A Study of WiMax QoS Mechanisms

RMOB Project

By:

Masood KHOSROSHAHY Vivien NGUYEN

April 2006

Project supervisor:

Prof. Philippe Godlewski

Table of Contents

A Study of WiMax QoS Mechanisms

Abstract: **In this study, we first give a brief overview of the WiMax/IEEE 802.16 technology covering almost all aspects which are discussed in the standard. Next, as the main focus of the study, we introduce the QoS mechanisms that have been proposed in the past few years and are available in the open literature. This study, being the first of its kind, tries to enlighten the reader regarding the mechanisms available, so that they can compare the pros and cons of each and adopt the method most suitable to any given case. The QoS mechanisms have been categorized based on the layer, MAC or PHY, in which they propose their architectures.**

Keywords: **WiMax, IEEE 802.16, QoS, Scheduling, MAC layer, Physical layer**

I. INTRODUCTION

The standards for Broadband Wireless Access Systems have been developed by the Institute of Electrical and Electronics Engineers (IEEE) which insures a global and open process that enjoys worldwide participation. So far, IEEE 802.11 (short-range: ~100 m), which is a standard for Wireless Local Area Networks, often called "Wi-Fi" for "Wi-Fi Alliance" and IEEE 802.16 (long-range: ~10 km), which is a standard for Wireless Metropolitan Area Networks, often called "WiMAX" for "WiMAX Forum", have been developed.

IEEE 802: **The LAN/MAN Standards Committee**
Wired: – 802.3 (Ethernet) – 802.17 (Resilient Packet Ring)
Wireless: – 802.11: Wireless LAN • Local Area Networks – 802.15: Wireless PAN • Personal Area Networks – 802.16: WirelessMAN • Metropolitan Area Networks – 802.20: • Vehicular Mobility

WiMax is considered as the "Last Mile" solution which provides fast local connection to the network. Compared to high-capacity cable/fiber, it is less expensive to deploy. The WiMax standardization process has had the milestones that are mentioned in the following table.

802.16 Projects
IEEE 802.16-2001 •MAC •10-66 GHz PHY
802.16c (Profiles)
802.16a 2-11 GHz PHY
802.16-2004
802.16e •Mobile Amendment

The WiMax forum has the mission of promoting deployment of Broadband Wireless Access (BWA) by using a global standard and certifying interoperability of products and technologies. It supports IEEE 802.16 standard and proposes and promotes access profiles for it. The forum certifies interoperability levels and hence achieves global acceptance.

In the table below, properties of the IEEE 802.16 standard are provided, in order to give a quick and short overview of the range of issues considered in the standard.

Properties of IEEE Standard 802.16
• Broad bandwidth –Up to 134 Mbit/s in 28 MHz channel (in 10-66 GHz air interface)
• Supports multiple services simultaneously with full QoS – Efficiently transport IPv4, IPv6, ATM, Ethernet, etc.
• Bandwidth on demand (frame by frame)
•MAC designed for efficient use of spectrum
• Comprehensive, modern, and extensible security
• Supports multiple frequency allocations from 2-66 GHz –OFDM and OFDMA for non-line-of-sight applications
• TDD and FDD
• Link adaptation: Adaptive modulation and coding – Subscriber by subscriber, burst by burst, uplink and downlink
• Point-to-multipoint topology, with mesh extensions
• Support for adaptive antennas and space-time coding
• Extensions to mobility

Adapted from [Mar-04]

II. MEDIUM ACCESS CONTROL LAYER

The IEEE 802.16 MAC is a scheduling one where the subscriber station (SS) that wants to attach to the network, has to compete once when it initially enters the network. A time slot allocation is made by the base station (BS) which can be enlarged and constricted. It remains assigned to the subscriber station meaning that other stations are not supposed to use the same resources. This scheduling algorithm has its advantages since it remains stable under overload and oversubscription, has more bandwidth efficiency and also allows the base station to control QoS, meaning that it is balancing the resources among the needs of the subscriber stations.

802.16 MAC: Overview
• Point-to-Multipoint
• Metropolitan Area Network
• Connection-oriented
• Supports difficult user environments – High bandwidth, hundreds of users per channel – Continuous and burst traffic – Very efficient use of spectrum
• Protocol-Independent core (ATM, IP, Ethernet, …)
• Balances between stability of contentionless and efficiency of contention-based operation
• Flexible QoS offerings – CBR, rt-VBR, nrt-VBR, BE, with granularity within classes
• Supports multiple 802.16 PHYs

Adapted from [Mar-04]

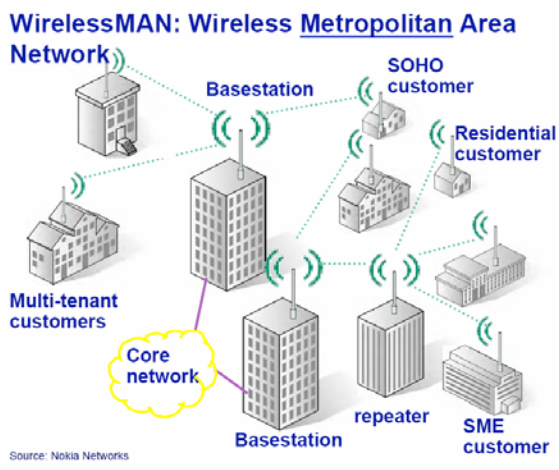
The main goal of the MAC layer is to manage the resources of the air interface efficiently. Indeed, access and bandwidth allocation algorithms must serve hundreds of terminals per channel. Those terminals can eventually be shared by multiple end users. To support a large variety of services such as voice, data or internet connection, the 802.16 MAC must accommodate both continuous and bursty traffic. The issues concerning the transport efficiency are also addressed at the interface between the MAC and the PHY layers. The modulation and coding schemes are specified in a burst profile adjusted in function of each burst sent to each SS. The MAC can make use of bandwidth-efficient burst profiles under favorable link conditions and shift to more reliable and robust ones if the opposite is the case even though the spectral efficiency will be lower. [EMSW-02]

The MAC includes service-specific convergence sublayers (ATM and Packet) that interface to layers above. At the core, there is the MAC common part sublayer that carries out the key MAC functions. Below the common part sublayer is the privacy sublayer. Extensive bandwidth allocation and QoS mechanisms are provided, but the details of scheduling and reservation management have not been specified in the standard. The functions of the common part sublayer will be discussed more in the following paragraphs. [EMSW-02]

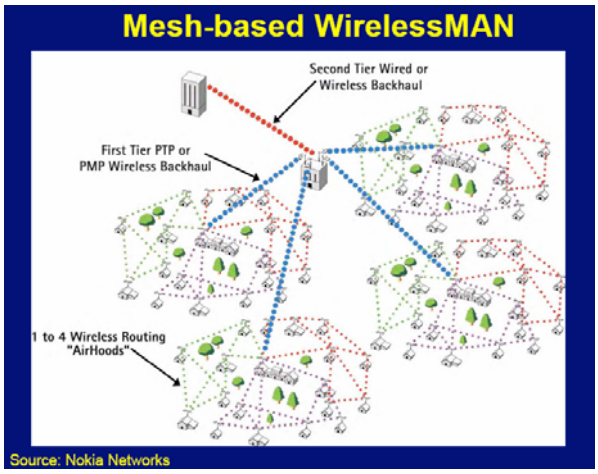
On the downlink, data to Subscriber Stations (SSs) are multiplexed in TDM fashion. The uplink is shared between SSs in TDMA fashion. The 802.16 MAC is connection-oriented, according to which, all services, including connectionless services, are mapped to a connection. This provides a mechanism for requesting bandwidth, associating QoS and traffic parameters, transporting and routing data to the appropriate convergence sublayer, and all other actions associated with the contractual terms of the service. Connections are referenced with 16-bit connection identifiers (CIDs) and may require continuously granted bandwidth or bandwidth on demand. Upon entering the network, the SS is assigned three management connections in each direction. These three connections reflect the three different QoS requirements used by different management levels. The first of these is the basic connection, which is used for the transfer of short, time-critical MAC and radio link control (RLC) messages. The primary management connection is used to transfer longer, more delay-tolerant messages such as those used for authentication and connection setup. The secondary management connection is used for the transfer of standards-based management messages such as Dynamic Host Configuration Protocol (DHCP), Trivial File Transfer Protocol (TFTP), and Simple Network Management Protocol (SNMP). In addition to these management connections, SSs are allocated transport connections for the contracted services. Transport connections are unidirectional to facilitate different uplink and downlink QoS and traffic parameters. [EMSW-02]

The MAC builds the downlink subframe starting with a frame control section containing the DL-MAP and UL-MAP messages. These indicate PHY transitions on the downlink as well as bandwidth allocations and burst profiles on the uplink. The advanced technology of the 802.16 PHY requires equally advanced radio link control (RLC), particularly the capability of the PHY to transition from one burst profile to another. The RLC must control this capability as well as the traditional RLC functions of power control and ranging. [EMSW-02]

Burst profiles for the downlink are each tagged with a Downlink Interval Usage Code (DIUC). Those for the uplink are each tagged with an Uplink Interval Usage Code (UIUC). Burst profile determines the modulation and FEC and is dynamically assigned according to link conditions. It is determined burst by burst and per subscriber station. There is always a trade-off between capacity and robustness in real time. Their utilization has roughly doubled capacity for the same cell area. Burst profile for downlink broadcast channel is well-known and robust, but other burst profiles can be configured “on the fly”. SS capabilities are recognized at registration. [EMSW-02] [Mar-04]



Point-to-Multipoint mode



Mesh Mode (Subscriber-to-Subscriber communications)

Each connection in the uplink direction is mapped to a *scheduling service*. Each scheduling service is associated with a set of rules imposed on the BS scheduler responsible for allocating the uplink capacity and the request-grant protocol between the SS and the BS. The detailed specification of the rules and the scheduling service used for a particular uplink connection is negotiated at connection setup time. The uplink scheduling types include Unsolicited Grant Service (UGS), for real-time flows or periodic fixed size packets (e.g. VoIPor ATM CBR), Real-Time Polling Service (rtPS), for real-time service flows or periodic variable size data packets (e.g. MPEG), Non-Real-Time Polling Service (nrtPS), for non real-time service flows with regular variable size bursts (e.g. FTPor ATM GFR) and Best Effort (BE), for best effort traffic(e.g. UDPor ATM UBR). [EMSW-02] [Mar-04]

The IEEE 802.16 MAC accommodates two classes of SS, differentiated by their ability to accept bandwidth grants simply for a connection or for the SS as a whole. Both classes of SS request bandwidth per connection to allow the BS uplink scheduling algorithm to properly consider QoS when allocating bandwidth. In Bandwidth Grant per Subscriber Station (GPSS), base station grants bandwidth to the subscriber station (SS), and SS in turn may re-distribute bandwidth among its connections, maintaining QoS and service-level agreements. This method is suitable for many connections per terminal by off-loading base station's work. It also allows more sophisticated reaction to QoS needs. It has low overhead but requires intelligent subscriber station. The GPSS is mandatory for P802.16 10-66 GHz PHY. In contrast, in Bandwidth Grant per Connection (GPC) mode, base station grants bandwidth to a connection. This method is mostly suitable for few users per subscriber station. It has higher overhead, but allows simpler subscriber station. [EMSW-02] [MEK-01]

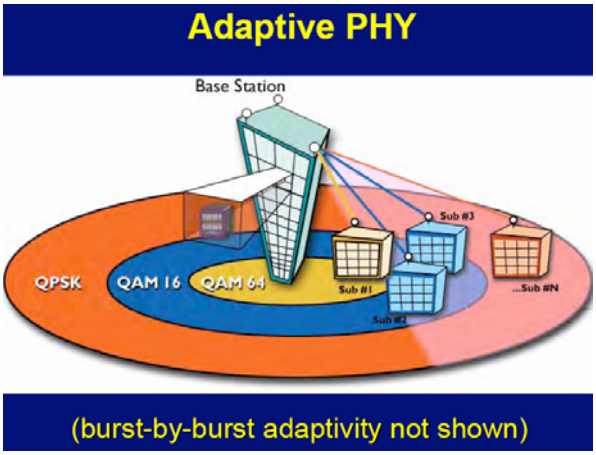
There are several methods for bandwidth request: Implicit requests (UGS), in which there is no actual message and the bandwidth is negotiated at connection setup; BW request messages, which uses the special BW request header and it can request up to 32 KB with a single message; Piggybacked request (for non-UGS services only), which can be up to 32 KB per request for the CID and Poll-Me bit (for UGS services only), which is used by the SS to request a bandwidth poll for non-UGS services. [MEK-01]

Maintaining QoS in GPSS follows a semi-distributed approach. In which, the BS sees the requests for each connection; based on this, grants bandwidth (BW) to the SSs (maintaining QoS and fairness). On the other side, the SS scheduler maintains QoS among its connections and is responsible to share the BW among the connections (maintaining QoS and fairness). It's worth mentioning that algorithms in BS and SS can be very different; SS may use BW in a way unforeseen by the BS. [MEK-01]

Subscriber Station (SS) initialization has several steps: At first, the SS scans for downlink channel and establishes synchronization with the BS. Then, it obtains transmit parameters. As next step, it performs ranging and negotiating basic capabilities. Then it is authorized by the BS and performs key exchange. Afterwards, it performs the registration and IP connectivity establishment. Then it's the turn of time of day establishment and the transfer of operational parameters. At the end, it sets up the connections. [MEK-01]

III. PHYSICAL LAYER

802.16a PHY Alternatives:
• OFDM (WirelessMAN-OFDM Air Interface)
256-point FFT with TDMA (TDD/FDD)
• OFDMA (WirelessMAN-OFDMA Air Interface)
2048-point FFT with OFDMA (TDD/FDD)
• Single-Carrier (WirelessMAN-SCa Air Interface)
TDMA (TDD/FDD)
BPSK, QPSK, 4-QAM, 16-QAM, 64-QAM, 256-QAM
[Mar-04]



Adapted from [Mar-04]

In this section, we introduce the techniques that are used in the physical layer: Orthogonal Frequency Division Multiplexing (OFDM) and Orthogonal Frequency Division Multiple Access (OFDMA).

Those techniques have been developed for the last few years to deliver broad band services that can be compared to those of wired services in terms of data rates. The main issue addressed for the PHY layer is to allocate the resources efficiently by assigning a set of subcarriers and by determining the number of bits to be transmitted for each subcarrier in an OFDMA system. An optimal algorithm has to be chosen to obtain a certain level of performance by considering some constraints such as delays, the total number of connected SS and the total power. [ECV-03]

OFDM will be first studied in order to understand OFDMA used to share the physical resources among each SS.

3.1 OFDM Physical Layer

Definition and advantages of OFDM

Orthogonal frequency-division multiplexing (OFDM) is a transmission technique that is based on the same idea as frequency-division multiplexing (FDM). In FDM, multiple signals are sent out at the same time, but on different frequencies. It actually divides a broadband channel into many narrowband subchannels. In OFDM, a single transmitter transmits on several different orthogonal frequencies. This technique, associated with the use of advanced modulation techniques on each component, give a transmitted signal with high resistance to multi-path interference and a much higher spectral efficiency is obtained.

As the chose of the manufacturers, the WiMAN OFDM PHY layer is the most commonly used because of the reasons previously quoted. It was also selected, rather than other techniques such as single-carrier (SC) or CDMA, due to its superior non line-of-sight (NLOS) performance. This multiplexing technique allows important equalizer design simplification to support operations in multipath propagation environments and overcome channel fading quite efficiently (Rayleigh channel model). [Intel-04]

Subchannelization

The OFDM PHY layer supports UpLink (UL) subchannelization, with the number of subchannels being 16. This feature is particularly useful when a power-limited platform such as a laptop is considered in the subscriber station in an indoor environment. With a sub-channelization factor of 1/16, a 12-dB link budget enhancement can be achieved. Sixteen sets of 12 subcarriers each, are defined, where one, two, four, eight or all of the sets can be assigned to a subscriber station in the uplink. Eight pilot carriers are used when more than one set of sub-channels are allocated. [Intel-04]

Multiplexed channel

This multiplexing technique supports Time Division Duplexing (TDD), in which the uplink and downlink share a channel but do not transmit simultaneously, and Frequency Division Duplexing

(FDD), in which the uplink and downlink operate on separate channels, possibly simultaneously, with support for FDD and also Half-Duplex FDD (H-FDD). In both TDD and FDD modes, the length of the frame can vary (under the control of the BS scheduler) per frame. In TDD mode, the division point between uplink and downlink can also vary per frame, allowing asymmetric allocation of an air time between uplink and downlink if required. The point-to-multipoint architecture forces the BS to transmit TDM signals in which time slots are allocated serially for each individual SS's. [Intel-04]

Error correcting codes

The specification defines as a requirement for the error correcting code, a combined variable-rate Read-Solomon (RS) and Convolutional Coding (CC) scheme, supporting code rates of 1/2, 2/3, 3/4, and 5/6, although variable-rate Block Turbo Code (BTC) and Convolutional Turbo Code (CTC) are optional. Also, the standard supports multiple types of modulation, including Binary Phase Shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK), 16-Quadrature Amplitude Modulation (QAM) and 64-QAM. [Intel-04]

Diversity

Finally, the PHY layer supports optionally the diversity transmission technique in which the information-carrying signal is transmitted along different propagation paths. It uses multiple transmitting and receiving antennas. In the Downlink (DL), it is using Space Time Coding (STC) and Adaptive Antenna Systems (AAS) with Spatial Division Multiple Access (SDMA). [Intel-04]

3.2 OFDMA Physical Layer

Multiple access

OFDMA is referred as Multiuser-OFDM as it uses OFDM as multiple access method especially for the 4G wireless networks.

Subchannelization

The OFDMA PHY layer is based on OFDM and supports subchannelization in both the UL and (DownLink) DL. Five different sub-channelization schemes in total are supported. The OFDMA PHY layer supports both TDD and FDD operations. CC is the required coding scheme by the specification and the code rates are the same as the ones supported by the OFDM PHY layer. BTC and CTC coding schemes are optionally supported. [Intel-04]

Multiplexing and diversity

The same signal modulations are supported as in OFDM. STC and AAS with SDMA are supported, as well as Multiple Input, Multiple Output (MIMO). MIMO regroup a certain number of techniques for using multiple antennas at both the BS and SS in order to increase the channel capacity and to decrease the Bit Error Rate (BER). [Intel-04]

To transmit information at a high data rate, the BWA has to provide efficient and flexible resource allocation. It has been shown that frequency hoping and adaptive modulation used in subcarrier allocation permit to have higher performances. The frequency hoping technique allows to compensate the negative effect of channel fading. [ECV-03]

But ODMA was chosen as it is a promising multiple access scheme having the same advantages as OFDM, that is to say, inter symbol interference immunity as well as frequency selective fading immunity while having a higher spectral efficiency.

3.3 Issues concerning the resource allocation methods

In OFDM, only a single user can transmit on the entire spectrum at a given time. Time division or frequency division multiple access is therefore needed. This scheme is not optimum. That is why OFDMA is preferred because of the fact that is allows multiple users to transmit simultaneously by sharing the sub-channels among the users. [WSEA-04]

The two studied papers have different approaches on the resource allocation methods for OFDMA systems:

- "The problem of assigning subcarriers and power to the different users in an OFDMA system has recently been an area of active research. In [WCLM-99], the margin-adaptive resource allocation problem was tackled, wherein an iterative subcarrier and power allocation algorithm was proposed to minimize the total transmit power given a set of fixed user data rates and bit error rate (BER) requirements. In [JL-03], the rate-adaptive problem was investigated, wherein the objective was to maximize the total data rate over all users subject to power and BER constraints. It was shown in [JL-03] that in order to maximize the total capacity, each subcarrier should be allocated to the user with the best gain on it, and the power should be allocated using the waterfilling algorithm across the subcarriers. However, no fairness among the users was considered in [JL-03]. This problem was partially addressed in [YL-00] by ensuring that each user would be able to transmit at a minimum rate, and also in [RC-00] by incorporating a notion of fairness in the resource allocation through maximizing the minimum user's data rate. In [SAE-03], the fairness was extended to incorporate varying priorities. Instead of maximizing the minimum user's capacity, the total capacity was maximized subject to user rate proportionality constraints. This is very useful for service level differentiation, which allows for flexible billing mechanisms for different classes of users. However, the algorithm proposed in [SAE-03] involves solving non-linear equations, which requires computationally expensive iterative operations and is thus not suitable for a cost-effective real-time implementation." [WSEA-04]
- "In multiuser environment, a good resource allocation scheme leverages multiuser diversity and channel fading [VTL-02]. It was shown in [KH-95] that the optimal solution is to schedule the user with the best channel at each time. Although in this case, the entire bandwidth is used by the scheduled user, this idea can also be applied to OFDMA system, where the channel is shared by the users, each owing a mutually disjoint set of subcarriers, by scheduling the subcarrier to a user with the best channel among others. Of course, the procedure is not simple since the best subcarrier of the user may also be the best subcarrier of another user who may not have any other good subcarriers. The overall strategy is to use the peaks of the channel resulting from channel fading. Unlike in the traditional view where the channel fading is considered to be an impairment, here it acts as a channel randomizer and increases multiuser diversity [VTL-02]. The resource allocation problem has been recently considered in many studies. Almost all of them define the problem as a real time resource allocation problem in which Quality of Service (QoS) requirements are fixed by the application. QoS requirement is defined as achieving a specified data transmission rate and bit error rate (BER) of each user in each transmission. In this regard, the problem differs from the water-pouring schemes wherein the aim is to achieve Shannon capacity under the power constraint [WCLM-99]." [ECV-03]

IV. QUALITY OF SERVICE ARCHITECTURES

Although the IEEE 802.16 standard is already a very mature and a sophisticated one, nevertheless, researchers around the world have not stopped proposing amendments to the standard. For example, in [GWAC-05], the authors have suggested the utilization of spatial multiplexing and multi-user OFDM to maximize the achieved throughput. They have also suggested the use of interference cancellation of dominant interferers and hybrid ARQ in order to increase the range and robustness of the system.

The various protocol mechanisms described in the standards, [Std-04] and [Std-05], may be used to support QoS for both uplink and downlink through the SS and the BS. The requirements for QoS include the following:

a) A configuration and registration function for pre-configuring SS-based QoS service flows and traffic parameters.
b) A signaling function for dynamically establishing QoS-enabled service flows and traffic parameters.
c) Utilization of MAC scheduling and QoS traffic parameters for uplink service flows.
d) Utilization of QoS traffic parameters for downlink service flows.
e) Grouping of service flow properties into named Service Classes, so upper-layer entities and external

applications (at both the SS and BS) may request service flows with desired QoS parameters in a globally consistent way.

The principal mechanism for providing QoS is to associate packets traversing the MAC interface into a service flow as identified by the transport CID. A service flow is a unidirectional flow of packets that is provided a particular QoS. The SS and BS provide this QoS according to the QoS Parameter Set defined for the service flow. The primary purpose of the QoS features defined in the standards is to define transmission ordering and scheduling on the air interface. However, these features often need to work in conjunction with mechanisms beyond the air interface in order to provide end-to-end QoS or to police the behavior of SSs. Service flows exist in both the uplink and downlink direction and may exist without actually being activated to carry traffic. All service flows have a 32-bit SFID; admitted and active service flows also have a 16-bit CID.

A service flow is a MAC transport service that provides unidirectional transport of packets either to uplink packets transmitted by the SS or to downlink packets transmitted by the BS. A service flow is characterized by a set of QoS Parameters such as latency, jitter, and throughput assurances. In order to standardize operation between the SS and BS, these attributes include details of how the SS requests uplink bandwidth allocations and the expected behavior of the BS uplink scheduler. It is useful to think of three types of service flows:

1) Provisioned: This type of service flow is known via provisioning by, for example, the network management system. Its AdmittedQoSParamSet and ActiveQoSParamSet are both null.
2) Admitted: This type of service flow has resources reserved by the BS for its AdmittedQoSParamSet, but these parameters are not active (i.e., its ActiveQoSParamSet is null). Admitted Service Flows may have been provisioned or may have been signalled by some other mechanism.
3) Active: This type of service flow has resources committed by the BS for its ActiveQoSParamSet, (e.g., is actively sending maps containing unsolicited grants for a UGSbased service flow). Its ActiveQoSParamSet is non-null.

The service flow is the central concept of the MAC protocol. It is uniquely identified by a 32-bit (SFID). Service flows may be in either the uplink or downlink direction. There is a one-to-one mapping between admitted and active service flows (32-bit SFID) and transport connections (16-bit CID). Outgoing user data is submitted to the MAC SAP by a CS (Convergence sublayer) process for transmission on the MAC interface. The information delivered to the MAC SAP includes the CID identifying the transport connection across which the information is delivered. The service flow for the connection is mapped to MAC transport connection identified by the CID. A Classifier Rule uniquely maps a packet to its transport connection.

The Service Class serves the following purposes: It allows operators to move the burden of configuring service flows from the provisioning server to the BS. Operators provision the SSs with the Service Class Name; the implementation of the name is configured at the BS. This allows operators to modify the implementation of a given service to local circumstances without changing SS provisioning. Also, it allows higher-layer protocols to create a service flow by its Service Class Name. The Service Class Name is “expanded” to its defined set of parameters at the time the BS successfully admits the service flow. The Service Class expansion can be contained in the following BS-originated messages: DSA-REQ (Dynamic Service Addition), DSC-REQ (Dynamic Service Change), DSA-RSP, and DSC-RSP.

Mobile networks require common definitions of service class names and associated AuthorizedQoSParamSets in order to facilitate operation across a distributed topology. Global service class names shall be supported to enable operation in this context. In operation, global service class names are employed as a baseline convention for communicating AuthorizedQoSParamSet or AdmittedQoSParamSet. Global service class name is similar in function to service class name except that 1) Global service class name use may not be modified by a BS, 2) Global service class names remain consistent among all BS, and 3) Global service class names are a rules-based naming system whereby the global service class name itself contains referential QoS Parameter codes. In practice, global service class names are intended to be accompanied by extending or modifying QoS Param Set defining parameters, as needed, to provide a complete and expedited method for transferring

Authorized- or AdmittedQoSParamSet information. Global service flow class name parameters: Uplink/Downlink indicator, Maximum sustained traffic rate, Traffic Indication Preference, Maximum traffic burst, Minimum reserved traffic rate, Maximum latency, SDU indicator and Paging Preference.

In the above paragraphs, a summary of the information provided in the QoS section (6.3.14) of the standard was given. But, as mentioned on several occasions, the details of actual QoS provisioning are not given in the standard. In what follows, we introduce, in a short format, the QoS architectures that have been proposed so far, which are available in the open literature. Specifically, in the MAC layer, three QoS architectures, which seem to offer more promising results, have been selected and described. Also, another five studies in this layer have been introduced briefly.

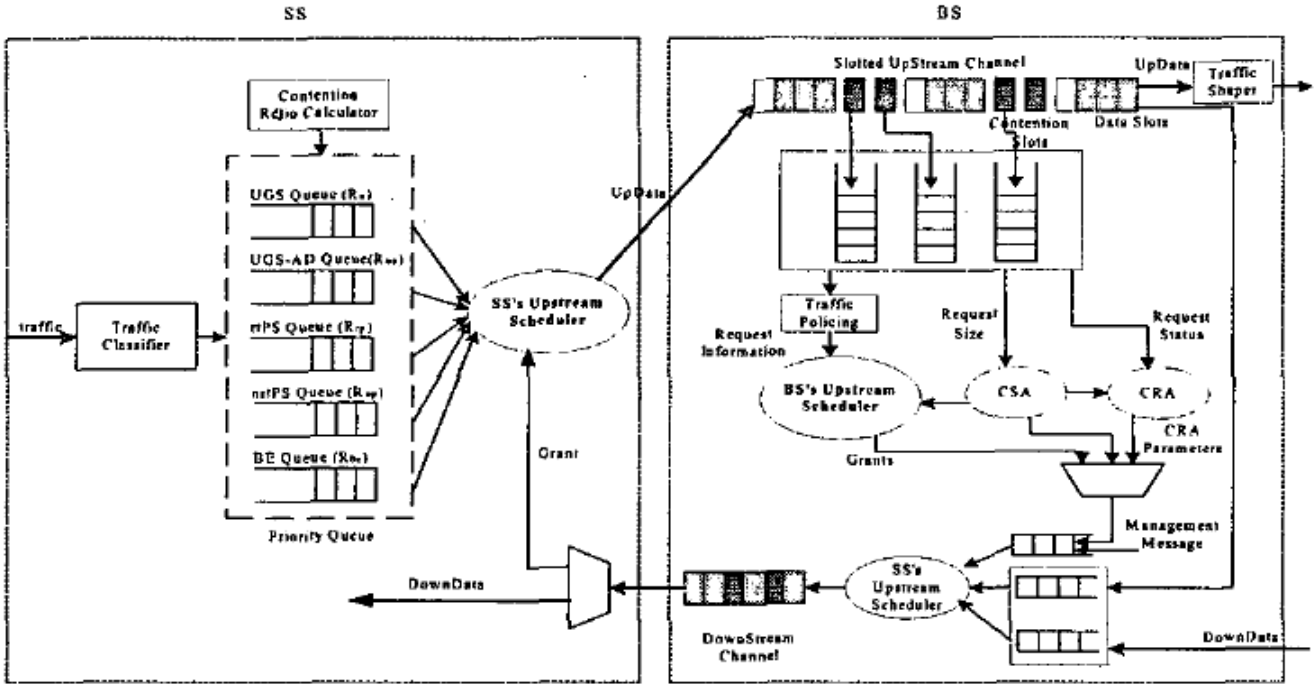
4.1 MAC LAYER QOS ARCHITECTURES

A . "A QoS Architecture for the MAC Protocol of IEEE 802.16 BWA System" [CWM-02]

In this research, the authors have proposed a QoS architecture based on priority scheduling (A scheduling strategy for the schedulers) and dynamic bandwidth allocation.

The have mentioned the following points as the responsibilities of the implementers: 1) A method for efficiently combining bandwidth request strategy and bandwidth allocation (Scheduling) strategy to maintain QoS and fairness for different traffic. 2) A method for choosing different algorithms considering implementation and computation complexity. 3) A method for recognizing high priority requests during first access.

The proposed QoS architecture, the following figure, has the following modules as its building blocks: Traffic Classifier, SS's Upstream Scheduler, BS's Upstream and Downstream Schedulers.



Adapted from [CWM-02]

The Contention Slot Allocator (CSA) is used by the BS to dynamically adjust the ratio of the bandwidth allocated to the contention slots and reservation slots: too few contention slots increase the chances of bandwidth request collision, on the other hand, too many of them reduces the bandwidth left for data transmission. So there is a trade-off in the design of the CSA, the output of which defines the bandwidth resources allocated to the CRA (Collision Resolution Algorithm)[GSS-99] and BS's upstream grants scheduler. The CRA defines the utilization of the contentions slots and the rules used to resolve contention.

In the Upstream Scheduler implementation, GPSS (Grant Per SS) mode is preferred. Since only the information about the overall required bandwidth is needed to inform the BS's upstream scheduler,

a small amount of bandwidth is needed to update this information. Furthermore, in this way, the resulting problems due to the time lag in receiving updated information from SSs can be resolved. For example, even if the BS does not know that an urgent packet arrives a the SS, this packet can still overtake other packets and thus provide tight guarantees as long as a good scheduling algorithm is used at the SS. In this way, the BS's upstream scheduler does not require detailed information on every connection at the SS.

At the SS, when traffic generates, the traffic classifier guides them into different priority queues based on traffic priority. Then the CRC (Contention Ratio Calculator) dynamically assigns different competitive ratio parameters Ru, Rua, Rrp, Rnp and Rbe to UGS (Unsolicited Grant Service), UGS-AD (UGS with Activity Detection), rtPS, nrtPS and BE queues, respectively. The traffic scheduler allocates the granted upstream slots to different connections with different QoS requirement.

According to the information of the request of the SS, the BS's upstream scheduler schedules grants to the SS. Then, the SS's upstream scheduler is responsible for the selection of appropriate packets from the respective UGS-AD, rtPS, nrtPS and BE queues and sends them through the upstream data slots granted by the BS's upstream scheduler.

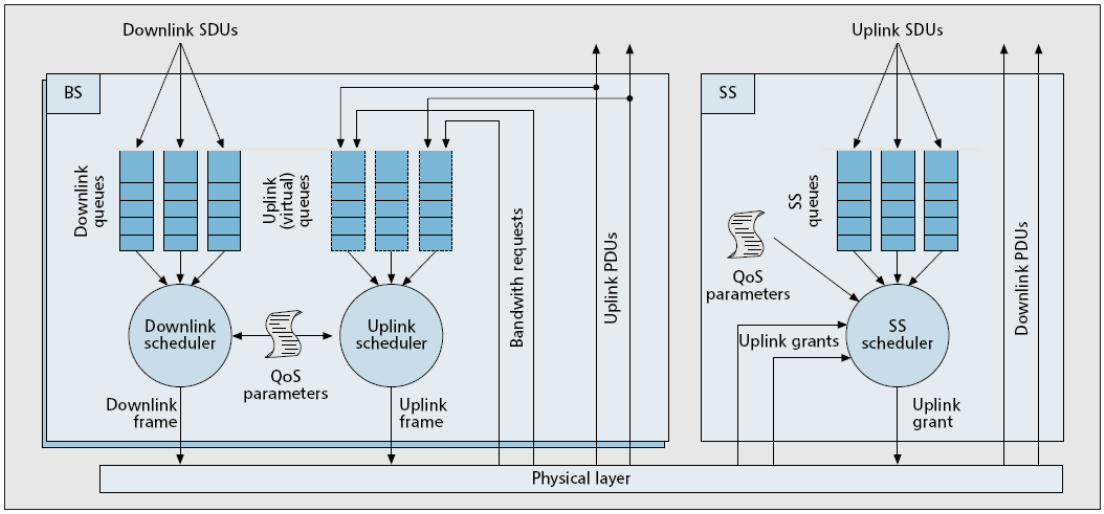
When implementing the SS's upstream scheduler, one can use MPFQ(Multiclass Priority Fair Queuing)[MLK-00] scheduler with some modification. For each service category, there is a priority class in MPFQ. Each of the priority classes has its own packet scheduler that determines the packet order it its class. One can use a Wireless Fair Queuing (WFQ) [LBS-97] Policy in the higher priorities in order to provide a lower delay bound for this service class, Weighted Round Robin (WRR) [PGa-93] scheduling for the middle priorities since the delay requirements are not that tight for these classes and WRR is simple, and FIFO scheduling at the lower priorities.

When implementing the BS's upstream scheduler, one finds that WRR is more suitable because the exact arrival time of a packet is not involved in the computing the virtual finishing time. This actually is the real situation where BS has limited information on the traffic generated at SS, and hence computing the time for transmission and bandwidth allocation just based on the bandwidth requests is more appropriate.

The practical issues such as performance and stability associated with the QoS architecture and also comparison of the different implementation approaches for each block in the architecture, including different scheduling algorithms, are left for future studies.

B. "Quality of Service Support in IEEE 802.16 Networks" [CLME-06]

In this article, the authors have reviewed and analyzed the mechanisms for supporting QoS at the IEEE 802.16 MAC layer. Then, they have analyzed by simulation the performance of IEEE 802.16 in two application scenarios, which consist of providing last-mile Internet access for residential and SME (Small and Medium-sized Enterprises) subscribers, respectively. Their analysis is aimed at showing the effectiveness of the 802.16 MAC protocol in providing differentiated services to applications with different QoS requirements, such as VoIP and Web.



Adapted from [CLME-06]

The previous figure shows the blueprint of the functional entities for QoS support, which logically reside within the MAC layer of the BS and SSs. Each downlink connection has a packet queue at the BS (represented with solid lines). In accordance with the set of QoS parameters and the status of the queues, the BS downlink scheduler selects from the downlink queues, on a frame basis, the next service data units (SDUs) to be transmitted to SSs. On the other hand, uplink connection queues (represented in the figure by solid lines) reside at SSs. Since the BS controls the access to the medium in the uplink direction, bandwidth is granted to SSs on demand. Bandwidth requests are used on the BS for estimating the residual backlog of uplink connections. In fact, based on the amount of bandwidth requested (and granted) so far, the BS uplink scheduler estimates the residual backlog at each uplink connection (represented in the figure as a virtual queue by dashed lines), and allocates future uplink grants according to the respective set of QoS parameters and the virtual status of the queues. However, although bandwidth requests are per connection, the BS nevertheless grants uplink capacity to each SS as a whole. Thus, when an SS receives an uplink grant, it cannot deduce from the grant which of its connections it was intended for by the BS. Consequently, an SS scheduler must also be implemented within each SS MAC, in order to redistribute the granted capacity to all of its own connections.

Here, the performance of 802.16 in two of the most promising application scenarios envisaged by the WiMAX forum is assessed. They consist in providing last-mile Internet access for residential and SME subscribers. Since a minimum reserved rate is the basic QoS parameter negotiated by a connection within a scheduling service, the class of latency-rate [SVa-98] scheduling algorithms is particularly suited for implementing the schedulers in the 802.16 MAC. Specifically, within this class, they selected deficit round robin (DRR) [ShV-96] as the downlink scheduler to be implemented at the BS, since it combines the ability of providing fair queuing in the presence of variable length packets with the simplicity of implementation. In particular, DRR requires a minimum rate to be reserved for each packet flow being scheduled. Therefore, although not required by the 802.16 standard, BE connections should also be guaranteed a minimum rate. This fact can be exploited in order to both avoid BE traffic starvation in overloaded scenarios, and let BE traffic take advantage of the excess bandwidth which is not reserved for the other scheduling services. On the other hand, DRR assumes that the size of the head-of-line packet is known at each packet queue; thus, it cannot be used by the BS to schedule transmissions in the uplink direction. In fact, with regard to the uplink direction, the BS is only able to estimate the overall amount of backlog of each connection, but not the size of each backlogged packet. Therefore, they selected weighted round robin (WRR) [KSC-91] as the uplink scheduler in their 802.16 simulator. Like DRR, WRR belongs to the class of rate-latency scheduling algorithms. Finally, they decided to implement DRR as the SS scheduler, because the SS knows the sizes of the head-of-line packets of its queues.

The metrics used for assessing the performance of 802.16 are the average packet-transfer delay and the delay variation. The simulations were carried out by means of a prototypical simulator of the IEEE 802.16 protocol. The simulator was event-driven and was developed using the C++ programming language. Specifically, the MAC layers of the SSs and the BS were implemented, including all functions for uplink/downlink data transmission.

Residential Scenario

The Residential scenario consists of a BS providing Internet access to its subscribers, by means of a variable number of BE connections evenly distributed among the SSs. Internet traffic is modeled as a Web traffic source. Since the BS knows the current status of downlink queues, as soon as a downlink packet is enqueued at the BS, it is immediately eligible for transmission to its intended SS. As long as the system is underloaded, the connection queues are almost always empty. Thus, the average delay of downlink packets is almost constant. However, the average delay increases sharply as soon as the system starts to get overloaded, because the BS is not able to fully serve the backlog of downlink connections before new packets are enqueued. The average delay of uplink traffic is higher than that of downlink traffic. In fact, any SS has to request bandwidth to the BS, in order to receive an uplink grant to transmit its backlog. The average time needed by an SS to request bandwidth by using contention increases with the offered load. The maximum achievable throughput decreases when the number of SSs increases, for both downlink and uplink curves.

SME Scenario

The SME scenario involves a BS providing several enterprise customer premises with three different types of services: VoIP, videoconference, and data. It is assumed that each SS has four VoIP sources multiplexed into an rtPS connection, two videoconference sources multiplexed into an rtPS connection, and a BE connection loaded with data traffic. Also, it is assumed that the BS grants a unicast poll to each VoIP and videoconference connection every 20 ms. Finally, for data traffic, the same Web traffic source model is used as in the previous scenario.

When the system is underloaded, there is no service differentiation between connections with data and multimedia traffic. However, when the system becomes overloaded, the average delay of data traffic increases much more sharply than that of multimedia traffic. This is due to the way in which capacity has been provisioned to the different connections. Specifically, scheduling algorithms have been configured so that rtPS connections have a reserved rate equal to the mean rate of VoIP and videoconference applications, respectively. On the other hand, the reserved rate for videoconference connections accounts for the two videoconference sources multiplexed into the same connection. Finally, BE connections are reserved a rate of 10 B/s. Note that the rate guaranteed to BE connections is negligible with respect to the rate guaranteed to rtPS connections, and this justifies the different performance of the BE and rtPS connections, respectively. With regard to uplink connections, the average delay of rtPS connections is almost constant when the number of SSs increases. This is because the BS grants a unicast poll to each rtPS connection every 20 ms. On the other hand, SSs have to request bandwidth for BE connections on a contention basis. Thus, as soon as the system gets overloaded, the average delay of BE connections increases remarkably, whereas the average delay of rtPS connections remains low.

Since VoIP and videoconference are interactive multimedia applications, a relevant performance index is also the 99^{th} percentile of the delay variation. Considering the performance of downlink traffic, when the number of SSs increases, the delay variation increases smoothly from 10 to 20 ms. Under these conditions the system is underloaded. When the number of SSs increases further, the BS downlink scheduler is not able, on average, to schedule each VoIP packet before the next one is generated from the same application. Hence, the delay variation of multimedia traffic increases sharply. In the study, the performance of uplink multimedia traffic and the bandwidth-request mechanisms used by rtPS and BE connections have also been discussed.

In summary, they have assessed the performance of an IEEE 802.16 system under two traffic scenarios. The first one (residential scenario) dealt with data (non-QoS) traffic only, and was thus managed by the BE scheduling service. The results have shown that the average delay of the uplink traffic is higher than that of the downlink traffic. Furthermore, the former increases more sharply than the latter with the offered load. This behavior can be explained by means of both the bandwidth-request mechanism and the overhead introduced by physical preambles. In the second scenario (SME scenario), on the other hand, they have shown the service differentiation, in terms of delay, between data (served via BE) and multimedia traffic (served via rtPS). This is achieved because scheduling in 802.16 is controlled by the BS in both the downlink and uplink directions. Therefore, it is possible to employ scheduling algorithms that have been proposed for wired environments, which are able to provide QoS guarantees. In the simulations, they have evaluated the DRR and WRR scheduling algorithms as possible candidates for algorithms to be implemented in a production system.

C. "Providing integrated QoS control for IEEE 802.16 broadband wireless access systems" [CJG-05]

In this study, the authors propose a new integrated QoS architecture for IEEE 802.16 Broadband Wireless MAN in TDD mode. A mapping rule for providing DiffServ between IP layer and MAC layer is given and a fast signaling mechanism is designed to provide cross layer integrated QoS for Point to Multi-Point (PMP) mode. Since IP network service is based on a connectionless and best-effort model, this service model is not adequate for many applications that normally require assurances on QoS performance metrics. So, a number of enhancements have been proposed to enable offering different levels of QoS in IP networks including the integrated services (IntServ) architecture, the differentiated service (DiffServ) architecture.

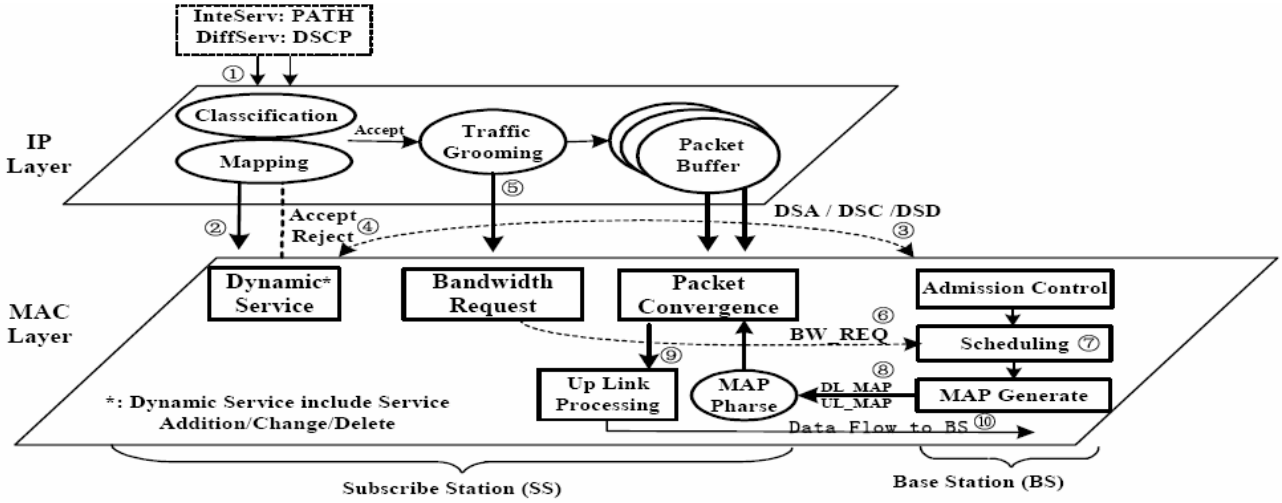
IntServ is implemented by four components: the signaling protocol (e.g. RSVP), the admission control, the classifier and the packet scheduler. Applications requiring guaranteed service or

controlled-load service must set up the paths and reserve resources before transmitting their data. The admission control routines will decide whether a request for resources can be granted. After classification of packets in a specific queue, the packet scheduler will then schedule the packet to meet its QoS requirement. In [Std-04] and [Std-05], some rules to classify DiffServ IP packets into different priority queues are also proposed based on IP QoS indication bits in IP header. So, in general, the QoS architecture of IEEE 802.16 under PMP mode can support both IntServ and DiffServ.

Two ways for providing cross layer QoS control via WirelessMAN technology may be candidates: For the first one, the traditional RSVP is used to provide cross layer QoS control. RSVP signaling message will be regarded as the traffic data by convergence sub-layer in the MAC layer. RSVP signaling message can be classified into a special high priority queue by a protocol-specific packet matching criteria. This RSVP queue will be transmitted in the second management connection.

In summary, the QoS provision procedure will consist of the following two part: on one hand, the secondary management connection will be used for RSVP to provide the layer 3 QoS; on the other hand, the primary management connection will be used for DSA/DSC/DSD to provide the layer 2 QoS. Since the second management connection is defined for delay tolerant traffic and there are many other IP protocol related message (DHCP, SNMP, TFTP, etc) sharing the same queue, the whole QoS provision will be rather slow. Furthermore, considering the RSVP signaling has a periodical refreshing procedures, which consume a lot of bandwidth, it is not efficient to use the secondary management connection for the RSVP. The second one is the proposed way. Since there are so many similarities between providing Internet IntServ using RSVP and MAC layer QoS using DSA/DSC/DSD, naturally, the mechanism will be superior to the first one in high efficiency and fastness by mapping between the cross layer QoS control and MAC QoS in IEEE802.16 network.

The message exchange for DSA and DSC can be deployed to achieve QoS guarantees through end-to-end resource reservation for packet flows and to perform per-flow scheduling which IntServ services require. For DiffServ services, on the other hand, a number of per-hop behaviors (PHBs) for different classes of aggregated traffic can be mapped into different connections directly. They propose an integrated QoS control architecture as shown in the following figure, which implements a cross layer traffic-based prioritization mechanism in a comprehensive way.



Adapted from [CJG-05]

Step 1 and 2 in the figure show when a new service flow arrives in IP layer, it will be firstly parsed according to the definition in PATH message (for InteServ) or Differentiated Services Code Point (DSCP for DiffServ); then classified and mapped into one of four types of services (UGS, rtPS, nrtPS or BE). In step 3, the dynamic service model in SS will send request message to the BS, then the admission control in BS will determine whether this request will be approved or not. If not, the service module will inform upper layer to deny this service in step 4; if yes, admission control will notify scheduling module to make a provision in its basis scheduling parameter according to the value shown in the request message and the accepted service will transfer into traffic grooming module in step 5.

According to the traffic grooming result, SS will send Bandwidth Request message to BS in step 6. The scheduling module in BS will retrieve the requests (step 7) and generate UL-MAP and DL-MAP message (step 8) following the bandwidth allocation results. Finally, the SS will package SDUs from IP layer into PDUs and upload them in its uplink slot to BS (step 9-10).

Traffic Classification and Mapping Strategies for IntServ Services

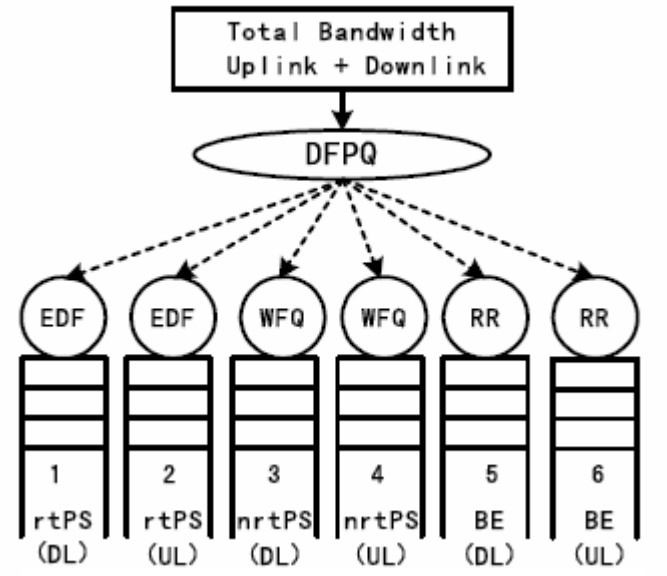
The sender will send a PATH message including traffic specification (TSpec) information. The parameters such as up/bottom bound of bandwidth, delay and jitter can be easily mapped into parameters in DSA message such as Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Tolerated Jitter and Maximum Latency. According to the response of DSA message, the provisioned bandwidth can be also freely mapped into reserved specification (RSpec) into RESV message. Four rules are defined to map IP layer service into MAC layer services.

Traffic Classification and Mapping Strategies for DiffServ Services

For DiffServ services, DSCP code is deployed for classification. There are three definitions of per-hop behavior (PHB) to specify the forwarding treatment for the packet. Expedited forwarding (EF) is intended to provide a building block for low delay, low jitter and low loss services by ensuring that the EF aggregate is served at a certain configured rate. Assured Forwarding (AF) PHB group is to provider different levels of forwarding assurances for IP packets. Four AF classes are defined, where each AF class is allocated a certain amount of forwarding resources (buffer space and bandwidth). Four rules are defined to map IP layer service into MAC layer services.

Admission Control and Scheduling in BS

It will collect all the DSA/DSC/DSD requests and update the estimated available bandwidth based on bandwidth change. The hierarchical structure of the bandwidth allocation in BS is shown in the following figure. In this architecture, two-layer scheduling is deployed. Six queues are defined according to their direction (uplink or downlink) and service classes (rtPS, nrtPS and BE). Since service of UGS will be allocated fixed bandwidth (or fixed time duration) in transmission, these bandwidths will be cut directly before each scheduling.



Adapted from [CJG-05]

The algorithm of the first layer scheduling is called Deficit Fair Priority Queue (DFPQ) proposed in [CJW-05], which is basically based on priority queue. Two policies of initial priority are defined as following:
Service class based priority:
rtPS > nrtPS> BE
Transmission direction based priority:
Downlink >Uplink.

In the second layer scheduling, three different algorithms are assigned to three classes of service to match its requirements. They have applied earliest deadline first (EDF) for rtPS [GGP-94], which means packets with earliest deadline will be scheduled first. The information module determines the packets' deadline and the deadline is calculated by its arrival time and maximum latency. Weight fair queue (WFQ) [DKS-89] is deployed for nrtPS services. This type of packets is scheduled based on the weight (ratio between a connection's nrtPS Minimum Reserved Traffic Rate and the total sum of the Minimum Reserved Traffic Rate of all nrtPS connections). The remaining bandwidth is allocated to each BE connection by round robin (RR).

Comparison between two ways of RSVP

The negotiation of QoS parameters for one traffic will be processed two times. For the first time, the parameters are carried in RSVP messages and transmitted through the Secondary Management connection. For the second time, the same parameters are mapped in MAC message and transmitted through the Primary Management Connection. With the utilization of the new mapping rule, the RSVP

signaling messages are mapped directly into the MAC messages, and then transmitted through the Primary Management Connection. In this way, the messages are transmitted only once, reducing the delay.

They have also developed a simulation platform for the proposed integrated QoS architecture and the behavior of the architecture has been verified by simulation.

D. Brief introduction of five other studies

D.1. "A Quality of Service Architecture for IEEE 802.16 Standards" [AMY-05]

In this paper, the authors have introduced an architecture to support Quality of Service in IEEE 802.16 standards. They have also proposed a design approach to implement such architecture. Simulation result shows the performance of their architecture for all types of traffic classes defined by the standard. They have developed some compatible methods for specific modules such as Scheduler, Traffic Shaper, and Request and Grant Manager to optimize Delay, Throughput and Bandwidth Utilization metrics. The simulation results show that the proposal meets these objectives.

According to them, considering the fact that IEEE 802.16 standards are connection-oriented, each user first sends a Connection Establishment Request to BS. The request is then analyzed in Call Admission Control and if accepted, attributes of QoS and also two identifiers for each direction of this connection are registered in Service Flow Data Bases. In order to perform QoS process, packets are classified according to their mentioned identifiers in MAC entrance point by Classifiers. For further details, please refer to their article.

D.2. "Quality of service scheduling in cable and broadband wireless access systems" [HPE-02]

In this article, the authors have presented a new and efficient scheduling architecture to support bandwidth and delay QoS guarantees for both DOCSIS and IEEE 802.16. Their design goals are simplicity and optimum network performance. The developed architecture supports various types of traffic including constant bit rate, variable bit rate (real-time and non-real-time) and best effort.

The architecture supports tight delay guarantees for UGS traffic and minimum bandwidth reservations for rtPS, nrtPS and BE flows. They remind that vendor-specific QoS parameters can also be used in DOCSIS and IEEE 802.16 and this means that users can also request QoS delay bounds for their rtPS and nrtPS service flows. They mention that since they are using a fair queuing algorithm in their scheduler, providing such guarantees is feasible and can be implemented easily given that the service flows feeding the scheduler are properly policed.

They also introduced a dynamic minislot allocation scheme that should improve the performance of the scheduling algorithm under varying load conditions. It speeds up the contention phase by providing extra bandwidth for contending packets. They argue that the loss of throughput due to this operation should not be a severe one. This is mainly due to the fact that request packets are much smaller than actual data packets, and because their algorithm allocates fewer contention minislots as the load on the system increases. For further details, please refer to their article.

D.3. "Exploiting MAC flexibility in WiMAX for media streaming" [SCGI-05]

In this article , the authors have studied the media access control (MAC) layer of WiMAX and have exploited its flexible features to dynamically construct the MAC packet data units (MPDU). The sizes of the MPDUs are constantly modified based on the channel state information. The desired payload is obtained either by aggregation or fragmentation of the upper layer data units. The robustness of MPDUs is also made tunable by means of cyclic redundancy code bits. They have considered the both scenarios- with and without feedback. They have adhered to the 802.16 specifications and proposed adapting the MPDU length for streaming media for better performance. Three metrics are defined: restore probability, goodput and dropping probability. Simulation experiments are conducted which show the performance enhancements of the proposed ARQ-enabled adaptive algorithm in terms of these three metrics.

They have focused on the MAC common part sublayer to explore its rich set of features. This sublayer controls the on-air timing based on consecutive frames that are divided into time slots. The size of these frames and the size of the individual slots within these frames can be varied on a frame-

by-frame basis. This allows effective allocation of on-air resources and they have applied this mechanism on the MPDUs that are to be transmitted. Depending on the feedback received from the receiver and on-air physical layer slots, they have exploited the feature of the common part sublayer that modifies the size of the MPDUs by changing the size of the payload.

The optimal size of the MPDU must be matched to the channel conditions so as to obtain a desired level of performance. Since packets often get lost being corrupted during transmission in error prone wireless channels, ARQ mechanism is usually used to identify and possibly recover the missing frames. In their case, ARQ plays a crucial role in estimating the channel condition and the fate of the MPDUs that have been transmitted. As a result, the round trip time (RTT) becomes crucial in determining the size of the MPDUs.

In summary, they have studied the problem of streaming media over WiMAX and exploited the flexible features present in the MAC layer of 802.16a. They proposed that the size of MAC packet data units be made adaptive to the instantaneous wireless channel state condition. Based on the type of feedback received, variable size MPDUs were constructed either by aggregation or fragmentation of MAC service data units. They conducted simulation experiments to verify the validity of their proposed scheme. Packet restore probability, goodput, and dropping probability of MPDUs were defined as the performance metrics. Simulation results demonstrate the effectiveness and performance improvement of the proposed scheme. For further details, please refer to their article.

D.4 "Algorithms for routing and centralized scheduling to provide QoS in IEEE 802.16 mesh networks" [SSh-05]

In this article, the authors have considered the problem of routing and centralized scheduling for IEEE 802.16 mesh networks. They first have fixed the routing, which reduces the network to a tree. Then, they have presented scheduling algorithms which provide per flow QoS guarantees to real and interactive data applications while utilizing the network resources efficiently. The algorithms are also scalable: they do not require per flow processing and queuing and the computational requirements are minimal. They have also discussed admission control policies which ensure that sufficient resources are available. They have handled UDP and TCP traffic separately at first and then jointly. They have verified their algorithms via extensive simulations. For further details, please refer to their article.

D.5. "Modeling and performance analysis of the distributed scheduler in IEEE 802.16 mesh mode" [CMZ-05]

In this paper, the authors have presented an analytical model for the distributed scheduling algorithm in the IEEE 802.16 mesh mode. The medium access control (MAC) layer of the IEEE 802.16 has point-to-multipoint (PMP) mode and mesh mode. In the mesh mode, all nodes are organized in an ad hoc fashion and use a pseudo-random function to calculate their transmission time based on the scheduling information of the two-hop neighbors. In this paper, they have developed a stochastic model for the distributed scheduler of the mesh mode. With this model, they have analyzed the scheduler performance under various conditions, and the analytical results match with the ns-2 simulation results. The analytical model developed in this paper is instrumental in optimizing the IEEE 802.16 mesh mode system performance.

In the mesh mode, every node competes for the channel access and tries to broadcast its scheduling information periodically. The channel contention result is correlated with the total node number, exponent value and network topology. Their model assumes that the transmit time sequences of all the nodes in the control subframe form statistically independent renewal processes. Based on this assumption, they have developed methods for estimating the distributions of the node transmission interval and connection setup delay, which are instrumental for evaluating performance, like throughput and delay. Comparisons with ns-2 simulation results show that the model is quite accurate in typical scenarios. Since the detail reservation scheme for the data subframe of the IEEE 802.16 mesh mode is left unstandardized, their model also sheds some light on the data subframe reservation scheme. For example, based on their analysis, the nodes with real time traffic shall have smaller holdoff exponents because they can have more chance to obtain data channel. However, too many nodes with small exponent value generate intensive competition that wastes system resource. Then the nodes can adjust their exponent values adaptively according to the competition node number variation to meet the connection QoS requirements. A good reservation scheme should guarantee the bandwidth

allocation fairness and improve the channel utilization at the same time. Such a reservation scheme needs the information like the connection setup time and success probability provided by their model. For further details, please refer to their article.

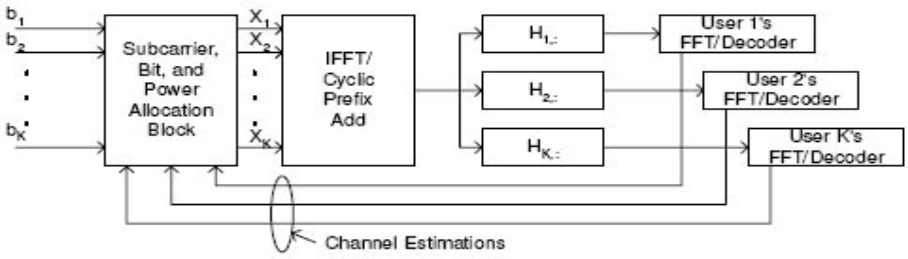
4.2 PHYSICAL LAYER QOS ARCHITECTURES

Two different techniques proposed by two different papers will be presented. The concepts will be explained but the different parts will be exempted of the equations and all the algorithms details as it is not the goal of this paper.

A. Proposed solutions in "A Low Complexity Algorithm for Proportional Resource Allocation in OFDMA Systems" [WSEA-04]

In this paper, a subcarrier allocation scheme is developed. It aims to linearize the power allocation problem while attaining approximate rate proportionality.
The goal is to reduce significantly the complexity and maintaining a reasonable performance. [WSEA-04]

Orthogonal Frequency Division Multiple Access System Model



[WSEA-04] OFDMA system block diagram for K users. Each user is allocated different set of subcarriers by the basestation.

In the above picture is represented, a block diagram for an OFDMA downlink system from the BS to the SSes. The BS transmitter emits some bit streams for each users k on different subcarriers with an index n ranging from 1 to N (N being the total number of subcarriers), allocated to a user k with a certain power p(k, n). The subcarriers are assumed not to be shared by several users. The transmitted signal (user's bits) is modulated into QAM symbols with a Gray bit mapping before being combined into an OFDMA symbol using an IFFT block.

The transmission channel is modeled by a slowly time varying, selective Rayleigh channel of bandwidth B.

At the reception, each SS decodes the bits on their assigned subcarriers only. This information is provides thru a control channel.

Also a channel estimation operation is performed. The estimation results are sent back from the receiver to the emitter (feedback), and give some information for the resource allocation algorithms. The slowly time varying nature of the channel is essential in this case otherwise resource allocation, realized thanks to the feedback, would not be efficient. [WSEA-04]

High Complexity Algorithm

"In [SAE-03], the approach was to first determine the subcarrier allocation, followed by the power allocation. The subcarrier allocation was determined by allowing each user to take turns choosing the best subcarrier for him. In each turn, the user with the least proportional capacity gets the priority to choose his best subcarrier. After the subcarrier allocation, the power allocation is then simplified into a maximization over continuous variables p(k,n)." They refer to this method of subcarrier and power allocation as *root-finding*. But the complexity of the calculations, used in these algorithms, makes them impractical for real-time systems. That is the reason why an approach called *linear* was proposed for reducing the complexity as well as maintaining good performances. [WSEA-04]

Reduced Complexity Algorithm

The different steps of this proposed solution involve the determination of the number of subcarriers N(k) assigned for each user, the assignment of the number of subcarriers distributed for each user k to guaranty a certain proportionality, the assignment of total power P(k) allocated to a given user k to maximize the channel capacity under the constrain of the proportionality, and finally the assignment of power p(k, n) for each subcarrier assigned to one user. [WSEA-04]

The steps are roughly described without any equation or algorithm. Please refer to [WSEA-04] for more details. The different steps involve:

- Step1 Number of subcarriers per use
  "This initial step is based on the reasonable assumption also made in [YL-00] that the proportion of subcarriers assigned to each user is approximately the same as their eventual rates after power allocation, and thus would roughly satisfy the proportionality constraints."
  The total number of subcarriers N is distributed equally among the users and N^* unallocated subcarriers are left.
- Step2 Subcarrier assignment
  This step allocates the per user set of subcarriers N(k) and then the remaining N^* unallocated subcarriers in a way that maximizes the overall capacity while maintaining rough proportionality.
- Step3 Power allocation among users
  The output of the previous two steps is a subcarrier allocation for each user that reduces the resource allocation problem to the finding of an optimal power allocation.
- Step4 Power allocation across subcarriers per user
  The previous step gives the total power P(k) for each user k, which are then used in this final step to perform waterfilling across the subcarriers for each user. The waterfilling process aims to attribute power to each subcarrier to increase capacity having a constant total power.

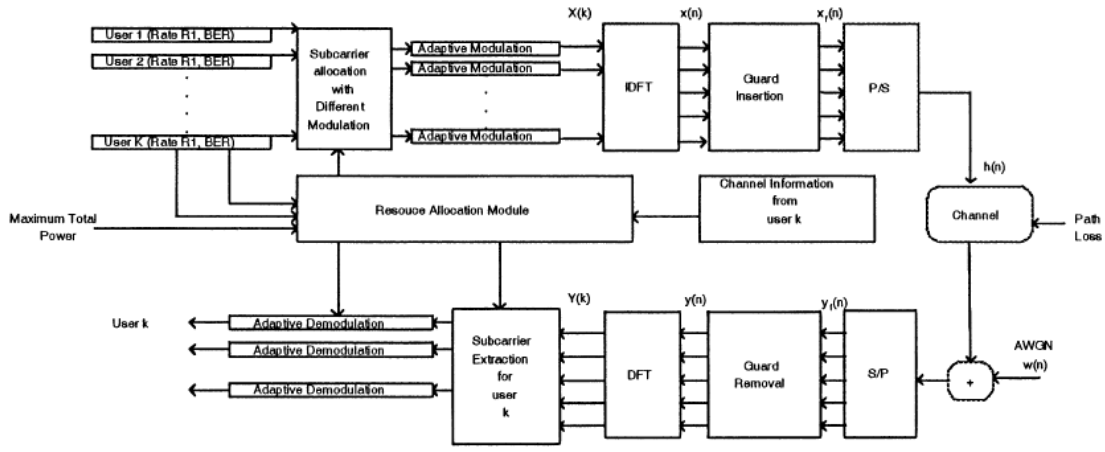
B. Proposed solutions in "QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems" [ECV-03]

In this paper, an iterative multi-user bit and power allocation scheme are introduce to fulfill QoS requirements to maintain reasonable performance for each user. The objective is to minimize the total transmitted power while allocating the subcarriers to the users. It is also to determine the bit rate transmitted on each subcarrier. An adaptive modulation was previously considered in [WCLM-99], [KLKL-01]. The scheme is simple and sufficiently fair to meet real time applications criteria in which a quick scheme is needed to allocate subcarriers before the channel changes and a fair scheme is needed to treat each user. [ECV-03]

A continuous allocation scheme is also proposed where the allocator uses the previous channel information per user for the current allocation. An extension of the point-to-point version of proportional fair scheduling (as in [KLKL-01]) to a point-to-multipoint version is proposed. In this scheme there is no fixed requirements per symbol, the aim is to maximize capacity. [ECV-03]

Orthogonal Frequency Division Multiple Access System

The difference between an OFDM and an OFDMA system is that OFDMA involve several users that must share several subcarriers and therefore it involves the need of an FFT block. Otherwise, the rest of the system is similar with OFDM. Each user is allocated a set of non overlapping subcarriers. A guard insertion is needed in order to prevent Inter Symbol Interference (ISI) at the emission, at the reception, after the sampler, the bits corresponding to the guard time are discarded. Each set of subcarriers contain a given user's bit. The modulation used to code the different bit stream for each user is adaptive. [ECV-03]



Orthogonal frequency division multiple access system.

[ECV-03]

"In a perfectly synchronized system, the allocation module of the transmitter assigns subcarriers to each user according to some QoS criteria. QoS metrics in the system are rate and bit error rate (BER). Each user's bit stream is transmitted using the assigned subcarriers and adaptively modulated for the number of bits assigned to the subcarrier. The power level of the modulation is adjusted to overcome the fading of the channel. The transmission power for AWGN channel can be predicted. In addition the channel gain of subcarrier to the corresponding user should be known." [ECV-03]

In this study, the channel is assumed to be known at both transmitter and receiver. The transmitter and receiver will be able to estimate the channel as long as the channel variation over time is slow. The resource allocation should be done within the coherence time prior to this statement. Also this channel property is required to apply a continuous allocation, where the allocator uses the previous channel information per user for the current allocation. "With the channel information, the objective of resource allocation problem can be defined as maximizing the throughput subject to a given total power constraint regarding the user's QoS requirements". [ECV-03]

The resource allocation problem is formulated with a total transmission power constraint. This transmission power is function of the required received power with unity channel gain for a reliable reception of the modulated symbols. The maximum BER(max) is predefined and the required BER(k) for each user should below BER(max). The data rate for each user should be equal to the required data rate R(k). Therefore the modulation type and the BER get involved in the resource allocation decision process. [ECV-03]

Optimal Solution

Several solutions have been described. Only the optimal one in this paper gives the exact solution of the problems mentioned above. As the previous paper from an implementation point of view it is not realistic because the allocation algorithm would not be executed fast enough as the channel is varying in time and also because the Integer Programming that would be used increase the complexity exponentially with the number of constraints. [ECV-03]

Suboptimal Solution

In most attempts to simplify the resource allocation problem, the problem is decomposed into two procedures: A subcarrier allocation with fixed modulation, and bit loading. Subcarrier allocation with fixed modulation deals with one matrix with fixed and then by using bit loading scheme, the number of bits is incremented. [ECV-03]

The subcarrier allocation problem can be solved with Linear Programming (LP) or Hungarian algorithms. Although the Hungarian algorithm is proposed as an optimal solution for resource allocation with a fixed modulation in [PJ-02], [WTCL-99], it is considered as a suboptimal solution for adaptive modulation. Linear programming is investigated in [KLKL-01].

The bit loading algorithm (BLA) appears after the subcarriers are assigned to users that have at least a certain number of bits assigned. Bit loading procedure is as simple as incrementing bits of the assigned subcarriers of the users until the total power is less or equal to the upper limit of the total transmission power. It allows to convert the fixed modulation scheme into adaptive modulation scheme for each subcarrier. [ECV-03]

Iterative Solution

"The GreedyLP and GreedyHungarian methods both first determine the subcarriers and then increment the number of bits on them according to the rate requirements of users. This may not be a good schedule in some certain cases: For instance, consider a user with only one good subcarrier and low rate requirement. The best solution for that user is allocating its good carrier with high number of bits. But if GreedyLP or Greedy-Hungarian is used, user may have allocated more than one subcarrier with lower number of bits and in some cases, its good subcarrier is never selected. Consider another scenario where a user does not have any good subcarrier (i.e. it may have a bad channel or be at the edge of the cell). In this case, rather than pushing more bits and allocating less subcarriers as in GreedyLP and GreedyHungarian, the opposite strategy is preferred since fewer bits in higher number of subcarriers give better result. Another difficulty arises in providing fairness. Since GreedyLP and GreedyHungarian are based on greedy approach, the user in the worst condition usually suffers. In any event, these are complex schemes and simpler schemes are needed to finish the allocation 366 IEEE TRANSACTIONS ON BROADCASTING, VOL. 49, NO. 4, December 2003 within the coherence time. To cope with these challenges, we introduce a simple, efficient and fair subcarrier allocation scheme with iterative improvement." [ECV-03]

The proposed scheme is composed of two modules called *scheduling* and *improvement modules*. For *scheduling*, bits and subcarriers are distributed to the users and passed to the *improvement module*. The allocation is then improved iteratively by bit swapping and subcarrier swapping algorithms. [ECV-03]

Physical Layer QoS Architectures Summary

"The paper [WSEA-04] presents a new method to solve the rate-adaptive resource allocation problem with proportional rate constraints for OFDMA systems. It improves on the previous work in this area [SAE-03] by developing a subcarrier allocation scheme that achieves approximate rate proportionality while maximizing the total capacity. This scheme was also able to exploit the special linear case in [SAE-03], thus allowing the optimal power allocation to be performed using a direct algorithm with a much lower complexity versus an iterative algorithm. It is shown through simulation that the proposed method performs better than the previous work in terms of significantly decreasing the computational complexity, and yet achieving higher total capacities, while being applicable to a more general class of systems." [WSEA-04]

"In [ECV-03] is considered the problem of resource allocation for adaptive modulation in OFDMA systems. Two different approaches are introduced. One maximizes the capacity and the other one satisfies fixed QoS criteria (i.e the rate and bit error rate requirements) in each symbol. Recent work has focused on developing algorithms to meet the QoS criteria [KT-01], [WCLM-99]–[RC-00]. In an OFDMA system, subcarriers are distributed among users and number of bits transmitted in each subcarrier is adjusted according to the rate requirements of users to minimize total transmit power. It has been shown that resource allocation can be optimized by Integer Programming [KLKL-01]. However, the optimal solution can not be implemented in real time. A simple suboptimal solution that fairly allocates and efficiently converges close to optimal meeting the QoS criteria per symbol was proposed. The algorithm showed good performance in terms of tight power control, iterative betterment and fair scheduling among users when compared with the optimal solution and previously proposed suboptimal schemes. The proposed solution can also be applied to the uplink when there is

perfect synchronization. We also considered a possible resource allocation scheme when the objective is to maximize capacity, based on proportional fair scheduling algorithm for point-to-point communication introduced in [VTL-02].” [ECV-03]

Consequently, both papers proposed efficient ways to solve the rate-adaptive resource allocation problem to meet the required Quality of Service criteria for OFDMA systems.

REFERENCES

[AMY-05] H.S.Alavi, M.Mojdeh, N.Yazdani, “A Quality of Service Architecture for IEEE 802.16 Standards”, 2005 Asia-Pacific Conference on Communications, 03-05 Oct. 2005

[CJG-05] Jianfeng Chen, Wenhua Jiao, Qian Guo, “Providing integrated QoS control for IEEE 802.16 broadband wireless access systems”, IEEE 62nd Vehicular Technology Conference, 25-28 Sept., 2005

[CJW-05]* Jianfeng Chen, Wenhua Jiao, Hongxi Wang, “A Fair Scheduling for IEEE 802.16 Broadband Wireless Access Systems”, ICC2005, May 16-20, Souel, Kerea

[CLME-06] C.Cicconetti, L.Lenzini, E.Mingozzi, C.Eklund, “Quality of service support in IEEE 802.16 networks”, IEEE Network, Vol. 20, Issue 2, Mar./April 2006

[CMZ-05] M.Cao, W.Ma, Q.Zhang, X.Wang and W.Zhu, “Modelling and performance analysis of the distributed scheduler in IEEE 802.16 mesh mode”, Proceedings of the 6th ACM international symposium on Mobile ad hoc networking and computing, Urbana-Champaign, IL, USA, May 25-27, 2005

[CWM-02] G.Chu, D.Wang, S.Mei, “A QoS Architecture for the MAC Protocol of IEEE 802.16 BWA System”, IEEE 2002 International Conference on Communications, Circuits and Systems and West Sino Expositions, 29 June-1 July 2002

[DKS-89]* Demers A, Keshav S, Shenker S. “Analysis and Simulation of a Fair Queuing Algorithm”. SIGCOMM CCR 19 1989;

[ECV-03] Mustafa Ergen, Sinem Coleri, and Pravin Varaiya “QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems”, IEEE Trans. Broadcast, vol. 49, no. 4, Dec. 2003.

[EMSW-02] C.Eklund, R.B.Marks, K.L.Stanwood and S.Wang, “IEEE Standard 802.16: A Technical Overview of the WirelessMAN™ Air Interface for Broadband Wireless Access”, IEEE Communications Magazine, June 2002.

[GGP-94]* Georgiadis L, Guerin R, Parekh A. “Optimal Multiplexing on a Single Link: Delay and Buffer Requirements”. Proceedings of IEEE INFOCOM 94; vol. 2, 1994; 524–532.

[GSS-99]* N.Golmie, Y.Saintillan, and D.H.Su, “A Review of Contention Resolution Algorithms for IEEE 802.14 Networks”, IEEE Commun. Surverys, 1st Quarter 1999

[GWAC-05] A.Ghosh, D.R.Wolter, J.G.Andrews and R.Chen “Broadband Wireless Access with WiMax/802.16: Current Performance Benchmarks and Future Potential”, IEEE Communications Magazine, Feb. 2005.

[HPE-02] M. Hawa, D.W. Petr, “Quality of service scheduling in cable and broadband wireless access systems”, Tenth IEEE International Workshop on Quality of Service, 15-17 May 2002

[Intel-04] G. Nair et al. “IEEE 802.16 Medium Access Control and Service Provisioning”, Intel Technology Journal, vol.08, issue 03, Aug. 20 2004

[JL-03]* J. Jang and K. B. Lee, “Transmit Power Adaptation for Multiuser OFDM Systems,” IEEE J. Select. Areas Commun., vol. 21, pp. 171–178, February 2003.

[KH-95]* R. Knopp and P. Humblet, “Information capacity and power control in single cell multiuser communications,” in ICC ’95, vol. 1, 1995, pp. 331–335.

[KLKL-01]* I. Kim, H. L. Lee, B. Kim, and Y. H. Lee, “On the use of linear programming for dynamic subchannel and bit allocation in multinser OFDM”, in IEEE GLOBECOM ’01, vol. 6, 2001, pp. 3648–3652.

[KSC-91]* M. Katevenis, S. Sidiropoulos, and C. Courcoubetis, “Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip,” IEEE JSAC, vol. 9, no. 8, Oct. 1991, pp. 1265–79.

[KT-01]* I. Koutsopoulos and L. Tassiulas, “Channel state-adaptive techniques for throughput enhancement in wireless broadband networks,” in INFOCOM 2001, vol. 2, 2001, pp. 757–766.

[LBS-97]* Songwu Lu, Vaduvur Bharghavan, and R.Srikant, “Fair scheduling in wireless packet networks”, Proceedings of ACM SIGCOMM 1997

[Mar-04] Roger B. Marks, (US) National Institute of Standards and Technology, IEEE 802.16 Working Group Chair, IEEE Computer Society Distinguished Visitors Program Presentation, IEEE Oregon Section, Beaverton, OR, USA, 14 July 2004

[MEK-01] Roger B. Marks, Carl Eklund et al. Presentation "The 802.16 WirelessMAN™ MAC: It's Done, but What Is It?" November 2001

[MLK-00]* Jay R.Moorman, John Lockwood, and Sung-Mo Kang, "Wireless quality of service using multiclass priority fair queuing", IEEE JSAC, Aug.2000

[PGa-93]* Abhay K.Parekh and Robert G.Gallage, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case", IEEE/ACM Transactions on Networking 1(3) Jun. 1993.

[PJ-02]* S. Pietrzyk and G. J. M. Janssen, "Multiuser subcarrier allocation for QoS provision in the OFDMA systems," in Proc. VTC 2002, vol. 2, 2002, pp. 1077–1081.

[RC-00]* W. Rhee and J. M. Cioffi, "Increase in Capacity of Multiuser OFDM System Using Dynamic Subchannel Allocation," in Proc. IEEE Vehic. Tech. Conf., Tokyo, Japan, May 2000, pp. 1085–1089.

[SAE-03]* Z. Shen, J. G. Andrews, and B. L. Evans, "Optimal Power Allocation in Multiuser OFDM Systems," in Proc. IEEE Global Communications Conference, San Francisco, CA, Dec. 2003, pp. 337–341.

[SCGI-05] S. Sengupta, M. Chatterjee, S. Ganguly, R. Izmailov, "Exploiting MAC flexibility in WiMAX for media streaming", Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, 13-16 June 2005

[ShV-96]* M. Shreedhar and G. Varghese, "Efficient Fair Queuing using Deficit Round Robin," IEEE Trans. Net., vol. 4, no. 3, June 1996, pp. 375–85.

[SSh-05] H.Shetiya and V.Sharma, "Algorithms for routing and centralized scheduling to provide QoS in IEEE 802.16 mesh networks", Proceedings of the 1st ACM workshop on Wireless multimedia networking and performance modeling, Montreal, Quebec, Canada, Oct.13, 2005

[Std-04] IEEE 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks — Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Oct. 1, 2004.

[Std-05] IEEE 802.16e-2005, "IEEE Standard for Local and Metropolitan Area Networks — Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems," Feb. 28, 2006.

[SVa-98]* D. Stiliadis and A. Varma, "Latency-Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms," IEEE/ACM Trans. Net., vol. 6, Oct. 1998, pp. 675–89.

[VTL-02]* P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," IEEE Trans. Information Theory, , vol. 48, no. 6, pp. 1277–1294, June 2002.

[WCLM-99]* C. Y. Wong, R. S. Cheng, K. B. Lataief, and R. D. Murch, "Multiuser OFDM System with Adaptive Subcarrier, Bit, and Power Allocation", IEEE J. Select. Areas Commun., vol. 17, pp. 1747–1758, Oct 1999.

[WSEA-04] Ian C. Wong, Zukang Shen, Brian L. Evans, and Jeffrey G. Andrews , "A Low Complexity Algorithm for Proportional Resource Allocation in OFDMA Systems", 2004

[WTCL-99]* C. Y.Wong, C. Y. Tsui, R. S. Cheng, and K. B. Letaief, "A real-time subcarrier allocation scheme for multiple access downlink OFDM transmission", in Proc. VTC 1999, vol. 2, 1999, pp. 1124–1128.

[YL-00]* H. Yin and H. Liu, "An Efficient Multiuser Loading Algorithm for OFDM-based Broadband Wireless Systems," in Proc. IEEE Global Telecommunications Conference, vol. 1, 2000, pp. 103–107.

* This reference has not been consulted by the authors of this article; rather, it had been referenced by the authors of one of the articles that were studied in this research.