Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems

Jonny SUN, Yanling YAO, Hongfei ZHU
Advanced System Technology Telecom Lab (Beijing)
STMicroelectronics
Beijing CHINA
jonny.sun@st.com; yanling.yao@st.com; hongfei.zhu@st.com

***Abstract*-IEEE 802.16, the specification for fixed, portable and even mobile Broadband Wireless Access (BWA) systems, is promising to support heterogeneous classes of traffic with differentiated Quality of Service (QoS). The proposed Medium Access Control (MAC) protocol defines a wide variety of mechanisms for bandwidth allocation and QoS provision. However, the details of how to schedule traffic are left unspecified so that the vendor may differentiate their product through implementation. In this paper, a new and efficient QoS scheduling strategy based on the hierarchical and distributed architecture is proposed for 802.16 BWA systems. Analytical and simulation results show that the proposed scheduling architecture can provide QoS guarantees for all types of traffic as defined in the standard.**

Keywords-802.16; MAC; QoS; Scheduling Policy

I. Introduction

Broadband Wireless Access (BWA) system is emerging as an integral part of the next generation wireless access infrastructure in recent years. It is originally developed to cater for the continued increasing demand on always-on high-speed Internet access. However, with the rapid growth trend in the use of wireless data services and multimedia applications, providing Quality of Service (QoS) in BWA systems becomes a very important and challenging issue. Therefore, the IEEE 802.16 standard, which is proposed as a BWA solution, is promising to provide differentiated levels of QoS for heterogeneous classes of services. Four types of scheduling service classes and a variety of bandwidth allocation mechanisms are defined in the Medium Access Control (MAC) protocol to support various types of traffic including CBR (Constant Bit Rate), real-time VBR (Variable Bit Rate), non-real-time VBR, and BE (Best Effort). However, the proposed MAC protocol does not include a complete solution for how to efficiently schedule traffic to fulfill the QoS requirements. In this paper, we will propose a new and efficient QoS scheduling strategy based on the hierarchical and distributed architecture. This architecture includes two layers of schedulers, i.e. Base Station (BS) scheduler and Subscriber Station (SS) scheduler. The BS scheduler grants bandwidth to SSs according to the bandwidth request and reservation, then SS scheduler should re-distribute the received transmission opportunities among all of its connections.

The remainder of this paper is organized as follows. Section II briefly introduces the IEEE 802.16 BWA systems. Some of the QoS-related features are described. Then the proposed QoS scheduling architecture and associated scheduling algorithms are introduced in Section III. In Section IV, the simulation model based on NS2 (Network Simulator) and corresponding simulation results are presented. Finally, Section V concludes the study and elaborates some of the future work directions.

II. IEEE 802.16 BWA Systems

IEEE 802.16 family is designed to evolve as a suite of air interfaces for fixed, portable and even mobile BWA systems. The first version, know as 802.16, was completed in October 2001. It specified a Single Carrier (SC) air interface for fixed point-to-multipoint (PMP) BWA systems operating between 10-66 GHz. The second amendment, 802.16a, was ratified in January 2003. It extends the physical environment towards lower frequency bands below 11 GHz. To optimize the deployment and operation in these bands, two OFDM-based air interfaces, 256-carrier Orthogonal Frequency Division Multiplex (OFDM) and 2048-carrier Orthogonal Frequency Division Multiple Access (OFDMA) are amended in this version. Furthermore, the MAC protocol is enhanced to support an optional MESH topology in addition to the mandatory PMP architecture. The recently approved version is 802.16d, which is published in June 2004 and also known as 802.16-2004. It incorporates all the previous versions to provide fixed BWA. Currently, IEEE is undertaking the standardization of 802.16e, which is expected to support full mobility up to 70-80m/s [1]. In this article, we study the 802.16d BWA systems employing PMP architecture.

As defined in 802.16, the PMP architecture is made up of a central BS and multiple independent SSs connected to the BS. In downlink, from BS to SS, the transmission is relatively simple because BS is the unique transmitter broadcasting to all the SSs without coordination with other stations. However, in the other direction, the uplink channel to BS is shared by all the SSs on a demand basis. Coordination among multiple SSs is necessary for uplink transmission and BS therefore holds the responsibility for controlling uplink system access and resources allocation. Depending on the QoS agreement between BS and SSs, the BS may issues data grants periodically as a result of bandwidth reservation for a particular SS or dynamically on receipt of the request from SS. To eliminate the overhead and delay of acknowledgements, the bandwidth request-grant mechanism utilized by 802.16 is self-

0-7803-9392-9/06/$20.00 (c) 2006 IEEE

correcting. The SS who needs to transmit over uplink should firstly request transmission opportunities from BS. BS then collects the requests from all the SSs and periodically broadcast a UL_MAP (Uplink Map) message over downlink to describe the uplink bandwidth allocation for a certain period. After receiving the UL-MAP, SS will transmit data as indicated. According to the UL_MAP message, some of the uplink bandwidth is dedicated for particular SSs to transmit and some is available for all the SSs to contend.

Since each SS is allowed to have multiple end-users, the MAC operation is designed to be connection-oriented, which enables end-to-end QoS for different end-users. At creation, each connection is associated with a unidirectional service flow that characterized by a set of QoS parameters. All the packets traversing the MAC interface should be mapped onto a connection and service flow so that they can be treated differently according to the QoS requirements. In 802.16, there are four types of scheduling services defined for different traffic models, i.e. Unsolicited Grant Service (UGS), real-time Polling Service (rtPS), non-real-time Polling Service (nrtPS) and Best Effort (BE).

- UGS

UGS is provided for real-time CBR or CBR-like traffic that generates fixed-sized data packets at periodic intervals, such as T1/E1 and VoIP without silence suppression. To eliminate overhead and latency of the request-grant process, SS is prohibited from using any explicit requests for UGS and BS should allocate unsolicited data grants periodically.

The key service parameters for UGS service are: Maximum Sustained Traffic Rate, Maximum Latency, Tolerated Jitter and Request/Transmission Policy.

- rtPS

RtPS is designed to support real-time VBR service flow that generate variable-sized data packets periodically, such as MPEG video, or VoIP with silence suppression. Since these applications have specific bandwidth requirements as well as a tight delay bound, BS should ensure periodic dedicated request opportunities for rtPS to request bandwidth dynamically. Due to the predictable signal delay of collision-free request, the bandwidth demand is guaranteed to be received by BS in time.

The key service parameters for rtPS service are: Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Maximum Latency, and Request/Transmission Policy.

- nrtPS

NrtPS is used for non-real-time VBR application that requires minimum bandwidth guarantee but can tolerate longer delay, such as bandwidth-intensive FTP. For nrtPS, BS offers dedicated request opportunities less frequently than rtPS. In case of the dedicated request opportunities can not satisfy the flow's bandwidth requirements, the contention request opportunities are allowed to be used as well.

The key parameters for nrtPS service are Maximum Sustained Traffic Rate, Minimum Reserved Traffic Rate, Traffic Priority, and Request/Transmission Policy.

- BE

BE is handled exactly in the same way as nrtPS except that the availability of dedicated request opportunities depends on the system load. Generally, BE service is required to contend for bandwidth during contention request opportunities. Therefore, it is suitable for such traffic as telnet and WWW, where no throughput or delay guarantees are demanded.

The key parameters for BE service are Maximum Sustained Traffic Rate, Traffic Priority, and Request/Transmission Policy.

III. QoS Scheduling for 802.16 Systems

Although the generic mechanisms for bandwidth allocation and QoS management have been defined in 802.16 standards, the details of how to efficiently schedule different types of traffic is left unspecified so that product differentiations may be achieved through different vendor implementations. In this section, we will propose a QoS scheduling strategy based on the hierarchical and distributed architecture.

A. *QoS Scheduling Architecture for 802.16 Systems*

According to the 802.16 standard, BS is responsible for the uplink bandwidth allocation based on the requests from SSs. Because a SS may have multiple connections at the same time, the bandwidth request messages should report the bandwidth requirement of each connection in SS. However, in response to the per connection requests, the allocated bandwidth is pooled together and granted to all the connections belonging to the SS. Then SS should re-distribute the received transmission opportunities among its connections. This allows a more sophisticated reaction to QoS needs, which may be useful for real-time applications that require a faster response from system. Hereunder, we proposed a hierarchical and distributed scheduling architecture that is compatible to the standard.

Fig. 1 illustrates the proposed architecture. It includes two layers of schedulers: BS scheduler and SS scheduler. The corresponding scheduling process is also divided into two steps.

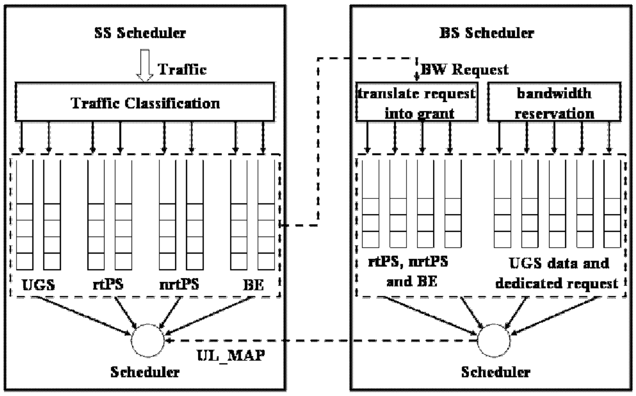

Figure 1. QoS scheduling architecture for 802.16 systems

The first step is performed in BS. Since the bandwidth allocation can be either a result of bandwidth reservation or a response for bandwidth request, we define two types of queues in BS scheduler, which are referred as Type I and Type II queues. The detailed scheduling mechanisms and algorithms are described in section B.

After individual SSs are assigned certain portion of uplink bandwidth, it comes to SS's turn to subdivide its assigned bandwidth sharing among all traffic connections belonged to it. Since each traffic connection is associated with one of the four scheduling services that represent different QoS requirements, SS scheduler should take these QoS requirements into account when it selects packets to be transmitted from respective UGS, rtPS, nrtPS and BE queues. The detailed scheduling mechanisms and algorithms are presented in section C.

B. *Scheduling Algorithm at BS*

In PMP topology, a central BS handles multiple independent SSs that may have several traffic connections. For the fairness of connections between different SSs, the bandwidth allocation procedure at BS is based on the requirement of each connection. Since the BS scheduler has limited information on the traffic generated at SS, the computing of bandwidth allocation should just consider the bandwidth request and reservation for each connection. Meanwhile, to provide differentiated QoS to varied kinds of connections, the amount of allocated bandwidth is determined by connection weight, which can be pre-assigned according to the scheduling service type.

- Scheduling algorithm for Type I queues

Type I queues are used to schedule data grants for UGS and allocate dedicated request opportunities for rtPS and nrtPS. Because the grants in these queues are periodically generated by BS independently, it is possible to control the timing of grant generation and hence achieve a strict QoS. To guarantee the generated grants to be scheduled without interruption, a First-in First-out (FIFO) discipline is employed.

Since the bandwidth allocated to Type I queues is reserved during connection setup, BS should process them prior to Type II queues.

- Scheduling algorithm for Type II queues

Type II queues are used to schedule data grants for rtPS, nrtPS and BE based on the information contained in the bandwidth request messages. To guarantee the minimum bandwidth for each service flow and ensure fairness in distributing excess bandwidth among all connections, we propose a fair queuing algorithm that consists of two phases: minimum reserved bandwidth assurance and excess bandwidth distribution.

Assuming that each connection is guaranteed with a minimum reserved bandwidth, the BS scheduler should firstly satisfy this part of requirement, which has been negotiated during connection establishment.

Let BiMIN denote the minimum reserved bandwidth for connection i, and BRi represent the bandwidth currently demanded by the connection. Since the connection will never get more resources than it has requested, the bandwidth actually allocated during this phase is

$$b_i^{MIN} = \min\{B_i^{MIN}, BR_i\}. \qquad (1)$$

For rtPS and nrtPS, BiMIN is specified by the QoS parameter termed Minimum Reserved Traffic Rate. For BE, since the QoS level is not guaranteed, BiMIN is set to zero.

Clearly, to guarantee the contracted bandwidth, the sum of minimum reserved bandwidth for all the connections should not exceed the available bandwidth B.

After each connection gets its guaranteed bandwidth, if there is still excess uplink bandwidth remained, BS scheduler should distribute the residual bandwidth in proportion to the pre-assigned connection weight. The algorithm in this phase can be described as:

$$B^{EX} = B - \sum_i b_i^{MIN} , \qquad (2)$$

$$b_i^{EX} = B^{EX} \times w_i / \sum_k w_k , \qquad (3)$$

where BEX is the excess bandwidth, biEX is the amount of excess bandwidth allocated to connection queue i and wi is the weight of connection queue i.

However, it is probable that a connection does not need so much bandwidth as its share. So the proposed algorithm allows the empty connection queue to contribute its unused portion to the next round of excess bandwidth allocation. This process of excess bandwidth allocation continues until all bandwidth is used up or all connection queues are empty.

While the scheduling algorithm in BS scheduler aims at individual connection, the bandwidth is granted to SS as an entity. SS should calculate the overall bandwidth allocated to all of its connections and proceed to next step described below.

C. *Scheduling Algorithm at SS*

The scheduler inside the BS may have only limited or even outdated information about the current state of each uplink connection due to the large Round Trip Delay (RTD) and possible collision occurred in the uplink channel transmission [2]. So we need an additional scheduler in each SS to reassign the received transmission opportunities among different connections. Since the uplink traffic is generated at SS, the distributed scheduler is able to arrange the transmission based on the up-to-date information and then provide tight QoS guarantee for its connections.

To provide differentiated and flexible QoS support for different scheduling services, the queuing algorithm proposed for SS scheduler should be tailored to the requirement of each service flow type. The priority of different scheduling services are specified in TABLE I.

TABLE I. PRIORITY OF SCHEDULING SERVICE

Scheduling Service	**Priority**
BE	1
nrtPS	2
rtPS	3
UGS	4

SS scheduler will select the packet to be transmitted from the highest priority queue that is not empty. Therefore, for packets in lower priority queues, their transmission requirements will be postponed until there is no packet available to send in a higher priority queue.

- Scheduling algorithm for UGS queues

UGS service has a critical delay and delay jitter requirement. Its transmission can not be deferred or interrupted by other flows. So SS scheduler will firstly guarantee the bandwidth for UGS queues.

- Scheduling algorithm for rtPS queues

For rtPS service, the scheduling algorithm should meet a tight delay bound. Each packet entering the rtPS queues should be marked with a delivery deadline equal to t + tolerated_delay, where t is the arrival time and tolerated_delay is the Maximum Latency for such a service flow. Then SS scheduler will schedule all of its rtPS packets based on the deadline stamp. The packet with smaller deadline will be transmitted earlier. This greatly reduces the end-to-end delay of rtPS service.

- Scheduling algorithm for nrtPS queues

The proposed algorithm for nrtPS queues targets at maintaining throughput. The specific method is similar to that for rtPS except that for this service we associate a virtual time with each packet [3]. When a new packet arrives in, the virtual time must be calculated at first. The virtual timestamp Vik associated with the kth packet of connection queue i is calculated as:

$$V_i^1 = t \qquad , (k = 1), \qquad (4)$$
$$V_i^k = \max(t, V_i^{k-1}) + L_i^k / r_i \, , (k > 1), \qquad (5)$$

where t is the packet arrival time, Lik is the length of this packet and ri is the guaranteed bandwidth share of connection i. With this algorithm employed, SS scheduler can guarantee the minimum bandwidth for every nrtPS connection and hence maintain an acceptable throughput.

- Scheduling algorithm for BE queues

For BE queues, since there is no QoS guarantee required, a simple FIFO mechanisms is applied.

IV. Simulation Model AND Result Analysis

To evaluate the effectiveness and efficiency of the proposed scheduler, we model the IEEE 802.16 MAC layer protocol using NS2. A number of simulations are conducted in this section. At first, we will describe the simulation environment and parameters. And then the simulation results will be presented for discussion.

A. *Simulation Environment and Parameters*

As mentioned before, we focus on the PMP MAC operation in this article. A TDD-OFDM system is used in our simulation and the network is configured as consists of one BS and multiple SSs.

Table II lists the PHY layer configuration parameters.

TABLE II. PHY Layer Configuration Parameters

PHY specification	OFDM
NFFT	256
bandwidth	7MHz
frame duration	4ms
symbol duration	0.33ms
duplex	TDD
modulation	QPSK
coding rate	½

B. *Simulation Result and Disscussion*

The first simulated 802.16 network consists of one BS and twenty SSs with different traffic patterns. The first SS is configured with all types of traffic flows nominated as UGS_1, rtPS_1, nrtPS_1 and BE_1, the second SS has the same application configuration as the first one, the third SS only generates rtPS_2 traffic, the fourth SS runs nrtPS_2, the fifth contains BE_2 and the other SSs are set to run BE flows acting as the background traffic. Two scenarios - with or without SS scheduler - are simulated to study the effect of SS scheduler. Here, "without SS scheduler" means that only the first step of our proposed scheduling process will be performed and BS scheduler will designate bandwidth to individual connection.

Fig. 2 and Fig. 3 display the end-to-end delay of different services with and without SS scheduler. The curves show that after SS scheduling, low priority service suffered longer delay. From UGS, rtPS, nrtPS to BE, the end-to-end delay increased with the service priority decreased. The fundamental requirement of QoS scheduling for 802.16 systems is achieved.

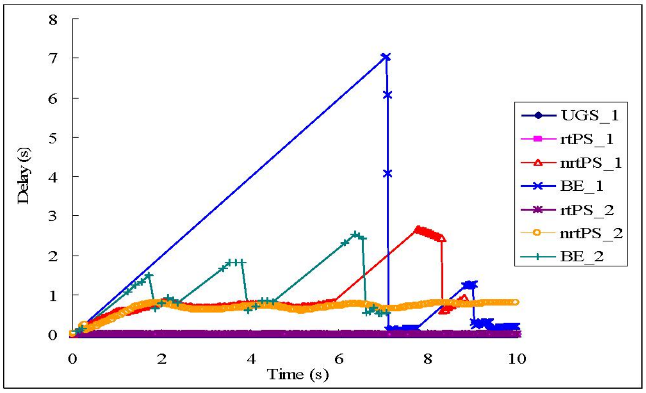

Figure 2. Service delay with SS scheduler (SS number=20)

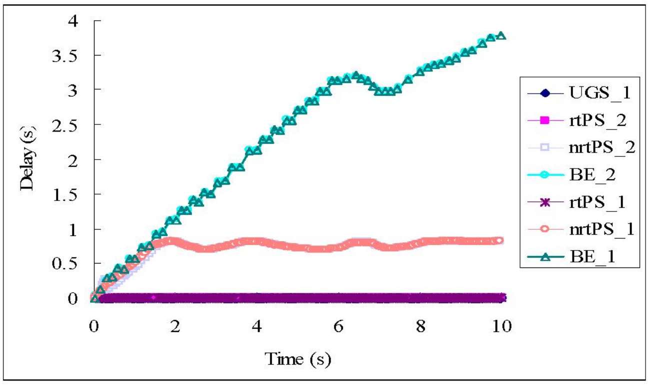

Figure 3. Service delay without SS scheduler (SS number=20)

Because of the relative small delay of UGS and rtPS flows, it is hardly to observe their performance in Fig. 2 and Fig. 3. So the UGS_1 and rtPS_1 flow's delays are separately presented in Fig. 4 and Fig. 5. From Fig. 4, we can see that the delay of UGS flow is far below its Maximum Latency. Moreover, Fig. 5 compares delay of rtPS service with and without SS scheduler. It is clearly that SS scheduler reduces the delay of rtPS flow.

To further demonstrate this benefit, we simulated the rtPS performance under different number of background SS. From Fig. 6, we can see that the SS scheduler can effectively reduce the QoS violation rate of rtPS service flow. Here, the QoS violation rate is defined as the amount of packets whose delay is larger than the Maximum Latency to the total amount of packets that have been received from network interface.

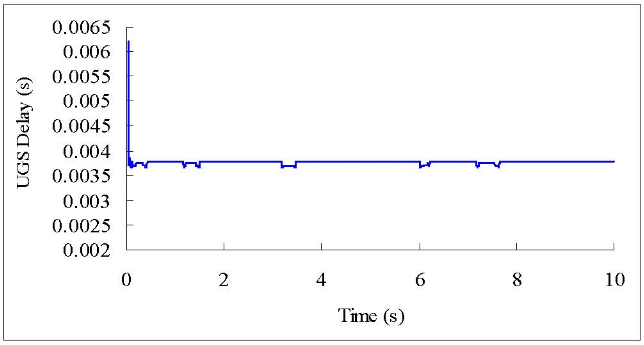

Figure 4. UGS_1 delay with SS scheduler (SS number=20)

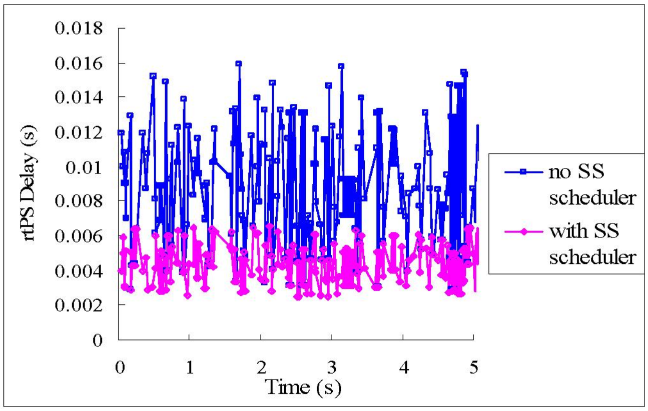

Figure 5. rtPS_1 delay (SS number=20)

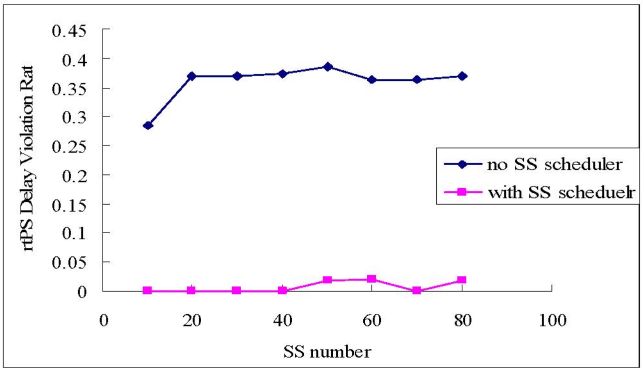

Figure 6. Violation rate of rtPS under different number of SS

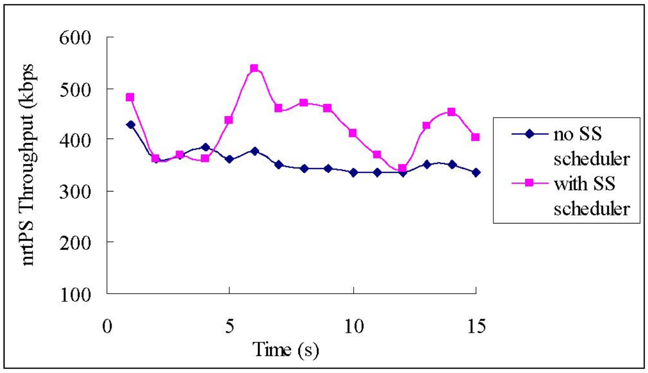

Figure 7. Throughput of nrtPS_1 (SS number=20)

By introducing virtual timestamp for nrtPS and scheduling all services on priority order, SS scheduler can help to increase the throughput of nrtPS service with impact on lower priority traffic - BE. The simulation result is shown in Fig. 7.

V. CONCLUSION

In this paper, the QoS scheduling mechanisms for 802.16 BWA systems are studied. A new and efficient QoS scheduling strategy based on the hierarchical and distributed architecture is proposed to provide differentiated levels of QoS guarantees to various applications. To evaluate the performance of proposed scheduling mechanism, the IEEE 802.16 MAC model is established based on NS2. Simulation results prove that the BS scheduler can guarantee the minimum bandwidth for each service flow and ensure fairness and QoS in distributing excess bandwidth among all connections. At the same time, the scheduler in SS can provide differentiated and flexible QoS support for all of the four scheduling service types. It can both reduce the delay of real-time applications and guarantee the throughput of non-real-time applications. Therefore, the proposed QoS scheduling architecture can provides QoS guarantees for all types of traffic classes as defined in the standard.

REFERENCES

[1] Arunabha Ghosh, David R. Wolter, Jeffrey G. Andrews, Runhua Chen, "Broadband wireless access with WiMAX/802.16: current performance benchmarks and future potential," IEEE Communication Magazine, pp. 129 - 136, Feb. 2005.

[2] CuoSong Chu, Deng Wang, Shunliang Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA system," IEEE Communications, pp. 435-439, Jul. 2002.

[3] H.Tayyar and H.Alnuweiri, "The Complexity of Computing Virtual-Time in Weighted Fair Queuing Schedulers," IEEE Communications, pp. 1996-2002, Jun. 2004.

[4] IEEE P802.16-REVd/D5-2004, "Draft IEEE Standard for Local and Metropolitan Area Network – Part 16: Air Interface for Fixed Broadband Wireless Access System," May 2004.

[5] IEEE P802.16e/D5, "Draft IEEE Standard for Local and Metropolitan Area Network – Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access System," Sep. 2004.

[6] IEEE P802.16e/D6, "Draft IEEE Standard for Local and Metropolitan Area Network – Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access System," Feb. 2005.