness [68]	Long-term fairness	Lack of short-term fairness	nrtPS and BE
M-LWDF [66]	Support both realtime and non-realtime traffic	Difficult to choose the optimal constant for each type of traffic	rtPS, nrtPS and BEt
A generalized weighted fairness [65]	Weighted fair for temporal and throughput fairness	Issue on scalability	nrtPS and BE

M-LWDF with channel gain concept [67]	Derive channel gain as a allocation metric	N/A	BE
UEPS [69]	Maximize nrtPS throughput with assure rtPS QoS Higher throughput than PF and MLWDF	N/A	rtPS and nrtPS
A modified version of M-LWDF [73]	Support Multiple traffic classes	Trade-off: system throughput and QoS (delay guarantee)	rtPS and BE
PLFS [75]	Apply packet loss as a priority	N/A	rtPS
Exponential higher bandwidth allocation [76]	Maximize system throughput	Unfairness (favors users with better channel)	BE
Throughput Maximization [77]	Maximize system throughput	Unfairness, starvation	BE

Power Constraint: The purpose of this class of algorithms is not only to optimize the throughput but also to meet the power constraint. In general, the transmitted power is at a MS is limited. As a result, the maximum power allowable is introduced as one of the constraints. Least amount of transmission power is preferred for mobile users due to their limited battery capacities and also to reduce the radio interference.

Link-Adaptive Largest-Weighted-Throughput (LWT) algorithm has been proposed for OFDM systems [78]. LWT takes the power consumption into consideration. If assigning nth subcarrier to kth user at power $p_{k,n}$ results in a slot throughput of $b_{k,n}$, the algorithm first determines the best assignment that maximizes the link throughput ($max \sum b_{k,n}$). The bit allocation is derived from the approximation function of received SNR, transmission power and instantaneous channel coefficient. Then, the urgency was introduced in terms of the difference between the delay constraint and the waiting time of HOL packets. After that, the scheduler selects the HOL packet with the minimum value of the transmission time and the urgency. The main assumption here is that the packets are equal length.

Integer Programming (IP) approach has also been used to assign subcarriers [79]. However, IP complexity increases exponentially with the number of constraints. Therefore a suboptimal approach was introduced with fixed subcarrier allocation and bit loading algorithm. The suboptimal Hungarian or Linear Programming algorithm with adaptive modulation is used to find the subcarriers for each user and then the rate of the user is iteratively incremented by a bit loading algorithm, which assigns one bit at a time with a greedy approach to the subcarrier. Since this suboptimal and iterative solution is greedy in nature, the user with worse channel condition will mostly suffer.

A better and fairer approach could be to start the allocation with the highest level of modulation scheme. The scheduler has to try to find the best subcarriers for the users with the highest number of bits. This is also a greedy algorithm in a sense of the algorithm is likely to fill the un-allocated subcarriers to gain the power reduction. To minimize the transmit power, a horizontal and vertical swapping technique can also be used. The bits can be shifted horizontally among subcarriers of the same user if the power reduction is needed. Or, the swapping can be done vertically (swap subcarriers between users) to achieve the power reduction.

IEEE 802.16e standard [5, 6] defines Power Saving Class type II (PSC II) for real-time traffic for an energy-saving in which during the silent period in VoIP traffic the mobile user must wake up periodically in case it might have some data to send or receive; however, it may not always be the case. Thus, a hybrid energy-saving scheme was proposed in [80] using a truncated binary exponential algorithm to decide sleep cycle length for VoIP with silence suppression (voice packets are generated periodically during talk-spurt but not generated at all during the silent period).

6 Conclusions and Open Research Issues

In this chapter, we discussed several scheduling proposals for WiMAX and discussed key issues and design factors. The scheduler designers need to be thoroughly familiar with WiMAX characteristics such as the physical layer, frame format, registration process and so on as described in Section 1. The goals of the schedulers are basically to meet QoS guarantees for all service classes, to maximize the system goodput, to maintain the fairness, to consume less power, to have as less a complexity as possible and finally to ensure the system scalability. To meet all these goals is quite challenging since achieving one may require that we have to sacrifice the other.

Moreover, we explained how to compute the capacity of a WiMAX system and account for various overheads. We illustrated the methodology using three sample workloads consisting of Mobile TV, VoIP, and data users. Analysis such as the one presented in this chapter can be easily programmed in a simple program or a spread sheet and effect of various parameters can be analyzed instantaneously. This can be used to study the sensitivity to various parameters so that parameters that have significant impact can be analyzed in detail by simulation. This analysis can also be used to validate simulations.

We classified recent scheduling disciplines based on the channel awareness in making the decision. Well-known scheduling discipline can be applied for each class such as EDF for rtPS and WFQ for nrtPS and WRR for inter-class. With the awareness of channel condition and with knowledge of applications, schedulers can maximize the system throughput or support more users.

Optimization for WiMAX scheduler is still in ongoing research topic. There are several holes to fill in, for example, polling mechanism, backoff optimization, overhead optimization and so on. WiMAX can support reliable transmission with Automatic Retransmission Request (ARQ) and Hybrid ARQ (HARQ). Future research on scheduling should consider the use of these characteristics. The use of Multiple Input Multiple Output with multiple antennas to increase the bandwidth makes the scheduling problem even more sophisticated. Also, the multi-hops scenario also needs to be investigated for end-to-end service guarantees. With user mobility, future schedulers need to handle base station selection and hand off. All these issues are still open for research and new discoveries.

7 Bibliography

1. C. So-In, R. Jain, and A. Al-Tamimi, “Scheduling in IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey,” Submitted to IEEE Journal on Selected Areas in Communications (JSAC), January 2008.
2. R. Jain, C. So-In, and A. Al-Tamimi, “System Level Modeling of IEEE 802.16e Mobile WiMAX Networks: Key Issues,” IEEE Wireless Communications Magazine, November 2008.
3. C. So-In, R. Jain, and A. Al-Tamimi, “Capacity Estimations in IEEE 802.16e Mobile WiMAX networks”, Submitted to IEEE Wireless communication magazine, April 2008.
4. C. So-In, R. Jain, and A. Al-Tamimi, “eOCSA: An Algorithm for Burst Mapping with Strict QoS Requirements in IEEE 802.16e Mobile WiMAX Networks,” Submitted to WCNC 2008, September 2008.
5. IEEE P802.16Rev2/D2, “DRAFT Standard for Local and metropolitan area networks, Part 16: Air Interface for Broadband Wireless Access Systems,” 2094 pp, December 2007.
6. IEEE Std 802.16 - 2004, “Air Interface for Fixed Broadband Wireless Access Systems 895 pp, October 2004.
7. WiMAX Forum, “WiMAX System Evaluation Methodology V2.1,” 230 pp, July 2008, available at www.wimaxforum.org/technology/documents/

8. 3GPP2-TSGC5, HTTP and FTP Traffic Model for 1xEV-DV Simulations, 3GPP2-C50-EVAL-2001022-0xx, 2001.
9. L-E. Jonsson, G. Pelletier, and K. Sandlund, "Framework and four profiles: RTP, UDP, ESP, and uncompressed," RFC 3095, July 2001.
10. G. Pelletier, K. Sandlund, L-E. Jonsson, and M. West, "RObust Header Compression (ROHC): A Profile for TCP/IP (ROHC-TCP)," RFC 4996, January 2006.
11. C. Eklund, R-B. Marks, S. Ponnuswamy, K-L. Stanwood, and N-V. Waes, "WirelessMAN Inside the IEEE 802.16 Standard for Wireless Metropolitan Networks," 400 pp, May 2006.
12. G. Jeffrey, J. Andrews, A. Arunabha Ghosh, and R. Muhamed, "Fundamentals of WiMAX Understanding Broadband Wireless Networking," 496 pp, March 2007.
13. L. Nuaymi, "WiMAX: Technology for Broadband Wireless Access," 310 pp, March 2007.
14. H. Yaghoobi, "Scalable OFDMA Physical Layer in IEEE 802.16 WirelessMAN," Intel Technology Journal, vol. 8, no. 3, August 2004.
15. R. Srinivasan, T. Papathanassiou, and S. Timiri, "Mobile WiMAX VoIP Capacity System Level Simulations," WiMAX Forum, March 2007.
16. D. Ozdemir and F. Retnasothie, "WiMAX Capcity Estimation for Triple Play Services including Mobile TV, VoIP, and Internet," WiMAX Forum June 2007.
17. C. Desset, E.B. de Lima Filho, and G. Lenoir, "WiMAX Downlink OFDMA Burst Placement for Optimized Receiver Duty-Cycling", ICC 2007, pp. 5149-5154, June 2007.
18. T. Ohseki, M. Morita, and T. Inoue, "Burst Construction and Packet Mapping Scheme for OFDMA Downlinks in IEEE 802.16 Systems", GLOBECOM 2007, pp. 4307-4311, November 2007.
19. A. Bacioccola, C. Cicconetti, L. Lenzini, E.A.M.E. Mingozzi, and A.A.E.A. Erta, "A downlink data region allocation algorithm for IEEE 802.16e OFDMA", 6th International Confernece on Information, Communications & Signal Processing, pp. 1-5, December 2007.
20. Y. Ben-Shimol, I. Kitroser, and Y. Dinitz, "Two-dimensional mapping for wireless OFDMA systems", IEEE Transactions on Broadcasting,vol. 52, no. 3, pp. 388-396, September 2006.
21. Lodi, A., Martello, S., Monaci, M., "Two-dimensional packing problems: A survey", European Journal of Operational Research, vol. 141, no. 2, pp. 241-252, September 2002.
22. F. Clautiaux, J. Carlier, and A. Moukrim, "A new exact method for the two-dimensional orthogonal packing problem", European Journal of Operational Research, vol. 127, no. 3, pp. 1196-1121, December 2007.
23. E. Hopper and B.C.H Turton, "A Review of the Application of Meta-Heuristic Algorithms to 2D Strip Packing Problems", Artif. Intell. Rev. Journal, vol. 16, no. 4, pp. 257-300, December 2001.
24. M-R. Garey and D-S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness",. W.H. Freeman, 340pp, January 1979.
25. I. Schiermeyer, "Reverse-Fit: A 2-Optimal Algorithm for Packing Rectangles," In Proceedings of the Second Annual European Symposium on Algorithms, pp. 290 - 299, 1994.
26. S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," IEEE/ACM Transactions on Networking, vol. 7, no. 4, pp. 473-489, August 1999.
27. N. H. Vaidya, P. Bahl, and S. Gupta, "Distributed fair scheduling in a Wireless LAN," Transactions on Mobile Computing, vol. 4, no. 6, pp. 616-629, December 2005.
28. L. Tassiulas and S. Sarkar, "Maxmin fair scheduling in wireless networks," The 21st Annual Joint Conference of the IEEE Computer and Communications Societies, 2002. INFOCOM 2002, vol. 2, pp. 763-772, New York, NY, June 2002.
29. P. Bhagwat, P. Bhattacharya, A. Krishna, and S. K. Tripathi, "Enhancing throughput over Wireless LANs using channel state dependent packet scheduling," The 15th Annual Joint Conference of the IEEE Computer Societies, 1996. INFOCOM 1996, vol. 3, pp. 1133-1140, San Francisco, CA, March 1996.
30. S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," ACM/Baltzer Wireless Networks, vol. 8, no. 1, pp. 13-26, January 2002.
31. E. Jung and N. H. Vaidya, "An energy efficient MAC protocol for Wireless LANs," The 21st Annual Joint Conference of the IEEE Computer and Communications Societies, 2002. INFOCOM 2002, vol. 3, pp. 1756-1764, New York, NY, June 2002.
32. X. Zhang, Y. Wang, and W. Wang, "Capacity analysis of adaptive multiuser frequency-time domain radio resource allocation in OFDMA systems," IEEE International Symposium on Circuits and Systems, 2006. ISCAS 2006, pp. 4-7, Greece, May 2006.
33. M. Hawa and D. W. Petr, "Quality of service scheduling in cable and broadband wireless access systems," The 10th IEEE International Workshop on Quality of Service, 2002. IWQoS 2002, pp. 247-255, Miami Beach, MI, May 2002.
34. Q. Ni, A. Vinel, Y. Xiao, A. Turlikov, and T. Jiang, "Wireless Broadband Access: WiMAX and Beyond - Investigation of Bandwidth Request Mechanisms under Point-to-Multipoint Mode of WiMAX Networks," IEEE Communications Magazine, vol. 45, no. 5, pp. 132 -138, May 2007.
35. L. Lin, W. Jia, and W. Lu, "Performance Analysis of IEEE 802.16 Multicast and Broadcast Polling based Bandwidth Request," IEEE Wireless Communications and Networking Conference, 2007. WCNC 2007, pp. 1854-1859, Hong Kong, March 2007.
36. B. Chang and C. Chou, "Adaptive Polling Algorithm for Reducing Polling Delay and Increasing Utilization for High Density Subscribers in WiMAX Wireless Networks," The 10th IEEE Singapore International Conference on Communication systems, 2006. ICCS 2006, pp. 1-5, Singapore, October 2006.
37. P. Rastin, S. Dirk, and M. Daniel, "Performance Evaluation of Piggyback Requests in IEEE 802.16," IEEE 66th Vehicular Technology Conference, 2007. VTC 2007-Fall, pp. 1892-1896, Baltimore, MD, October 2007.

38. V. Alexey, Z. Ying, N. Qiang, and L. Andrey, "WLC22-4: Efficient Request Mechanism Usage in IEEE 802.16," IEEE Global Telecommunications Conference, 2006. GLOBECOM 2006, pp. 1-5, San Francisco, CA, November 2006.
39. O. Alanen, "Multicast polling and efficient voip connections in ieee 802.16 networks", The 10th ACM Symposium on Modeling, analysis and simulation of wireless and mobile systems, 2007. MSWiM 2007, pp. 289-295, Crete Island, Greece, October 2007.
40. A. Doha, H. Hassanein, and G. Takahara, "Performance Evaluation of Reservation Medium Access Control in IEEE 802.16 Networks," The 4th ACS/IEEE International Conference on Computer Systems and Applications, 2006. AICCSA 2006, pp.3 69-374, Dubai, UAE, March 2006.
41. A. Sayenko, O. Alanen, and T. Hamalainen, "Adaptive Contention Resolution for VoIP Services in the IEEE 802.16 Networks," IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007, pp. 1-7, Helsinki, Finland, June 2007.
42. J. Yan and G. Kuo, "Cross-layer Design of Optimal Contention Period for IEEE 802.16 BWA Systems," IEEE International Conference on Communications, 2006. ICC 2006, vol. 4, pp. 1807-1812, Istanbul, Turkey, June 2006.
43. H. Lee, T. Kwon, and D. Cho, "An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system," IEEE Communications Letters, vol. 9, no.8, pp. 691-693, August 2005.
44. H. Lee, T. Kwon, and D. Cho, "Extended-rtPS Algorithm for VoIP Services in IEEE 802.16 systems," IEEE International Conference on Communications, 2006. ICC 2006, vol. 5, pp. 2060-2065, Istanbul, Turkey, June 2006.
45. P. T. Brady, "A model for generating on-off speech patterns in two-way conversation," Bell System Technical Journal, pp. 2445-2472, September 1969.
46. O. Yang and J. Lu, "New scheduling and CAC scheme for real-time video application in fixed wireless networks," The 3rd IEEE Consumer Communications and Networking Conference, 2006. CCNC 2006, vol. 1, pp. 303-307, Las Vegas, NV January 2006.
47. M. Andrews, "Probabilistic end-to-end delay bounds for earliest deadline first scheduling," The 19th Annual Joint Conference of the IEEE Computer and Communications Societies, 2000. INFOCOM 2000, vol. 2, pp. 603-612, Israel, March 2000.
48. A. L. Stolyar and K. Ramanan, "Largest Weighted Delay First Scheduling: Large Deviations and Optimality," Annals of Applied Probability, vol. 11, no. 1, pp. 1-48, 2001.
49. C. E. Koksal, H. I. Kassab, and H. Balakrishnan, "An analysis of short-term fairness in wireless media access protocols", The 2000 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, pp. 118-119, Santa Clara, CA, June 2000.
50. C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," IEEE Network, vol. 20, no. 2, pp. 50-55, April 2006.
51. A. Sayenko, O. Alanen, J. Karhula, and T. Hamaainen, "Ensuring the QoS Requirements in 802.16 Scheduling," The 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems, 2006. MSWiM 2006, pp. 108-117, Terromolinos, Spain, October 2006.
52. A. Sayenko, O. Alanen, and T. Hamaainen, "Scheduling solution for the IEEE 802.16 base station," The International Journal of Computer and Telecommunications Networking, vol. 52, no.1, pp. 96-115, January 2008.
53. L. F. M. de Moraes and P. D. Jr. Maciel, "Analysis and evaluation of a new MAC protocol for broadband wireless access," International Conference on Wireless Networks, Communications and Mobile Computing, 2005. WIRELESSCOM 2005, vol. 1, pp. 107-112, Kaanapali Beach Maui, Hawaii, June 2005.
54. D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation for polling services in 802.16 broadband wireless networks," IEEE Global Telecommunications Conference, 2005. GLOBECOM 2005, vol. 6, pp. 5-9, St. Louis, MO, December 2005.
55. J. Chen, W. Jiao, and H. Wang, "A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode," IEEE International Conference on Communications, 2005. ICC 2005, vol. 5, pp. 3422-3426, Seoul, Korea, May 2005.
56. Y. Mai, C. Yang, and Y. Lin, "Cross-Layer QoS Framework in the IEEE 802.16 Network," The 9th International Conference on Advanced Communication Technology, 2007. ICACT 2007, vol. 3, pp. 2090-2095, Seoul, Korea, February 2007.
57. A. Iera, A. Molinaro, S. Pizzi, and R. Calabria, "Channel-Aware Scheduling for QoS and Fairness Provisioning in IEEE 802.16/WiMAX Broadband Wireless Access Systems," IEEE Network, vol. 21, no. 5, pp. 34-41, October 2007.
58. N. Liu, X. Li, C. Pei, and B. Yang, "Delay Character of a Novel Architecture for IEEE 802.16 Systems," The 6th International Conference on Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005, pp. 293-296, Dalian, China, December 2005.
59. K. Wongthavarawat and A. Ganz, "Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems," International Journal of Communication Systems 2003, vol. 16, no. 1, pp. 81-96, February 2003.
60. A. Sayenko, T. Hamalainen, J. Joutsensalo, and J. Siltanen, "An adaptive approach to WFQ with the revenue criterion," The 8th IEEE International Symposium on Computers and Communication, 2003. ISCC 2003, vol. 1, pp. 181-186, Kemer-Antalya, Turkey, July 2003
61. D. H. Kim and C. G. Kang, "Delay Threshold-based Priority Queueing Packet Scheduling for Integrated Services in Mobile Broadband Wireless Access System," High Performance Computing and Communications, 2005, LNCS 3726. HPCC 2005, pp. 305-314, Sorrento, Italy, 2005.
62. J. M. Ku, S. K. Kim, S. H. Kim, S. Shin, J. H. Kim, and C. G. Kang, "Adaptive delay threshold-based priority queueing scheme for packet scheduling in mobile broadband wireless access system," IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006, vol. 2, pp. 1142-1147, Las Vegas, NV, April 2006.
63. K. Wongthavarawat and A. Ganz, "IEEE 802.16 based last mile broadband wireless military networks with quality of service support," IEEE Military Communications Conference, 2003. MILCOM 2003, vol. 2, pp. 779-784, Boston, MA, October 2003.

64. A. K. F. Khattab and K. M. F. Elsayed, "Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks," International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2006. WoWMoM 2006, pp. 10-19, Buffalo, NY, June 2006.
65. C. So-In, R. Jain, and A. Al-Tamimi, "Generalized Weighted Fairness and its support in Deficit Round Robin with Fragmentation in IEEE 802.16 WiMAX", Submitted to Broadband Wireless Access Workshop, July 2008.
66. M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," IEEE Communications Magazine, vol. 39, no. 2, pp. 150-154, February 2001.
67. P. Parag, S. Bhashyam, and R. Aravind, "A subcarrier allocation algorithm for OFDMA using buffer and channel state information," IEEE 62nd Vehicular Technology Conference, 2005. VTC 2005-Fall, pp. 622-625, Dallas, TX, September 2005.
68. P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A Bandwidth-Efficient High-Speed Wireless Data Service for Nomadic Users," IEEE Communications Magazine, vol. 38, no. 7, pp. 70-77, July 2000.
69. S. Ryu, B. Ryu, H. Seo, and M. Shi, "Urgency and efficiency based wireless downlink packet scheduling algorithm in OFDMA system," IEEE 61st Vehicular Technology Conference, 2005. VTC 2005-Spring, vol. 3, pp. 1456-1462, Stockholm, Sweden, June 2005.
70. F. Hou, P. Ho, X. Shen, and A. Chen, "A Novel QoS Scheduling Scheme in IEEE 802.16 Networks," IEEE Wireless Communications and Networking Conference, 2007. WCNC 2007, pp. 2457-2462, Hong Kong, March 2007.
71. H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," IEEE Communications Letters, vol. 9, no. 3, pp. 210-212, March 2005.
72. Q. Liu, X. Wang, and G. B. Giannakis, "Cross-layer scheduler design with QoS support for wireless access networks," The 2nd International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, 2005. QSHINE 2005, pp. 8-15, Lake Buena Vista, FL, August 2005.
73. L. Wan, W. Ma, and Z. Guo, "A Cross-layer Packet Scheduling and Subchannel Allocation Scheme in 802.16e OFDMA System," IEEE Wireless Communications and Networking Conference, 2007. WCNC 2007, pp. 1865-1870, Hong Kong, March 2007.
74. S. A. Filin, S. N. Moiseev, M. S. Kondakov, A. V. Garmonov, D. H. Yim, J. Lee, S. Chang, and Y. S. Park, "QoS-Guaranteed Cross-Layer Transmission Algorithms with Adaptive Frequency Subchannels Allocation in the IEEE 802.16 OFDMA System," IEEE International Conference on Communications, 2006. ICC 2006, vol. 11, pp. 5103-5110, Istanbul, Turkey, June 2006.
75. S. J. Shin and B. H. Ryu, "Packet Loss Fair Scheduling Scheme for Real-Time Traffic in OFDMA Systems," ETRI Journal, vol. 26, no. 5, pp. 391-396, October 2004.
76. S. Shakkottai, R. Srikant, and A. Stolyar, "Pathwise Optimality and State Space Collapse for the Exponential Rule," In Proceeding of IEEE Int. Symposium on Inf. Theory 2002, pp. 379, 2002.
77. V. Singh and V. Sharma, "Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks," IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006, vol. 2, pp. 984-990, Las Vegas, NV, April 2006.
78. Y. J. Zhang and S. C. Liew, "Link-adaptive largest-weighted-throughput packet scheduling for real-time traffics in wireless OFDM networks," IEEE Global Telecommunications Conference, 2005. GLOBECOM 2005, vol. 5, pp. 5-9, St. Louis, MO, December 2005.
79. M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," IEEE Transactions on Broadcasting, vol. 49, no. 4, pp. 362-370, December 2003.
80. H. Choi and D. Cho, "Hybrid Energy-Saving Algorithm Considering Silent Periods of VoIP Traffic for Mobile WiMAX," IEEE International Conference on Communications, 2007. ICC 2007, pp. 5951-5956, Glasgow, Scotland, June 2007.
81. C. Jiang and T. Tsai, "Token bucket based CAC and packet scheduling for IEEE 802.16 broadband wireless access networks," The 3rd IEEE Consumer Communications and Networking Conference, 2006. CCNC 2006, vol. 1, pp. 183-187, Las Vegas, NV, January 2006.
82. D. Niyato and E. Hossain, "A Queuing-Theoretic and Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks," Transactions on Computers, vol. 55, no. 11, pp. 1473-1488, November 2006.
83. H. Wang, B. He, and D. P. Agrawal, "Admission control and bandwidth allocation above packet level for IEEE 802.16 wireless MAN," The 12th International Conference on Parallel and Distributed Systems, 2006. ICPADS 2006, vol. 1, pp. 6-13, Minneapolis, MN, July 2006.
84. W. S. Jeon and D. G. Jeong, "Combined Connection Admission Control and Packet Transmission Scheduling for Mobile Internet Services," IEEE Transactions on Vehicular Technology, vol. 55, no. 5, pp. 1582-1593, September 2006.
85. D. Niyato and E. Hossain, "Joint Bandwidth Allocation and Connection Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks," IEEE International Conference on Communications, 2006. ICC 2006, vol. 12, pp. 5540-5545, Istanbul, Turkey, June 2006.
86. C Qin, G. Yu, Z. Zhang, H. Jia, and A. Huang, "Power Reservation-Based Admission Control Scheme for IEEE 802.16e OFDMA Systems," IEEE Wireless Communications and Networking Conference, 2007. WCNC 2007, pp. 1831-1835, Hong Kong, March 2007.
87. B. Chang, Y. Chen, and C. Chou, "Adaptive Hierarchical Polling and Cost-Based Call Admission Control in IEEE 802.16 WiMAX Networks," IEEE Wireless Communications and Networking Conference, 2007. WCNC 2007, pp. 1954-1958, Hong Kong, March 2007.
88. O. Yang and J. Lu, "New scheduling and CAC scheme for real-time video application in fixed wireless networks," The 3rd IEEE Consumer Communications and Networking Conference, 2006. CCNC 2006, vol. 1, pp. 303-307, Las Vegas, NV, January 2006.

89. H. Wang, W. Li, and D. P. Agrawal, “Dynamic admission control and QoS for 802.16 wireless MAN,” Wireless Telecommunications Symposium, 2005. WTS 2005, pp. 60-66, Pomona, CA, April 2005.
90. B. Rong, Y. Qian, and K. Lu, “Downlink Call Admission Control in Multiservice WiMAX Networks,” IEEE International Conference on Communications, 2007. ICC 2007, pp. 5082-5087, Glasgow, Scotland, June 2007.
91. 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.
92. C. Cicconetti, A. Erta, L. Lenzini, and E. A. M. 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, November 2006.