1. Introduction

The major challenge for the second decade of this century is the implementation of access to high-speed internet, known as broadband internet. With the popularization of access to the global network, it is evident that there will be a reduction of physical barriers to the transmission of knowledge, as well as in transaction costs, and will be instrumental in fostering competitiveness, especially for developing countries. But, wired access to broadband internet has a very high cost and is sometimes impracticable, since the investment needed to deploy cabling throughout a region often outweighs the reduces 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 of deployment (Gosh et al., 2005).

Motivated by the growing need ubiquitous high speed access, ie, the use of computers everywhere, embedded in the structures of our lives, wireless technology is an option to provide a cost-effective solution that may be deployed quickly and with easily, providing high bandwidth connectivity in the last mile, ie, in places where business or residential customers, also known as the tail of the distribution network of services. As wireless networks have an ease of deployment and low maintenance cost, ease of configuration and mobility of their devices, there are challenges that must be overcome in order to further advance the widespread use of this type of network.

Thus, the IEEE (Institute of Electrical and Electronics Engineers) has developed a new standard for wireless access, called IEEE 802.16 (802.16-2004, 2004). Also known as WiMAX (Worldwide Interoperability for Microwave Access), it is an emerging technology for next generation wireless networks which supports a large number of users, both mobile and nomadic (fixed), distributed across a wide geographic area. Furthermore, this technology
provides strict QoS (Quality of Service) guarantees for data, voice and video applications (Camargo et al., 2009). As a service provider, WiMAX will create new alternatives for applications such as telephony, TV broadcasts, broadband Internet access for residential users, and commercial, industrial and university centers. This is a new market niche that is revolutionizing telecom companies and interconnection equipment manufacturers (Eklund et al., 2002). Moreover, WiMAX enables broadband connection in areas which are inaccessible or lacking in infrastructure, since it requires no installation or complex physical connections via cables and traditional technologies (WiMAX Forum, 2011). The increasing deployment of wireless infrastructure is enabling 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), among others, which demand real-time data delivery (Sun et al., 2005).

Based on these assumptions and considering that the standard leaves open certain issues related to network resources management and mechanisms for packet scheduling, is that several researchers have presented proposals to resolve issues related to scheduling mechanisms and QoS architectures for Broadband Wireless Access (BWA). However, many of these solutions only address the implementation or addition of a new QoS architecture to the IEEE 802.16 standard.

Thus, this work presents a new scheduler with call admission control to a WiMAX Base Station (BS). An analytical model, based on Latency-Rate (LR) server theory (Stiliadis & Varma, 1998) is developed, from which an ideal frame size, called 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 is considered when is calculated the duration of each time slot. After the developed 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, optimizing the bandwidth.

This work is structured as follows: in Section 2, a brief overview of IEEE 802.16 standard is presented. In Sections 3, 4 and 5 we present the concepts about the operation of scheduling mechanisms, call admission control and QoS, respectively. In Section 6 the related research in this area are discussed. Our proposal for a new scheduler, with its implementation and evaluation, is explained in Section 7. Finally, in Section 8 the conclusion are briefly described.

2. Overview of IEEE 802.16 fixed standard

The basic topology of an 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 (Dosciatti et al., 2010). The BS is the central node, responsible to coordinate communications and provide connectivity to the 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, as shown in 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. So, all stations are peers which can act as routers and forward packets to neighboring nodes (Akyildiz & Wang, 2005). This work only considers the PMP topology, since it is implemented by first-generation WiMAX devices, 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 SSs (WiMAX Forum, 2011).

Fig. 1. IEEE 802.16 network architecture

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 to perform handoffs/handovers, which allow a client station to switch to another base station without losing connectivity. In this case, SSs are instead called mobile stations (MSs). The functionality of handoff/handover was included in the IEEE 802.16 standard in early 2006 with the publication of the IEEE 802.16e (802.16e-2005, 2006), which quickly received the name of "IEEE 802.16 mobile".

WiMAX is designed to leverage wireless broadband metropolitan area networks and, it is obtained performance comparable to traditional cable and xDSL technology, with the following main advantages:

- The ability to provide services in areas with poor infrastructure deployment;
- Elimination unnecessary expenses with facilities;
- The ability to overcome physical boundaries, such as walls or buildings;
- High scalability; and
- Low update and maintenance costs.
WiMAX technology can reach a theoretical maximum distance of 50 km (Tanenbaum, 2003). Data transmission rates can vary from 50 to 150 Mbps, depending on channel frequency bandwidth and modulation type (Intel, 2005). 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 operate in different frequencies using Frequency Division Duplexing (FDD) mode or at different times using Time Division Duplexing (TDD) mode.

In TDD mode, 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 (Wongthavarawant & Ganz, 2003) shows the structure of a TDD frame. In this work, the system was operated in TDD mode with the OFDM (Orthogonal Frequency Division Multiplexing) air interface, as determined by the standard.

In TDD mode, 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 (Wongthavarawant & Ganz, 2003) shows the structure of a TDD frame. In this work, the system was operated in TDD mode with the OFDM (Orthogonal Frequency Division Multiplexing) air interface, as determined by the standard.

![IEEE 802.16 frame structure](image)

Figure 2 (Hoymann, 2005) shows an example OFDM frame structure in TDD mode. As seen earlier, each frame has a DL subframe followed by an UL subframe. In this structure, the system supports frame-based transmission, in which variable frame lengths can be adopted. These subframes consist of a fixed number of OFDM symbols. Details of the OFDM symbol structure may be found in (Gosh et al., 2005).

The DL subframe starts with a long preamble (two OFDM symbols) through which SSs can synchronize with the network and check the duration of the current frame. Instantly after the DL long preamble, the BS transmits the Frame Control Header (FCH), which consists of an OFDM symbol and is used by SSs to decode MAC control messages transmitted by the BS.
Scheduling Mechanisms with Call Admission Control (CAC) and an Approach with Guaranteed Maximum Delay for Fixed WiMAX Networks

Fig. 3. OFDM frame structure with TDD

The UL subframe consists of contention intervals for initial ranging and bandwidth request purposes and one or several UL transmission bursts, each from a different SS. The initial ranging slots allow an SS to enter the system, by adjusted its power level and frequency offsets and correctness of its time offset. Bandwidth request slots are used by SSs to transmit bandwidth request headers.

Two gaps separate the DL and UL subframes: the Transmit/Receive Transition Gap (TTG) and the Receive/Transmit Transition Gap (RTG). These gaps allow the BS to switch from transmit to receive mode, and vice versa.

3. Scheduling mechanisms

Scheduling mechanisms were intentionally left outside the scope of the IEEE 802.16 standard. The diversity of service offered combined with scheduling mechanisms is an important area for differentiation in the development of research in both industry and academia. However, some concepts are common to all implementations, and some ideas were intended even while not explicitly made a part of the standard.

The scheduling mechanism in the WiMAX MAC layer is designed to efficiently deliver broadband data services such as voice, video and other data related to change of broadband wireless channel. Scheduling is the main component of the MAC layer that helps assure QoS to various service classes. The scheduler is located at each BS to enable rapid response to traffic requirements and channel conditions. So, the scheduler works as a distributor to allocate the resources among SSs. The scheduling mechanism is provided for both DL and UL traffic. The allocated resource can be defined as the number of slots and then these slots are mapped into a number of subchannels (each subchannel is a group of multiple physical subcarriers) and time duration (OFDM symbols). These allocated resources are delivered in MAP messages at the beginning of each frame. Therefore, the resource allocation can be changed from frame to frame in response to traffic and channel conditions. The amount of resources in each allocation can range from one slot to the whole frame. The MAC scheduler handles data transport on a connection-by-connection basis. Each connection is associated
with a single scheduling mechanism that is determined by a set of QoS parameters that quantify aspects of its behaviour.

Scheduling has also been studied intensively in many disciplines, such as CPU task scheduling in operating systems, service scheduling in a client-server model, and events scheduling in communication and computer networks. Thus a lot of scheduling algorithms have been developed. However, compared with the traditional scheduling problems, the problem of scheduling at the MAC layer of WiMAX networks is unique and worth study by four reasons described below.

1. The total bandwidth in a WiMAX network is adaptive by Adaptive Modeling and Coding (AMC) and is deployed at the physical layer and the number of bytes each time slot can carry depends on the coding and modulation scheme.
2. Multiple service types have been defined and their QoS requirements need to be satisfied at the same time. How to satisfy various QoS requirements of different service types simultaneously has not been addressed by any other wireless access standard before.
3. The time complexity of WiMAX scheduling algorithm must be simple, since in real-time services require a fast response from the central controller in the BS.
4. The frame boundary in the WiMAX MAC layer also serves as the scheduling boundary, which makes the WiMAX scheduling problem different from the continuous time scheduling problem.

To implement a scheduler, the following aspects, as defined in IEEE 802.16, must be taken into consideration:

- The distribution of resources should be made based on the bandwidth requests sent by the SSs and QoS parameters of each connection, and different connections use the same type of service, different values for the same QoS parameter may occur.
- Bandwidth allocation should allow not only the transmission of data, but also the transmission of bandwidth requests in accordance with the request mechanism established for each type of service.
- All QoS parameters defined by the standard should be guaranteed.

In addition, the scheduler must efficiently use the available bandwidth so that a greater number of users can be admitted, thus resulting in high levels of network utilization.

Although the scheduler is implemented at the MAC layer, the technology used at the PHY layer can influence its project. When used a WirelessMAN-SC (Single Carrier) PHY, there is only one carrier frequency and the whole is given to an SS. This PHY layer requires line of sight (LOS) communication. Rain attenuation and multipath also affect reliability of the network at these frequencies. To allow non-line of sight (NLOS) communication, IEEE 802.16 designed the Orthogonal Frequency Division Multiplexing (OFDM) PHY, popularly known as IEEE 802.16d and used in this work, and designed for wireless fixed stations. In this PHY layer, multiple subcarriers form a physical slot, but because they are transparent to the MAC layer, the subcarriers can be seen as a logical channel from the point of view of the scheduler. However, each subchannel may use different modulations so that different SSs may have different transmission rates. How OFDM is a multicarrier transmission in which thousands of subcarriers are transmitted, each user is given complete control of all
subcarriers. The scheduling decision is simply to decide what time slots should be allocated to each SS. For mobile users, it is better to reduce the number of subcarriers and to have higher signal power per SS. Therefore, multiple users are allowed to transmit on different subcarriers at the same time slot. The scheduling decision is then to decide which subcarriers and what time slots must be allocated to a given user. This combination of time division and frequency division multiple access in conjunction with OFDM is called Orthogonal Frequency Division Multiple Access (OFDMA). Then, the OFDMA requires allocation of resources in two dimensions: frequency and time. In other words, the scheduler must decide not only on the allocation of slots, but also of subcarriers for each user. Since more than one user can use the channel at the same instant of time, it is essential that the scheduling algorithms consider the characteristics of the OFDMA physical layer. Figure 4 (So-In et al., 2009) illustrates a schematic view and the differences between the three types of the 802.16 PHYs, discussed above. The details of these interfaces can be found in (802.16-Rev2/D2, 2007).

Fig. 4. IEEE 802.16 PHYs: SC, OFDM and OFDMA (the letters A, B, C, D, E, F and G represent different users)

The scheduler for WirelessMAN-SC can be fairly simple because only time domain is considered. The entire frequency channel is given to the BS. For OFDM, it is more complex since each subchannel can be modulated differently, but it is still only in time domain. On the other hand, both time and frequency domains need to be considered for OFDMA. The OFDMA scheduler is the most complex because each SS can receive some portions of the allocation for the combination of time and frequency so that the channel capacity is efficiently utilized.

A scheduling mechanism needs to consider the allocations logically and physically. Logically, the scheduler should calculate the number of slots based on QoS service classes. Physically, the scheduler needs to select which subchannels and time intervals are suitable for each user. The goal is to minimize power consumption, to minimize bit error rate and to maximize the total throughput.
The scheduling mechanism in IEEE 802.16 includes the scheduling of downlink traffic, carried out by BS, and the scheduling of uplink traffic, performed by two schedulers, one at the BS and another at the SSs as shown in Figura 5. To perform the allocation of resources, the schedulers use information about the QoS requirements and the status of the queues of the connections.

At the BS, packets from the upper layer are put into different queues. However, the optimization of queue can be done and the number of required queues can be reduced. Then, based on the QoS parameters and some extra information such as the channel state condition, the DL-BS scheduler decides which queue to service and how many service data units (SDUs) should be transmitted to the SSs. Since the BS controls the access to the medium, the second scheduler (the UL-BS scheduler) makes the allocation decision based on the bandwidth requests from the SSs and the associated QoS parameters. Finally, the third scheduler is at the SS. Once the UL-BS grants of the bandwidth for the SS are allocated, the SS scheduler decides which queues should use that allocation. While the requests are per connections, the grants are per subscriber and the subscriber is free to choose the appropriate queue to service. The SS scheduler needs a mechanism to allocate the bandwidth in an efficient way, what is Call Admission Control and is treated in the next section.

Fig. 5. Components of the schedulers at BS and SSs
4. Call Admission Control - CAC

Typically, a Call Admission Control (CAC) procedure is also implemented at the BS that ensures the load supplied by the SSs can be handled by the network.

While the scheduling mechanism ensures that the required amount of resources is allocated to the connections, so that the QoS requirements are met, the admission control mechanism limits the number of connections to the network so the network is not overloaded by many users.

Whenever an user wants to establish a new connection, a request sent to the BS for the admission control mechanism to decide whether the new connection may or may not be accepted. To make this decision, the admission control must ensure that there are sufficient resources to meet the QoS requirements of the new connection without compromising the minimum QoS requirements of ongoing connections.

The choice of admission control policy to be adopted in an IEEE802.16 network is strongly associated with the scheduling mechanism used. For example, by the adoption of an admission control mechanism that estimates the resources available from the difference between the total capacity of the link and the sum of the minimum rate requirements of already admitted connections, you should ensure that the scheduler will not allocate more than the minimum rate for a connection when other connections have not yet met their minimum requirements. In addition, the integration of scheduling and admission control can result in simple solutions, because when one the mechanisms is able to guarantee the fulfillment of a requirement for the same QoS guarantees need not be implemented by other mechanism.

Thus, CAC restricts the access to the network in order to prevent network congestion or service degradation for already accepted users. It can prevent the system from be overloaded. CAC has been characterized as the decision maker for the network as is shown in Figure 6.

![CAC policy process](image)

Fig. 6. CAC policy process

However, before the user is allocated in the CAC policies, a bandwidth request is required. Thus, to design a CAC algorithm, we must also worry about the bandwidth allocation algorithm. For this, the following points should be highly valued (Mei et al., 2010):

- The fairness of the bandwidth allocation: Different terminals carry different data transmission business, so the bandwidth requirement of different business is varying.
This means that to allocate the bandwidth, QoS should not simply be the average allocation. Meanwhile, it would be not appropriate if bandwidth quantity between terminals varies dramatically.

- Data transmission delay: Adjust the delay time of data transmission of terminals or connections through allocation of bandwidth, and to keep the delay in a reasonable range tolerated by terminals or connections.

- The data throughput of the system.

- Combined with protocol construction of WiMAX, the ways of the bandwidth allocation request and the existence mechanism of QoS assurance mechanism.

In this work, we only consider the PMP mode architecture of IEEE 802.16 BWA networks, where transmission only occurs between a BS and SSs and the BS controls all the communications between BS and SSs. The connection can be either downlink (from BS to SS) or uplink (from SS to BS) as it is depicted in Figure 1. In PMP architecture, two modes are defined: Grant-Per-Connection (GPC) and Grant-Per-Subscriber-Station (GPSS). Under GPC, the CAC algorithm considers each individual connection arriving from an SS, while for GPSS each SS manages admission of its own individual connections before sending a single bandwidth (BW) request to the BS.

The IEEE 802.16 standard does not define policies for admission control, which has encouraged researchers from academia and industry to investigate solutions to this problem. There are already several proposals in the literature, such as (Chen et al., 2005; Guo et al., 2007; Masri et al., 2009; Rong et al., 2008; Wang H et al., 2007; Wang L et. al. 2007). In Section 7 is introduced a new solution to the problem.

5. Quality of Service - QoS

IEEE 802.16 can support multiple communication services (data, voice, video) with different Quality of Service (QoS) requirements organized into different connections. The MAC layer of IEEE 802.16 standard defines mechanisms to provide QoS and control of data transmission between the BS and SSs. The main QoS mechanism is the association of packets that pass through the MAC layer to a service flow. The service flow is a MAC layer service that provides unidirectional message transport. During connection establishment, these service flows are created and activated by the BS and SSs. Each service flow must define its own set of QoS parameters, among them maximum delay, minimum bandwidth and type of service scheduling.

Within this context, the IEEE 802.16 standard defines four service classes associated with traffic flows, and each such class has different QoS requirements, which are fulfilled has a scheduler allocated bandwidth to the SSs under a set of rules (802.16-2004, 2004). The four service classes defined by standard are:

- Unsolicited Grant Service (UGS): this service is designed to support real-time applications that generate fixed-size data packets on a periodic basis, such as T1/E1 and Voice over IP (VoIP) without silence suppression. UGS does not need the SS to explicitly request bandwidth, thus eliminating the overhead and latency associated with bandwidth request. Because UGS connections never request bandwidth, the amount of bandwidth to allocate to such connections is computed by the BS based on the minimum reserved traffic rate defined in the service flow of that connection.
• real-time Polling Services (rtPS): this service is designed to support real-time services that generate variable-size data packets at periodic intervals, such as moving pictures expert group (MPEG) video and VoIP with silence suppression. Unlike UGS connections, rtPS connections must inform the BS of their bandwidth requirements. Therefore the BS must periodically allocate bandwidth for rtPS connections specifically for the purpose to request bandwidth. In this service class, the BS provides unicast polling opportunities for the SS to request bandwidth. The unicast polling opportunities are frequent enough to ensure that latency requirements of real-time services are met. This service requires more request overhead than UGS does but is more efficient for service that generates variable-size data packets or has a duty cycle less than 100 percent.

• non-real-time Polling Services (nrtPS): this service is designed to support delay-tolerant applications such as FTP (File Transfer Protocol) for which a minimum amount of bandwidth is required. Also, this service is very similar to rtPS except that the SS can also use contention-based polling in the uplink channel to request bandwidth. In nrtPS, it is allowable to have unicast polling opportunities, but the average duration between two such opportunities is in the order of few seconds, which is large compared to rtPS. All SSs that are part of the group can also request resources during the contention-based polling opportunity, which can often result in collisions and additional attempts.

• Best-Effort (BE): this service provides very little QoS support and is applicable only for services that do not have strict QoS requirements, such as HTTP (Hypertext Transfer Protocol) and SMTP (Simple Mail Transfer Protocol). Data is sent whenever resources are available and not required by any other scheduling-service classes.

Each of these service classes should be treated differently by the MAC layer packet scheduling mechanism. Thus, each type of application can be included in a class of service. However, the WiMAX standard does not define nor specify a scheduler, and one of the premises needed to guarantee QoS in WiMAX networks is the application of scheduling in both the uplink and downlink directions, which should translate the QoS requirements of SSs to appropriate slot allocation. When the BS makes a scheduling decision, it informs its decision to all SSs using the messages at the beginning of each frame. These messages explicitly define which slots are allocated to each SS in both directions, uplink and downlink.

This work focuses on packet scheduling in the uplink direction, because it guarantees optimization of physical network rate and ensures the delay requested by the user, therefore to maximize the number of users transmitting data in each frame. The classes of services described above, has its own QoS parameters such as minimum throughput requirement and delay/jitter constraints.

6. Related research

Several scheduling algorithms and QoS architectures for Broadband Wireless Access (BWA) have been proposed in the literature, since the standard only provides 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 wait list and is one of the mechanisms responsible for the distribution of bandwidth between multiple streams (through the attribution of each stream bandwidth that was needed and available). In these proposals, there are often no analytical models for ensuring maximum delay and to maximize the number of SSs allocated in the system, which are represented accurately by certain performance metrics of the medium access protocol such as delay.

In (Wongthavarawant & Ganz, 2003), a packet scheduler for IEEE 802.16 uplink channels based on a hierarchical queue structure was proposed. A simulation model was developed to evaluate the performance of the proposed scheduler. However, despite presenting simulation results, the authors overlooked the fact that the complexity to implement this solution is not hierarchical, and did not define clearly how requests for bandwidth are made.

In (Chu et al., 2002) authors proposed a QoS architecture to be built into the IEEE 802.16 MAC sublayer, which significantly impacts system performance, but did not present an algorithm that makes efficient use of bandwidth.

In (Cicconetti et al., 2007), authors presented a simulation study of the IEEE 802.16 MAC protocol operating with an OFDM air interface and full-duplex stations. They evaluated system performance under different traffic scenarios, with the variation of a set of relevant system parameters. About the data traffic, it was observed that the overhead due to the physical transmission of preambles increases with the number of stations.

In (Iyengar et al., 2007), 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. They 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 for delay guarantees.

In (Cho et al., 2005), authors proposed a QoS architecture in which the scheduler is based on packet lifetime for each type of flow. In this paper, authors considered the process of data communication between BS and SS 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, in spite to present simulation results, the work neglects performance by not adequately address the functional blocks of the proposed architecture and by not clearly how lifetime is calculated for each packet.

In (Kim & Yeom, 2007), the scheduling algorithm handles traffic with Best Effort (BE), and concludes that it is difficult to estimate the amount of bandwidth required due to dynamic changes in traffic transmission rate. The purpose of this algorithm is to keep fair bandwidth allocation between 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 (Maheshwari, 2005) 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 the allocation uplink bandwidth to each SS, although the decision is taken based on the following aspects: bandwidth required by each SS for uplink data transmission, periodic
bandwidth needs for UGS flows in SSs and bandwidth required to make requests for additional bandwidth.

Given 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 getting QoS guarantees, with no concerns for maximize 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 (Stiliadis & Varma, 1998), from 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.

7. Proposed scheduler: Implementation and evaluation

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. In this context, some features of the system related to the subject were discussed earlier in this work. To achieve this, this section presents an analytical model of the new scheduler and an analytical description of its call admission control.

7.1 System description

Figure 7 (Foronda et al., 2007) illustrates a wireless network that use the new proposed scheduler with call admission control, which is based on a modified LR scheduler (Stiliadis & Varma, 1998) and uses the token bucket algorithm. The basic approach consists of the token bucket to limit incoming traffic and the LR scheduler provide 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 LR scheduler.

The behavior of an LR scheduler is determined by two parameters for each session $i$: latency $\theta_i$ and allocated rate $r_i$. The latency $\theta_i$ 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 $\theta_i$ is a TF period, which is the time needed to transmit a 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 $\theta_i$ of is given by

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

(1)

where $T_{TTG}$ and $T_{RTG}$ are the DL and UL subframe gap durations, $T_{TL}$ 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 $r_i$ for each session $i$ may be determined, and how to optimize $TF$ in order to increase the number of connections accommodated with Call Admission Control (CAC).

### 7.2 CAC description

An LR scheduler can provide a bounded delay if input traffic is shaped by a token bucket. A token bucket (Gosh et al., 2005) is a non-negative counter which accumulates tokens at a constant rate $\rho_i$ until the counter reaches its capacity $\sigma_i$. Packets from session $i$ can be released into the queue only after remove the required number of tokens from the token bucket. In an LR scheduler, if the token bucket is empty, packets that arrive are dropped; however, our model ensures that there will always be tokens in the bucket and that no packets are dropped, as described in next section. If the token bucket is full, a maximum burst of $\sigma_i$ 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 $\sigma_i$, 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 $L_{\text{max},i}$ and required maximum allowable delay $D_{\text{max},i}$, 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 $i$ ($i = 1, \ldots, N$) passes through a token bucket inside the user terminal during the time interval $(0, t)$, as shown in Figure 8 (Dosciatti et al., 2010).

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

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

(2)
where $\sigma_i$ is the bucket size and $\rho_i$ 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 (Stiliadis & Varma, 1998) and (Parekh & Gallager, 1993), it is shown that if input traffic $A_i$ 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 $D_i$:

$$D_i \leq \frac{\sigma_i}{r_i} + \theta_i - \frac{L_{\text{max},i}}{r_i}$$

where $\sigma_i$ is the token bucket size, $r_i$ is the service rate, $\theta_i$ is the scheduler latency, $L_{\text{max},i}$ is the maximum size of a packet and, $(\sigma_i / r_i) + \theta_i - (L_{\text{max},i} / r_i)$ is the bound on the delay, $D_{\text{bound}}$.

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. $D_{\text{bound}}$ should be smaller than or equal to the maximum allowable delay:

$$\frac{\sigma_i}{r_i} + \theta_i - \frac{L_{\text{max},i}}{r_i} \leq D_{\text{max},i}$$

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

$$\frac{(\sigma_i' - L_{\text{max},i}')TF}{r_i'TF - \Delta R + L_{\text{max},i}'} + TF + \frac{L_{\text{max},i}'}{R} + T_{RTG} + T_{RTG} \leq D_{\text{max},i}$$
where $\sigma_i'$ is the token bucket size with overhead, $L_{\text{max},i}'$ is the maximum size of a packet with overhead, $TF$ is the time frame, $r_i'$ is the rate allocated by the server with overhead, $R$ is the outgoing link capacity, $T_{\text{TTC}}$ is the gap between downlink and uplink subframes, $T_{\text{RTG}}$ is the gap to between uplink and downlink subframes, $D_{\text{max},i}$ is the maximum allowable delay and $\Delta$ is the sum of initial ranging and BW request, which is the uplink subframe overhead and whose value will be discussed when evaluated their performance.

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 that must be smaller than the service rate to place a bound on delay. Thus, the constraint condition is

$$\rho_i + \frac{\Delta R + L_{\text{max},i}'}{TF} \leq r_i'$$  \hspace{1cm} (6)

where $\rho_i$ is the bucket rate, $\Delta$ is the uplink subframe overhead, $R$ is the outgoing link capacity, $L_{\text{max},i}'$ is the maximum packet size with overhead, $TF$ is the time frame and $r_i'$ is the rate allocated by the service with overhead.

Figure 9 shows a frame structure with TDD allocation formulas as described by Equation (5). 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).

![Frame structure with TDD allocation formulas of Equation (5)](image)

Previous schedulers do not provide any mechanisms to estimate the TF need 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, $D_{\text{bound}}$. 
Different optimization techniques may be used to solve this problem. In this work, we have used a step-by-step approach, which does not change the scheduler's essential operation. We start with a small TF, 2.5 ms, calculate $r'_i$ and repeat this process every 0.5 ms until we find the minimum $r'_i$ that satisfies both equations.

7.3 Performance evaluation

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 analytical model proposed is validated, 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.

7.3.1 Calculation of optimal time frame

In this work, 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 find 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. All PHY and MAC layer parameters used in simulation are summarized in Table 1 and can be seen in Figure 9.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>OFDM Symbol Duration</td>
<td>13.89 μs</td>
</tr>
<tr>
<td>Delay</td>
<td>5, 10, 15 and 20 ms</td>
</tr>
<tr>
<td>Δ (Initial Ranging and BW Request) → 9 OFDM Symbols</td>
<td>125.10 μs</td>
</tr>
<tr>
<td>TTG + RTG → 1 OFDM Symbol</td>
<td>13.89 μs</td>
</tr>
<tr>
<td>UL Subframe (preamble + pad) → 10% OFDM Symbol</td>
<td>1.39 μs</td>
</tr>
<tr>
<td>Physical Rate</td>
<td>70 Mbps</td>
</tr>
<tr>
<td>DL Subframe</td>
<td>1% TF</td>
</tr>
</tbody>
</table>

Table 1. PHY and MAC parameters

Performance of the new scheduler with call admission 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 to define the token bucket parameters, which are estimated in accordance with the characteristics of incoming traffic and are listed on Table 2. It is important to note that the details about the incoming traffic must be known in advance. This is normal for various applications such as audio, CBR and video on demand.

<table>
<thead>
<tr>
<th></th>
<th>Audio</th>
<th>VBR video</th>
<th>MPEG4 video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token Size (bits)</td>
<td>3000</td>
<td>18000</td>
<td>1000</td>
</tr>
<tr>
<td>Token Rate (Kb/s)</td>
<td>64</td>
<td>500</td>
<td>4100</td>
</tr>
</tbody>
</table>

Table 2. Token bucket parameters

Thus, the optimal TF value is estimated according to the PHY and MAC layer's parameters (see Table 1), token bucket parameters (see Table 2), 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 7.2, the simulation starts with a step-by-step approach, with the value of TF estimated at 2.5 ms. The $r_i$ is calculated and the procedure is repeated every 0.5 ms until that the minimum $r_i$ that satisfies both equations be found.

The graph in Figure 10 shows the optimal TF value, for four delay values required by users (5, 10, 15 and 20 ms). For example, in the graph, 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 10.5 ms. Finally, for a requested delay of 20 ms, the optimal TF is 15 ms.

![Fig. 10. Optimal TF](www.intechopen.com)
Next, we show the number of SSs assigned to each traffic type. 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 11 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 11 should be highlighted:

- 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.
- 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 increase of the 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 comply 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.
7.3.2 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 12 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 12. It may be noticed, then, that to choose 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 and ensures an upper bound on maximum delay, as discussed next.

![Fig. 12. Users assigned as a function of TF for audio traffic](image)

7.3.3 Guaranteed maximum delay

In this work, only UL traffic is considered. To test the new scheduler's performance, we have carried out simulations of an IEEE 802.16 network consists 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 13. In this topology, six SSs transmit on-off CBR audio traffic (64 kb/s), six transmit CBR MPEG4 video traffic (3.2 Mb/s) and six transmit VBR video traffic.
Table 3 summarizes the different types of traffic used in this simulation.

<table>
<thead>
<tr>
<th>Node</th>
<th>Application</th>
<th>Arrival Period (ms)</th>
<th>Packet size (max) (bytes)</th>
<th>Sending rate (kb/s) (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 → 6</td>
<td>Audio</td>
<td>4.7</td>
<td>160</td>
<td>64</td>
</tr>
<tr>
<td>7 → 12</td>
<td>VBR video</td>
<td>26</td>
<td>1024</td>
<td>≈200</td>
</tr>
<tr>
<td>13 → 18</td>
<td>MPEG4 video</td>
<td>2</td>
<td>800</td>
<td>3200</td>
</tr>
</tbody>
</table>

Table 3. Description of traffic types

On Figure 14, 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.
7.3.4 Comparison with other schedulers

The New Scheduler with Call Admission Control was compared to those of (Iyengar et al., 2005), here called Scheduler_1, and (Wongthavaraewat & Ganz, 2003), here called Scheduler_2. The comparison was accomplished through the ability to allocate users in a particular time frame (TF).

Table 4 shows the parameters used in the comparisons.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>OFDM symbol duration</td>
<td>13.89 μs</td>
</tr>
<tr>
<td>Delay requested by user</td>
<td>Dependent of each comparison</td>
</tr>
<tr>
<td>Maximum data rate</td>
<td>70 Mbps</td>
</tr>
<tr>
<td>Traffic type</td>
<td>Audio</td>
</tr>
</tbody>
</table>

Table 4. Parameters used in the comparisons
In the graph of Figure 15, we compare the New Scheduler with the Scheduler_1. A maximum delay of $0.12 \text{ ms}$ was requested by the user, and the duration of each frame (TF) was set at $5 \text{ ms}$, as in Scheduler_1. Other parameters are listed in Table 4. 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.

![Fig. 15. Comparison between the New Scheduler and Scheduler_1](image)

In the graph of Figure 16, we compare the New Scheduler with the Scheduler_2. A maximum delay of $20 \text{ ms}$ was requested by the user, and the duration of each frame (TF) was set at $10 \text{ ms}$, as in Scheduler_2. Other parameters are listed in Table 4. The comparison was extended by also considering frame duration values of $7.00 \text{ ms}$, $8.00 \text{ ms}$ and $9.00 \text{ ms}$ to demonstrate the efficiency of the new scheduler. For a TF of $10 \text{ ms}$, 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.
Fig. 16. Comparison between the New Scheduler and Scheduler_2

8. Conclusion

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. 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 and demonstrated that an optimal TF was obtained with a guaranteed maximum delay in accordance with 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.

9. References


This book has been prepared to present state of the art on WiMAX Technology. It has been constructed with the support of many researchers around the world, working on resource allocation, quality of service and WiMAX applications. Such many different works on WiMAX, show the great worldwide importance of WiMAX as a wireless broadband access technology. This book is intended for readers interested in resource allocation and quality of service in wireless environments, which is known to be a complex problem. All chapters include both theoretical and technical information, which provides an in depth review of the most recent advances in the field for engineers and researchers, and other readers interested in WiMAX.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following: