Uplink and downlink scheduling for point to multipoint WiMAX Networks

Mohammad Hayajneh, Najah Abu Ali*,** and Issa Khalil

College of IT, UAEU, Al-Ain, United Arab Emirates

Summary

IEEE 802.16 standard lacks a Media Access Control (MAC) scheduling architecture to enforce Quality of Service (QoS) requirements of different services classes. In this paper, we propose an opportunistic and optimized scheduler that meets connections’ QoS requirements while it pledges fairness among admitted connections. Our approach involves separating the scheduling problem into two sub-problems. In the first sub-problem, which addresses interclass time-slots allocation, the proposed scheduler calculates the optimum number of time-slots in each time frame corresponding to the service classes with the objective of minimizing the blocking probability of each class. In the second sub-problem, the intraclass slot allocation problem, time-slots for each class are allocated using an integrated cross-layer priority functions (PFs) that provide proportional fairness among a class connections. The simulation results reveal that the proposed scheduler realizes our objectives, and provides efficient QoS scheduling without starving the connections of the BE class. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: WiMax; scheduling; RRM; QoS; access networks

1. Introduction

Answering to an ever increasing demand for last mile high speed Internet access, based on IEEE 802.16 standard, Broadband wireless access (BWA) networks provide a large bandwidth sufficient to support multiple users sessions of voice, video, and data applications. Naturally, different applications demand different QoS requirements or levels such as delay, jitter, packet loss, and flow rate. For example, real time applications such as telemedicine, strictly require a nullified, deterministic delay and jitter. Other real time applications, like streaming multimedia, have soft bounds on delay. To meet these requirements, IEEE 802.16 is designed to support different service classes over Uplink (UL) and Downlink (DL) transmission. It defines the concept of service flow, where all traffic is carried on a connection, even for flows that implement connectionless protocols, such as Internet protocol (IP) [1]. For the purpose of transporting a service flow’s traffic, connections are identified by a connection identifier (CID) and mapped into one service flow once a subscriber station (SS) joined a network and a connection is established. Service flows provide a mechanism for UL and DL QoS management, mainly the bandwidth allocation process. Normally, bandwidth is allocated to a SS by a base station (BS) as a response to a per connection request from the SS. Bandwidth allocation may be constant depending on the type of service, for example, T1 unchannalized services, or it may

*Correspondence to: Najah Abu Ali, College of IT, UAEU, Al-Ain, United Arab Emirates.
**E-mail: najah@uaeu.ac.ae

Copyright © 2008 John Wiley & Sons, Ltd.
be adaptive such as that granted for the IP bursty services. IEEE 802.16-2004 [1] defines four service flow classes:

- Unsolicited grant service (UGS): supports real time services with constant bit data rate, such as voice over IP (VoIP) without silence suppression and T1/E1 emulation. The mandatory QoS service flow parameters for this service are minimum reserved traffic rate (MRTR), which equals to the maximum sustained traffic rate (MSTR), minimum latency (ML), tolerated jitter (TJ), and request transmission policy (RTP).
- Real time polling service (rtPS): supports real-time services with variable size data on a periodic basis, such as MPEG and VoIP with silence suppression. The mandatory QoS service flow parameters for this service are MRTR, MSTR, ML, and RTP.
- Non-real-time polling service (nrtPS): supports non-real-time services that require variable size data grant bursts on a regular basis, such as FTP. The mandatory QoS service flow parameters for this service are MRTR, MSTR, traffic priority, and RTP.
- Best effort (BE): for data streams for which no minimum service level is required and therefore, may be handled on space-available bases. The mandatory QoS service flow parameters for this service are MSTR, traffic priority, and RTP.

IEEE 802.16e [2] defines new scheduling mechanism, the extended rtPS (ertPS), which is based on UGS and rtPS services. ertPS is similar to UGS by providing unicast grants, thus saving the delay incurred for requesting the bandwidth. On the other hand, ertPS allocations are dynamic as rtPS while UGS allocations are fixed. The ertPS is introduced to support real-time service flows that generate periodical variable-sized data packets. Thus, ertPS is especially important to support VoIP, since it allows for managing traffic rates and improves latency and jitter.

Even though IEEE 802.16 defines different service flows and their associated QoS parameters, the standard did not define a packet scheduler that enforces the required QoS parameters to various service flows associated with users’ applications. Therefore, a scheduling algorithm is required to efficiently provide for the different applications’ QoS requirements. We propose a scheduling algorithm for IEEE 802.16 point to multipoint (PMP) mode that enforces the QoS requirements for different service flows, mainly bandwidth and delay bounds requirements. The problem is divided into two sub-problems. First, interclass slots allocation problem, where slots are optimally calculated to satisfy each service class QoS requirements, if this is applicable. Second, intraclass slot allocation problem, where connections are selected based on a fair opportunistic priority function (PF). The proposed scheduling scheme utilizes not only the multiuser diversity [3] through the instantaneous channel conditions of different users, but also the users’ QoS requirements and long term fairness among them.

Performance evaluations of IEEE 802.16 can be found in References [4–9] where authors implemented well-known schedulers such as weighted fair queuing (WFQ), deficit round robin (DRR), and earliest deadline first (EDF). Normally, implementing different schedulers for the different service classes and on a per connection bases is not a trivial task for the following reasons: (1) IEEE 802.16 standard supports changing the connection QoS descriptors during connections life time. Thus, the schedulers are required to adapt for these changes by reassigning the slots. For instance, in case of WFQ, dynamic weights may be the solution, which considerably increases the scheduler complexity and render its implementation challenging. (2) In IEEE 802.16 frame size is constant, consequent, and based on each service class load, some of time-slots may not be allocated, which makes IEEE 802.16 behavior closer to the non-conserving schedulers, meanwhile the proposed hierarchy of schedulers is work-conserving. To address these issues, some authors based their work on some assumptions. For example, in Reference [9] and for the sake of simplicity, the authors implement DRR as DL scheduler, irrespective of the connection scheduling service, and weighted round robin (WRR) as UL scheduler with constant weights, without clarifying how the weights are chosen or provide a justification for adopting constant weights. On the other hand, some authors instead of designing a hierarchical scheduler, focused on scheduling of one service class for certain applications (e.g., VoIP in Reference [10]).

Another group of schedulers is based on cross layer optimization [11–13]. The work in Reference [11] presents a QoS architecture for IEEE 802.16 broadband wireless MAN in time division duplex (TDD) mode to provide for InterServ, DiffServ, and signaling under PMP and mesh mode. The paper is toward proposing a QoS architecture rather than focusing on scheduling. The scheduler in Reference [12] utilizes users diversity, however, it does not provide for fairness among users since it allocates slots for one connection in the frame after satisfying the UGS connections requirements. The connections are selected based on a PF that
SCHEDULING FOR POINT TO MULTIPOINT WIMAX NETWORKS

depicts the QoS requirements of the connections and prioritizes the different service classes through using constant weights.

The third category of schedulers includes the schemes which implement fairness approaches such as References [13,14]. The scheduler in References [13] utilizes users diversity for bandwidth allocation, however, authors propose a heuristic algorithm for bandwidth allocation instead of an optimal bandwidth allocation algorithm since the complexity of their proposed optimal bandwidth allocation algorithm may be prohibitive from an implementation point of view.

Our work is different from the aforementioned literature by supporting all service classes, optimally calculates number of slots in each frame such that the blocking probability of each class is minimized, provides isolation between classes, and integrating a proportional fairness scheme to SS channel quality information to allocate time-slots among connections of the same class. We proceed by introducing the mathematical model in Section 2, discuss the performance evaluation and results in Section 3, and conclude the paper in Section 4.

2. Scheduler Approach

Before proceeding in presenting the mathematical model of the scheduler, we introduce some preliminaries essential for the mathematical model.

2.1. Preliminaries

In this paper, we propose a scheduler for an IEEE 802.16 wireless networks. We consider a single cell consists of one BS with several SSs in a PMP mode. Frequency division duplex (FDD) or TDD multiplexing can be used to divide transmission time into UL and DL periods. On DL, data to SSs are multiplexed in time division multiplexing (TDM) fashion and generally broadcasted to all SSs capable to listen to the DL frame. Each SS checks the connection ID in the protocol data units (PDUs) and retains the PDUs addressed to it. The UL is shared between SSs implementing a medium access mechanism such as TDMA or OFDMA on demand bases. Thus, the BS allocates the total number of time-slots of UL subframe to the SSs based on their admitted connections’ number and to which service class they belong. Afterwards, the SS allocates these granted time-slots to its connections in an optimal and a fair manner.

Table I. Modulation and coding schemes for IEEE 802.16.

<table>
<thead>
<tr>
<th>Modulation (coding)</th>
<th>Info bits/symbol</th>
<th>Required SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPSK(1/2)</td>
<td>0.5</td>
<td>6.4</td>
</tr>
<tr>
<td>QPSK(1/2)</td>
<td>1</td>
<td>9.4</td>
</tr>
<tr>
<td>QPSK(3/4)</td>
<td>1.5</td>
<td>11.2</td>
</tr>
<tr>
<td>16QAM(1/2)</td>
<td>2</td>
<td>16.4</td>
</tr>
<tr>
<td>16QAM(3/4)</td>
<td>3</td>
<td>18.2</td>
</tr>
<tr>
<td>64QAM(2/3)</td>
<td>4</td>
<td>22.7</td>
</tr>
<tr>
<td>64QAM(3/4)</td>
<td>4.5</td>
<td>24.4</td>
</tr>
</tbody>
</table>

Through a perfect and robust feedback channel, a SS conveys its received signal to noise power ratio (SNR) from the BS. Therefore, it is assumed that the BS has complete channel state information of all connections, this information is not changing over one time frame, since the wireless channel between each SS and BS is assumed to undergo a flat fading that is fixed over one frame period. The adaptive modulation coding (AMC) module divides the range of the received SNR into seven non-overlapping regions where the dividing thresholds are evaluated based on a target prescribed bit error rate (BER). According to the received SNR at a specific SS, the module in the PHY layer of BS decides the suitable transmission mode for each SS as shown in Table I. Each transmission mode consists of modulation and coding pair aims at efficiently using the bandwidth while satisfying a prescribed BER.

2.2. Mathematical Model

The proposed scheduler is designed to solve two sub-problems: inter-class provisioning and intra-class scheduling. The former allocates time slots among the service classes. The later allocates time slots within one class among the class active connections. Thus, the scheduler isolates these service classes by allocating $K_i$ time-slots for each class $i$ out of $K$ total time-slots in a given time frame.

In IEEE 802.16 time frames are divided into constant number of time slots with same time-slot duration, however, the number of bits per slot may change on frame bases depending on the AMC mode selected from Table I as a result of the channel condition. The total $K$ time-slots are completely partitioned among the classes such that the blocking probability of each class is minimized. Consequently, we model each class as a $M/M/1/K_i$ queue. $K_i$ represents the number of connections that will be served simultaneously over $K_i$ time-slots in a given time frame period, given that no more than one connection can be allocated to one time-slot. The blocking probability of a $M/M/1/K_i$ queue
is given by
\[
p(K_i) = \frac{(1 - \rho_i)\rho_i^{K_i}}{(1 - \rho_i^{K_i+1})}
\]  
(1)

where \( \rho_i = \lambda_i/\mu_i \) is the ratio of the arrival rate (\( \lambda_i \)) to the departure rate (\( \mu_i \)) of the connections of class \( i \). Intuitively, the number of time-slots allocated for a given service class should depend on the requested bandwidth of all connections belong to this class and their channel status represented by their attainable rates.

We define the following nonlinear mathematical program to optimally calculate the required time-slots for each service class such that the blocking probability of each class is minimized.

\[
\min_{K_i, 1 \leq i \leq 4} \sum_{i=1}^{4} \beta_i p(K_i)
\]  
(2)

subject to

\[
K \leq \sum_{i=1}^{4} K_i
\]  
(3)

\[
K_i^{\min} \leq K_i \leq K_i^{\max}
\]  
(5)

\[
K_1^{\min} \leq K_1^{\max}
\]  
(6)

\[
K_2^{\min} \geq n
\]  
(7)

\[
K_3^{\min} \geq 0
\]  
(8)

\[
K_4^{\min} \geq 0
\]  
(9)

where \( n \) is the number of admitted rtPS connections.

The strictly convex cost function in Equation (2) represents the total sum of all weighted blocking probabilities of all classes. The weights \( \beta_i = \sum_{k=1}^{4} BW_k^{\text{req}}/\sum_{k=1}^{4} BW_k^{\text{req}} \) are introduced to reflect the priority of the different classes. \( BW_k^{\text{req}} \) is the total requested bandwidth of each class \( k \). The constraint in Equation (4) is to satisfy that total number of time-slots in a time frame is fixed and equal to \( K \). Constraints in Equation (5) are to guarantee that all connections of all classes will be assigned enough number of time-slots to achieve at least their minimum required bandwidth taking into account their priorities. \( K_1^{\min} \) is a constraint to guarantee that UGS active admitted connections requirements are met. \( K_1^{\max} \) is to guarantee that the QoS requirements of admitted active connections in addition to new UGS connections arrivals are met, if enough capacity is available, otherwise some or all new arrivals are rejected. Note that UGS maximum sustainable rate is equal to the minimum reserved rate, thus \( K_1^{\min} \) is the summation of constant data rates of all admitted UGS connections, while, \( K_1^{\max} \) is the summation of all admitted and new UGS connections. For rtPS and nrtPS minimum reserved rate and maximum sustainable rate are denoted by \( K_2^{\min}, K_2^{\max}, K_3^{\min}, \) and \( K_3^{\max} \), respectively, given that the bandwidth requests for these two classes are adaptive on a per frame basis, these values convey different information than \( K_1^{\min} \) and \( K_1^{\max} \) as clarified below. In the UL case, the minimum value of \( K_2^{\min} \) should be one time-slot for each SS to be used for adapting bandwidth requests in case that the SS has nothing to send at the current time frame, because rtPS class is not allowed to contend for the channel during the contention time slots [1]. On the other hand, in the DL case, \( K_2^{\min} = 0 \). For nrtPS the minimum value of \( K_3^{\min} \) may be 0 ——there are no backlogged packets—— since nrtPS SSs are allowed to participate in the contention period. \( K_4^{\min} \) is equal to one time-slot and \( K_3^{\max} \) is equal to the maximum sustainable rate for all BE connections, because we do not want to allocate more time-slots for any BE connection than its maximum sustainable rate. Mapping the requested bandwidth to a maximum and minimum number of time-slots is based on the fact that IEEE 802.16 standard defines fixed size of time frames and their corresponding number of frames sent per second as shown in Table II. Thus, given a requested bandwidth requirement per second for each connection \( n \), the requested bandwidth requirement per frame is equal to the requested bandwidth per second divided by the number of frames per second. Thus, given minimum reserved bandwidth per frame, \( (\hat{B}_n^{\min}) \), and maximum sustainable bandwidth per frame \( (\hat{B}_n^{\max}) \), number of time-slots for each connection \( n \) is given by

\[
\hat{K}_n^{\min} = \frac{\hat{B}_n^{\min}}{r_n}
\]  
(10)

\[
\hat{K}_n^{\max} = \frac{\hat{B}_n^{\max}}{r_n}
\]  
(11)

### Table II. OFDM frame duration and number of frames per second.

<table>
<thead>
<tr>
<th>Code</th>
<th>Frame duration (ms)</th>
<th>Frames per second</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.5</td>
<td>400</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>250</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>125</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>12.5</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>50</td>
</tr>
</tbody>
</table>
where \( r_n \) is the number of bits attainable in one time slot. The value \( r_n \) can be used as the lowest AMC level mode for the new connections. For admitted connections, \( r_n \) can be used as an averaged value that reflects the channel quality over a predefined window size using a proper prediction algorithm. Then, the minimum and maximum number of time slots of service class \( i \) are: \( K_i^{\text{min}} = \sum_{k=1}^{\left| C_i(t) \right|} K_i^{\text{min}}_k \) and \( K_i^{\text{max}} = \sum_{k=1}^{\left| C_i(t) \right|} K_i^{\text{max}}_k \), respectively. \( \left| C_i(t) \right| \) is the number of connections of class \( i \).

The problem in Equations (1–5) is a nonlinear convex programming problem in a standard form and it is equivalent to finding the minimum of the following cost function:

\[
L(\{K_i\}_1^4, \alpha_1, \alpha_2) = \sum_{i=1}^{4} \beta_i p(K_i) + \alpha_1^{5}(K - \sum_{i=1}^{4} K_i)
\]

\[
+ \sum_{i=1}^{4} \alpha_1^{4}(K_i - K_i^{\text{min}})
\]

\[
+ \sum_{i=1}^{4} \alpha_2^{4}(K_i^{\text{max}} - K_i)
\]  \hspace{1cm} (12)

where \( \alpha_1 = (\alpha_1^1, \ldots, \alpha_1^4) \) and \( \alpha_2 = (\alpha_2^1, \ldots, \alpha_2^4) \) are the Lagrange multipliers. Differentiating 12 with respect to \( K_m \) and setting the derivative to zero we obtain:

\[
\frac{\partial L}{\partial K_1} = \beta_1 \frac{(1 - \rho_1) \log(\rho_1) K_1^1}{(1 - \rho_1 K_1^{1+1})^2} - \alpha_1^5 + \alpha_1^4 - \alpha_2^4 = 0
\]  \hspace{1cm} (13)

and

\[
\frac{\partial L}{\partial K_m} = \beta_m \frac{(1 - \rho_m) \log(\rho_m) K_m^m}{(1 - \rho_m K_m^{m+1})^2} - \alpha_1^5 + \alpha_1^4 - \alpha_2^4 = 0
\]  \hspace{1cm} (14)

where \( m = 2, 3, 4 \). Using both Equations (13) and (14) and setting \( X_m = \rho_m K_m^m \) we have

\[
X_m \left( 1 - \rho_m X_m \right)^2 = \frac{\beta_1 (1 - \rho_1) \log(\rho_1) K_1^1}{\beta_m (1 - \rho_m) \log(\rho_m) (1 - \rho_1 X_1)^2}
\]  \hspace{1cm} (15)

Let

\[
\chi_{1,m} = \frac{\beta_1 (1 - \rho_1) \log(\rho_1) X_1}{\beta_m (1 - \rho_m) \log(\rho_m) (1 - \rho_1 X_1)^2}
\]  \hspace{1cm} (16)

Then, Equation (15) is rewritten as

\[
\rho_m^2 X_m^2 - X_m \left( 2 \rho_m + \frac{1}{\chi_{1,m}} \right) + 1 = 0
\]  \hspace{1cm} (17)

which is a quadratic equation in \( X_m \) that has one feasible solution given as

\[
X_m \triangleq f(K_1, \rho_1, \rho_m, \beta_1, \beta_m)
\]

\[
= \frac{1}{2 \rho_m^2} \left( 2 \rho_m + \frac{1}{\chi_{1,m}} + \sqrt{4 \rho_m + \left( \frac{1}{\chi_{1,m}} \right)^2} \right)
\]  \hspace{1cm} (18)

Now, we can express \( K_m, m = 2, 3, 4 \) as a function of just \( K_1 \) as follows:

\[
K_m = \frac{\log(f(K_1, \rho_1, \rho_m, \beta_1, \beta_m))}{\log(\rho_m)}
\]  \hspace{1cm} (19)

Using the constraints in Equation (2), one can find \( K_1 \) as the solution of

\[
K_1 + \sum_{m=2}^{4} \frac{\log(f(K_1, \rho_1, \rho_m, \beta_1, \beta_m))}{\log(\rho_m)} = K
\]  \hspace{1cm} (20)

Using an iterative technique, Equation (20) converges to a unique solution in few iterations.

After optimally evaluating \( \{K_i\}_{i=1}^4 \) and isolating the four QoS classes, the connections of each class competes for the available time-slots. Given that we may have new arrivals at this time frame, the main concern is to satisfy the currently admitted connections in each class, then the residual time-slots, if any, is calculated to admit some or all new connections. For UGS class connections, since the minimum reserved bandwidth is equal to the maximum reserved bandwidth, we define \( \bar{K} \) as \( K_1(t) - K_1(t - 1) \), where \( t \) is the current time frame. If \( \bar{K} > 0 \), new UGS connections will be admitted, otherwise new connections are rejected because their is not enough capacity to admit them all. For rtPS and nrtPS the current active connections will contend for time-slots and any excess time-slots will be utilized to admit new connections because we choose the new arrivals such that the summation of the minimum reserved bandwidth of the new arrivals in addition to that of the admitted connections will not exceed the optimal calculated time slots \( K_2 \) and \( K_3 \) for rtPS and nrtPS, respectively.

DOI: 10.1002/wcm

Copyright © 2008 John Wiley & Sons, Ltd.
In the following section, we propose the service classes PFs of the four service classes supported by the scheduler.

2.2.1. Downlink scheduling

- **UGS**: the new connections are selected according to the following proposed PF, \( u_m(t) \), \( \forall m \in C_1(t) \); \( \hat{C}_1(t) \) is the set of all UGS connections at time frame \( t \), which is given by

\[
u_m(t) = \begin{cases} \frac{r_m^1(t)}{r_m^1(t) + \Delta t_m^1}, & \Delta t_m^1 > 0, r_m^1(t) \neq 0 \\ \infty, & \Delta t_m^1 = 0, r_m^1(t) \neq 0 \\ 0, & r_m^1(t) = 0 \end{cases} \quad (21)
\]

where \( r_m^1(t) \) is the \( m \)th connection attainable bandwidth at time frame \( t \) and it reflects the channel quality between the BS and SS of connection \( m \). \( \hat{r}_m^1(t) \) is the average throughput for connection \( m \) at time \( t \) estimated over a window size \( 1/\tau \) and it is updated as follows:

\[
r_m^1(t + 1) = \begin{cases} \hat{r}_m^1(t)(1 - \tau), & m \notin \hat{C}_1^b(t) \\ \hat{r}_m^1(t)(1 - \tau) + \tau r_m^1(t), & m \in \hat{C}_1^b(t) \end{cases} \quad (22)
\]

The ratio \( r_m^1(t)/\hat{r}_m^1(t) \) is used to provide fairness among users. If a user suffered deep fading channel over multiple frames such that the value \( \hat{r}_m^1(t) \) and the channel condition has improved at the current frame, this part of the PF will allow the user to have higher value of the function \( u_m^1(t) \) which means higher priority to compensate for his low transmission rates in previous frames due to his deep fade channel conditions. The value \( \frac{1}{\Delta t_m^1} \) is used to prioritize users who spend longer time in the system waiting for the service as long as their waiting time does not exceed their delay bound requirement. \( \hat{C}_1^b(t) \subseteq C_1^b(t) \) is the subset of connections that were selected to be served at the current time frame. The residual time to reach the latency bound of connection \( m \) is expressed as \( \Delta t_m^1 \).

- **rtPS**: the metrics for this QoS class are minimum reserved bandwidth and latency bound, however rtPS tolerates more delay than UGS. Therefore, we propose the following PF \( u_m^2(t) \), \( \forall m \in C_2(t) \) to schedule rtPS connections:

\[
u_m^2(t) = \begin{cases} \frac{r_m^2(t) \frac{1}{\Delta t_m^2}}{q_m^2}, & \Delta t_m^2 > 0, r_m^2(t) \neq 0, q_m^2 \neq 0 \\ \infty, & \Delta t_m^2 = 0, r_m^2(t) \neq 0, q_m^2 \neq 0 \\ 0, & r_m^2(t) = 0 \text{ or } q_m^2 = 0 \end{cases} \quad (23)
\]

here, \( C_2(t) \) is the set of all rtPS connections at time frame \( t \), and \( \hat{C}_2^b(t) \subseteq C_2(t) \) is the set of the served connections at time frame \( t \) based on the PF \( u_m^2(t) \), that is, \( |\hat{C}_2^b(t)| = K_2 \). The size of the queue of connection \( m \) is denoted as \( q_m^2 \) and \( q_2^2 = \max_{m \in \hat{C}_2^b(t)} q_m^2 \). This is to take the amount of backlogged packets waiting for transmission into consideration; that is, prioritize connections with longer queues. The ratios \( r_m^2(t)/\hat{r}_m^2(t) \) and \( r_m^2(t) \) serve the same purpose as explained above.

- **nrtPS**: the priority metric for this class is the minimum reserved bandwidth, henceforth, we propose the following PF, \( u_m^3(t) \), \( \forall m \in \hat{C}_3(t) \) such that

\[
u_m^3(t) = \frac{r_m^3(t) q_m^3}{\hat{r}_m^3(t) q_3^2}, \forall m \in \hat{C}_3(t) \quad (24)
\]

where \( q_3^2 \) is introduced to guarantee that no connection is scheduled if it has no packets to transmit even if it has a good channel quality.

- **BE**: in this class, there are no QoS requirements, therefore we propose a proportional fair PF \( u_m^4(t) \), \( \forall m \in \hat{C}_4(t) \), expressed as

\[
u_m^4(t) = \frac{r_m^4(t)}{\hat{r}_m^4(t)}, \forall m \in \hat{C}_4(t) \quad (25)
\]

in which \( r_m^4(t) \), the attainable bandwidth of connection \( m \), captures the channel quality. The connection with better channel quality will have higher priority to increase the system throughput, while the average bandwidth allocated over a window size \( 1/\tau \) in the PF is to provide fairness for all connections.

2.2.2. Uplink scheduling

To discriminate the UL class allocated slots from the DL ones, we assume that a SSi was assigned \( M_i \) timeslots for UL transmission, with \( M_i = M_i^{ugs} + M_i^{rt} + M_i^{nrt} + M_i^{be} \). Where \( M_i^{ugs}, M_i^{rt}, M_i^{nrt}, M_i^{be} \) are the number of timeslots assigned (by the BS) to the UGS, rtPS, nrtPS, and BE class connections, respectively. Denote the number of UGS, rtPS, nrtPS, and BE admitted
connections at SS, that will be allowed to compete for the remaining time-slots by \( NC_i \), \( NC_i^{rt} \), \( NC_i^{pr} \), and \( NC_i^{be} \), respectively. The number of time-slots \( M_i \) allocated to SS are assumed to be enough to support the \( NC_i \) connections, that is, \( NC_i \leq M_i \) (only one connection can send data in one time-slot).

To exploit the idea that some connections can tolerate to stay idle for a given time frame, while other connections may require to send data that cannot be sent in one time-slot, we propose the following optimization problem. After satisfying the demand of all UGS connections, the remaining time-slots are allocated to the rtPS, nrtPS, and BE connections according to the following objective function:

\[
J \left( \{ B_{i,m} \}_{m=1}^{NC_i^*} \right) = \sum_{m=1}^{NC_i^*} \log(1 + B_{i,m}Z_{i,m}) + \gamma \left( M_i^* - \sum_{m=1}^{NC_i^*} B_{i,m} \right) \tag{28}
\]

where \( M_i^* \) and \( NC_i^* \) are located to SS, respectively. The number of time-slots allocated to SS are assumed to be enough to support the \( NC_i \) connections, that is, \( NC_i \leq M_i \) (only one connection can send data in one time-slot).

We implemented in house simulator module consisting of a TDD cell of one BS and several SSs. We did not make use of the NS2 simulator since the extension of NS2 for WiMAX is based on OFDMA and the physical layer channel model is based on block-fading model suitable for slow varying channel conditions. Our simulator is based on Nakagami-m
channel model which is adopted to accurately describe the statistical variation of the channel gains between the BS and the SSs based on OFDM channel multiplexing. On the other hand, our simulator is similar to NS2 in the sense it implements the convergence sublayer (CS), the MAC common part sublayer (CPS), and the PHY layer. The model implements functionalities such as ranging, MAC management, IP-service flow ID (IP-SFID) mapping and ID-transport connection ID (SFID-TCID) mapping. IP-SFID mapping is used to record the characteristics of the packets coming from the upper layers while SFID–TCID is used to map the QoS characteristics of the SSs to one of the four traffic classes (UGS, rtPS, nrtPS, and BE).

Connections from each service class arrive following exponential distribution process with an exponential holding time. However, the packets arrival within each connection is implemented as a poisson process. For the connections’ traffic model, we implemented part of the traffic model presented in Reference [15], since it is specifically designed and tested for WiMAX simulation. This model implements VoIP traffic for the UGS class, video streaming traffic for the rtPS class, FTP traffic for the nrtPS class, and background traffic for the BE class. Table III shows the traffic model used in simulation. For each simulation scenario, experiments are repeated 20 times and results are obtained with 95% confidence interval lower than 0.0022.

The DL bandwidth is simulated as 20 MHz. The channel quality of each SS remains constant per frame, but is allowed to vary from frame to frame. Therefore, the channel quality is captured using a single parameter, the instantaneous SNR, which remains constant for the duration of the frame. Based on AMC and as defined by the IEEE 802.16 standard, we implemented the SNR to be mapped into six transmission modes as shown in Table IV. These values are in conformance with the IEEE 802.16-2004 standard for channel bandwidth of 20 MHz.

Nodes are placed in random over a simulation grid of 5000 × 5000 m². Number of SSs varies from 1 to 30. Each SS can have different types of connections at the same time; that is voice, video, FTP, or background traffic. The frame size is fixed at a value of 10 ms equally divided between UL and DL traffic, the symbol duration is 12.5 μs, and the rate of frames is 100 frames/s. Time-slots, which are directly related to OFDM symbols, are allocated for active flows at the beginning of each frame, where one connection is the maximum number of connections allocated per one time-slot.

3.1. Downlink Scheduling

We evaluated the performance of the scheduler under heavy loaded network, where the queues at the BS are always backlogged. Backlogged queues are due to backlogged packets and/or backlogged connections. Figures 1 and 2 show the performance evaluation of the scheduler to meet the delay bounds of the UGS and rtPS service classes during simulation time. The Figures show the delay over a truncated interval of the simulation time to allow Figures to be readable. Figure 1 shows the average delay practiced by all connections of UGS class with a delay bound of 80 ms. The Figure shows that the max delay practiced by the UGS flow packets does not exceed 65 ms and the minimum is around 0. Moreover, we observe that this variation is almost constant since the UGS traffic has a constant number of slots to meet its bandwidth requirement.

![Fig. 1. Simulation time versus the average delay bound of all UGS class connections.](image-url)
SCHEDULING FOR POINT TO MULTIPOINT WIMAX NETWORKS

Figure 2 shows the performance of the scheduler in meeting the delay bound of all rtPS connections. The maximum delay practiced by rtPS connections is 155 ms, while the delay bound requirement is 180 ms. The delay variation is not constant as in the UGS case, since the requested bandwidth is adaptive and the number of time-slots allocated for each flow within the class is varying over the different frames.

Figure 3 shows the mean of the moving average throughput, \( \phi_i \), of all connections of rtPS class. \( \phi_i \) is calculated over a window size, \( \tau \), which is a multiple number of a frame duration; that is, \( \frac{\tau}{10\text{ ms}} \) is the last frame in the window interval. The moving average, \( \phi_i \), of a rtPS connection is given by

\[
\phi_i(t) = \left(1 - \frac{1}{\tau}\right)\phi_i(t - 1) + \frac{1}{\tau} R_i(t),
\]

\[ t = 1, 2, \ldots, \frac{\tau}{10\text{ ms}} \]  \hspace{1cm} (35)

if connection \( i \) is scheduled during the current time frame, \( t \). \( R_i(t) \) is the attainable rate in the current frame of the SS of connection \( i \). If the connection is not scheduled in the current frame then \( \phi_i(t) = \left(1 - \frac{1}{\tau}\right)\phi_i(t - 1) \).

The Figure shows that the scheduler is capable of meeting the QoS parameters of rtPS connections given in Table III. Note that during the window interval 78 and 82 some of the SS suffer bad channel quality which results in decreasing the average throughput, while the opposite is noticeable during window interval 45. Figure 4 shows the average throughput of all nrtPS connections, again the proposed scheduler was able to honor the QoS requirements of nrtPS connections.

Figure 5 shows the throughput of BE traffic. Note that the BE connections are not starved over the simulation time. This is because we considered the BE as part of our mathematical formulation to minimize the blocking probability. The throughput of the BE traffic is the smallest because in the mathematical formulation we assume that the minimum number of slots assigned for the BE is one slot and the maximum number of slots is the number of slots necessary to meet the maximum sustainable rate, thus, the number of slots assigned for the BE traffic is the least.

Figure 6 shows the packet loss of the rtPS and UGS flows with and without implementing the PFs as the
number of connections increases. Even though in the two cases under study we implemented the inter-class provisioning part of the algorithm; that is, interclass provisioning, the performance of the algorithm implementing the PF in the intraclass stage is better than that without implementing the PFs. In the former case, packet loss is decreased by more than the half. Thus, the algorithm in its two phases; interclass and intraclass is capable of lowering the cost of the packet loss.

Figure 7 shows the throughput of the nrtPS and rtPS traffic with and without implementing the first phase of the algorithm, interclass provisioning, as the number of connections in the cell increase. In this experiment, we study the two classes of traffic under the same conditions where both classes’ flows require the same MRTR, 0.5 Mbps, and MSTR, 4 Mbps. The Figure shows that the performance of the algorithm is better implementing the first phase than that without it. Without the first phase the algorithm was not able to meet the minimum required bandwidth of the nrtPS traffic. The intraclass PF scheduling is implemented in the two cases.

3.2. Uplink Scheduling

The performance of the UL scheduler is the same as that for DL scheduler because of the similarities between the two schedulers. However, since the UL scheduler uses the excess time slots of the other classes based on their QoS requirements and PFs, we further investigate this merit in this section. The PF of rtPS is similar to that of nrtPS—the difference is in the delay parameter of rtPS traffic. Thus, without loss of generality, we consider two services classes, BE and rtPS, to evaluate the UL scheme, Figure 8 shows the performance of the scheduler in meeting the delay bound of the rtPS connections. The figure shows that non of the packets practiced a delay a round zero because the scheduler monitors the delay bounds of packets, if packets can tolerate the delay until the next frame, the scheduler retain the corresponding time slot to be assigned to BE traffic to avoid its starvation. Otherwise, The packets are scheduled in the current frame to provide delay bound guarantees for the connections.

Figures 9 and 10 show the performance of the scheduler serving the BE connections with a maximum sustained rate of 2 Mbps. Figure 9 shows that the average
throughput of the BE flows served by the algorithm is higher than that served by allocating the BE class slots only, Figure 10. This is due to the fact that the excess time-slots of rtPS is allocated to the BE flows which increases their average throughput without degrading the rtPS QoS performance.

4. Conclusions

We presented an opportunistic cross layer based scheduler to support all QoS classes in IEEE 802.16 standard. Our main contribution is that the scheduler optimally calculates the number of time-slots required to meet the QoS requirements of the connections within each class such that the blocking probabilities of connections are minimized. The proposed scheduler selects connections within each class that have the highest value of a PF which facilitate the long-term fairness among connections. The PF is based on the connections average throughput, current attainable transmission rate and the connection QoS requirements. In the UL scheduler, since large proportion of traffic is expected to be BE, the algorithm satisfies the rtPS connections first then any excess time slots of the rtPS class are assigned to BE. Performance evaluation of the scheduler shows that the scheduler is capable of meeting all connections QoS requirements without starving the BE class connections by increasing its share of time-slots.

References

2. IEEE Standard 802.16 Working Group. IEEE 802.16-2005e Standard for Local and Metropolitan Area Networks: Air Inter-

SCHEDULING FOR POINT TO MULTIPORT WIMAX NETWORKS

Fig. 10. Simulation time versus the average transmission rate of BE without using excess slots time.

Mohammad Hayajneh (SM’01, M’04) received his B.Sc. and M.Sc. in Electrical Engineering with majors in Electronics and Communications from Jordan University of Science and Technology in 1995, 1998, respectively. He received his Ph.D. in Electrical Engineering from University of New Mexico (Albuquerque, U.S.A.) in 2004. In September 2004 he joined the College of Information Technology, United Arab Emirates University (Al Ain, UAE) as an Assistant Professor with the Computer System Engineering track. Before joining the Ph.D. program at University of New Mex-

Authors’ Biographies

Mohammad Hayajneh (SM’01, M’04) received his B.Sc. and M.Sc. in Electrical Engineering with majors in Electronics and Communications from Jordan University of Science and Technology in 1995, 1998, respectively. He received his Ph.D. in Electrical Engineering from University of New Mexico (Albuquerque, U.S.A.) in 2004. In September 2004 he joined the College of Information Technology, United Arab Emirates University (Al Ain, UAE) as an Assistant Professor with the Computer System Engineering track. Before joining the Ph.D. program at University of New Mex-

Dr. Hayajneh held several positions in the academia: he was lecturer and head of the Telecommunications Department in the Institute of Science for Telecommunications and Technology in Riyadh, KSA from 2000–2001. And from 1998–2000 he was an instructor at Princess Sumaya University of Technology, (PSUT) Amman, Jordan. Dr. Hayajneh research interests include: power control for wireless data networks, WiMAX cross-layer scheduling, resource allocation and pricing in wireless data networks, MIMO systems, performance analysis, and modeling of OFDMA-based wireless networks.

Najah Abu Ali received her B.Sc. and M.Sc. degrees in Electrical Engineering in 1989 and 1995, respectively from University of Jordan, Amman, Jordan and her Ph.D. degree in 2006 in Computer Networks from Electrical Engineering department at Queen’s University, Kingston, Canada. She joined the College of Information Technology, United Arab Emirates University (Al Ain, UAE), as an Assistant Professor with the Computer Networks Engineering track. Before that she held a post doctoral position at the school of computing at Queen’s University, Kingston, Canada. She worked as an instructor and the head of the engineering department at Queen Alia College in Jordan from 1995–2003. Her research interests comprise wired and wireless communication networks. Specifically, analytical and measurement-based network performance management and quality of service, and resource management of single and multihop wireless networks.

Issa Khalil received the B.Sc. degree in Computer Engineering from Jordan University of Science and Technology (JUST), Jordan, in 1994, the MS degree in computer engineering from JUST in 1996, and the Ph.D. degree in computer engineering from Purdue University, USA, 2006. Immediately thereafter he worked as a post-doctoral researcher at the Dependable Computing Systems Lab of Purdue University until he joined the College of Information Technology (CIT) of the United Arab Emirates University (UAEU) as an assistance professor in 2007. His research interests span the areas of wireless and wire-line communication networks. He is especially interested in security, routing, and performance of wireless sensor, ad hoc, WiMAX, and mesh networks. He has worked as an instructor and a director of the computer and communication center of Alquds Open University, West Bank, for more than 6 years.