Scheduling in IEEE 802.16e Mobile WiMAX Networks: Key Issues and a Survey

Chakchai So-In, *Student Member, IEEE,* Raj Jain, *Fellow, IEEE,* and Abdel-Karim Tamimi, *Student Member, IEEE*

***Abstract*—Interest in broadband wireless access (BWA) has been growing due to increased user mobility and the need for data access at all times. IEEE 802.16e based WiMAX networks promise the best available quality of experience for mobile data service users. Unlike wireless LANs, WiMAX networks incorporate several quality of service (QoS) mechanisms at the Media Access Control (MAC) level 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 implementer differentiation. Scheduling is, therefore, of special interest to all WiMAX equipment makers and service providers. This paper discusses the key issues and design factors to be considered for scheduler designers. In addition, we present an extensive 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.**

***Index Terms*—IEEE 802.16e, Mobile WiMAX, QoS, Resource Allocation, Scheduling, WiMAX.**

I. INTRODUCTION

IEEE 802.16 is 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 Fig. 1. It covers a metropolitan area of several kilometers and is also called WirelessMAN. Theoretically, a WiMAX base station can provide broadband wireless access in range 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 [1], [2] compared to 802.11a with 54 Mbps up to several hundred meters, EDGE (Enhanced Data Rates for Global Evolution) with 384 kbps to a few kms, or CDMA2000 (Code-Division Multiple Access 2000) with 2 Mbps for a few kms.

IEEE 802.16 standards group has been developing a set of 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 paper, the terms WiMAX and the IEEE 802.16 are used interchangeably.

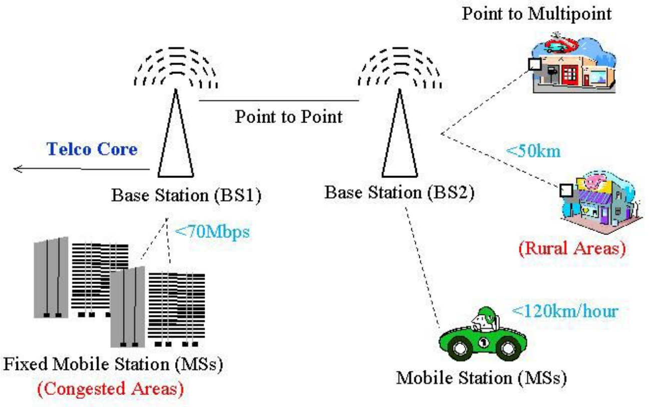

Fig. 1. WiMAX Deployment Scenarios

A. 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 and 8. Multiple QoS classes

Manuscript received 15 January 2008; revised 28 October 2008. This work was sponsored in part by a grant from Application Working Group of WiMAX Forum. "WiMAX," "Mobile WiMAX," "Fixed WiMAX," "WiMAX Forum," "WiMAX Certified," "WiMAX Forum Certified," the WiMAX Forum logo and the WiMAX Forum Certified logo are trademarks of the WiMAX Forum.

C. So-In, R. Jain and A. Tamimi is with the Department of Computer Science and Engineering, Washington Univeristy in St.Louis, St.Louis, MO, 63130 (e-mail: cs5, jain and aa7@cse.wustl.edu).

Digital Object Identifier 10.1109/JSAC.2009.090207.

suitable not only for voice but designed for a combination of data, voice and video services.

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. Thus, TDD will be the main focus of this paper. 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 (Media Access Control) 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 a longer distance, multipath and fading effects are also put into consideration. The Request/Grant mechanism is used for mobile stations (MSs) to access the media with a centralized control at base stations (BSs). WiMAX is a connection-oriented technology (with 16 bits connection identifier 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 for certain time slots reserved specifically for contention-based access.

To meet QoS requirements especially for voice and video transmission with the delay and delay jitter constraints, the key issue is how to allocate resource among the users not only to achieve those constraints but also to maximize goodput, 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 [3], Distributed Fair Scheduling [4], MaxMin Fair Scheduling [5], Channel State Dependent Round Robin (CSD-RR) [6], Feasible Earliest Due Date (FEDD) [7] and Energy Efficient Scheduling [8], 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 paper is to both provide a survey of recently proposed scheduling algorithms and give detailed information about WiMAX characteristics that need to be considered in developing a scheduler. Scheduler designers need to know all key issues and design decisions related to their designs. In the remainder of this section, we briefly describe the key issues that affect the scheduling decision. For example, in Section I.B, we provide a brief introduction to various WiMAX physical layers (PHYs) while we focus on the OFDMA based PHY in the rest of the paper. Section I.C gives an overview of WiMAX frame structure, downlink map (DL-MAP) and uplink map (UL-MAP) for OFDMA and some issues related to WiMAX frame. WiMAX QoS service classes and application service classes are discussed in Sections I.D and I.E. Finally, the Request/Grant mechanism and issues are explained for each QoS class in Section I.F. In Section II, we introduce the concepts of downlink (DL) and uplink (UL) schedulers and survey several recently proposed scheduling techniques. We classify these proposals based on the use of channel state information in Section III. Finally, the conclusions and the potential research on the scheduling techniques are presented in Section IV.

B. 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 mobile 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). Fig. 2 illustrates a schematic view of the three 802.16 PHYs discussed above. The details of these interfaces can be found in [1].

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 other hand, 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 [9]. 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 paper have been actively participating in the

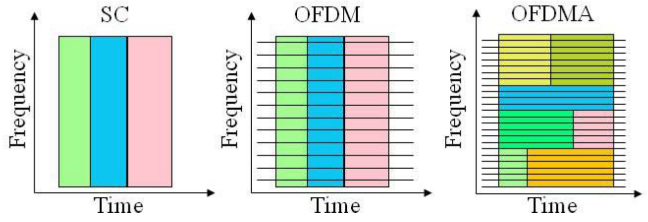

Fig. 2. IEEE 802.16 PHYs: SC, OFDM and OFDMA

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 paper.

C. WiMAX Frame Structure

IEEE 802.16 standard defines a frame structure as depicted logically in Fig. 3 and a mapping from burst to MPDU in Fig. 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. In Fig. 4, each burst can contain 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 and packing subheaders, grant management (GM) subheader (2 bytes) 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. Fig. 3 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. We discuss five main issues related to the frame structure below; namely, number of bursts, two dimensional rectangular mapping for downlink subframe, MPDU size, fragmentation and packing considerations.

First, number of bursts 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

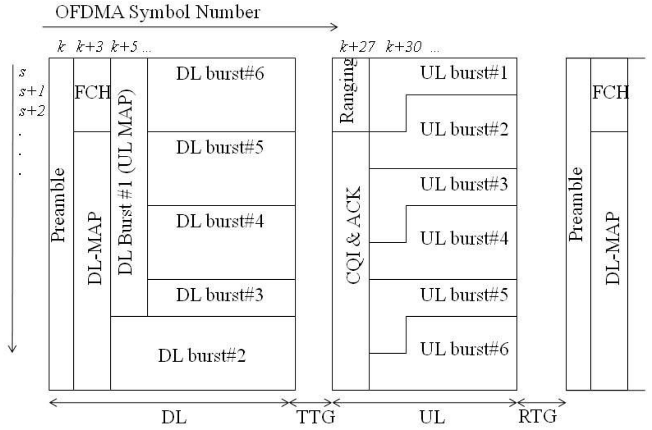

Fig. 3. A Sample OFDMA Frame Structure

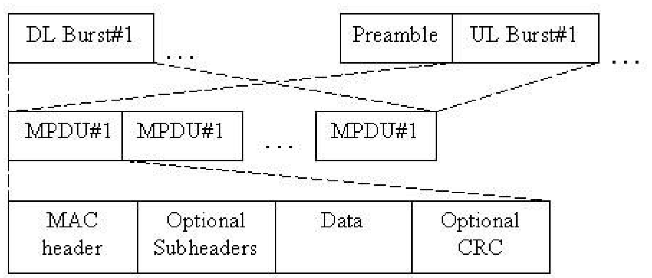

Fig. 4. MPDU frame format

transmission starts at a particular slot and continues until the end of UL subframe. Then it continues on the next subchannel horizontally. 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, it increases DL-MAP overhead. The standard also allows more than one connection packed into one burst with the increased DL-MAP IE size. It is even 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. 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.

Second, 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 another given shape.

The mapping problem in WiMAX is different from the original bin packing in that: first there are no fixed length and width limitations. Instead only bin sizes are given. Second, with increasing number of bursts (number of bins), the other end of the big bin (left side of the WiMAX frame) in which

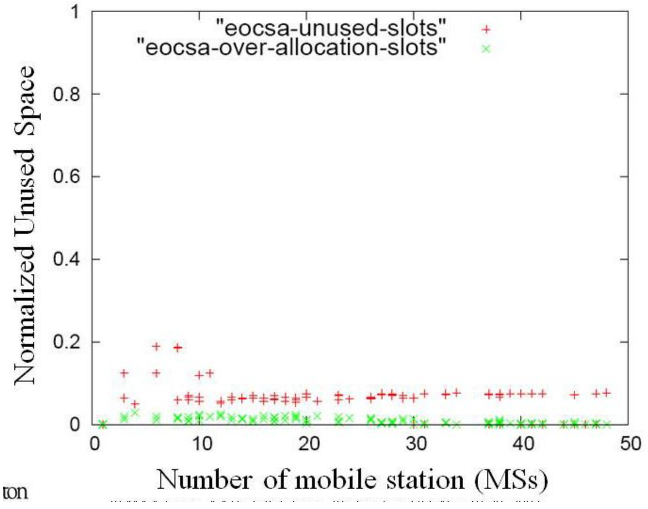

Fig. 5. Normalized unused space vs. number of MSs for eOCSA [15]

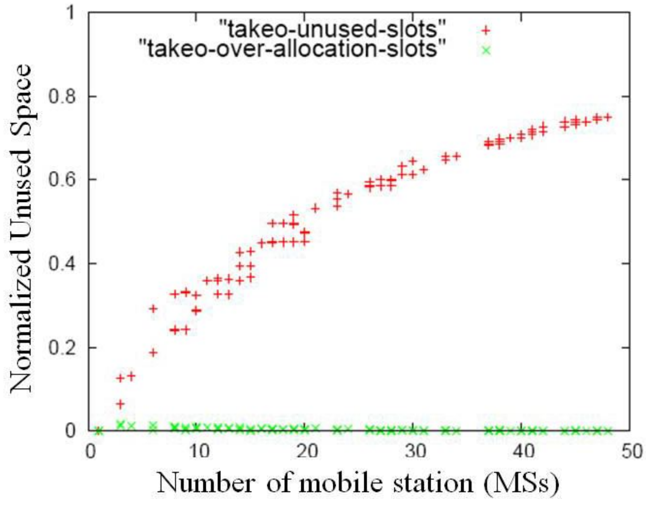

Fig. 6. Normalized unused space vs. number of MSs for Takeo Ohseki *et al*'s algorithm

small bins are fitted also changes to allow increasing size of the variable part of DL_MAP.

In Table I, we compare and summarize several proposed mapping algorithms for WiMAX networks. Notice that each algorithm has its own pros and cons and complexity trade-offs. Also, the performance trade-off of increasing DL_MAP overhead vs. number of bursts has not yet been studied in the literature.

With rectangular mapping, a subscriber is usually allocated more slots than its demand. Also, some left-over spaces are too small to allocate to any users. These two types of wasted slots are called over-allocation and unused slots, respectively. We present simulation results comparing an algorithm called "eOCSA" that we have developed with that proposed by Takeo Ohseki *at el* [10]. The comparison is limited to these two algorithms for various reasons. For example, Yehuda Ben-Shimol *et al*. [11] provide no details of how to map the resources to unused spaces if their sizes are over multiple rows. Andrea Bacioccola *et al*. [12], assume that it is possible to have more than one burst per subscriber. This violates our goal of minimizing burst overhead. The binary-tree full search can support only 8 subscribers [13] and so it is not of any practical use.

With Partially Used Subchannelization (PUSC) mode, 10 MHz channel, and DL:UL ratio of 2:1, the DL frame consists of 14 columns of 30 slots each or 420 slots [14]. Assuming we reserve the first two columns for DL/UL MAPs, we can allocate the remaining 12 columns resulting in 360 slots per frame for the users. We also assume that each MS needs one burst. The number of MSs is randomly chosen from 1 to 49. The resource demand for each MS is also randomly generated so that the total demand is 360 slots. The over allocations and unused slots are averaged over 100 trials.

The results for eOCSA are shown in Fig. 5 in terms of the normalized over allocations and unused slots versus the number of MSs. The normalization is done by dividing the total space required to map the demands. On average, the normalized over allocation and unused slots are 0.0088 and 0.0614, respectively.

Fig. 6 shows the corresponding results for the algorithm by Takeo Ohseki *et al*. On the average, the normalized over allocation slots and unused slots are 0.0029 and 0.5198, respectively. Notice that they have significantly higher unused slots than eOCSA because they do not allocate unused spaces below or above an allocated user's burst. On the other hand, eOCSA has a slightly higher over-allocation because we try to fit rectangles in these small unused spaces. More details on this other tradeoffs in burst mapping are presented in [15].

Third, number of MPDUs in a burst and their sizes are important. Each MPDU has 6 bytes MAC header (See Fig. 3). One can have large MPDU, but then the MPDU loss probability due to bit errors is higher. On the other hand, the MPDU header is significant if there are many small MPDUs. Note that in [16], the estimation of optimal MPDU size (L) was drawn.

$$L = \frac{O}{2} - \frac{\sqrt{(O\ln(1-E))^2 - 4BO\ln(1-E)}}{2\ln(1-E)}$$

Here, O is the overhead measured by the number of bytes in headers, subheaders and CRC. E is the block error rate (BLER) after forward error correction (FEC). B stands for FEC block size in bytes.

Depending on number of retransmission or loss, dynamic change of MPDU size was introduced in [17] by typically adding more FEC for MPDU in poor channel situation.

Notice that WiMAX also supports fragmentation and packing. Their overheads should be also taken into account. Consider fragmentation. Deficit round robin with fragmentation was brought up in [18]. Without the fragmentation consideration, the WiMAX frame is underutilized since it may be possible that within a particular frame, all full packets can not be transmitted. In [14], we have shown that with proper packing especially for small packets such as voice packets, the number of users can be increased significantly; however, packet delays can also increase.

TABLE I
TWO-DIMENSIONAL RECTANGULAR MAPPING FOR DOWNLINK

	Algorithm Descriptions	Pros	Cons	Complexity
Takeo Ohseki *et al.* [10]	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. Do not consider a variable part of DL-MAP	$O(N)$ + $O(Searching)$ + $O(Compaction)$
Yehuda Ben-Shimol *et al.* [11] (Raster Algorithm)	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 Do not consider a variable part of DL-MAP	N/A
Andrea Bacioccola *et al.* [12]	Allocate from right to left and bottom to top	Optimize frame utilization Consider a variable part of DL-MAP	They map a single allocation in to multiple rectangular areas that may result in increased DL MAP elements overhead	N/A
Claude Desset *et al.* [13]	Binary-tree full search algorithm	Optimize frame utilization	Only 8 users at maximum can be supported Do not consider a variable part of DL-MAP	N/A
Chakchai So-In *et al.* [15]	Allocate from right to left and bottom to top with the least width first vertically and the least height first horizontally for each particular burst.	Optimize frame utilization Consider a variable part of DL-MAP	Lacks of detail simulation	$O(N^2)$
Ting Wang *et al.* [19]	Apply the less flexibility first (LFF) allocation (select the area with the least free space edge)	Consider all possible mapping pair	Fixed resource reserved for DL-MAP	$O(N^2)$

D. 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 minimum throughput requirement and delay/jitter constraints. Table II 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.

Note that for non-real-time traffic, traffic priority is also one of the QoS parameters that can differentiate among different connections or subscribers within the same service class.

Consider bandwidth request mechanisms for uplink. UGS, ertPS and rtPS are real-time traffic. UGS has a static allocation. ertPS is a combination of UGS and rtPS. Both UGS and ertPS can reserve the bandwidth during setup. Unlike UGS, ertPS allows all kinds of bandwidth request including contention resolution. rtPS can not participate in contention resolution. For other traffic classes (non real-time traffic), nrtPS and BE, several types of bandwidth requests are allowed such as piggybacking, bandwidth stealing, unicast polling and contention resolution. These are further discussed in Section I.F.

E. Application Traffic Models

WiMAX Forum classifies applications into five categories as shown in Table III. Each application class has its own characteristics such as the bandwidth, latency and jitter constraints

TABLE II
COMPARISON OF WIMAX QOS SERVICE CLASSES

QoS	Pros	Cons
UGS	No overhead. Meet guaranteed latency 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.

TABLE III
WIMAX APPLICATION CLASSES [2]

Classes	Applications	Bandwidth Guideline		Latency Guideline		Jitter Guideline		QoS Classes
1	Multiplayer Interactive Gaming	Low	50 kbps	Low	< 25 ms	N/A		rtPS and UGS
2	VoIP and Video Conference	Low	32-64 kbps	Low	< 160 ms	Low	< 50 ms	UGS and ertPS
3	Streaming Media	Low to high	5 kbps to 2 Mbps	N/A		Low	< 100 ms	rtPS
4	Web Browsing and Instant Messaging	Moderate	10 kbps to 2 Mbps	N/A		N/A		nrtPS and BE
5	Media Content Downloads	High	> 2 Mbps	N/A		N/A		nrtPS and BE

in order to assure a good quality of user experience. The traffic models for these applications can be also found in [2].

F. Request/Grant Mechanism

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 lengths and packet sizes to help make the scheduling decisions.

For uplink traffic, the MSs need to send 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 BWR: 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 IV and V. 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 VI provides a comparison of these mechanisms. The optimal way to request the bandwidth for a given QoS requirement is still in open research area [20], [21], [22], [23], [24], [25], [26], [27], [28], [29].

We briefly discuss the issue of bandwidth request mechanisms for each QoS class. Obviously there is a trade-off between the flexibility of resource utilization and QoS requirements. For example, unicast polling can guarantee the delay; however, resources 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.

First consider UGS. There is no polling (static allocation) but 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 2500 bytes are allowed in one frame, all 10 flows can not start in the same frame. The scheduler needs to rearrange (phase) these flows in order to meet the delay-jitter while maximizing frame utilization. The problem gets more difficult when the UGS flows dynamically join and leave.

Consider the delay requirements. 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 33 ms [30] because video frames are generated every 30-40 ms. Without the arrival information of packets, it is difficult for BS to guarantee the delay requirements. As a result, the polling optimization is still in an open research topic.

Second, consider rtPS. There is a strict or loose requirement of delay. If any packets are over the deadline, those packets will be dropped.

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 phase new connections so that the new connection's I-frames do not overlap with the exiting connections' I-frames [31].

Third, consider ertPS. This service is used for VoIP traffic which has active and silent periods. As an example, if Adaptive Multi-Rate (AMR) coding is used, only 33 bytes are sent every 20 ms during the active periods and 7 bytes during silent periods. The silent period can be up to 60% [32], [33], [34]. Schedulers for voice users need to be aware of these 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 [32], or send a management message directly to the BS.

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

There is also a provision for a contention region and for 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.

In addition, recent research shows how to optimize the backoff algorithm including backoff start and stop timer [28]. In fact, the efficiency is just 33% with the random binary exponential backoff [35].

Fourth, nrtPS. The only constraint for nrtPS is the minimum guaranteed throughput. Polling is allowed for this service. Some proposed schemes recommend polling intervals of over 1 second [30]. The polling should be issued if and only if the average rate which is calculated from Proportional Fairness (PF) is less than the minimum reserved rate [36]. We will describe PF in Section III.B.1.

Finally, best effort. All bandwidth request mechanisms are allowed for BE but contention resolution is most commonly used. The main issue for BE is fairness. 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. Also, the scheduler should prevent starvation.

As can be seen from this discussion, with the combination of different types of traffic and many types of bandwidth request mechanisms, WiMAX scheduler design is complicated.

II. Scheduler

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 the 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 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 [1].

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 Fig. 7. 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 second scheduler - the UL-BS scheduler - makes the allocation decision based on the bandwidth requests from the MSs and the

TABLE IV
IMPLICIT BANDWIDTH REQUEST MECHANISMS

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 non UGS connections	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 from those set aside for bandwidth requests.	Six subchannels over 1 OFDM symbol for up to 256 codes	nrtPS and BE

TABLE V
EXPLICIT BANDWIDTH REQUEST MECHANISMS

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.	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.	BWR (6 bytes) per group	ertPS, rtPS, nrtPS and BE

TABLE VI
COMPARISONS OF BANDWIDTH REQUEST MECHANISMS

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, Broadcast and Group Polling	Reduced polling overhead	Some MSs may not get a chance to request bandwidth; need contention resolution technique.

associated QoS parameters. Several ways to send bandwidth requests were described earlier in Section I.F. Finally, the third scheduler is at the MS. Once the UL-BS grants the bandwidth for the MS, the MS scheduler decides which queues should use 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.

A. 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 III.

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) [37] is an example of a technique

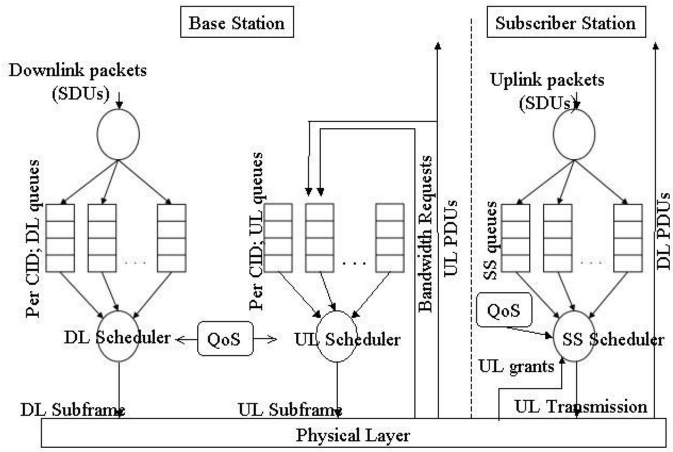

Fig. 7. Component Schedulers at BS and MSs

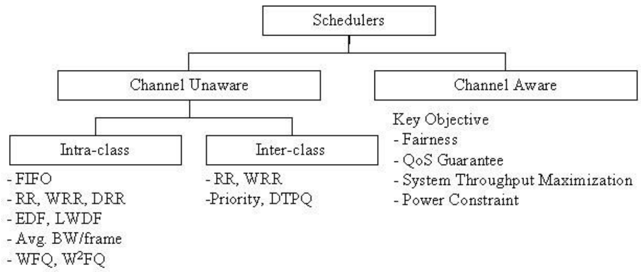

Fig. 8. Classifications of WiMAX schedulers

used to guarantee the delay requirement. Similarly, Largest Weighted Delay First (LWDF) has been used to guarantee the minimum throughput [38].

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 [39].

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 [1], 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.

III. 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 is involved. Once the QoS can be assured, how to split the allocated bandwidth among the connections depends on the MS scheduler.

Recently published scheduling techniques for WiMAX can be classified into two main categories: channel-unaware schedulers and channel-aware schedulers as shown in Fig. 8. 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 II.A 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.

A. Channel-Unaware Schedulers

This type of schedulers makes no use of channel state conditions 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 VII and also the mappings between the scheduling algorithms and the QoS classes are shown in Table VIII.

1) Intra-class Scheduling: Intra-class scheduling is used to allocate the resource within the same class given the QoS requirements.

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 any 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 10th frame.

Since RR cannot assure QoS for different service classes, RR with weight, Weighted Round Robin (WRR), has been applied for WiMAX scheduling [40], [41], [42]. The weights can be used to adjust for the throughput and delay requirements. Basically the weights are in terms of queue length and packet delay or the number of slots. The weights are 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 [40]. 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.

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^{2}Q) was introduced. Similar to WFQ, WF^{2}Q uses a virtual time concept. The virtual finish time is the time GPS would have finished sending the packet. WF^{2}Q, 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 [43]. Note that in [43], the authors also introduced the concept of flow compensation with leading and lagging flows.

In achieving the QoS assurance, procedure to calculate the weight plays an important role. The weights 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) [44]. 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 [45]. The minimum reserved rate can be used as the weight [35]. The pricing can be also used as a weight [46]. Here, the goal is to maximize service provider revenue.

Delay-based algorithms: 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) [38] chooses the packet with the largest delay to avoid missing its deadline.

Delay Threshold Priority Queuing (DTPQ) [47] 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 [48].

Note that variants of RRs, WFQs and delay based algorithms can resolve some of the QoS requirements. However, there are no published papers considering the tolerated delay jitter in the context of WiMAX networks. Especially for UGS and ertPS, the simple idea is to introduce a zero delay jitter by the fragmentation mechanism. Basically, BS transfers the last fragmented packet at the end of period. However, this fragmentation increases the overhead and also requires fixed buffer size for two periods. Compared to EDF, this simple technique may require more bursts. This needs to be investigated further.

2) Inter-class Scheduling: As shown in Fig. 8, RR, WRR and priority-based mechanism have been applied for inter-class scheduling in the context of WiMAX networks. The main issue for inter-class is whether each traffic class should be considered separately, that is, have its own queue. For example, in [49] rtPS and nrtPS are put into a single queue and moved to the UGS (highest priority) queue once the packets approach their deadline. Similarly in [50] UGS, rtPS and ertPS queues are combined to reduce the complexity. Another issue here is how to define the weights and/or how much resources each class should be served. There is a loose bound on service guarantees without a proper set of weight values.

Priority-based algorithm (PR): In order to guarantee the QoS to different classes of service, priority-based schemes can be used in a WiMAX scheduler [50], [51], [52]. 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 the priority level, e.g., more bandwidth is allocated to connections with longer queues [53].

The direct negative effect of priority is that it may starve some connections of lower priority service classes. 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 [54]. 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 [55].

To sum up, 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. Each class has its own distinct characteristics such as the hard-bound delay for rtPS and ertPS. Most proposed algorithms have applied some basic algorithms proposed in wired/wireless networks to WiMAX networks such as variations of RR and WFQ. For example, to schedule within a class, RR and WFQ are common approaches for nrtPS and BE and EDF for UGS and rtPS [52], [56]. The priority-based algorithm is commonly used for scheduling between the classes. For example, UGS and rtPS are given the same priority which is also the highest priority [44].

Moreover, "two-step scheduler [57]" 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, especially in WiMAX networks (OFDMA-based) 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 be to optimize the total goodput, to maintain the fairness, to minimize the power and to optimize delay and jitter.

B. 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 a high loss rate. In this section, we discuss the use of channel state conditions in scheduling decisions.

The channel aware schemes can be classified into four classes based on the primary objective: fairness, QoS guarantee, system throughput maximization, or power optimization. A comparison of the scheduling disciplines is presented in Table VIII.

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. Also, it is assumed that the channel information is known at both the transmitter and the 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 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 schedulers also need to consider other users' QoS requirements such as the minimum reserved rate and may need to 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.

Consider the compensation issue. Unlike the wireless LAN networks, WiMAX users pay for their QoS assurance. Thus, in [18] 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 (GWF) concept, which equalizes a weighted sum of the slots and the bytes, was introduced in [18]. 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. The GWF can be illustrated as an equation below:

$$Total\_Slots = \sum_{i=1}^{N} S_i$$

$$wS_i + \frac{(1-w)B_i}{M} = wS_j + \frac{(1-w)B_j}{M}$$

$$B_i = b_i \times S_i$$

For all subscriber i and j in N. Here, S_i and B_i are total number of slots and bytes for subscriber i. b_i is the number of bytes per slot for subscriber i. N is the number of active subscribers. M is the highest level MCS size in bytes. w is a general weight parameter.

It has been observed that allowing unlimited compensation to meet the QoS requirements may lead to bogus channel information to gain resource allocations [58]. The compensation needs to be taken into account with leading/lagging mechanisms [59]. The scheduler can reallocate 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 some 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. When the error rate is high, a credit history can be built based on the lagging flows and the scheduler can allocate the bandwidth based on the ratio of their credits to theirs minimum reserved rates when the

TABLE VII
COMPARISON OF CHANNEL-UNAWARE SCHEDULERS

Scheduling	Pros	Cons
FIFO	Fast and Simple	Unfair and cannot meet QoS requirements
RR	Very simple	Unfair (variable packet size), cannot meet QoS requirements
WRR	Simple; meets the throughput guarantee	Unfair (variable packet size)
DRR/DWRR	Simple, supports variable packet sizes	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

error rate is acceptable [60]. In either case, if and how the compensation mechanism should be put into consideration are still open questions.

1) Fairness: This metric mainly applies for the Best Effort (BE) service. One of the commonly used baseline schedulers in published research is the Proportional Fairness Scheme (PFS) [61], [62]. The objective of PFS 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. Note that defining T_i affects the fluctuation of the throughput. There are several proposals that have applied and modified the PFS. For example, T_i derivation with delay considerations is described in [63]. In [36], given 5 ms frame duration, setting T_i to 50 ms is shown to result in an average rate over 1 second instead of 10 seconds with T_i = 1000 ms. In [64], the moving average was modified to not update when a user queue is empty. A starvation timer was introduced in [65] to prevent users from starving longer than a predefined threshold.

2) QoS Guarantee: Modified Largest Weighted Delay First (M-LWDF) [66] can provide QoS guarantee by ensuring a minimum throughput guarantee and also to maintain delays smaller than a predefined threshold value with a given probability for each user (rtPS and nrtPS). And, it is provable that the throughput is optimal for LWDF [38]. 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 for traffic class i.

There are several proposals that have used or modified M-LWDF. For example, in [67], the scheduler selects the users on each subcarrier during every time slot. For each subcarrier k, the user (i) selection for the subcarrier is expressed by

$$\max[channel\_gain(i,k) \times HOL\_delay(i) \times \frac{a(i)}{d(i)}]$$

In this equation, a is the mean windowed arrival and d is mean windowed throughput. "a" and "d" are averaged over a sliding-window. HOL_delay is the head of line delay. The channel state information is indirectly derived from the normalized channel gain. Note that 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 HOL_delay, a and d.

Similar to M-LWDF, Urgency and Efficiency based Packet Scheduling (UEPS) [68] was introduced to make use of the efficiency of radio resource usage and the urgency (time-utility as a function of the delay) as the two factors for making the scheduling decision. 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:

$$i = \max|U_i'(t)| \times \frac{R_i(t)}{R_i'(t)}$$

Another modification of M-LWDF has been proposed to support multiple traffic classes [69]. 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. This modification handles QoS traffic and BE traffic separately. The HOL packet's waiting time is used for QoS traffic and the queue length for BE traffic.

3) System Throughput Maximization: A few schemes, e.g., [70], [71], [72], focus on maximizing the total system throughput. In these, Max C/I (Carrier to Interference) is used to opportunistically assign resources to the user with the highest channel gain.

Another maximum system throughput approach is the exponential rule [71] in that 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. 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.

Another modification for maximum throughput was proposed in [72] using a heuristic approach of allocating a subchannel to the MS so that it can transmit the maximum amount of data on the subchannel. Suppose a BS has n users and m subchannels, let 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 subchannel j, the goal of scheduling is to minimize the unsatisfied demand, that is,

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

subject to the following constraints: $\sum_{1\leq i\leq n} N_{ij} < N'_j$ and $\sum_{1\leq j\leq m} R_{ij}N_{ij} \leq \lambda_i$

Here, N'_j is the total number of slots available for data transmission in the j^{th} 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 MSs that can transmit maximum amount of data.

4) 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 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 [73]. LWT takes the power consumption into consideration. If assigning n^{th} subcarrier to k^{th} 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 is 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 [73]. 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 [74] 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 [1] defines Power Saving Class (PCS) type I, II and III. Basically PSC I increases the sleep window size by a power of 2 every time there is no packet (similar to binary backoff). Sleep window size for PSC type II is constant. PSC III defines a pre-determined long sleep interval without the existence of the listen period.

Most of the proposals on this topic concentrate on constructing the analytical models for the sleep time; to figure out the optimal sleep time with guaranteed service especially delay (the more the sleep time, the more the packet delay and the more the buffer length). The models basically are based on the arrival process such as in [75] Possion distribution is used for arrival process. Hyper-Erlang distribution is used for self-similarity of web traffic in [76].

In order to reduce waking period for each MS, Burst scheduling was proposed in [77]. A rearrangement technique for unicast and multicast traffic is used so that a MS can wake up and received both type of traffic at once if possible [78].

In [79] a hybrid energy-saving scheme was proposed by 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).

IV. Conclusion

In this paper, we provided an extensive survey of recent 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 I. 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 minimize power consumption, 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 others.

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

TABLE VIII
COMPARISON OF CHANNEL-AWARE SCHEDULERS

Category	Scheduling Algorithms	Pros/Cons	Traffic Classes
Fairness	Variation of PFS [36], [61], [62], [63], [64]	Achieve long term fairness but can not guarantee the delay constraint	BE
QoS Guaranteed (minimum throughput and delay)	Variation of M-LWDF [66], [67], [68], [69]	Meet the throughput and delay guarantee with threshold probability	ertPS, rtPS and nrtPS
System throughput maximization	Variation of maximum C/R [70], [71], [72]	Maximize the total system throughput but can not meet QoS requirement especially delay as well as unfairness	BE
Power constraint	LWT [74], Linear Programming [73], [74]	Minimize the power consumption but can not meet QoS requirement especially delay as well as unfairness	BE

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 an 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) [80], [81]. 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.

ACKNOWLEDGMENT

We would like to thank Mark C. Wood, Michael Roche and Ritun Patney, who participated earlier in the WiMAX Scheduling research at the Washington University, for their direct and indirect contributions to our understanding of the issues discussed here.

REFERENCES

[1] IEEE P802.16Rev2/D2, "DRAFT Standard for Local and metropolitan area networks," Part 16: Air Interface for Broadband Wireless Access Systems, Dec. 2007, 2094 pp.
[2] WiMAX Forum, "WiMAX System Evaluation Methodology V2.1," Jul. 2008, 230 pp. Available: http://www.wimaxforum.org/technology/documents/
[3] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Networking*, vol. 7, pp. 473-489, Aug. 1999.
[4] N. H. Vaidya, P. Bahl, and S. Gupta, "Distributed fair scheduling in a Wireless LAN," *IEEE Trans. Mobile Comput.*, vol. 4, pp. 616-629, Dec. 2005.
[5] L. Tassiulas and S. Sarkar, "Maxmin fair scheduling in wireless networks," in *Proc. IEEE Computer Communication Conf.,* 2002, New York, NY, vol. 2, pp. 763-772.
[6] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. K. Tripathi, "Enhancing throughput over Wireless LANs using channel state dependent packet scheduling," in *Proc. IEEE Computer Communication Conf.,* San Francisco, CA, 1996, vol. 3, pp. 1133-1140.
[7] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," *ACM/Baltzer Wireless Networks.*, vol. 8, pp. 13-26, Jan. 2002.
[8] E. Jung and N. H. Vaidya, "An energy efficient MAC protocol for Wireless LANs," in *Proc. IEEE Computer Communication Conf.,* New York, NY, 2002, vol. 3, pp. 1756-1764.
[9] X. Zhang, Y. Wang, and placeW. Wang, "Capacity analysis of adaptive multiuser frequency-time domain radio resource allocation in OFDMA systems," in *Proc. IEEE Int. Symp. Circuits and Systems.,* Greece, 2006, pp. 4-7.
[10] T. Ohseki, M. Morita, and T. Inoue, "Burst Construction and Packet Mapping Scheme for OFDMA Downlinks in IEEE 802.16 Systems," in *Proc. IEEE Global Telecommunications Conf.,* Washington, DC, 2007, pp. 4307-4311.
[11] Y. Ben-Shimol, I. Kitroser, and Y. Dinitz, "Two-dimensional mapping for wireless OFDMA systems," *IEEE Trans. Broadcast.*, vol. 52, pp. 388-396, Sept. 2006.
[12] 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," in *Proc. 6th Int. Conf. Information, Communications & Signal Processing.*, Singapore, 2007, pp. 1-5.
[13] C. Desset, E. B. de Lima Filho, and G. Lenoir, "WiMAX Downlink OFDMA Burst Placement for Optimized Receiver Duty-Cycling," in *Proc. IEEE Int. Conf. Communications.*, Glasgow, Scotland, 2007, pp. 5149-5154.
[14] C. So-In, R. Jain, and A. Al-Tamimi, "Capacity Estimations in IEEE 802.16e Mobile WiMAX networks," Submitted for publication, *IEEE Wireless Comm. Mag.*, April 2008. Available: http://www.cse.wustl.edu/~jain/papers/capacity.htm
[15] 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 for publication, *IEEE Wireless Communication and Networking Conf.,* 2008. Available: http://www.cse.wustl.edu/~jain/papers/eocsa.htm
[16] H. Martikainen, A. Sayenko, O. Alanen, and V. Tykhomyrov, "Optimal MAC PDU Size in IEEE 802.16," *Telecommunication Networking Workshop on QoS in Multiservice IP Networks.*, Venice, Italy, 2008, pp. 66-71.
[17] S. Sengupta, M. Chatterjee, and S. Ganguly, "Improving Quality of VoIP Streams over WiMAX," *IEEE Trans. Comput.*, vol. 57, pp 145-156, Feb. 2008.
[18] 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 for publication, *IEEE Sarnoff Symp.*, 2009, Dec. 2008. Available: http://www.cse.wustl.edu/~jain/papers/gwf.htm
[19] T. Wand, H. Feng, and B. Hu, "Two-Dimensional Resource Allocation for OFDMA System," in *Proc. IEEE Int. Conf. Communications Workshop.*, Beijing, China, 2008, pp. 1-5.
[20] M. Hawa and D. W. Petr, "Quality of service scheduling in cable and broadband wireless access systems," in *Proc. IEEE Int. Workshop Quality of Service*, Miami Beach, MI, 2002, pp. 247-255.
[21] 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 Commun. Mag.,* vol. 45, pp. 132-138, May 2007.
[22] L. Lin, W. Jia, and W. Lu, "Performance Analysis of IEEE 802.16 Multicast and Broadcast Polling based Bandwidth Request," in *Proc. IEEE Wireless Communication and Networking Conf.,* Hong Kong, 2007, pp. 1854-1859.
[23] B. Chang and C. Chou, "Analytical Modeling of Contention-Based Bandwidth Request Mechanism in IEEE 802.16 Wireless Network," *IEEE Trans. Veh. Techol.*, vol. 57, pp 3094-3107, Sept. 2008.

[24] P. Rastin, S. Dirk, and M. Daniel, “Performance Evaluation of Piggyback Requests in IEEE 802.16,” in *Proc. IEEE Vehicular Technology Conf.*, Baltimore, MD, 2007, pp. 1892-1896.

[25] V. Alexey, Z. Ying, N. Qiang, and L. Andrey, “Efficient Request Mechanism Usage in IEEE 802.16,” in *Proc. IEEE Global Telecommunications Conf.,* San Francisco, CA, 2006, pp. 1-5.

[26] O. Alanen, “Multicast polling and efficient voip connections in ieee 802.16 networks,” in *Proc. Int. Workshop Modeling Analysis and Simulation Wireless and Mobile Systems.*, Crete Island, Greece, 2007, pp. 289-295.

[27] A. Doha, H. Hassanein, and G. Takahara, “Performance Evaluation of Reservation Medium Access Control in IEEE 802.16 Networks,” in *Proc. ACS/IEEE Int. Cont. Computer Systems and Applications.*, Dubai, UAE, 2006, pp. 369-374.

[28] A. Sayenko, O. Alanen, and T. Hamalainen, “On Contention Resolution Parameters for the IEEE 802.16 Base Station,” in *Proc. IEEE Global Telecommunications Conf.,* Washington, DC, 2007, pp. 4957-4962.

[29] J. Yan and G. Kuo, “Cross-layer Design of Optimal Contention Period for IEEE 802.16 BWA Systems,” in *Proc. IEEE Int. Conf. Communications.*, Istanbul, Turkey, 2006, vol. 4, pp. 1807-1812.

[30] C. Cicconetti, A. Erta, L. Lenzini, and E. A. M. E. Mingozzi, “Performance Evaluation of the IEEE 802.16 MAC for QoS Support,” *IEEE Trans. Mobile Comput.*, vol. 6, pp. 26-38, Nov. 2006.

[31] O. Yang and J. Lu, “New scheduling and CAC scheme for real-time video application in fixed wireless networks,” in *Proc. IEEE Consumer Communications and Networking Conf.*, Las Vegas, NV, 2006, vol. 1, pp. 303-307.

[32] 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 Commun. Lett.*, vol. 9, pp. 691-693, Aug. 2005.

[33] H. Lee, T. Kwon, and D. Cho, “Extended-rtPS Algorithm for VoIP Services in IEEE 802.16 systems,” in *Proc. IEEE Int. Conf. Communications.*, Istanbul, Turkey, 2006, vol. 5, pp. 2060-2065.

[34] P. T. Brady, “A model for generating on-off speech patterns in two-way conversation,” *Bell System Technical Journal.*, pp. 2445-2472, Sept. 1969.

[35] M. Hawa and D. W. Petr, “Quality of service scheduling in cable and broadband wireless access systems,” in *Proc. IEEE Int. Workshop Quality of Service.*, Miami Beach, MI, 2002, pp. 247-255.

[36] J. Wu, J. Mo, and T. Wang, “A Method for Non-Real-Time Polling Service in IEEE 802.16 Wireless Access Networks,” in *Proc. IEEE Vehicular Technology Conf.,* Baltimore, MD, 2007, pp 1518-1522.

[37] M. Andrews, “Probabilistic end-to-end delay bounds for earliest deadline first scheduling,” in *Proc. IEEE Computer Communication Conf.,* Israel, 2000, vol. 2, pp. 603-612.

[38] A. L. Stolyar and K. Ramanan, “Largest Weighted Delay First Scheduling: Large Deviations and Optimality,” *Annals of Applied Probability.*, vol. 11, pp. 1-48, 2001.

[39] C. E. Koksal, H. I. Kassab, and H. Balakrishnan, “An analysis of short-term fairness in wireless media access protocols,” in *Proc. ACM SIGMETRICS Performance Evaluation Review.*, Santa Clara, CA, 2000, vol. 28, pp. 118-119.

[40] C. Cicconetti, L. Lenzini, placeE. Mingozzi, and C. Eklund, “Quality of service support in IEEE 802.16 networks,” *IEEE Network*, vol. 20, pp. 50-55, April 2006.

[41] A. Sayenko, O. Alanen, J. Karhula, and T. Hamaainen, “Ensuring the QoS Requirements in 802.16 Scheduling,” in *Proc. Int. Workshop Modeling Analysis and Simulation Wireless and Mobile Systems.*, Termolinos, Spain, 2006, pp. 108-117.

[42] A. Sayenko, O. Alanen, and T. Hamaainen, “Scheduling solution for the IEEE 802.16 base station,” *Int. J. Computer and Telecommunications Networking*, vol. 52, pp. 96-115, Jan. 2008.

[43] 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, pp. 34-41, Oct. 2007.

[44] N. Liu, X. Li, C. Pei, and B. Yang, “Delay Character of a Novel Architecture for IEEE 802.16 Systems,” in *Proc. Int. Conf. Parallel and Distributed Computing, Applications and Technologies.*, Dalian, China, 2005, pp. 293-296.

[45] K. Wongthavarawat and A. Ganz, “Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems,” *Int. J.Communication Systems.*, vol. 16, pp. 81-96, Feb. 2003.

[46] A. Sayenko, T. Hamalainen, J. Joutsensalo, and J. Siltanen, “An adaptive approach to WFQ with the revenue criterion,” in *Proc. IEEE Int. Symp. Computers and Communication.*, 2003, vol. 1, pp. 181-186.

[47] D. H. Kim and C. G. Kang, “Delay Threshold-based Priority Queueing Packet Scheduling for Integrated Services in Mobile Broadband Wireless Access System,” in *Proc. IEEE Int. Conf. High Performance Computing and Communications*, Kemer-Antalya, Turkey, 2005, pp. 305-314.

[48] 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,” in *Proc. IEEE Wireless Communication and Networking Conf.,* Las Vegas, NV, 2006, vol. 2, pp. 1142-1147.

[49] J. Borin and N. Fonseca, “Scheduler for IEEE 802.16 Networks,” *IEEE Commun. Lett.,* vol. 12, pp. 274-276, April 2008.

[50] Y. Wang, S. Chan, M. Zukerman, and R.J. Harris, “Priority-Based fair Scheduling for Multimedia WiMAX Uplink Traffic,” in *Proc. IEEE Int. Conf. Communications.*, Beijing, China, 2008, pp. 301-305.

[51] L. F. M. de Moraes and P. D. Jr. Maciel, “Analysis and evaluation of a new MAC protocol for broadband wireless access,” in *Proc. Int. Conf. Wireless Networks, Communications, and Mobile Computing.*, Kaanapali Beach Maui, Hawaii, 2005, vol. 1, pp. 107-112.

[52] W. Lilei and X. Huimin, “A new management strategy of service flow in IEEE 802.16 systems,” in *Proc. IEEE Conf. Industrial Electronics and Applications.*, Harbin, China, 2008, pp 1716-1719.

[53] D. Niyato and E. Hossain, “Queue-aware uplink bandwidth allocation for polling services in 802.16 broadband wireless networks,” in *Proc. IEEE Global Telecommunications Conf.,* St. Louis, MO, 2005, vol. 6, pp. 5-9.

[54] J. Chen, W. Jiao and H. Wang, “A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode,” in *Proc. IEEE Int. Conf. Communications.*, Seoul, Korea, 2005, vol. 5, pp. 3422-3426.

[55] J. Chen, W. Jiao, and Q. Quo, “An Integrated QoS Control Architecture for IEEE 802.16 Broadband Wireless Access Systems,” in *Proc. Global Telecommunications Conf.*, St. Louis, MO, 2005, pp. 6-11.

[56] K. Wongthavarawat and A. Ganz, “IEEE 802.16 based last mile broadband wireless military networks with quality of service support,” in *Proc. IEEE Military Communications Conf.*, Boston, MA, 2003, vol. 2, pp. 779-784.

[57] A. K. F. Khattab and K. M. F. Elsayed,“Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks,” in *Proc. Int. Symp. World of Wireless Mobile and Multimedia Networks.*, Buffalo, NY, 2006, pp. 10-19.

[58] Z. Kong, Y. Kwok, and J. Wang, “On the Impact of Selfish Behaviors in Wireless Packet Scheduling,” in *Proc. IEEE Int. Conf. Communications.,* Beijing, China, 2008, pp. 3253-3257.

[59] 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,” in *Proc. IEEE Int. Conf. Communications.*, Istanbul, Turkey, 2006, vol. 11, pp. 5103-5110.

[60] W. K. Wong, H. Tang, S. Guo, and V. C. M. Leung, “Scheduling algorithm in a point-to-multipoint broadband wireless access network,” in *Proc. IEEE Vehicular Technology Conf.,* Orlando, FL, 2003, vol. 3, pp. 1593-1597.

[61] P. Bender, P. Black, M. Grob, R. Padovani, placeN. Sindhushayana, and A. Viterbi, “CDMA/HDR: A Bandwidth-Efficient High-Speed Wireless Data Service for Nomadic Users,” *IEEE Commun. Mag.*, vol. 38, pp. 70-77, Jul. 2000.

[62] H. Kim and Y. Han, “A proportional fair scheduling for multicarrier transmission systems,” *IEEE Commun. Lett.*, vol. 9, pp. 210-212, Mar. 2005.

[63] F. Hou, P. Ho, X. Shen, and A. Chen, “A Novel QoS Scheduling Scheme in IEEE 802.16 Networks,” in *Proc. IEEE Wireless Communication and Networking Conf.,* Hong Kong, 2007, pp. 2457-2462.

[64] N. Ruangchaijatupon and Y. Ji, ’“Simple Proportional Fairness Scheduling for OFDMA Frame-Based Wireless Systems,” in *Proc. IEEE Wireless Communication and Networking Conf.,* Las Vegas, NV, 2008, pp. 1593-1597.

[65] J. Qiu and T. Huang, “Packet scheduling scheme in the next generation high-speed wireless packet networks,” in *Proc. IEEE Int. Wireless and Mobile Computing, Networking and Communications.*, Montreal, Canada, 2005, pp. 224-227.

[66] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, “Providing quality of service over a shared wireless link,” *IEEE Commun. Mag.*, vol. 39, pp. 150-154, Feb. 2001.

[67] P. Parag, S. Bhashyam, and R. Aravind, “A subcarrier allocation algorithm for OFDMA using buffer and channel state information,”.in *Proc. IEEE Vehicular Technology Conf.,* Dallas, TX, 2005, vol. 1, pp. 622-625.

[68] S. Ryu, B. Ryu, H. Seo, and M. Shi, “Urgency and efficiency based wireless downlink packet scheduling algorithm in OFDMA system,” in

Proc. IEEE Vehicular Technology Conf., Stockholm, Sweden, 2005, vol. 3, pp. 1456-1462.
[69] W. Park, S. Cho, and S. Bahk, "Scheduler Design for Multiple Traffic Classes in OFDMA Networks," in *Proc. IEEE Int. Conf. Communications.*, Istanbul, Turkey, 2006, vol. 2, pp. 790-795.
[70] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp 1277-1294, Jun. 2002.
[71] S. Shakkottai, R. Srikant, and A. Stolyar, "Pathwise Optimality and State Space Collapse for the Exponential Rule," in *Proc. IEEE Int. Symp. Information Theory.,* 2002, pp. 379.
[72] V. Singh and V. Sharma, "Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks," in *Proc. IEEE Wireless Communication and Networking Conf.,* Las Vegas, NV, 2006, vol. 2, pp. 984-990.
[73] Y. J. Zhang and S. C. Liew, "Link-adaptive largest-weighted-throughput packet scheduling for real-time traffics in wireless OFDM networks," in *Proc. IEEE Global Telecommunications Conf.,* St. Louis, MO, 2005, vol. 5, pp. 5-9.
[74] Z. Liang, Y. Huat Chew, and C. Chung Ko, "A Linear Programming Solution to Subcarrier, Bit and Power Allocation for Multicell OFDMA Systems," in *Proc. IEEE Wireless Communication and Networking Conf.,* Las Vegas, NV, 2008, pp. 1273-1278.
[75] Z. Yan, "Performance Modeling of Energy Management Mechanism in IEEE 802.16e Mobile WiMAX," in *Proc. IEEE Wireless Communication and Networking Conf.*, Hong Kong, 2007, pp. 3205-3209.
[76] X. Yang, "Performance analysis of an energy saving mechanism in the IEEE 802.16e wireless MAN," ," in *Proc. IEEE Consumer Communications and Networking Conf.*, Las Vegas, NV, 2006, pp. 406-410.
[77] J. Shi, G. Fang, Y. Sun, J. Zhou, Z. Li, and E. Dutkiewicz, "Improving Mobile Station Energy Efficiency in IEEE 802.16e WMAN by Burst Scheduling," in *Proc. IEEE Global Telecommunications Conf.,* San Francisco, CA, 2006, pp. 1-5.
[78] L. Tian, Y. Yang, J. Shi, E. Dutkiewicz, and G. Fang, "Energy Efficient Integrated Scheduling of Unicast and Multicast Traffic in 802.16e WMANs," in *Proc. IEEE Global Telecommunications Conf.,* Washington, DC, 2007, pp. 3478-3482.
[79] H. Choi and D. Cho, "Hybrid Energy-Saving Algorithm Considering Silent Periods of VoIP Traffic for Mobile WiMAX," in *Proc. IEEE Int. Conf. Communications.*, Glasgow, Scotland, 2007, pp. 5951-5956.
[80] A. Sayenko, O. Alanen, and T. Hamalainen, "ARQ Aware Scheduling for the IEEE 802.16 Base Station," in *Proc. IEEE Int. Conf. Communications.,* Beijing, China, 2008, pp. 2667-2673.
[81] F. Hou, J. She, and P. Ho, and X. Shen, "Performance Analysis of ARQ with Opportunistic Scheduling in IEEE 802.16 Networks," in *Proc. IEEE Global Telecommunications Conf.,* Washington, DC, 2007, pp 4759-4763.

Chakchai So-In received the B.Eng. and M.Eng. degrees in computer engineering from Kasetsart University, Bangkok, Thailand in 1999 and 2001 respectively. In 2003, He was an internetworking trainee in CNAP (CCNP) program at Nanyang Technological University, Singapore. He was also a student intern at mobile IP division, Cisco Systems, CA, USA in 2006. He is currently working toward a Ph.D. degree at the Department of Computer Science and Engineering, Washington University in St. Louis, MO, USA. His research interests include congestion control in high speed network, protocols to support network and transport mobility, and Quality of Service in broadband wireless access networks (WiMAX). He is an IEEE student member.

Raj Jain is a Professor of Computer Science and Engineering at Washington University in St. Louis. He was one of the Co-founders of Nayna Networks, Inc - a next generation telecommunications systems company in San Jose, CA. Previously, he was a Senior Consulting Engineer at Digital Equipment Corporation in Littleton, Mass and then a professor of Computer and Information Sciences at Ohio State University in Columbus, Ohio. Dr. Jain is a Fellow of IEEE, a Fellow of ACM and ranks among the top 50 in Citeseer's list of Most Cited Authors in Computer Science. He is the author of "Art of Computer Systems Performance Analysis," which won the 1991 "Best-Advanced How-to Book, Systems" award from Computer Press Association. His fourth book entitled " High-Performance TCP/IP: Concepts, Issues, and Solutions," was published by Prentice Hall in November 2003. He is also a winner of ACM SIGCOMM Test of Time award.

Abdel-Karim Tamimi is a PhD candidate in Computer Engineering at Washington University in St. Louis, MO. Abdel Karim received a BA degree in computer engineering from Yarmouk University in Jordan. His college education was supported by a full scholarship given to excellent students from the High Education Ministry [1999-2004]. During this period, he worked on several projects with different parties tackling problems related to AI, Networking, Computer Security and Digital Imaging. After graduating, he held a full time position as a teaching assistant for several courses in Yarmouk University [2004-2005]. Since then, he has been awarded a full scholarship to peruse his Master and PhD degrees in Computer Engineering at Washington University in St. Louis, where he obtained his master degree [2007].