Capacity Evaluation of a New Scheduler with Call Admission Control to Fixed WiMAX Networks with Delay Bound Guarantee

Eden Ricardo Dosciatti
GETIC-NATEC-UTFPR
Federal University of Technology
Pato Branco - Parana - Brazil
Email: edenrd@utfpr.edu.br

Walter Godoy Junior
NATEC-CPGEI-UTFPR
Federal University of Technology
Curitiba - Parana - Brazil
Email: godoy@utfpr.edu.br

Augusto Foronda
DAELN-NATEC-UTFPR
Federal University of Technology
Curitiba - Parana - Brazil
Email: foronda@utfpr.edu.br

***Abstract*—IEEE 802.16, also known as WiMAX, is a solution for mobile and fixed access to broadband networks, currently in development by the Working Group of the Institute of Electrical and Electronics Engineers - IEEE. The WiMAX Working Group focuses on the development of a standard for wireless broadband metropolitan area networks, whose main goal is to allow high-speed access to data, video and voice services. As a wireless broadband technology, WiMAX networks implement Quality of Service (QoS) mechanisms as a crucial element to satisfy users' demands for high data rates. QoS mechanisms and bandwidth allocation are covered by IEEE 802.16 standard. However, the exact details of scheduling and call admission control management, which guarantee QoS as required by multimedia applications, are left unspecified by the standard. In fact, the standard supports scheduling only for fixed-size real-time service flows. The choice of a scheduling algorithm for WiMAX systems is of major importance. A efficient, robust and fair WiMAX scheduling algorithm is still an open issue. Based on these facts, a new scheduler with call admission control with delay bound guarantee was proposed. The new scheduler calculates an optimal time frame, which allows the number of stations allocated in the system to be maximized and manages the delays required by each user. Properties of this algorithm are investigated both theoretically and through simulations. The results show that an upper bound on the delay can be achieved for a large range of network loads, with bandwidth optimization.**

Keywords-IEEE 802.16; WiMAX; QoS; Latency-Rate; scheduling; time frame; call admission control.

I. INTRODUCTION

The deployment of high-speed Internet access is often cited as an open challenge for the second decade of this century. Also known as broadband Internet, it is effective in reducing physical barriers to the transmission of information, as well as transaction costs, and is fundamental in fostering competitiveness. However, providing wired access to broadband Internet is costly and sometimes infeasible, since the investment needed to deploy cabling throughout a region often outweighs the service provider's financial gains. One of the possible solutions in reducing the costs of deploying broadband access in areas where such infrastructure is not present is to use wireless technologies, which require no cabling and reduce both implementation time and cost [2].

This was one of the motivations behind the development by the IEEE (Institute of Electrical and Electronics Engineers) of the 802.16 standard for wireless access [3], also known as Worldwide Interoperability for Microwave Access (WiMAX). It is an emerging technology for next-generation wireless networks, which provides supports for a large number of both mobile and nomadic (fixed) users distributed over a wide geographic area. Furthermore, this technology provides strict QoS (Quality of Service) guarantees for data, voice and video applications [4].

As a service provider, WiMAX creates new alternatives for applications such as telephony, TV broadcasts, broadband Internet access for residential users, and commercial, industrial and university centers. The development of this new market niche represents a revolution for telecommunications companies and interconnection equipment manufacturers [5]. Moreover, WiMAX enables broadband connection for areas, which are inaccessible or lacking in infrastructure, since it requires no complex physical installations of cable connections and traditional technologies [6].

Motivated by the growing need for ubiquitous, high-speed network access, wireless technology is an option to provide a cost-effective solution that may be deployed quickly and easily, providing high bandwidth connectivity in the last mile. However, despite its many advantages, such as low deployment and maintenance costs, ease of configuration, and device mobility, there are challenges that must be overcome in order to further advance its widespread use. The increasing deployment of wireless infrastructure enables a variety of new applications that require flexible, but also robust, support by the network, such as multimedia applications including video streaming and VoIP (Voice over Internet Protocol), which demand real-time data delivery [7].

To this purpose, the IEEE 802.16 standard introduces a set of mechanisms, such as service classes and several coding and modulation schemes that adapt themselves according to channel conditions. However, the standard leaves certain

issues pertaining to network resource management and scheduling algorithms open.

This paper presents a new scheduler with admission control of connections to a WiMAX Base Station (BS). We develop an analytical model based on Latency-Rate (LR) server theory [8], which an ideal frame size, called the Time Frame (TF), is estimated, with guaranteed delays for each user. At the same time, the number of stations allocated in the system is maximized. In this procedure, framing overhead generated by the MAC (Medium Access Control) and PHY (Physical) layers was taken into account when calculating the length of each time slot. After developing this model, a set of simulations is presented for constant bit rate (CBR) and variable bit rate (VBR) streams, with performance comparisons between situations with different delays and different TFs. The results show that an upper limit on the delay may be achieved for a wide range of network loads, thus optimizing bandwidth.

The paper is an extension to [1] and is structured as follows. In Section II, related research is described. In Section III, a brief description of the IEEE 802.16 standard is presented. Our analytical model for packet scheduling is proposed and explained in Section IV. Evaluation of the capacity of the new scheduler with Call Admission Control (CAC) is shown in Section V. Conclusions are presented in Section VI.

II. Related Work

Several scheduling algorithms and QoS architectures for Broadband Wireless Access (BWA) have been proposed in the literature [9-15], since the standard only specifies signaling mechanisms and no specific scheduling and admission control algorithms. However, many of these solutions only address the implementation or addition of a new QoS architecture to the IEEE 802.16 standard. A scheduling algorithm decides the next packet to be served on the queue and is one of the mechanisms responsible for distributing bandwidth among several streams (by assigning each flow the bandwidth that was required and available). In these proposals [9-15], there are often no analytical models for ensuring maximum delay and maximizing the number of SSs (Subscriber Stations) allocated in the system, which are represented accurately by certain performance metrics, such as the delay, of the medium access protocol.

In [9], a packet scheduler for IEEE 802.16 uplink channels based on a hierarchical queue structure is proposed. A simulation model is developed to evaluate the performance of the proposed scheduler. However, despite presenting simulation results, the authors overlooked the fact that the complexity of implementing this solution is not hierarchical, and do not define clearly how requests for bandwidth are made.

In [10], the authors propose a QoS architecture to be built into the IEEE 802.16 MAC sublayer, which significantly impacts system performance, but do not present an algorithm that makes efficient use of bandwidth.

In [11], the authors present a simulation study of the IEEE 802.16 MAC protocol operating with an OFDM (Orthogonal Frequency Division Multiplexing) air interface and full-duplex stations. System performance is evaluated under different traffic scenarios, by varying the values of a set of relevant system parameters. Regarding data traffic, it was observed that the overhead due to the physical transmission of preambles increases with the number of stations.

In [12], a polling-based MAC protocol is presented along with an analytical model to evaluate its performance, considering a system where the BS issues probes in every frame to determine bandwidth requirements for each node. The authors developed closed-form analytical expressions for cases in which stations are polled at the beginning or at the end of uplink subframes. It is not possible to know how the model may be developed to provide delay guarantees.

In [13], the proposal is of a QoS architecture in which the scheduler is based on packet lifetime for each type of flow. The process of data communication between BS and SS is considered from the start, that is, connection and negotiation of traffic parameters such as bandwidth and delay. The proposal features an architecture defined in well-structured blocks, which may make data flows and architecture actions inaccurate. However, despite presenting simulation results, the work neglects performance by not adequately addressing the functional blocks of the proposed architecture and by not specifying clearly how lifetime is calculated for each packet.

In [14], the scheduling algorithm handles traffic with Best Effort (BE), and it is concluded that there exists considerable difficulty in estimating the amount of bandwidth required due to dynamic changes in traffic transmission rate. The purpose of this algorithm is to ensure fairness in bandwidth allocation among BE flows and full bandwidth usage. The system measures the transmission rate for each flow and allocates bandwidth based on the average transmission rate.

Finally, in [15] the author presents a well-established architecture for QoS in the IEEE 802.16 MAC layer. The subject of this work is the component responsible for allocating uplink bandwidth to each SS, although the decision is taken based on the following aspects: the bandwidth required by each SS for uplink data transmission, periodic bandwidth needs for UGS flows in SSs and the bandwidth required for making requests for additional bandwidth.

Considering the limitations exposed above, these works form the basis of a generic architecture, which can be extended and specialized. However, in these studies, the focus is in achieving QoS guarantees, with no concerns for maximizing the number of allocated users in the network. This paper presents a scheduler with admission control of connections to the WiMAX BS. We developed an

analytical model based on Latency-Rate (LR) server theory [8], which an ideal frame size called Time Frame (TF) was estimated, with guaranteed delays for each user and maximization of the number of allocated stations in the system. A set of simulations is presented with constant bit rate (CBR) and variable bit rate (VBR) streams and performance comparisons are made for different delays and different TFs. The results show that an upper bound on the delay may be achieved for a large range of network loads with bandwidth optimization.

A. Latency-Rate Servers

Providing quality of service (QoS) guarantees in a packet network requires the use of traffic scheduling algorithms in the routers. The function of a scheduling algorithm is to select, for each outgoing link of the router, the packet to be transmitted next cycle from the available packets belonging to the sessions sharing the output link.

Since networks are unlikely to be homogeneous in the type of scheduling algorithms employed by the individual routers, a general model for the analysis of scheduling algorithms will be a valuable tool in the design and analysis of such networks.

In work [8] was developed a model to study the behavior of the worst-case of individual sessions in a network of schedulers where the schedulers may employ a broad range of scheduling algorithms. This approach allows to calculate tight bounds on the end-to-end delay of individual sessions and the buffer sizes needed to support them in an arbitrary network of schedulers. The basic approach consists in defining a general class of schedulers, called *Latency-Rate* servers [8], or simply *LR* servers. The theory of *LR* servers provides a means to describe the worst-case behavior of a broad range of scheduling algorithms in a simple and elegant manner. This theory is based on the concept of a busy period of a session, a period of time during which the average arrival rate of the session remains at or above its reserved rate r_i. For a scheduling algorithm to belong to the *LR* class, it is only required that the average rate of service offered by the scheduler to a busy session, over every interval starting at time θ from beginning of the busy period, is at least equal to its reserved rate. The parameter θ is called latency of the scheduler.

The behavior of an LR scheduler is determined by two parameters: the *latency* (θ) and the *allocated rate* (r_i). The latency of LR server may be seen as the worst-case delay seen by the first packet of the busy period of a session, which is a packet arriving when the queue is empty session. The latency of a particular scheduling algorithm may depend on its internal parameters, its transmission rate on the outgoing link, and the allocated rates of various sessions. However, the maximum end-to-end delay experienced by a packet in a network of schedulers can be calculated from only the latencies of the individual schedulers on the path of the session, and the traffic parameters of the session that generated the packet. Since the maximum delay in a scheduler increases directly in proportion to its latency, the model brings out the significance of using low-latency schedulers to achieve low end-to-end delays. Likewise, upper bounds on the queue size and burstiness of individual sessions at any point within the network can be obtained directly from the latencies of the schedulers.

III. The IEEE 802.16 standard

The basic topology of a IEEE 802.16 network includes two entities that participate in the wireless link: Base Stations (BS) and Subscriber Stations (SS), as shown in Figure 1 [16].

The BS is the central node, responsible for coordinating communication and providing connectivity to SSs. BSs are kept in towers distributed so as to optimize network coverage area, and are connected to each other by a backhaul network, which allows SSs to access external networks or exchange information between themselves.

Networks based on the IEEE 802.16 standard can be structured in two schemes. In PMP (Point-to-multipoint) networks, all communication between SSs and other SSs or external networks takes place through a central BS node. Thus, traffic flows only between SSs and the BS (see Figure 1). In Mesh mode, SSs communicate with each other without the need for intermediary nodes; that is, traffic can be routed directly through SSs. Thus, all stations are peers, which can act as routers and forward packets to neighboring nodes [17]. This article only considers the PMP topology, since it is implemented by first-generation WiMAX devices,

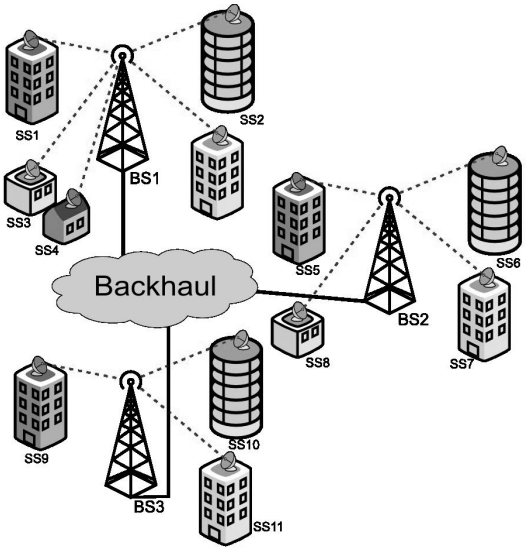

Figure 1. IEEE 802.16 Network Architecture

and also due to the strong trend towards its adoption by Internet providers because it allows them to control network parameters in a centralized manner, without the need to recall all subscriber stations [5].

Although it is referred to as fixed pattern, IEEE 802.16 allows stations to provide customers with low-speed mobility. A feature missing in this pattern and that justifies its designation as fixed is the possibility of performing handoffs/handovers, which allow a client station to switch to another base station without losing connectivity. In this case, subscriber stations are instead called mobile stations. The functionality of handoff/handover was included in the IEEE 802.16 standard in early 2006 with the publication of the IEEE 802.16e [18], which quickly received the name of "IEEE 802.16 mobile".

WiMAX technology can reach a theoretical maximum distance of 50 km [19]. Data transmission rates can vary from 50 to 150 Mbps, depending on channel frequency band width and modulation type [20]. Communication between a BS and SSs occurs in two different channels: uplink (UL) channel, which is directed from SSs to the BS, and downlink (DL) channel, which is directed from the BS to SSs. DL data is transmitted by broadcasting, while in UL access to the medium is multiplexed. UL and DL transmissions can be operated in different frequencies using Frequency Division Duplexing (FDD) mode or at different times using Time Division Duplexing (TDD) mode.

In TDD, the channel is segmented in fixed-size time slots. Each frame is divided into two subframes: a DL subframe and an UL subframe. The duration of each subframe is dynamically controlled by the BS; that is, although a frame has a fixed size, the fraction of it assigned to DL and UL is variable, which means that the bandwidth allocated for each of them is adaptive. Each subframe consists of a number of time slots, and thus both the SSs and the BS must be synchronized and transmit the data at predetermined intervals. The division of TDD frames between DL and UL is a system feature controlled by the MAC layer. Figure 2 [10] shows the structure of a TDD frame. In this paper, the system was operated in TDD mode with the OFDM (Orthogonal Frequency Division Multiplexing) air interface, as determined by the standard [3].

IV. Analysis of the Analytical Model

A minimum acceptable performance level should be sought throughout the development of any system, be it computer-related or not. This requires a measure or gauge of performance in these systems. To accomplish this, there exist design tools that provide the analyst with different metrics and measures. Within this scope, some related system characteristics are proposed and discussed in this article. To accomplish this, this section presents an analytical model of the new scheduler and an analytical description of its call admission control facility.

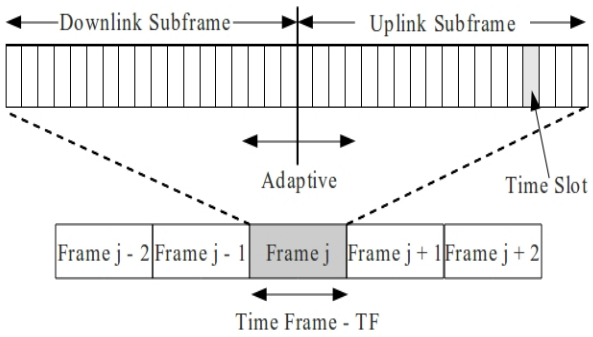

Figure 2. IEEE 802.16 Frame Structure

A. System Description

Figure 3 [21] illustrates a wireless network operating the newly proposed scheduler with connection admission control, which is based on a modified LR scheduler [8] and uses the token bucket algorithm.

The basic approach consists on the token bucket limiting input traffic and the LR scheduler providing rate allocation for each user. Then, if the rate allocated by the LR scheduler is larger than the token bucket rate, the maximum delay may be calculated.

A scheduler that provides guaranteed bandwidth can be modeled as an LR scheduler. The behavior of an LR scheduler is determined by two parameters for each session i: latency θ_i and allocated rate r_i. The latency θ_i of the scheduler may be seen as the worst-case delay and depends on network resource allocation parameters. In the new scheduler with call admission control, the latency θ_i is a TF period, which is the time needed to transmit a

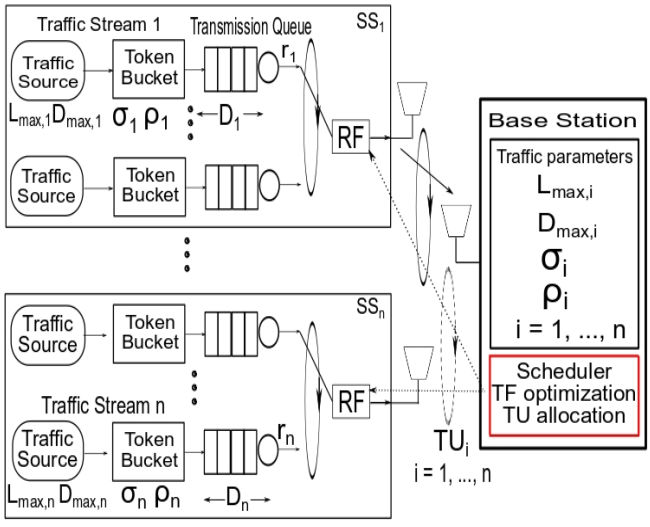

Figure 3. Wireless Network with New Scheduler

maximum-size packet and separation gaps (TTG and RTG) of DL and UL subframes. In the new scheduler, considering the delay for transmitting the first packet, the latency θ_i is given by

$$\theta_i = T_{TTG} + T_{RTG} + T_{DL} + T_{UL} + \frac{L_{max,i}}{R} \quad (1)$$

where T_{TTG} and T_{RTG} are the DL and UL subframe gap durations, T_{DL} and T_{UL} are the DL and UL subframe durations, $L_{max,i}$ is the maximum packet size and R is the outgoing link capacity.

Now, we show how the allocated rate r_i for each session i may be determined, and how to optimize TF in order to increase the number of connections accommodated.

B. CAC Description

An LR scheduler can provide a bounded delay if input traffic is shaped by a token bucket. A token bucket [2] is a non-negative counter, which accumulates tokens at a constant rate ρ_i until the counter reaches its capacity σ_i. The rate of incoming packets (ρ_i) is constant because the parameters of the token bucket, for all three types of traffic, that will be used for performance evaluation are constant, i.e., audio will be 64 kb/s, VBR video will be 500 kb/s, and MPEG4 video will be 4100 kb/s. Packets from session i can be released into the queue only after removing the required number of tokens from the token bucket. In an LR scheduler, if the token bucket is empty, arriving packets are dropped; however, our model ensures that there will always be tokens in the bucket and that no packets are dropped, as described in Section IV. If the token bucket is full, a maximum burst of σ_i packets can be sent to the queue. When the flow is idle or running at a lower rate as the token size reaches the upper bound σ_i, accumulation of tokens will be suspended until the arrival of the next packet. We assume that the session starts out with a full bucket of tokens. In our model, we consider IEEE 802.16 standard overhead for each packet. Then, as we will show below, the token bucket size will decrease by both packet size and overhead.

The application using session i declares the maximum packet size $L_{max,i}$ and requires maximum allowable delay $D_{max,i}$, which are used by the WiMAX scheduler to calculate the service rate for each session so as to guarantee the required delay and optimize the number of stations in the network. Incoming traffic $A_i(t)$ from session $i (i = 1, ..., N)$ passes through a token bucket inside the user terminal during the time interval $(0, t)$.

This passage of data traffic by the token bucket is bounded by

$$A_i(t) \leq \sigma_i + \rho_i t \quad (2)$$

where σ_i is the bucket size and ρ_i is the bucket rate.

Then, the packet is queued in the station until it is transmitted via the wireless medium. Queue delay is measured as the time interval between the receipt of the last bit of a packet and its transmission. In the new scheduler with call admission control, queuing delay depends on token bucket parameters, network latency and allocated rate. In [8] and [22], it is shown that if input traffic $A_i(t)$ is shaped by a token bucket and the scheduler allocates a service rate r_i, then an LR scheduler can provide a bounded maximum delay D_i:

$$D_i \leq \frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i} \quad (3)$$

where σ_i is the token bucket size, r_i is the service rate, θ_i is the scheduler latency, $\frac{L_{max,i}}{r_i}$ is the maximum size of a package and, $\frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i}$ is the bound on the delay, D_{bound}.

Equation (3) is an improved bound on the delay for LR schedulers. Thus, the token bucket rate plus the overhead transmission rate must be smaller than the service rate to provide a bound on the delay. The upper bound D_{bound} should be smaller than or equal to the maximum allowable delay:

$$\frac{\sigma_i}{r_i} + \theta_i - \frac{L_{max,i}}{r_i} \leq D_{max,i} \quad (4)$$

Therefore, three different delays are defined. The first is the maximum delay D_i, the second is the upper bound on the delay D_{bound} and the third is the required maximum allowable delay $D_{max,i}$. The relation between them is $D_i \leq D_{bound} \leq D_{max,i}$.

So, the delay constraint condition of the new scheduler is

$$\frac{(\sigma_i' - L_{max,i}')TF}{r_i'TF - \Delta R + L_{max,i}'} + TF + \\ + \frac{L_{max,i}'}{R} + T_{TTG} + T_{RTG} \leq D_{max,i} \quad (5)$$

where σ_i' is the token bucket size with overhead, $L_{max,i}'$ is the maximum size of a packet with overhead (preamble+pad), TF is the time frame, r_i' is the rate allocated by the server with overhead, R is the outgoing link capacity, T_{TTG} is the gap between downlink and uplink subframes, T_{RTG} is the gap between uplink and downlink subframes, $D_{max,i}$ is the maximum allowable delay and Δ is the sum of initial ranging and BW request, which is the uplink subframe overhead and whose value will be discussed when evaluating performance. Physical rate, maximum packet size and token bucket size are parameters declared by the application. However, TF and total allocated service rate must satisfy Equation (5).

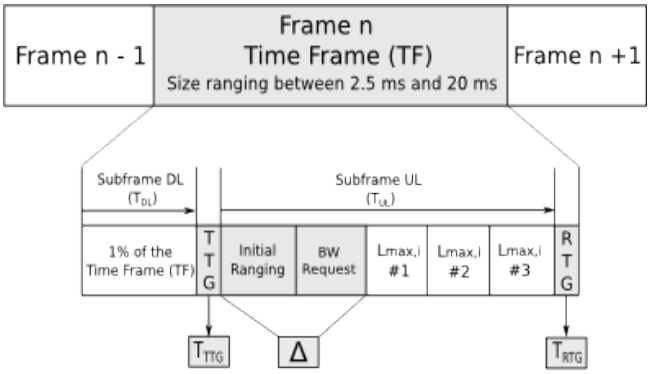

Figure 4. Frame structure with TDD allocation formulas of Equation (5)

Figure 4 shows a frame structure with TDD allocation formulas as described by Equation (5).

The second delay constraint condition to TF and service rate is that the token bucket rate plus the rate to transmit overhead and a maximum-sized packet must be smaller than the service rate to place a bound on delay. Thus, the second constraint condition is

$$\rho_i + \frac{\Delta R + L'_{max,i}}{TF} \leq r'_i \quad (6)$$

where ρ_i is the bucket rate, Δ is the uplink subframe overhead, R is the outgoing link capacity, $L'_{max,i}$ is the maximum packet size with overhead, TF is the time frame and r'_i is the rate allocated by the service with overhead.

Previous schedulers do not provide any mechanisms to estimate the TF needed to place a bound on delay or to maximize the number of stations, because each application requires a TF without the use of criteria to calculate the time assigned to each user. However, TF estimation is important because of a tradeoff. A small TF reduces maximum delay, but increases overhead at the same time. On the other hand, a large TF decreases overhead, but increases delay. Therefore, we must calculate the optimal TF to allocate the maximum number of users under both constraints. The maximum number of users is achieved when the service rate for each user is the minimum needed to guarantee the bound on the delay, D_{bound}.

To find the maximum number of users in each frame, we solve a problem of non-linear optimization. Solving non-linear problems is characterized by not having a single algorithm for solving their problems. The biggest difficulty with this approach is the uncertainty that the solution to this problem is really the best, and this is a fact inherent in the non-linear nature of the problem, whereas its great advantage is the scope, that is, once the mathematical model developed the problem, with its objective function and its constraints, usually no simplification is needed in terms of formulation.

So, in this work, nonlinear optimization makes use of search techniques using numerical information given in an iterative process, generating better solutions in the optimization process. These techniques allow us to use numerical methods to solve problems when there is no known analytical solution.

In the specific case of this work, an approach step-by-step was used, where a small initial value for the TF is determined, in this case, the value of 2.5 ms (lower reference value for the TF according to 802.16 standard [3]). After, the value of r'_i is calculated and the process is repeated with a certain step length, in this case, 0.5 ms, until the minimum value of r'_i is found, satisfying the constraints of Equations (5) and (6). The value of the step length can be determined randomly by the limit of 20 ms (maximum value of a frame in accordance with 802.16 standard [3]) and there will always be a solution because at every step the two constraints of Equations (5) and (6) will be confronted in order to verify that the minimum value r'_i found.

V. Performance Analysis

To analyze the IEEE 802.16 MAC protocol behavior with respect to the new scheduler with call admission control, this section presents numerical results obtained with the analytical model proposed in the previous section. Then, with a simulation tool, the proposed analytical model is validated by showing that the bound on the maximum delay is guaranteed. In this section, two types of delays are treated: required delay, in which the user requires the maximum delay, and the guaranteed maximum delay, which is calculated with the analytical model.

A. Calculation of Optimal Time Frame

In this paper, the duration of downlink subframes is fixed at 1% of the TF because our interest is only in the uplink subframe. In the simulation, after finding the optimal number of SSs per frame for each traffic flow, the header value of the uplink subframe is calculated at a rate of 10% of the value of an OFDM symbol [2].

All PHY and MAC layer parameters used in simulation are summarized in Table I.

Performance of the new scheduler with call admission

Table I
PHY and MAC parameters

Parameter	Value
Bandwidth	20 MHz
OFDM Symbol Duration	13,89 μs
Delay	5, 10, 15 and 20 ms
Δ (Initial Ranging and BW Request) = 9 OFDM Symbols	125,10 μs
TTG + RTG = 1 OFDM Symbol	13,89 μs
UL Subframe (preamble + pad) = 10% OFDM Symbol	1,39 μs
Physical Rate	70 Mbps
DL Subframe	1% TF

Table II
Token bucket parameters

	Audio	**VBR video**	**MPEG4 video**
Token Size (bits)	3000	18000	10000
Token Rate (kb/s)	64	500	4100

control is evaluated as the delay requested by the user and assigned stations. Station allocation results, in the system with an optimal TF, limited by the delay requested by the user, are described in sequence. The first step is defining token bucket parameters, which are estimated according to the characteristics of incoming traffic and are listed on Table II. It's worth noting that the details about the traffic must be known in advance. This is normal for various applications such as audio, CBR and video on demand.

Thus, the optimal TF value is estimated according to the PHY and MAC layer's parameters (see Table I), token bucket parameters (see Table II), required maximum allowable delay, physical rate and maximum packet size. With all parameters defined, and with the constraints set by Equations (5) and (6), described in Section IV-B, we use a step-by-step approach, starting with a small TF of 2.5 ms, calculating r_i' and repeating this process every 0.5 ms until the minimum r_i' that satisfies both equations is found. The graph in Figure 5 shows the optimal TF value, for four delay values required by users (5, 10, 15 and 20 ms):

- For a requested delay of 5 ms, the optimal TF is 3 ms.
- For a requested delay of 10 ms, the optimal TF is 6.5 ms.
- For a requested delay of 15 ms, the optimal TF is 10.5ms.
- For a requested delay of 20 ms, the optimal TF is 15 ms.

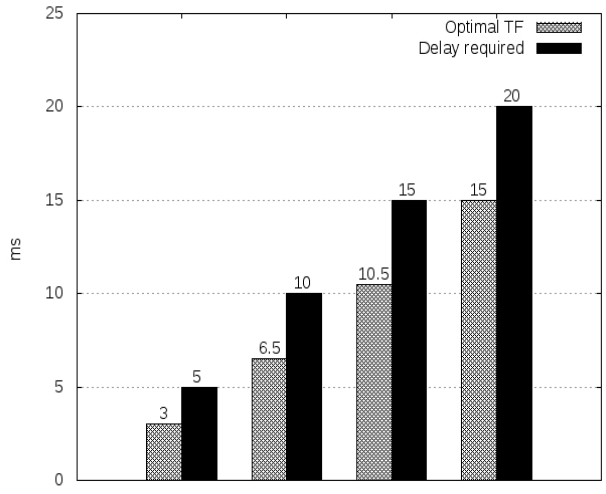

Figure 5. Optimal TF

Figure 6 shows the number of SSs assigned to each traffic type in each frame, through of the optimal TF calculated. The result shows the maximum number of SSs assigned to each range of optimal TF values for each traffic type. It should be noted that three traffic types were used: audio traffic, VBR video traffic and MPEG4 video traffic. For the simulation, the allocation of users is performed by traffic type; i.e., only one traffic at a time will be transmitted within each frame.

As an example, Figure 6d shows that when the user-requested delay is of 20 ms, an optimal TF of 15 ms is calculated and 50 users can be allocated for audio traffic, or 30 users for VBR video traffic, or 13 users for MPEG4 video traffic.

Two important observations from Figure 6d should be highlighted:

1) With a requested delay of 20 ms, we cannot choose a TF of less than 15 ms, since the restrictions placed by Equation (5) (which regards delay) and Equation (6) (which regards the token bucket) are not respected and thus no bandwidth allocation guarantees exist.
2) We also cannot choose a TF greater than 15 ms, even though it complies with Equations (5) and (6) with respect to guaranteed bandwidth, because there will be a decrease in the number of users allocated to each traffic flow due to increasing delay.

Thus, it is evident that since the IEEE 802.16 standard does not specify an ideal time frame (TF) duration, this approach becomes advantageous because, in addition to meeting the restrictions of the analytical model, it optimizes the allocation of users on the system. The same philosophy holds true for other delay values of 5, 10 and 15 ms.

B. Comparison of User Allocation and Optimal Time Frame

In this work, an optimal TF was reached, so that the number of SSs in the network may be optimized and a maximum delay may be guaranteed. To make a comparison of the results in this work, Figure 7 shows that, for an audio traffic and a requested delay of 15 ms, an optimal TF of 10.5 ms is obtained and 41 users can be allocated. When compared to other randomly-chosen TFs, it may be observed that the optimal TF yields a greater number of users. Thus, when an user requests a delay guarantee, an optimal TF is calculated in order to allocate the largest number of users in a given traffic flow, as seen in the example in Figure 7. It may be noticed, then, that choosing a non-optimal TF will lead to a decreased number of allocated SSs. Therefore, the new scheduler with call admission control proposed herein maximizes the number of SSs in place and ensures an upper bound on maximum delay, as discussed below.

C. Guaranteed Maximum Delay

In this article, only UL traffic is considered. To test the new scheduler's performance, we have carried out

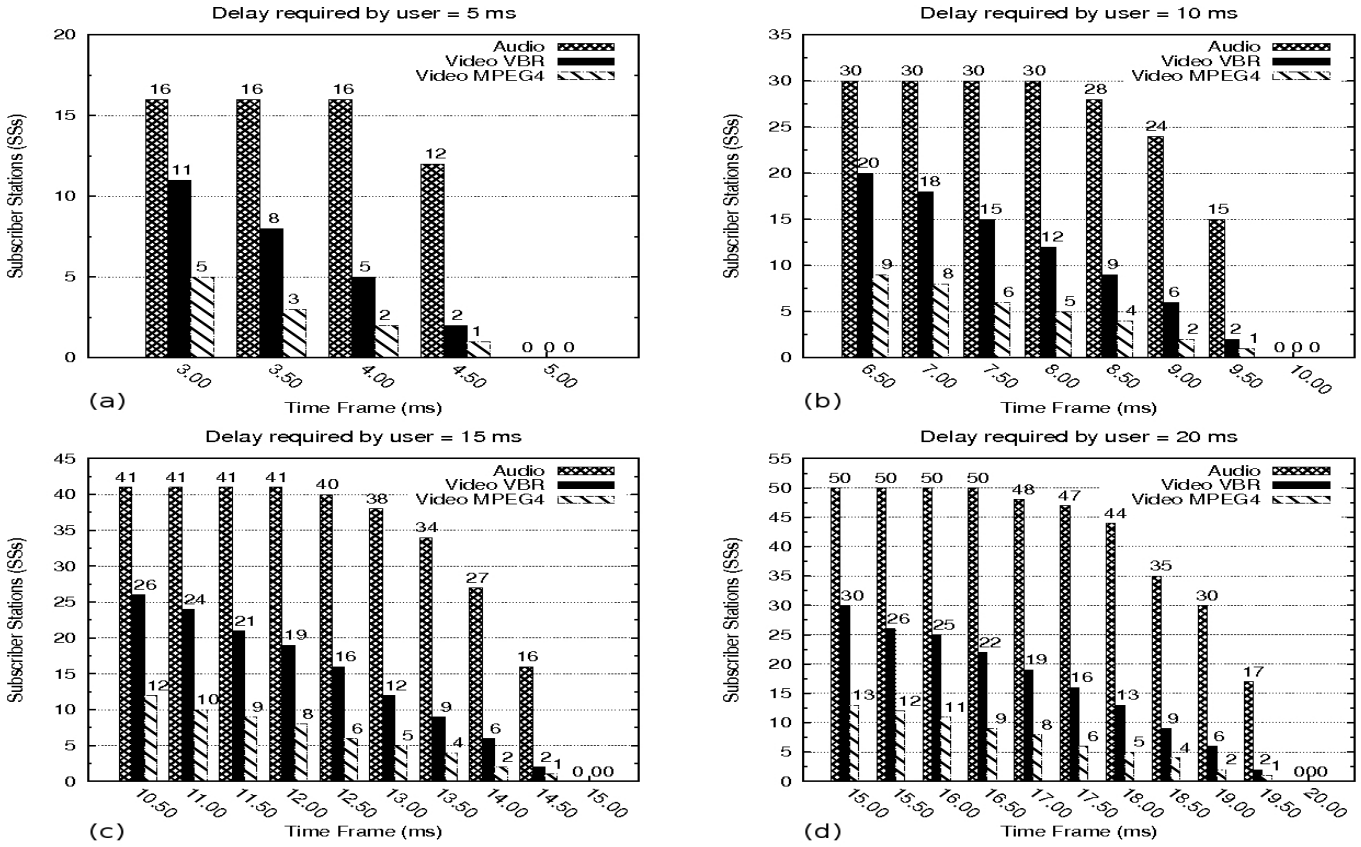

Figure 6. Number of subscriber stations for all delay required

simulations of an IEEE 802.16 network consisting of a BS that communicates with eighteen SSs, with one traffic flow type by SS and the destination of all flows being the BS, as shown in Figure 8. In this topology, six SSs transmit on-off CBR audio traffic (64 kb/s), six transmit CBR MPEG4 video

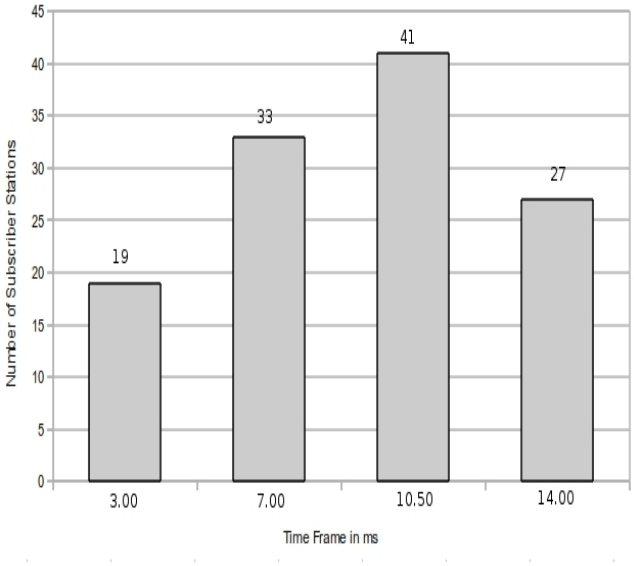

Figure 7. Users assigned as a function of TF, for audio traffic

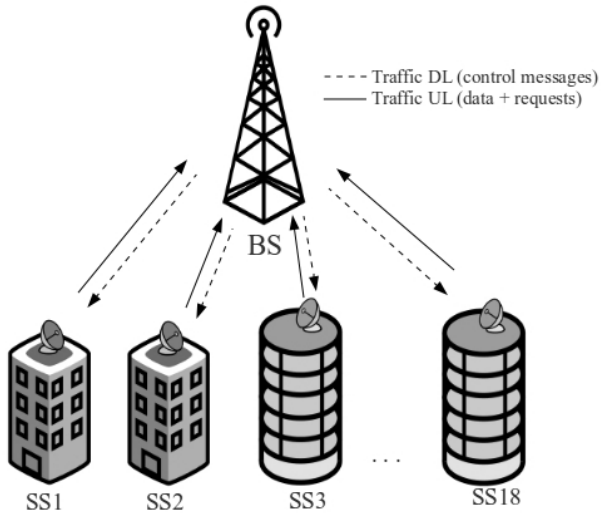

Figure 8. Simulation scenario

Table III
Description of traffic types

Node	**Application**	**Arrival Period** *(ms)*	**Packet Size** *(max) (B)*	**Sending Rate** *(kb/s) (mean)*
$1 \rightarrow 6$	Audio	4.7	160	64
$7 \rightarrow 12$	VBR video	26	1024	$\approx$ 200
$13 \rightarrow 18$	MPEG4 video	2	800	3200

traffic (3.2 Mb/s) and six transmit VBR video traffic. Table III summarizes the different types of traffic used in this simulation.

In Section V-D we have the algorithm of the simulator and its source code, developed in C programming language [23]

In Figure 9, with an optimal TF of 3 ms and an user-requested delay of 5 ms, the average guaranteed maximum delay for audio traffic is 1.50 ms. For VBR video traffic, whose packet rate is variable, the average maximum delay is 1.97 ms. For MPEG4 video traffic, the average maximum delay is 2.00 ms. Data that supports the stated maximum guaranteed delay values is listed in the tables below, which relate the number of packets read in each simulation to the resulting guaranteed maximum delay. A number of simulations were run for each type of traffic to keep results from varying too widely. Our choice of six simulations for each case produced values with noticeably little variation. After running simulations for each optimal TF and each traffic type, averages of resulting guaranteed maximum delays were taken and the graph of Figure 9 was constructed.

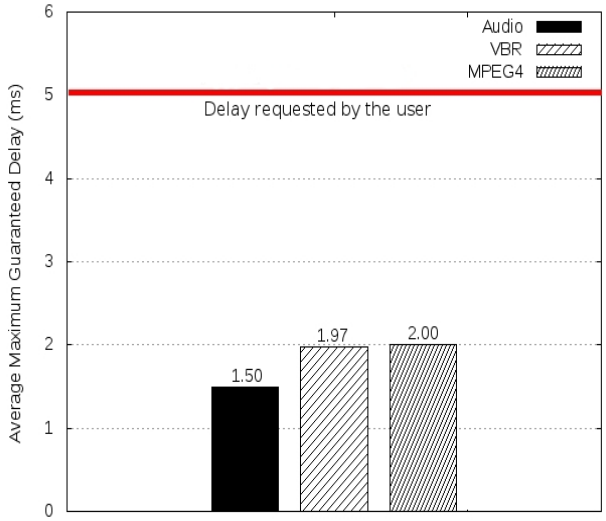

Figure 9. Guaranteed Maximum Delay

Table IV
Algorithm to compute the delay

Step 1	(**initialization**): Initialize the variables of total packet and time frame.
Step 2	(**perform**): **While** the total number of packets is smaller the number of packets in the system **and** the time frame is than smaller than the number of packets in the system, **do**: a. calculate the size of the package. b. calculate the time frame.
Step 3	(**testing**): **If** time frame is greater than the packet size, calculate the packet delay and the total delay, **Else** increment the time frame.
Step 4	(**results**): calculate the average delay and print the result on the screen.

D. Pseudocode of the Algorithm Simulator

In this section, we describe the structure of the simulator and the pseudocode of the simulator. The C programming language [23] was used to build the simulator that calculates the guaranteed maximum delay. In Table IV, is shown the pseudocode algorithm of the simulator to calculate the guaranteed maximum delay. This algorithm uses the parameters of Table III in Section V-C.

After reading the file with the amount of packets, variables that calculate the packet size and time frame, are used to perform the calculation of the delay of each packet, and if the value of packet size calculated is greater than the value of time frame calculated, there was a delay and it will be stored in variables to calculate the delay of each packet and total delay. In the end, the average delay is calculated and print the results on the screen.

This code is generic and is used to calculate the delay of all traffic used in this work.

The tables below show the result of using the simulation algorithm with the three traffic types used, namely audio, VBR video and MPEG4 video.

Table V shows the results of audio traffic simulations. Table VI shows the results of VBR video traffic simulations, whose packet rate is variable. Table VII shows the results of MPEG4 video traffic simulations.

Table V
Audio Traffic

Delay by the user	5 ms	10 ms	15 ms	20 ms
Optimal TF	3 ms	6.5 ms	10.5 ms	15 ms
Packages read *amount*	**Guaranteed Maximum Delay** (ms)			
1000	1.48	3.23	5.24	7.49
3000	1.49	3.24	5.24	7.50
5000	1.49	3.25	5.25	7.50
10000	1.50	3.25	5.25	7.50
30000	1.50	3.25	5.28	7.50
50000	1.50	3.35	5.29	7.51
Mean	1.50	3.25	5.26	7.50
Standard Deviation	0.00816	0.00837	0.02137	0.00632

Table VI
VBR Video Traffic

Delay by the user	5 ms	10 ms	15 ms	20 ms
Optimal TF	3 ms	6.5 ms	10.5 ms	15 ms
Packages read *amount*	**Guaranteed Maximum Delay** (ms)			
2176	2.06	3.48	5.50	7.98
1358	1.94	3.52	5.45	7.96
1177	1.97	3.48	5.59	8.07
1226	2.02	3.32	5.41	8.07
1159	1.87	3.33	5.57	8.08
1449	1.96	3.45	5.53	8.04
Mean	1.97	3.43	5.51	8.03
Standard Deviation	0.06573	0.08438	0.06940	0.05125

Table VII
MPEG4 Video Traffic

Delay by the user	5 ms	10 ms	15 ms	20 ms
Optimal TF	3 ms	6.5 ms	10.5 ms	15 ms
Packages read *amount*	**Maximum Guaranteed Delay** (ms)			
1000	2.00	3.50	5.51	8.01
3000	2.00	3.50	5.50	8.00
5000	2.00	3.50	5.50	8.00
10000	2.00	3.50	5.50	8.00
30000	2.00	3.50	5.50	8.00
50000	2.00	3.50	5.50	8.00
Mean	2.00	3.50	5.50	8.00
Standard Deviation	0.0	0.0	0.00408	0.00408

E. Comparison with other Schedulers

The new scheduler with call admission control, here called *New Scheduler*, was compared to those of [12], here called *Scheduler_1*, and [9], here called *Scheduler_2*. The comparison was accomplished through the ability to allocate users in a particular time frame (TF). Table VIII shows the parameters used for comparisons.

In the graph of Figure 10, we compare the *New Scheduler* with the *Scheduler_1*. A maximum delay of 0.12 ms was requested by the user, and the duration of each frame (TF) was set at 5 ms. Other parameters are listed in Table VIII. In comparison, the *New Scheduler* allocates 28 users in each frame, while the *Scheduler_1*, allocates 20 users. Thus, the *New Scheduler* presents a gain in performance of 40% when compared with the *Scheduler_1*.

In the graph of Figure 11, we compare the *New Scheduler* with the *Scheduler_2*. A maximum delay of 20 ms was

Table VIII
Parameters used for comparisons

Parameter	*Scheduler_1*	*Scheduler_2*
Bandwidth	20 MHz	20 MHz
OFDM symbol duration	13.89 μs	13.89 μs
Delay Requested by the user	0.12 ms	20 ms
Time Frame (TF)	5 ms	10 ms
Maximum Data Rate	70 $Mbps$	70 $Mbps$
Traffic type	Audio	Audio

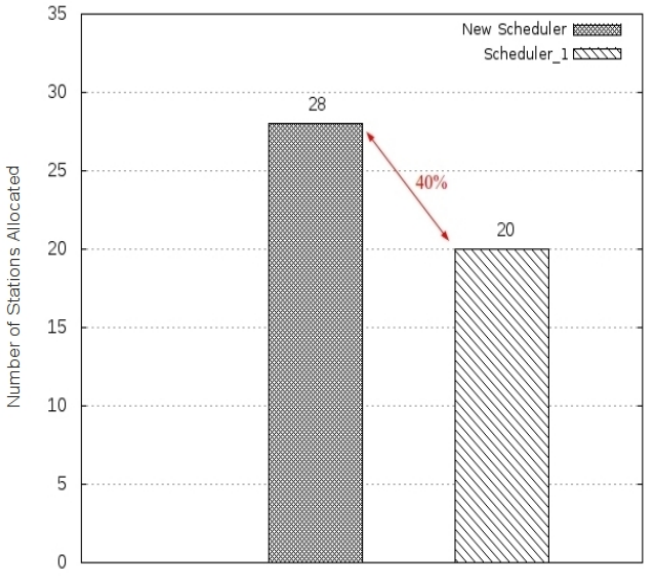

Figure 10. Comparison of user allocation with *Scheduler_1*

requested by the user, and the duration of each frame (TF) was set at 10 ms. Other parameters are listed in Table VIII.

The comparison was extended by also considering frame duration values of 7.00 ms, 8.00 ms and 9.00 ms to demonstrate the efficiency of the *New Scheduler*. For a TF of 10 ms, the *New Scheduler* allocates 41 users in each frame, while the *Scheduler_2* allocates only 33 users. This represents 24.24% better performance for the *New Scheduler*. Similarly, the *New Scheduler* also allocates more users per frame in comparison with the *Scheduler_2* for all other frame duration values.

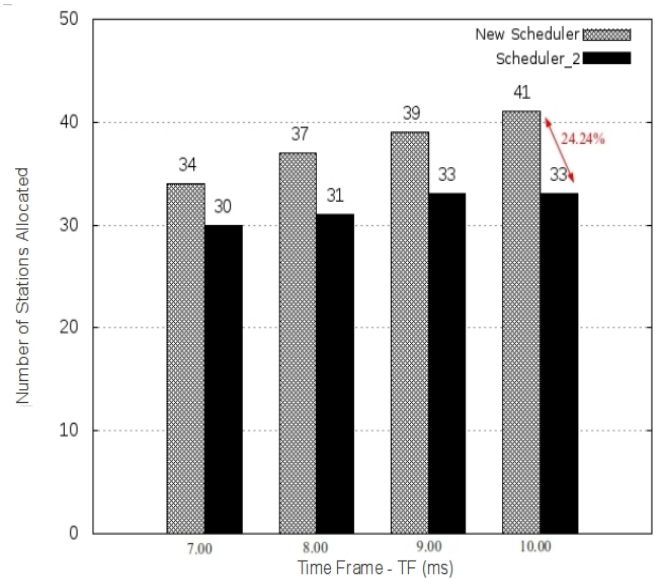

Figure 11. Comparison of user allocation with *Scheduler_2*

VI. Conclusion and Future Work

This work has presented the design and evaluation of a new scheduler with call admission control for IEEE 802.16 broadband access wireless networks (known worldwide as WiMAX) that guarantees different maximum delays for traffic types with different QoS requisites and optimizes bandwidth usage.

A. Conclusion

Firstly, we developed an analytical model to calculate an optimal TF, which allows an optimal number of SSs to be allocated and guarantees the maximum delay required by the user. Then, a simulator was developed to analyze the behavior of the proposed system.

To validate the model, we have presented the main results obtained from the analysis of different scenarios. Simulations were performed to evaluate the performance of this model, demonstrating that an optimal TF was obtained along with a guaranteed maximum delay, according to the delay requested by the user. Thus, the results have shown that the new scheduler with call admission control successfully limits the maximum delay and maximizes the number of SSs in a simulated environment.

B. Future Work

In a communication system with a wireless link, the channel effects can heavily degrade the system performance since the wireless link is time-varying and may experience multipath fading and interference. In future work, the effects of the channel will be treated.

Furthermore, most four improvements will be introduced in order to improve traffic in Fixed WiMAX Networks:

1) The loss of packets in the communication channel will be dealt with so we can get more accurate results.
2) The Call Admission Control (CAC) will use an optimization tool that can perform more efficiently, control of connections that will be served by the system.
3) To calculate the time frame (TF), Particle Swarm Optimization (PSO) [24] is used.
4) The Network Simulator NS-3 [25] is used for the simulations of performance evaluation of these new improvements.

In a communication system with a wireless link, the channel effects can heavily degrade the system performance since the wireless link is time-varying and may experience multipath fading and interference.

Acknowledgment

We thank all researchers and collaborators at the Advanced Nucleus of Communication Technology at UTFPR.

References

[1] E. R. Dosciatti, W. Godoy Jr., and A. Foronda, "New Scheduler with Call Admission Control (CAC) for IEEE 802.16 Fixed with Delay Bound Guarantee," in *Proc. of the First International Conference on Mobile Services, Resources, and Users (MOBILITY 2011)*, Barcelona, Spain, Oct. 2011, pp. 139-145.

[2] A. Gosh, D. Wolter, J. Andrews, and R. Chen, "Broadband wireless access with WiMAX/802.16: current performance benchmarks and future potential," in *IEEE Communications*, v. 43(2), Feb. 2005, pp. 129-136.

[3] IEEE 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems," *IEEE Std., Rev. IEEE Std802.16-2004*, New York, Oct. 2004.

[4] E. G. Camargo, C. B. Both, R. Kunst, L. Z. Granville, and J. Rochol, "Uma Arquitetura de Escalonamento Hierárquica para Transmissões Uplink em Redes WiMAX Baseadas em OFDMA," in *Proc. of Brazilian Symposium on Computer Networks and Distributed Systems (SBRC 2009)*, Recife, Brazil, May 2009, pp. 525-538. (in Portuguese).

[5] 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," in *IEEE Communications Magazine*, v. 40(6), June 2002, pp. 98-107.

[6] WiMAX Forum. *WiMAX Forum*. 2012. [Online]. Available: http://www.wimaxforum.org. [Accessed: Dec. 10, 2012].

[7] Y. Sun, I. Sheriff, E. M. Belding-Royer, and K. C. Almeroth, "Experimental Study of Multimedia Traffic Performance in Mesh Networks," in *Proc. of the International Workshop on Wireless Traffic Measurements and Modeling*, Seattle, EUA, June 2005, pp. 25-30.

[8] D. Stiliadis and A. Varma, "Latency-Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms," in *IEEE-ACM Transactions on Networking*, v. 6(5), Oct. 1998, pp. 611-624.

[9] K. Wongthavarawant and A. Ganz, "Packet Scheduling for QoS Support in IEEE 802.16 Broadband Wireless Access Systems," in *International Journal of Communications Systems*, v. 16, Feb. 2003, pp. 81-96.

[10] G. Chu, D. Wang, and S. Mei, "A QoS architecture for the MAC protocol of IEEE 802.16 BWA System," in *Proc. of IEEE Conference on Communications, Circuits, and Systems*, v. 1, Chengdu, China, June/July 2002, pp. 435-439.

[11] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support," in *IEEE Transactions on Mobile Computing - TMC07*, v. 6(1), Jan. 2007, pp. 26-38.

[12] R. Iyengar, P. Iyer, and B. Sikdar, "Delay Analysis of 802.16 Based Last Mile Wireless Networks," in *Proc. of IEEE Global Telecommunications Conference - GLOBECOM'05*, v. 5, St. Louis, EUA, Dec. 2005, pp. 1-5.

[13] D. Cho, J. Song, M. Kim, and K. Han, "Performance Analysis of the IEEE 802.16 Wireless Metropolitan Area Network," in *IEEE Computer Society, DFMA'05*, Feb. 2005, pp. 130-137.

[14] S. Kim and I. Yeom, "TCP-aware Uplink Scheduling for IEEE 802.16," in *IEEE Communications Letters*, v. 11(2), Feb. 2007, pp. 146-148.

[15] S. Maheshwari, "An Efficient QoS Scheduling Architecture for IEEE 802.16 Wireless MANs," *Master Degree*, K. R. School of Information Technology, Bombay, India, Jan. 2005.

[16] E. R. Dosciatti, W. Godoy Jr., and A. Foronda, "Scheduling Mechanisms with Call Admission Control (CAC) and an Approach with Guaranteed Maximum Delay for Fixed WiMAX Networks," in *Quality of Service and Resource Allocation in WiMAX*, INTECH, Croatia, Feb. 2012, pp. 59-84.

[17] I. F. Akyildiz and X. Wang, "A Survey on Wireless Mesh Networks," in *IEEE Communications Magazine*, v.43(9), Sept. 2005, pp. 523-530.

[18] IEEE 802.16e-2005, "IEEE Standard for Local and Metropolitan Area Networks. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands: Part 16: Air Interface for Fixed Broadband Wireless Access Systems," New York, Dec. 2005.

[19] A. S. Tanenbaum, *Computer Networks*. 4. ed. New Jersey: Prentice-Hall, 2003.

[20] INTEL. *Deploying License-Exempt WiMAX Solutions: White paper*. 16 p., Jan. 2005.

[21] A. Foronda, Y. Higuchi, C. Ohta, M. Yoshimoto, and Y. Okada, "Delay Guarantee and Service Interval Optimization for HCCA in IEEE 802.11e WLANs," in *IEEE Wireless Communications and Networking Conference - WCNC 2007*, v. 1, Hong Kong, Mar. 2007, pp. 2080-2085.

[22] A. Parekh and R. Gallager, " A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: the Single-Node Case," in *IEEE/ACM Transactions Networking*, v. 1(3), June 1993, pp. 344-357

[23] D. M. Ritchie and B. W. Kernighan, *The C Programming Language*. 2. ed. New Jersey: Prentice-Hall, 1988.

[24] J. Kennedy and R. C. Eberhart, "Particle Swarm Optimization," in *Proc. of IEEE International Conference on Neural Networks*, v. 4, Piscataway, New Jersey, Nov./Dec. 1995, pp. 1942–1948.

[25] NS-3. *Network Simulator-3*. 2012. [Online]. Available: http://www.nsnam.org. [Accessed: Dec. 10, 2012].