Scheduling and CAC in IEEE 802.16 Fixed BWNs: A Comprehensive Survey and Taxonomy

Ikbal Chammakhi Msadaa, Daniel Câmara, and Fethi Filali

EURECOM
Mobile Communications Department
2229 route des crêtes, BP 193 F-06560
Sophia-Antipolis, cedex, France

{msadaa, camara, filali}@eurecom.fr

Abstract

IEEE 802.16 technology has emerged as a competitive alternative to wireline broadband access solution. IEEE 802.16 can provide quality of service (QoS) guarantees for heterogeneous classes of traffic with different QoS requirements. The standard, however, leaves open the resource management and scheduling issues, which are crucial components to guarantee QoS performance. The main objective of this paper is to provide a better understanding of the missing components to ensure QoS support in IEEE 802.16 fixed broadband wireless networks (BWNs), namely scheduling and connection admission control (CAC) schemes. First, we highlight the key challenges in designing such schemes, for both point-to-multipoint (PMP) and mesh modes. Then, we survey, classify, and compare different scheduling and admission control mechanisms proposed in this work-in-progress area.

Keywords: IEEE 802.16, QoS, scheduling, CAC, PMP, mesh.

1 Introduction

The development of 802.16 standards for broadband wireless access (BWA[1]) technologies was motivated by the rapidly growing need for high-speed, ubiquitous and cost-effective access. Addressing these pervasive needs, the IEEE 802.16 technology has emerged as a competitive alternative to wireline broadband access.

The IEEE 802.16 standard supports heterogeneous classes of traffic with different QoS requirements and defines several signaling mechanisms to request and allocate resources. Also it offers the possibility of adapting the modulation and coding schemes (MCSs) based on the channel conditions and proposes a set of techniques such as packing and fragmentation to allow efficient use of the available bandwidth. The standard, however, leaves unstandardized the resource management and scheduling algorithms. The main objective of this paper is to provide a better understanding of the different technical issues that researchers are currently facing to ensure QoS support in IEEE 802.16 fixed broadband wireless networks (BWNs) and to give an insight into the new research interests in this field. Therefore, we first highlight the main challenges to address when designing a CAC and scheduling solution for IEEE 802.16 networks and then we summarize, classify, and compare the different mechanisms that have been proposed to solve this problem in both PMP and mesh modes. To the best of our knowledge, this is the first work surveying the different resource allocation mechanisms that have been proposed in this work-in-progress area.

The rest of this paper is organized as follows. In Section 2, we give an overview of the IEEE 802.16 standard with a main focus on media access control (MAC) QoS related issues. For a better understanding of the scheduling problem in 802.16 networks, we provide an insight into the main capabilities offered by the physical layer in terms of resource allocation. In the same section, both PMP and mesh media-sharing modes are presented and the context to which we restricted this survey is defined. Section 3 points out the necessary and desirable features to incorporate in a scheduling and CAC solution for the IEEE 802.16 networks. A survey and taxonomy of the different scheduling and CAC mechanisms presented in literature for both PMP and mesh modes are given in Sections 4 and 5, respectively. In order to understand how the issue of scheduling and CAC is tackled in real deployed networks, we show in Section 6 the main features supported by some examples of WiMAX equipment. Section 7 concludes the survey and gives directions for future research on the field of CAC and

[1] The abbreviations and acronyms used in this survey are listed in Appendix A.

scheduling in 802.16 networks. All the abbreviations and acronyms used in this survey are listed in Appendix A.

2 Overview on the IEEE 802.16 standard

The 2004 version of the IEEE 802.16 standard [1] defines the air interface for fixed BWA systems in the frequency ranges 10-66 GHz and sub 11 GHz. The standard covers both the media access control (MAC) and the physical (PHY) layers. The 802.16 MAC layer was designed to accommodate different PHYs and services, which address the needs of different environments. In this paper, systems of interest are those operating at frequencies below 11 GHz—where line-of-sight (LOS) is not required—and more precisely those using either single carrier (SC) or orthogonal frequency division multiplex (OFDM). We focus only on these two modulation modes because the use of orthogonal frequency division multiple access (OFDMA), in comparison with the two others, introduces a second dimension constraint—frequency—by adding subchannels allocation to the scheduling problem. Nevertheless, we mention in this survey some works, based on OFDMA modulation, that can be easily generalized to the OFDM case. More details about the specific case of OFDM-based physical layer are given in Section 2.1.

Nodes belonging to the same network, share the same wireless medium using one of the two modes specified in the IEEE 802.16 standard [1], [2]: the two-way PMP mode (mandatory) and the mesh mode (optional). The main difference between the two modes is that in mesh mode, subscriber stations (SSs) have the possibility to communicate with each other directly or through the base station (BS), depending on the transmission algorithm in use: distributed, centralized, or a combination of both. In PMP mode however, a central BS—corresponding in general to the Internet service provider (ISP)—receives and coordinates all the transmissions occurring between SSs, which represent the residential or business customers. Further details on the operation mode in both PMP and mesh are given in Sections 2.3 and 2.4, respectively.

2.1 OFDM physical layer

System Profile Identifier	Channel Bandwidth BW (MHz)	Sampling factor n
profP3_1.75	1.75	8/7
profP3_3	3	86/75
profP3_3.5	3.5	8/7
profP3_5.5	5.5	316/275
profP3_7	7	8/7

TABLE 1: WirelessMAN-OFDM System Profiles

OFDM PHY is designed for frequencies below 11 GHz where LOS is not necessary and where multipath may be significant. To collect multipath, a cyclic prefix (CP) is used. As depicted in Figure 1.a, this prefix corresponds to a copy of the last T_g of the useful symbol time T_b of an OFDM symbol T_{sym}. The OFDM symbol transmission time is then expressed as follows: $T_{sym} = T_g + T_b$; where the guard time T_g is given by: $T_g = g * T_b$. g corresponds to the ratio of CP time to useful time. The possible values of g are: 1/4, 1/8, 1/16, and 1/32 [2].

As for the frequency domain structure, an OFDM symbol, described by Figure 1.b, is composed of data subcarriers (for data transmission), pilot subcarriers (for estimation purposes) and null subcarriers such as guard subcarriers. The total number of subcarriers corresponds to the fast Fourier transform (FFT) size N_{fft}. According to [2], $N_{fft} = 256$. Let BW, n and F_s denote the nominal channel bandwidth, the sampling factor and the sampling frequency, respectively. The sampling frequency corresponds to: $F_s = n * BW$. The value of the sampling factor n depends on the channel bandwidth BW as it is illustrated by Table 1. The possible values of BW correspond to those specified in the system profiles proposed by the IEEE 802.16 standard [2] for systems operating with the WirelessMAN-OFDM air interface. As shown in Table 1, five PHY profiles are specified for these systems, each corresponding to a channel bandwidth. Suppose that $\triangle f$ stands for the subcarrier spacing, then: $\triangle f = F_s / N_{fft}$ and the useful time is given by: $T_b = 1/\triangle f$.

For a given system configuration (BW and g fixed), the duration of an OFDM symbol is fixed. However, in terms of data, the number of information bits per OFDM symbols varies depending on the modulation and coding scheme (MCS) in use. Indeed, the number of of information bits per symbol is computed as follows.

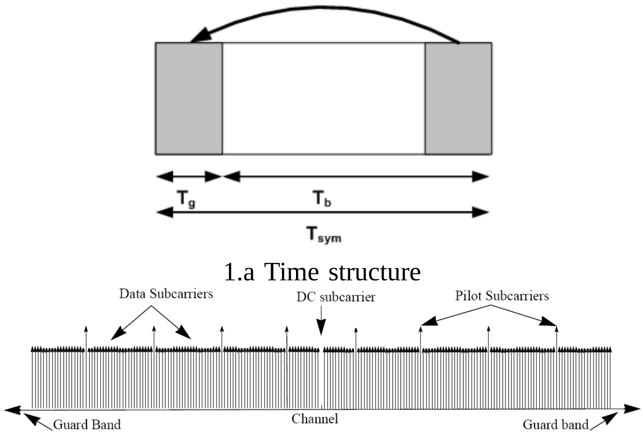
$$N_{MCS}^{bpsym} = N_{data-sub} * efficiency_{MCS} * codingrate_{MCS} - 8$$

where:

- $N_{data-sub}$ stands for the number of data subcarriers ($N_{data-sub} = 192$).
- $efficiency_{MCS}$ is the efficiency, also called repetition, of the MCS (x2, x4, or x6).
- $codingrate_{MCS}$ is the coding rate of the MCS (1/2, 2/3, or 3/4).
- The “-8” refers to the 0x00 tail byte at the end of each OFDM symbol.

For 16QAM 3/4, for instance, $N_{16QAM-3/4}^{bpsym} = 192 * 4 * 3/4 - 8 = 568$.

Note that, unlike OFDMA and multi-user OFDM [3], the OFDM scheme we are considering in this survey allows only one user to access the channel at any given time. Nevertheless, to accommodate multi-user access, OFDM can be combined with a time division multiple access (TDMA) scheme which allows multiple users to access the channel in separate time slots (cf. Section 2.3).



1.a Time structure

1.b Frequency description

Fig. 1: OFDM Symbol structure

2.2 QoS support in IEEE 802.16 networks

The standard defines a connection-oriented MAC protocol where all the transmissions occur within the context of a unidirectional connection. Each connection, identified by a unique Connection ID (CID), is associated to an admitted or active service flow (SF) whose characteristics provide the QoS requirements to apply for the protocol data units (PDUs) exchanged on that connection. There are three types of service flows: (a) provisioned service flows for which the QoS parameters are provisioned for example by the network management system, (b) admitted service flows for which resources—mainly bandwidth—are reserved and (c) active service flows which are activated to carry traffic using resources actually provided. Each service flow is uniquely identified by a service flow identifier (SFID). Service flows may be dynamically managed. They may be created, changed or deleted using Dynamic Service Addition (DSA), DS change (DSC), and DS delete (DSD) MAC management messages, respectively. As mentioned above, a service flow defines the QoS that should be provided to the packets traversing the MAC interface and which are associated to that SF. In order to facilitate the MAC service data units (SDUs) delivery with the appropriate QoS constraints, the IEEE 802.16 Standard defines a classification process by which a MAC SDU is mapped to the associated connection and so to the SF corresponding to that connection. The classification procedure is performed by classifiers consisting of a set of protocol-specific matching criteria.

Depending on the service to be tailored to each user application, a specific scheduling service is attributed to handle the flow. Based on that, a specific set of QoS parameters should be specified when creating a new service flow (like it is shown in Table2). Uplink flows however are associated, in addition to a scheduling service, to one of these request/grant scheduling types: unsolicited grant service (UGS), real-time polling service (rtPS), extended real-time polling service (ertPS)—introduced by the IEEE 802.16e-2005 standard [2], non-real-time polling service (nrtPS), and best effort (BE). Each scheduling service is designed to meet the QoS requirements of a specific applications category. More details about each request/grant scheduling type are given in the next paragraphs.

- **UGS** is designed to support real-time applications that generate fixed-size data packets at periodic intervals, such as T1/E1 and voice over IP (VoIP) without voice activity detection (VAD). The mandatory service flow QoS parameters for UGS service are listed in Table 2. This table summarizes, according to the scheduling service type, the QoS parameters that must be specified when establishing a new service flow. 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.
- **rtPS** is designed to support real-time applications that generate variable-size data packets at periodic intervals, such as moving pictures expert group (MPEG) video. 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 of requesting bandwidth. This corresponds to the polling bandwidth-request mechanism. This mechanism exists in three variants: unicast polling, multicast polling and broadcast polling. Only unicast polling can be used for rtPS connections.
- **Extended rtPS** is a new scheduling service introduced by the IEEE 802.16e-2005 standard [2] to support real-time service flows that generate variable size data packets on a periodic basis, such as Voice over IP services with silence suppression. Like in UGS, the BS shall provide unicast grants in an unsolicited manner which saves the latency of a bandwidth request. However, unlike UGS allocations that are fixed in size, ertPS allocations are dynamic like in rtPS. By default, the size of allocations corresponds to the current value of Maximum Sustained Traffic Rate at the connection. The SS however may request changing the size of the UL allocation.
- **nrtPS** is designed to support delay-tolerant applications such as FTP for which a minimum amount of bandwidth is required. The polling mechanism can be applied to nrtPS connections. However, unlike for rtPS, nrtPS connections are not necessarily polled individually—multicast and broadcast polling are possible—and the polling must be regular not necessarily periodic.
- **BE** is designed for applications that do not have any specific bandwidth or delay requirement, such as HTTP and SMTP. For BE connections, all forms of polling are allowed in order to request bandwidth.

The QoS parameters that must be specified when establishing a new service flow are listed in Table 2. The value of the Request/Transmission (Rx/Tx) Policy parameter offers the possibility to specify options for PDU formation. It

Traffic/Applications Characteristics	real-time, fixed-rate data, Fixed/Variable length PDUs		real-time, variable bit rates, requiring guaranteed data rate and delay		real-time, variable bit rates, requiring guaranteed data rate and delay		requiring guaranteed data rate, insensitive to delays		No rate or delay requirement	
Downlink (DL)/ Uplink (UL)	DL	UL	DL	UL	DL	UL	DL	UL	DL	UL
Maximum Sustained Traffic Rate	√	√	√	√	√	√	√	√	√	√
Minimum Reserved Traffic Rate	√	√	√	√	√	√	√	√	—	—
Maximum Latency	√	√	√	√	√	√	—	—	—	—
Tolerated Jitter	√	√	√	√	—	—	—	—	—	—
Request/Transmission Policy	√	√	√	√	√	√	√	√	√	√
Traffic Priority	—	—	√	√	√	√	√	√	—	—
Request/Grant Scheduling Type	—	√ (UGS)	—	√ (ertPS)	—	√ (rtPS)	—	√ (nrtPS)	—	√ (BE)
Unsolicited Grant Interval	—	√	—	√	—	—	—	—	—	—
Unsolicited Polling Interval	—	—	—	—	—	√	—	—	—	—
SDU Size (If fixed length SDU)	√	√	—	—	—	—	—	—	—	—
Example of application	T1/E1, VoIP without VAD		VoIP with VAD		MPEG video		FTP		HTTP, SMTP	

TABLE 2: Mandatory QoS parameters for each scheduling service

might define for instance a restriction on packing and fragmentation capabilities as well as attributes affecting the bandwidth request types.

2.3 PMP mode

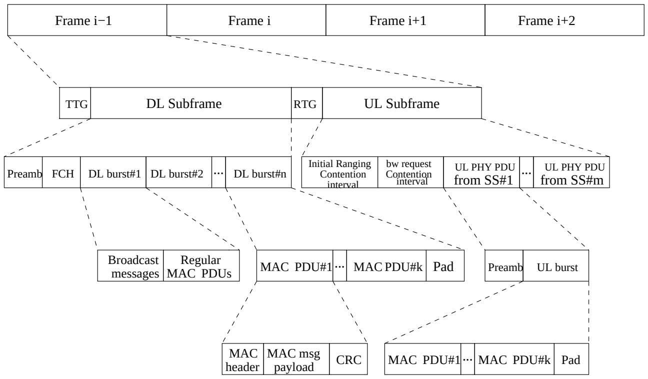


Fig. 2: PMP OFDM Frame Structure with TDD [1]

The basic topology of an IEEE 802.16-based network consists of one BS and one or more SSs.

In PMP, the SSs within a given antenna sector receive the same transmission broadcast by the BS—corresponding in general to the ISP—on the downlink channel (DL). Each SS is required to capture and process only the traffic addressed to itself (or to a broadcast or multicast group it is a member of). On the uplink channel (UL) however, the time division multiple access (TDMA) scheme is applied. Downlink and uplink channels are duplexed using one of the two following techniques: frequency division duplexing (FDD) and time division duplexing (TDD). The main difference between the two duplex modes is that in FDD, the DL and UL use different frequencies, while in TDD both channels use the same frequency in different time intervals. In this paper, we focus on 802.16 systems operating in TDD mode. Figure 2 shows an example of the OFDM frame structure in TDD mode.

In the IEEE 802.16, the channel consists of fixed-length frames, as shown in Figure 2. Each frame is divided into DL and UL subframes. [1] specifies that, when using TDD, the UL subframe and DL subframe durations shall vary within the same shared frame. The downlink subframe consists of one single PHY PDU while the uplink subframe consists of two contention intervals followed by multiple PHY PDUs, each transmitted by a different SS. The first contention interval is used for ranging which is the process of adjusting the radio frequency (RF). The second interval may be used by the SSs to request bandwidth since bandwidth is granted to SSs on demand. Two gaps separate the downlink and uplink subframes: transmit/receive transition gap (TTG) and receive/transmit transition gap (RTG). These gaps allow the BS to switch from the transmit to receive mode and vice versa.

The downlink PHY PDU consists of one or more bursts, each transmitted with a specific burst profile. A burst profile is a set of parameters describing the transmission properties (modulation type, forward error correction (FEC) type, etc.) corresponding to an interval usage code (IUC). Each SS is required to adapt the IUC in use (a DIUC for the downlink and an UIUC for the uplink) based on measurements on the physical layer. The length of each burst is set by the BS. Indeed, at the beginning of each frame, the BS schedules the uplink and downlink grants (by mechanisms that are outside the scope of the standard [1], [2]) and then broadcasts the downlink frame prefix (DLFP), the DL-MAP and the UL-MAP informing the SSs of its scheduling decisions. The DLFP describes the location and profile of the first downlink bursts (at most four). SSs using the same DIUC are advertised as a single burst. The DL-MAP, when sent, describes the location and profile of the other downlink bursts—if they exist. However, the IEEE 802.16 standard specifies that, at least one full DL-MAP must be broadcast within the Lost DL-MAP Interval even if there are less than five bursts. The UL-MAP should be transmitted in each frame. It contains information elements (IE) that indicate the types and the boundaries of the uplink allocations directed to the SSs. The profile of each downlink and uplink burst are specified in the downlink channel descriptor (DCD) and uplink channel descriptor (UCD), respectively. The BS broadcasts the DCD and the UCD messages periodically—every DCD/UCD Interval—in order to define the characteristics of the downlink and uplink physical channels. Referring to Figure 2, we note that each burst consists of one or more MAC PDUs. Each MAC PDU begins with a fixed-length MAC header followed by a payload and a cyclic redundancy check (CRC) field. The burst may also contain padding bytes since each burst must consist of an integer number of OFDM symbols. UL bursts begin with a preamble used for PHY synchronization.

2.4 Mesh mode

In the last few years Wireless Mesh Networks (WMNs) have been attracting a huge amount of attention from both, academia and industry. It has emerged as a promising technology for future broadband wireless access [4], [5]. One of the main reasons for this popularity is the inclusion of the mesh mode in many of the IEEE standards, especially the last version of the IEEE 802.16 [1]. The addition of the mesh mode to the IEEE 802.16 standard, not only extends this kind of network area coverage, but also brings a series of other advantages, among them, non-Line-of-Sight (NLOS) capacity, higher network reliability, scaling, throughput and availability [6].

In contrast to the PMP mode, in the Mesh mode, the traffic is not restricted to occur just between the BS and the SSs. In the mesh mode the communication may occur also between SSs, even without the knowledge of the BS. At the extreme case, even the existence of a BS in this kind of network is optional. Actually, the role of BSs is different in WiMAX PMP and mesh mode. Within a Mesh network, BS, or Mesh BS, is the term used to designate the station that has a direct connection to backhaul services outside the Mesh network [1]. In other words, BS is the station acting as a gateway between the mesh network and the rest of the world.

The scheduling problem for WMNs, just considering throughput and ignoring other QoS parameters, is already proved to be NP-hard [7], [8]. This means that if the number of nodes, or links, in the WMN increases it becomes computationally nonviable to find an optimal solution for the scheduling. So in this context suboptimal scheduling solutions, with lower complexity, are acceptable and even desired.

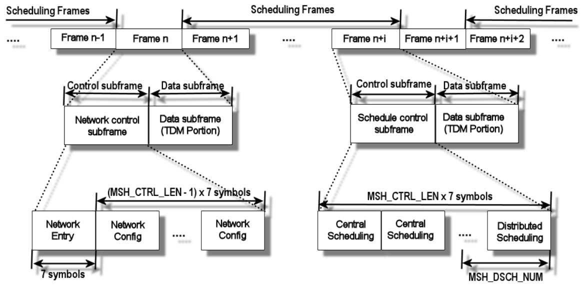


Fig. 3: Mesh frame structure [9]

The WiMAX mesh mode, introduced in the standard by the IEEE 802.16a amendment [10], supports two different physical layers: WirelessMAN-OFDM^{TM}, operating in a licensed band, and WirelessHUMAN^{TM}, operating in an unlicensed band. Both of them use 256 point FFT OFDM TDMA/TDM for channel access and operate in a frequency band below 11 GHz. Despite the fact that some researchers defend the use of Frequency Division Duplexing (FDD) for the upcoming standard of IEEE 802.16j [11] relay networks, the current standard version [1] allows only Time Division Duplex (TDD) in mesh mode. This means that the uplink and downlink transmissions share the same frequencies and must occur at different time slots.

The Mesh frame, depicted in Figure 3, is divided into control and data subframes. The control subframe, presenting 0 to 15 transmission opportunities, has two basic functions, the first one is the creation and maintenance of the structure of the network. The second function is to coordinate the scheduling of data transfers between stations. The length of the Control sub-frame, L_{CS} expressed as number of OFDM symbols, where MSH CTRL LEN is the number of transmission opportunities is defined as:

$$L_{CS} = MSH\_CRTL\_LEN\ x\ 7$$

The data subframe, consisting of up to 256 minislots, carries the MAC PDUs transmitted by different users. The MAC PDU consists of a generic MAC header, a Mesh subheader and optional data. The standard supports both centralized and distributed scheduling, and allows the co-existence of both at the same time in the network. The number of distributed scheduling messages is denoted as MSH_DSCH_NUM.

There are two types of control sub-frame: schedule control sub-frame and network control sub-frame. The network control sub-frame provides the basic functionality for network entry and topology management. The schedule control sub-frame controls the nodes transmissions. The scheduling is done by negotiating minislots ranges for the traffic demand of each link. All the communications are in terms of links established between nodes. All data transmissions, between two nodes, are done through one link and the QoS is provisioned over links on a message by message basis. Upper layer protocols make the traffic classification and flow regulation for new nodes.

In the mesh mode there is no clear differentiation between downlink and uplink subframes. Each station is free to communicate to any other node in the network, so the uplink and down link notion have no meaning in this context. However, in the typical expected case, there will be some nodes providing a backhaul connection to the network. In this case, these nodes, for the centralized schedule, will play nearly the same role the BS plays in the PMP mode, so centralized scheduling has the notion of uplink and downlink traffic. Table 3 presents the messages used for CAC and scheduling in the WiMAX mesh mode.

IEEE 801.16 mesh mode networks present three different scheduling mechanisms, Coordinated centralized scheduling, Coordinated distributed scheduling and Uncoordinated distributed scheduling. These three scheduling policies can be either used alone or together in the same network. Some works like [6], [12] suggested that centralized schedule should be used for external traffic and distributed schedule should be used for intra network traffic. This came from the fact that the centralized schedule trusts in a mesh BS, that in last instance, is a backhaul acting as gateway between the internal and external network traffic.

Message type	Name	Description	Connection
39	MSH-NCFG	Mesh Network Configuration	Broadcast
40	MSH-NENT	Mesh Network Entry	Basic
41	MSH-DSCH	Mesh Network Distributed Schedule	Broadcast
42	MSH-CSCH	Mesh Network Centralized Schedule	Broadcast
43	MSH-CSCF	Mesh Network Centralized Schedule Configuration	Broadcast

TABLE 3: Mesh MAC Management Messages

2.4.1 Centralized scheduling

For the Centralized Scheduling, the mesh BS schedules all SSs, and even BS, transmissions. The resource request and the BS assignments are both transmitted during the control portion of the frame. The centralized scheduling coordinates the transmissions and ensures they are all collision-free. Since the BS has the knowledge of the entire network, it is expected to be closer to the optimal usage of the spectrum than the distributed forms.

The MSH-CSCH message has two variants, MSH-CSCH Request and MSH-CSCH grant. With the MSH-CSCH Request each node estimates and reports the level of its own upstream and downstream traffic demand to its parent, it also computes the demands reported by the node children. With the MSH-CSCH Grant the BS propagates down, through the tree, the levels of flows and grants to each node in the network. Figure 4 shows an example of message flow for the centralized schedule.

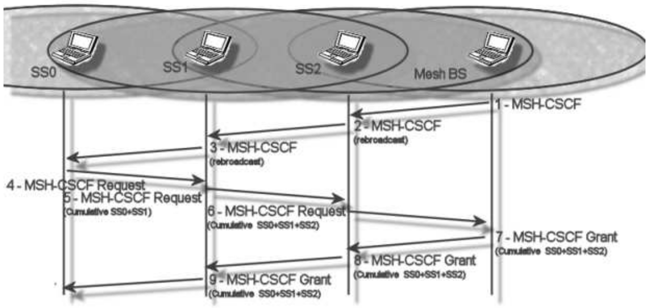


Fig. 4: A message flow example for the centralized scheme

All MSH-CSCH Grant messages contain information about all network grants, since all nodes need complete information for the schedule computation. Upon receiving any message in the current scheduling sequence, and assuming that the node has up-to-date scheduling configuration information, a node will be able to compute locally all the schedule of transmissions, including its own. Besides the BS, a node should never transmit any downstream centralized scheduling packet in a centralized scheduling sequence in which it has not yet received a MSH-CSCH message from a parent. Also, a node should not send any centralized scheduling packets if its MSH-CSCF information is outdated.

In terms of eligibility to send and receive MSH-CSCH messages, all nodes are eligible to retransmit the grant schedule, except those that have no children. For transmitting MSH-CSCH grant messages, all nodes with children are eligible. For transmitting MSH-CSCH request messages, all nodes, except the mesh BS are eligible.

2.4.2 Distributed scheduling

In both distributed scheduling mechanisms, coordinated and uncoordinated, all the stations in the two hop neighborhood must have their transmissions coordinated to avoid collision. The coordinated distributed scheduling uses the control part of the frame to transmit its own traffic schedule. The distributed schedule may work with the centralized schedule, at the same time, but does not rely neither on its operation nor in the existence of a mesh BS.

The uncoordinated distributed scheduling is a simpler version of the distributed scheduler and may be used for fast ad-hoc setup of schedules in a hop-by-hop basis. The uncoordinated schedule is basically an agreement between two nodes and should not cause collision with the data and control traffic scheduled by the coordinated schedules.

Both coordinated and uncoordinated distributed scheduling employ a three-way handshake to setup the connection. The first message in the three-way handshake is a MSH-DSCH Request, the transmission is scheduled using a random-access algorithm among the “idle” slots of the current schedule. If the attempt was unsuccessful a random backoff is used to avoid new collisions. Figure 5 shows schematically the messages in the three way handshake.

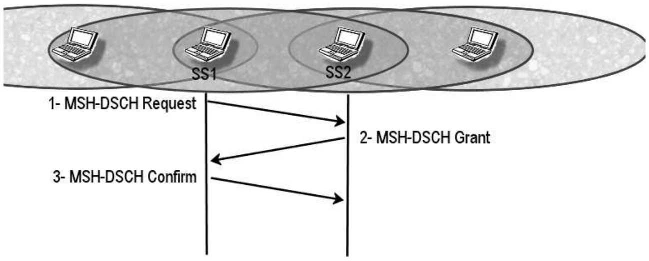


Fig. 5: Distributed Scheduling Three Way Hand Shake

The MSH-DSCH Grant can be issued by any neighbor that listen to the MSH-DSCH Request. The grant message contains the list with the subset of the resources awarded. The first node awarded with the grant may start its grant transmission in the immediately following base-channel idle minislot. More than one granter may respond to the request. The requesting node sends the same received MSH-DSCH Grant message in confirmation. Doing this the requester’s neighbors become aware of the grant awarded. The grant confirmation is sent in the first available minislots following the minislots reserved for the grant opportunity of the last potential granter.

2.4.3 Network configuration

Two more messages, responsible for creating and maintaining the network configuration, may be transmitted in the network control subframe: Mesh Network Configuration (MSH-NCFG) and Mesh Network Entry (MSH-NENT).

A new node that wishes to join the mesh network waits until receiving a MSH-NCFG message. When the new node receives this message it is able to establish the synchronization with the already established mesh network. More precisely, to decide which node is the best sponsor the new node may wait for more than one MSH-NCFG message to arrive. When the sponsor node is chosen, the new node uses it to send a MSH-NENT message to the BS with its registration information. The sponsor node then establishes a quick schedule, through the uncoordinated scheduler process, and communicates this schedule to the new node. The new node confirms the schedule and sends the required security information. Finally, in the last step, the sponsor node grants the new node access to the network.

3 Design challenges for IEEE 802.16 scheduling and CAC

The objective of this section is to provide a better understanding of the design challenges of a new scheduling and/or CAC solution for IEEE 802.16. A joint view of the two modes, PMP and Mesh, is a great asset to understand the problem as a whole. However, as we have seen in Sections 2.3 and 2.4, the two WiMAX modes are quite different. Therefore we address both the general and specific constraints.

- **Common constraints**
  - **Channel utilization:** The channel utilization is expressed in percentage of the available capacity and it represents the achieved throughput. It corresponds to the fraction of time used to transmit data packets. In the case of a PMP communication, this parameter is almost equal to the channel capacity. Nevertheless, to maximize the channel utilization, the scheduler should minimize the overhead by optimizing the bandwidth-request strategy and taking advantage of the concatenation, packing, and fragmentation mechanisms, proposed by the standard.
  - **QoS requirements guarantee:** The scheduler should satisfy the QoS requirements of the different types of service specified by the standard. Hence it has to monitor, for each connection, the required QoS parameters, presented in Table 2, and check if they are in line with what has been negotiated.
  - **Graceful service degradation:** It is an interesting characteristic for CAC and scheduling algorithms, when accepting new connections, to degrade the service of the ongoing over provisioned connections as gracefully as possible. Since radio resources are limited the use of this kind of strategy would compensate lagging flows and ensure fairness in radio resources management (RRM).
  - **Fairness:** One of the most challenging problems for RRM is to find a compromise between increasing the channel utilization— by serving flows with good channel conditions— and being fair to different flows. To estimate this parameter Jain’s fairness index [13] might be used:

    $$F_J = \frac{\left(\sum_{i=1}^{m} x_i\right)^2}{m.\sum_{i=1}^{m} x_i^2}$$

    Where m is the total number of flows and x_i is the proportion of received packets of flow i during run time. F_J is equal to 1 when all flows equally share the bandwidth, and equal to $1/m$ when a flow monopolizes the network.
  - **Implementation complexity:** Scheduling and CAC algorithms deal with many different constraints. Nevertheless, because they address— among others—real time flows, they need to be

fast and should not have a prohibitive implementation complexity.

- **Scheduling delay:** This parameter depends mainly on the bandwidth request strategy adopted by the scheduler since it corresponds to the time interval between when the bandwidth is requested and when it is allocated. The scheduling algorithm should try to minimize this time interval in order to meet the time constraints of delay-sensitive applications.
- **Scalability:** Scalability is the capability of the scheduling algorithm to handle growing number of flows, or nodes, in a graceful manner. Scalability is also important in the context of mobile WiMAX networks for mobility management.
- **Energy consumption:** Increasing the autonomy of mobile nodes is a common concern in wireless networks. Therefore the scheduler should adopt optimized power-saving strategies. This could consist, for instance, in keeping the SS awake only when it needs to send or receive data.
- **Bandwidth-request strategy:** Because the standard gives a choice among several bandwidth request and grant techniques, it is important for each scheduling solution to define its own bandwidth request strategy.
- **MAC-PHY cross-layer design:** This constraint consists mainly in considering the adaptive modulation and coding (AMC) capability defined by the standard. Indeed, it is important, when allocating resources at the MAC level, to take into account the burst profile in use at the PHY level.
- **SS scheduler:** The scheduling issue concerns not only the BS but also the SS. Indeed, since bandwidth allocation is made on a per-SS basis, a scheduler should be integrated in the MAC structure of an SS to share resources among uplink flows.

- **Mesh mode specific constraints**
  - **Spectral efficiency/Frequency reuse:** Reusing the same radio frequency in a different area for two or more different transmissions increases the network capacity and then the channel utilization. Nevertheless, interference should be avoided as much as possible.
  - **Routing:** Mesh mode networks imply the transfer of messages between peer nodes. Unlike PMP networks, at Mesh networks the Mesh BS does not necessarily take part in all communications. Thus, routing is a fundamental process to enable the communication among nodes inside the same mesh network.
  - **Topology construction:** To enable the routing inside the mesh network, nodes should be aware of the network topology. In this way, they would be able to build a consistent view of the network.
- **PMP mode specific constraints**
  - **Dynamic DL/UL assignment in TDD mode:** As far as PMP is concerned, when considering the TDD mode, the amount of bandwidth allocated for uplink and downlink should be dynamically adapted to the traffic on each direction.

Table 4 summarizes the importance of each constraint according to the mode in use. Three levels of importance have been defined: (1) *important* which refers to all the constraints that must be taken into account by a scheduler, (2) *desirable* to describe the optional features that could improve the scheduling procedure, and (3) *not applicable* when it is a constraint that is specific to another mode and does not apply to the considered one. The topology construction constraint, for instance, does not apply to the PMP mode since all the SSs communicate only with the BS in a point-to-multipoint fashion. To illustrate the difference between an *important* and a *desirable* constraint, in PMP mode, we can consider the difference between the graceful service degradation and the QoS requirements guarantee (cf. Table 4). Indeed, guaranteeing the mandatory QoS parameters listed in Table 2 for each SF, like insuring the minimum reserved traffic rate for nrtPS service flows, is an important issue and one of the main features that should be supported by a PMP scheduler. However, applying a graceful service degradation is just a desirable property that would decrease the blocking and/or the dropping rate.

Metric/Constraint		PMP	Mesh	
			Centralized scheduling	Distributed scheduling
Channel utilization		**	**	**
QoS requirements guarantee		**	**	**
Graceful service degradation		*	*	*
Fairness	among nodes	*	*	**
	among SFs	*	*	**
Implementation complexity		*	*	**
Scheduling delay		*	**	**
Scalability	nodes	*	**	**
	data traffic	*	*	*
	mobility	*	**	**
Energy consumption		*	**	**
Spectral reuse		N/A	**	**
Routing		N/A	**	**
Topology construction		N/A	**	**
Schedulers	BS and SS schedulers	**	**	N/A
	Only SS Schedulers	N/A	N/A	**
Bandwidth-request strategy		**	**	**
AMC (MAC-PHY cross-layer)		**	*	*
TDD: DL and UL dynamic assignments		*	N/A	N/A

** important   * desirable   N/A not applicable

TABLE 4: 802.16 PMP and Mesh modes: Scheduling Challenges

For the mesh mode, the two different scheduling modes—distributed and centralized—have some differences. For example, for the distributed scheme, paying attention to the algorithm complexity is more important than it is for the centralized one. Normally, the distributed scheduling has more stringent time constraints. So, if the scheduling process takes too long to deliver results, these results may even become useless. For the centralized scheduling, however, this complexity is not that important. First, because the mesh BS has a complete view of the concerned network. The scheduling is, in this way, easier to implement and the algorithms are expected to be simpler. Second, for the centralized mode, the process already considers a substantial time interval to spread the scheduling information among the nodes. This delay is expected to be bigger than the time the scheduler would take in order to run. Another point to observe is that the difference between BS and SS schedulers does not apply (N/A) to the mesh distributed scheduler, since there is no central BS for this scheme. The same line of thinking holds for the opposite case: the use of only one scheduling procedure does not apply to the mesh centralized scheduling. For this scheme, it is required to have different functions for the scheduling at mesh BSs and for the ones at mesh SSs.

4 PMP scheduling and CAC

As shown in Figure 6, the approaches adopted in literature when designing a scheduling solution can be divided into three main categories. (1) The first one is a queuing-derived strategy where the authors focus on the queuing aspect of the scheduling problem and try to find the appropriate queuing discipline that meet the QoS requirements of the service classes supported by the IEEE 802.16 standard [1], [2]. In this first category, two kinds of structures are proposed: either simple structures consisting in general in one queuing discipline applied for all the scheduling services [14], [15], [16] or hierarchical structures consisting in two or multiple layers reflecting different levels of scheduling like in [17], [18], [19], [20], [21], [22], [23], [24], [25]. (2) In the second category, the scheduling problem is formulated as an optimization problem whose objective is to maximize the system performance subject to constraints reflecting in general the QoS requirements of different service classes [26], [27], [28], [29], [30], [31], [32], [33], [34]. (3) The third category of scheduling mechanisms that can be found in literature is the cross-layer strategy. The scheduling schemes adopting this strategy are usually based on a cross-layer architecture. The objective of this architecture is to optimize the communication between two [35], [36], [37], [38], [39] or three different layers [40], [41] and thus improve the system performance. As we will see in Section 4.3, these schemes could be further classified based on the layers involved in the cross-layer design.

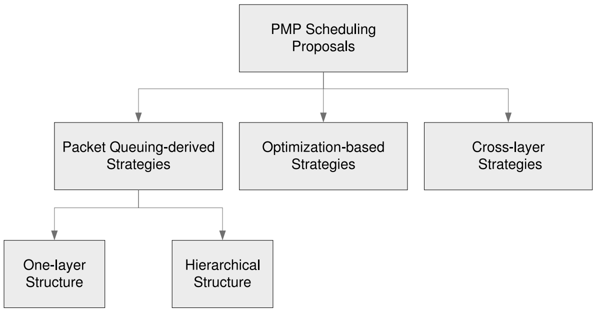


Fig. 6: Classification of the scheduling strategies of IEEE 802.16 PMP mode

4.1 PMP scheduling proposals: packet queuing-derived strategies

4.1.1 One-layer scheduling structures

Sayenko *et al* [16] consider that because there is not much time to do the scheduling decision, a simple one-level scheduling mechanism is much better than a hierarchical one. Therefore they propose a scheduling solution based on the round-robin (RR) approach. They argue that there is no need to use disciplines like fair queuing (FQ) since the weights in such algorithms are floating numbers while the number of allocated slots, in 802.16 networks, should have an integer value. They also try to outline the difference between the weighted round-robin (WRR) discipline and the 802.16 environment. They insist on the fact that WRR may lead to a waste of resources because of its work-conserving behavior that does not fit the fixed-size frame of 802.16 that implies a non-work conserving behavior.
Based on the above considerations, the authors proposed in [16] a scheduling solution that consists in four main steps:

- Allocating for each connection the minimum number of slots that ensure the minimum reserved traffic rate with respect to the used modulation and coding scheme,
- Distributing the free slots between rtPS and nrtPS connections and then assigning the remaining to BE connections,
- Ordering the slots in such a manner the delay and jitter values are decreased.
- Estimating the overhead for UGS, ertPS, and in some cases nrtPS connections. This is not possible for rtPS and BE connections since it is more likely that the SDU size varies.

Note that [16] is one of the rare research works in which the overhead resulting from the scheduling decision, and packing or fragmentation capability is taken into account. However it is also worth mentioning that the authors consider a grant per connection (GPC[2]) mechanism and when ordering slots, they apply an interleaved scheme that is in contradiction with the frame structure specified by the standard.

In [14], [15], Cicconetti *et al* conjecture that the class of latency-rate *(LR)* scheduling algorithms is particularly suited for implementing schedulers in 802.16 MAC since the basic QoS parameter required by a given connection is the minimum reserved traffic rate. Indeed the behavior of such algorithms is determined by two parameters which are the latency and the allocated rate [42]. From this class, the authors have chosen the deficit round robin (DRR) algorithm. DRR is simple to implement ($O(1)$ complexity if specific allocation constraints are met) and provides, according to [14], [15], fair queuing in presence of variable length packets[3]. It nevertheless requires a minimum rate to be reserved for each packet flow; so even BE connections should be guaranteed a minimum rate. Also since this algorithm assumes that the size of the head-of-line packet is known, it can not be applied by the BS to schedule uplink transmissions. For this reason the authors have made the choice of implementing it as SS scheduler and as a downlink scheduler at the BS, since both BS and SS know the head-of-line packet sizes of their respective queues. To schedule uplink transmissions at the BS—based on backlog estimation—they have selected the WRR algorithm which belongs, like DRR, to the class of *LR* algorithms.
The simulation study carried by Cicconetti *et al* [14] demonstrated that the performance of 802.16 systems, in terms of throughput and delay, depends on several metrics such as frame duration, the mechanisms used to request UL bandwidth, the offered load partitioning—how traffic is distributed among SSs, the connections within each SS, and the traffic sources within each connection.

4.1.2 Hierarchical scheduling structures

Wongthavarawat *et al.* [24], [25] are the first authors who introduced a hierarchical structure of bandwidth allocation for 802.16 systems. This hierarchical scheduling structure, shown in Figure 7, combines strict priority policy, among the service classes, and an appropriate queuing management discipline for each class: earliest deadline first (EDF) for rtPS, and weighted fair queuing (WFQ) for nrtPS. Fixed time duration is allocated to UGS connections and remaining bandwidth is equally shared among BE connections. In order to avoid starvation for lower priority connections, a policing module is included in each SS. It forces each connection to respect the traffic contract when demanding bandwidth. The proposed scheduling algorithm takes into account the queue size information and the service actually received by each connection. It also considers the arrival time and the deadline requirements of rtPS connections. However, the authors focused only on UL scheduling. They considered TDD mode and assumed that the durations of UL and DL subframes are dynamically determined by the BS but they did not specify how these proportions are fixed. The QoS architecture they proposed in [24] includes a token-bucket based admission control module that will be described in Section 4.4.

Most of the works that we will present in this section are "quite similar" to the scheduling model introduced by Wongthavarawat *et al.* in [24], [25]. Nevertheless, since more or less features are supported by each scheme, we have grouped them based on their main common contribution.

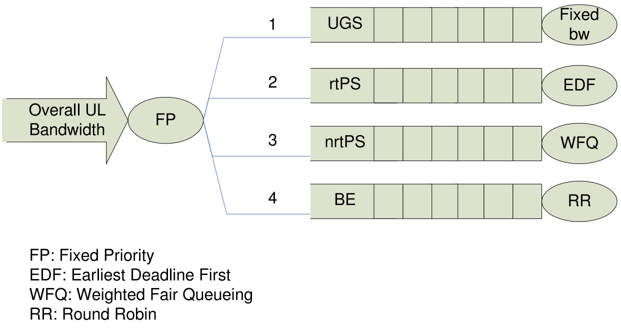


Fig. 7: Hierarchical structure for bandwidth allocation in WiMAX PMP mode [24], [25]

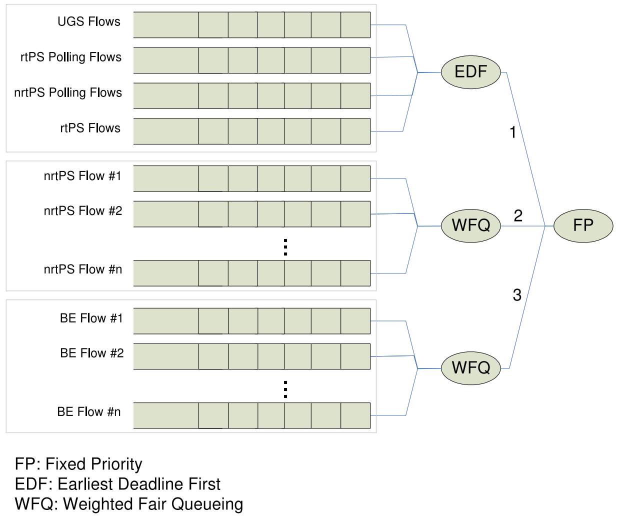
Delay-aware scheduling : In [23], Sun *et al.* proposed a two-layers scheduling structure composed of a BS scheduler and an SS scheduler. At BS scheduler, priority is given to schedule data grants for UGS connections and bandwidth request opportunities for rtPS and nrtPS connections. The amount of bandwidth allocated in this phase is reserved during connections setup. Data grants for rtPS, nrtPS are then scheduled taking into account the information contained into bandwidth request messages and their minimum requirements. Finally, the residual bandwidth, if any, is redistributed in proportion to pre-assigned connections weights. The proposed SS scheduler considers a fixed priority scheme—1, 2, 3 and 4 for BE, nrtPS, rtPS and UGS scheduling service, respectively. Bandwidth is firstly guaranteed for UGS connections. rtPS packets are then scheduled based on their respective deadline stamps—corresponding to their $arrival\_time + tolerated\_delay$. Each nrtPS packet is associated with a virtual time calculated to guarantee the minimum reserved bandwidth and hence maintain an acceptable throughput. A simple first-in-first-out (FIFO) mechanism is applied for BE queues.

Other scheduling schemes focusing on delay requirements were proposed in literature. In [19] for instance, three schedulers were combined to meet the QoS requirements of different classes (cf. Figure 8). Time sensitive traffic streams—namely UGS flows, rtPS flows and (n)rtPS polling flows—are served by Scheduler 1 that applies EDF algorithm. Minimum bandwidth reserving flows (nrtPS flows)

[2] This approach consists in allocating the bandwidth on a per connection basis. In contrast with GPC, the grant per subscriber station (GPSS) refers to the allocation of bandwidth per SS. Both concepts should have been disused since the publication of the IEEE 802.16a-2003 Standard [10]. Indeed, it is clearly specified in [1], [2] that bandwidth is requested on a per connection basis while grants are aggregated and addressed as a whole for each SS.

[3] This is in contradiction to what has been stated by Fattah and Leung in [43] where they qualify the fairness of DRR algorithm as "poor".

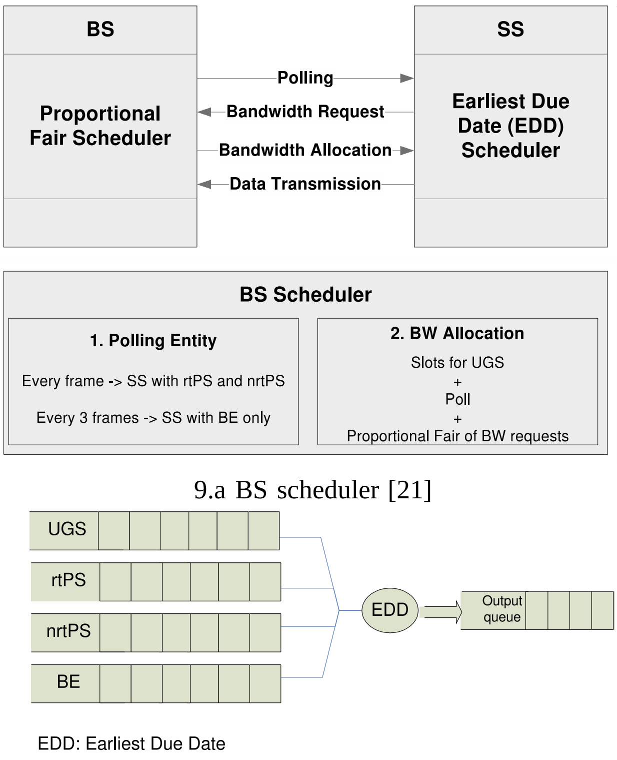
are scheduled by Scheduler 2 using WFQ. The weights correspond to the proportion of requested bandwidth. WFQ algorithm is also applied by Scheduler 3 to serve BE traffics; weights nevertheless correspond in that case to traffic priorities specified by each BE connection. Other components of the proposed architecture are then used to plan contention and reserved transmission opportunities according to the bandwidth availability and to the priorities assigned to each scheduler—the highest priority is assigned to Scheduler 1.



FP: Fixed Priority
EDF: Earliest Deadline First
WFQ: Weighted Fair Queueing

Fig. 8: 3 schedulers proposal for WiMAX PMP mode [19]

In [21], a multimedia supported uplink scheduler is proposed by Perumalraja *et al.*. It includes a proportional fair (PF) BS scheduler and an earliest due date (EDD) SS scheduler. The BS scheduler (Figure 9.a) allocates resources first for the UGS service and then to poll SSs having at least one non-UGS connection: one slot is allocated in each frame for each SS having rtPS or nrtPS connections and one slot every three frames is allocated for SSs having only BE service connections. Finally, remaining OFDMA resources are proportionally allocated for SSs based on the received bandwidth requests. As can be seen from Figure 9.b, the EDD SS scheduler serves packets from the four traffic queues (UGS, rtPS, nrtPS and BE) in the order of the deadline assigned to each packet regardless of their scheduling service type.


BS
Proportional Fair Scheduler
SS
Earliest Due Date (EDD) Scheduler
Polling
Bandwidth Request
Bandwidth Allocation
Data Transmission
BS Scheduler
1. Polling Entity
Every frame -> SS with rtPS and nrtPS
Every 3 frames -> SS with BE only
2. BW Allocation
Slots for UGS
+
Poll
+
Proportional Fair of BW requests

9.a BS scheduler [21]

9.b EDD SS scheduler [21]

Fig. 9: Multimedia supported uplink scheduler [21]

Asymmetric DL/UL scheduling: [18] is one of the rare research works that have proposed a scheduling algorithm considering simultaneously uplink and downlink bandwidth allocation in TDD mode. In first layer scheduling—of the two-layer hierarchical scheduling structure proposed in this work—Chen *et al* [18] have suggested the use of deficit fair priority queuing (DFPQ) algorithm instead of strict priority in order to avoid starvation for low priority classes. This first layer scheduling is based on two policies. The first one is a transmission direction-based priority where they chose to attribute to DL a higher priority than UL. The second policy is a service class-based priority applying the following scheme: rtPS>nrtPS>BE. As can be seen from Figure 12, the authors have combined these two policies using a strict priority scheme which assigns strict priority from highest to lowest to: $DL_{rtPS}, UL_{rtPS}, DL_{nrtPS}, UL_{nrtPS}, DL_{BE}, \text{and} UL_{BE}$. For DL and UL UGS connections, they have chosen to apply a fixed bandwidth allocation strategy. In second layer scheduling, three different algorithms were assigned to the other classes of services: EDF for rtPS, WFQ for nrtPS and RR for BE. nrtPS connections are scheduled based on weights corresponding to the ratio between the nrtPS connection minimum reserved traffic rate and the sum of the minimum reserved traffic rates of all nrtPS connections. A basic admission control algorithm is also proposed in this work. It accepts the connections for which the minimum reserved traffic rate does not exceed the available channel capacity; all BE connections are nevertheless accepted.

In order to take advantage of the DL/UL map of the 802.16d standard [1], Ma *et al.* propose in [20] a three-tier scheduling framework in which DL and UL respective loads could be unbalanced. Unlike in [18] however, the ratio of DL subframe with respect to the frame size is computed at the beginning of each frame. Indeed, a pre-scale dynamic resource reservation (PDRR) is used to allocate dynamically the overall frame bandwidth to DL and UL subframes with respect to a pre-scaled bound. The ratio of each subframe to the entire frame is computed based on the queues lengths and on the sizes of the bandwidth requests.

Scheduling proposal	**Layer/Phase**		**DL**	**UL**	**UGS**	**rtPS**	**nrtPS**	**BE**
[24], [25]	1^{st} layer				Fixed Priority			
	2^{nd} layer				Fixed Bandwidth	EDF	WFQ	Equally distributed
[23]	BS Scheduler	1^{st} phase		●	Fixed Bandwidth	Grant Bandwidth Request Opportunities		___
		2^{nd} phase			___	Guarantee the Minimum Reserved Rate		___
		3^{rd} phase			___	WFQ to distribute residual bandwidth		
	SS Scheduler			●	Fixed Priority			
					Fixed bandwidth	EDF	EDF (Virtual Time)	FIFO
[21]	BS Scheduler	1^{st} phase		●	Fixed Bandwidth	Unicast Polling		
		2^{nd} phase			___	Proportional Fair based on bandwidth Requests		
	SS Scheduler				EDD			
[22]	1^{st} layer		●		Fixed Priority			
	2^{nd} layer		●		Fixed Bandwidth	WRR		RR
[18]	1^{st} layer		●		DFPQ			
	2^{nd} layer		●		Fixed Bandwidth	EDF	WFQ	RR
[20]	Tier 1 (at BS)			●	Fixed Bandwidth	PQLW + MMFS among SSs		
	Tier 2 (at SS)			●	Fixed Bandwidth	SCFQ		WRR
	Tier 3 (per traffic flow)				___	EDF		SPLF
[19]	Scheduler 1				EDF (UGS + rtPS + Polling rtPS and nrtPS)		___	___
	Scheduler 2				___	___	WFQ (based on bandwidth requests)	___
	Scheduler 3				___	___	___	WFQ (based on traffic priority)

TABLE 5: WiMAX PMP mode hierarchical scheduling structures

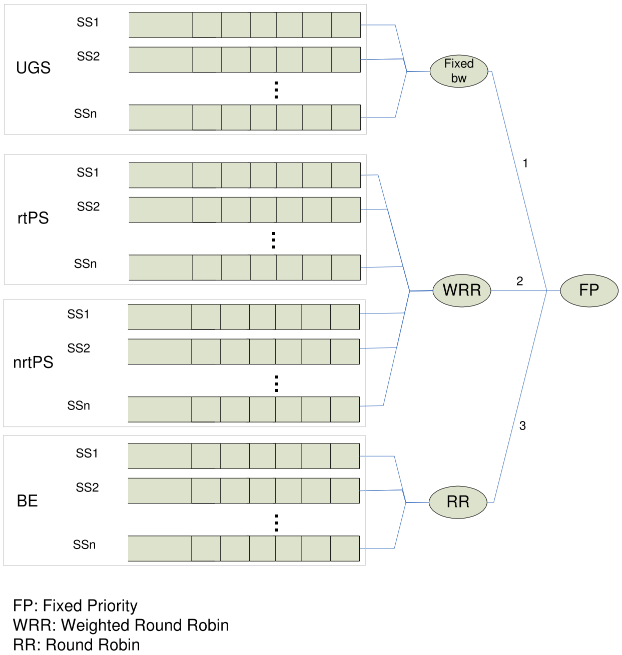


Fig. 10: Scheduler model for WiMAX PMP mode [22]

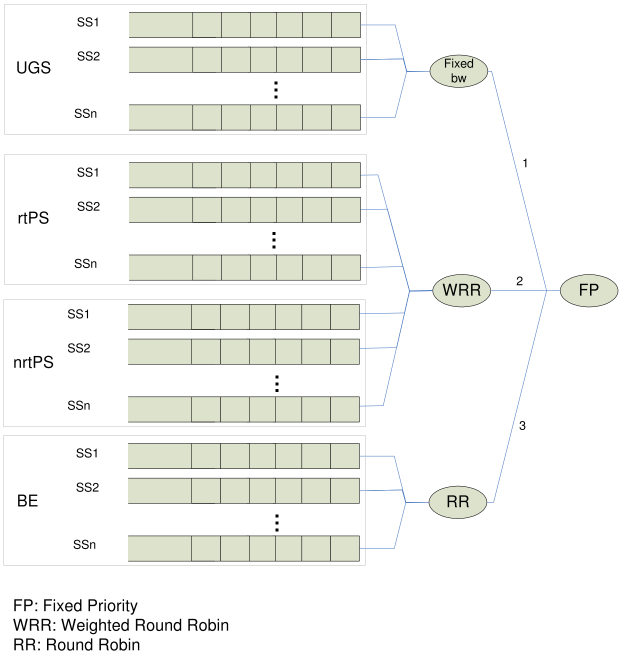


Fig. 11: Scheduler model for WiMAX PMP mode [22]

Packet-based scheduling: use of packing, fragmentation, PHS and AMC: Fragmentation, packing and PHS capabilities as well as their impact on the scheduling performance were considered in the packet-based scheduling strategy proposed in [22] by Settembre *et al.*. As can be seen

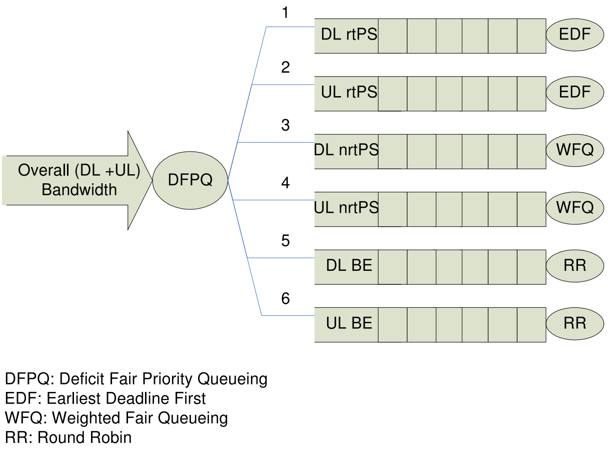


Fig. 12: Hierarchical structure of bandwidth allocation for WiMAX PMP mode [18]

from Figure 11, the proposed scheduler combines a strict priority policy among the different service categories and a specific queuing management discipline for each class: fixed bandwidth, WRR and RR for UGS, (n)rtPS and BE, respectively. For WRR discipline, weights are determined according to the guaranteed bandwidth.
Adaptive modulation and coding was also addressed in [22]. A preliminary WRR/RR allocation is achieved assuming the use of the most robust burst profile while bandwidth is allocated taking into account the actual burst profile! It is true that this way of proceeding guarantees enough bandwidth for existing flows even in the worst case. However, it might cause an unjustified high blocking rate and a low link utilization when the channel is good. Another shortcoming of [22] is that the admission control algorithm that manages the access of new connection—and based on which the minimum bandwidth requirements are guaranteed—is not described.

Table 5 summarizes the hierarchical scheduling proposals described above. In this table, we show whether DL connections are concerned or not by the proposed scheduling mechanism. Also, the table reflects the different steps of each scheduling process as well as the queuing discipline applied at each considered level of aggregation (per service type, per connection, etc.).

Satisfaction-based scheduling: In [17], an original two-tier scheduling algorithm (2TSA) was proposed to avoid starvation problem and to provide fair allocation of residual bandwidth. UGS connection is not concerned by the "2TSA" algorithm since it is allocated a fixed amount of bandwidth per frame. Each connection is classified into either "unsatisfied", "satisfied", or "over-satisfied" connection and is assigned a weight indicating its shortage or satisfaction degree—depending on its category. The connection is considered as:

- *"unsatisfied"* if the allocated bandwidth is less than its minimum requirement,
- a "*satisfied*" connection if the allocated bandwidth is between its minimum and maximum specified requirements,
- *"over-satisfied"* if it is granted more bandwidth than its maximum need.

The first-tier allocation algorithm is category-based and gives the highest priority to "unsatisfied" connections. For a specific category, the second-tier allocation algorithm is applied to share residual bandwidth based on weights. The flowchart of the proposed 2TSA is shown in Figure 13.

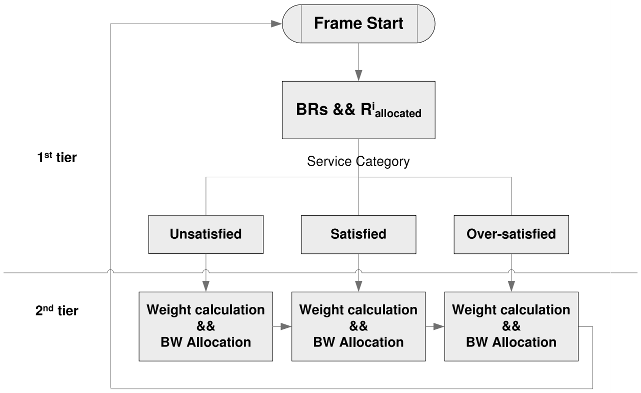

Frame Start
BRs && R^i_allocated
1st tier
Service Category
Unsatisfied
Satisfied
Over-satisfied
2nd tier
Weight calculation && BW Allocation
Weight calculation && BW Allocation
Weight calculation && BW Allocation

Fig. 13: Operation flowchart of 2TSA [17]

Compared to simple-structured scheduling solutions, the hierarchical scheduling mechanisms presented in this section combine in general an inter-service scheduling discipline with a specific queuing mechanism for each service class. Such structures lead to a high computational complexity that may be prohibitive from an implementation point of view and that may not fit the delay constraints of real-time scheduling services.

Service-specific scheduling: Regardless of the proposed scheduling structure, some service-specific scheduling solutions are presented in literature. Lee *et al.* for instance focused in [44] on VoIP services. They argued that both UGS and rtPS have some problems to support the VoIP services and proposed an enhanced scheduling algorithm to solve the mentioned problems. In fact, the fixed-size grants, assigned to UGS connections of voice users, cause a waste of uplink resources during silence periods. Moreover, the bandwidth request mechanism used by rtPS connections leads to MAC overhead and access delay which is not convenient for VoIP applications. Therefore the authors assumed that a voice activity detector (VAD) or silence detector (SD) is used by the SS in the higher layer and proposed an algorithm to be used by the SSs to inform the BS of their voice state transitions. In order to avoid MAC overhead, the proposed algorithm makes use of one of the reserved bits of the conventional generic MAC header of IEEE 802.16 [1] to do that. Simulation results showed that, compared to rtPS, the proposed algorithm decreases the MAC overhead and access delay. Also it can admit more voice users than UGS making more efficient use of uplink resources.

In a more recent work [45], they demonstrated, using the analysis of resource utilization efficiency, that the ertPS service introduced by the IEEE 802.16e standard [2] is more suitable than UGS and rtPS for VoIP services with variable data rate and silence suppression. Indeed they proved that ertPS not only solves the problems of resource wasting, delay, and overhead caused by the use of UGS and rtPS, respectively but also increases the number of voice users that can be supported by the network.

4.2 PMP scheduling proposals: optimization-based strategies

This second category of scheduling strategies consists in formulating the scheduling problem, in 802.16 environment, as an optimization problem aiming at optimizing the allocation of resources to different SSs. Table 6 presents the formulation of some examples of optimization problems proposed in literature.

To get an optimal solution to the optimization problem formulated in [34] (see Table 6), the authors need to use an NP-complete Integer Programming because the number of slots allocated per SS on a given channel should have an integer value. Relaxing this constraint, the authors proposed a second solution based on a linear programming approach that exhibits a complexity of $O(n^3.m^3.N)$ where *n*, *m*, and *N* denote the number of SSs, the number of subchannels and the total number of slots, respectively. However, because it is still a computationally demanding problem, the authors suggested the use of a heuristic algorithm whose computational complexity is $O(n.m.N)$. The authors then proved that the proposed algorithms optimize the overall system performance but may not be fair to different SSs. Therefore they modified them using the proportional-fair concept.
Based on the developed algorithms, they defined a scheduling algorithm for the BS and another one for the SS. The authors agree that considering a joint scheduling for uplink and downlink, at the BS, is more efficient. They nevertheless argue that it is not possible to do that when considering the context of OFDMA/TDD. Therefore they adopted a scheduling mechanism in which downlink and uplink are scheduled separately for all the classes. The priorities are assigned as follows. Allocations are made first for UGS, then rtPS, then for nrtPS just to guarantee the minimum requirements, and finally to satisfy the remaining demands. The choice of one of the proposed algorithms depends on the availability of resources and on the channel conditions. As for the SS, the authors took into account the overall system performance and fairness to different users. They proposed the same sequence followed by the BS but with two different models: a packet model, in which fragmentation is prohibited, for both UGS and rtPS and a byte model—fragmentation is possible—that may be used by nrtPS and BE services.

In [32], Niyato and Hossain considered systems operating in a TDMA/TDD access mode and using WirelessMAN-SC air interface. They defined a utility function that depends on the amount of allocated bandwidth, the average delay, the throughput, and the admission control decision for UGS, rtPS, nrtPS, and BE, respectively. Using these utility functions, they formulated the optimization problem illustrated in Table 6. The authors set a limit of the allocated bandwidth between b_{min} and b_{max} for each connection. They also defined a threshold for each service class since the total available bandwidth is shared using a threshold-based complete partitioning approach. To obtain the optimal threshold setting, an optimization-based scheme is proposed. To solve the proposed optimization problem, Niyato and Hossain suggested two solutions using an optimal approach and an iterative approach, respectively. The first solution has a complexity of $O(2^{M(\triangle b)})$ where M denotes the number of ongoing and incoming connections and $\triangle b = b_{max} - b_{min} + 1$. Since the complexity of the optimal algorithm may be prohibitive from an implementation point of view, the authors proposed an iterative approach based the water-filling mechanism. This solution is more implementation-friendly—its complexity is $O(C)$—while providing similar system performances.
To analyze the connection-level (such as the blocking probability) and packet-level (e.g. transmission rate) performance measures, the authors developed a queuing and an analytical model, respectively. The proposed connection-level model [32], [33] defines the connection blocking probability and the number of ongoing connections via a Continuous Time Markov Chain (CTMC) model. These parameters are then used to formulate an optimization problem (see Table 6) aiming at maximizing the system revenue while maintaining the blocking probability at the target level.

4.3 PMP scheduling proposals: cross-layer strategies

In Sections 4.1 and 4.2, corresponding to the two first scheduling strategies, we have seen some works (such as [16], [22]) that take into account the AMC capability which is also referred to as MAC-PHY cross layer capability. In those works, the cross-layer aspect is only one of the supported features. However, the scheduling schemes we are presenting in this section are totally found on a cross-layer architecture whose objective is to optimize the communication between different layers of the open systems interconnection (OSI) stack. We can further classify these schemes into: (1) MAC-PHY cross-layer schemes, (2) IP-MAC cross-layer schemes, and application-MAC-PHY cross-layer schemes.

4.3.0.1 MAC-PHY cross-layer schemes: The standard provides a link adaptation framework based on which the MCS can be adapted to the channel conditions. However, since no scheduler has been defined by the standard, the way of implementing this capability has been left undefined which explains the need for such MAC-PHY cross-layer design. This need has been explained and justified through preliminary simulation by Noordin *et al.* in [39]

Proposed Solution	**Cost Function (Minimize/Maximize)**	**Constraints (subject to)**
Joint Bandwidth Allocation and admission control [33]	Minimize The average delay	* The average delay meets the delay requirements of rtPS connections. * The transmission rate meets the transmission rate requirements of connections. * The amount of allocated bandwidth for each connection is between b_{min} and b_{max}. * The total amount of allocated bandwidth does not exceed the total available bandwidth.
Queuing theoretic and optimization-based model for resource management [32]	Maximize level of users’ satisfaction <=> Maximize Utility function	* The allocated bandwidth for UGS connections is equal to the required bandwidth * The delay requirements for rtPS connections (depending on the arrival rate, the average SNR and the allocated bandwidth) are met. * The transmission rate requirements of nrtPS connections (depending on the arrival rate, the average SNR and the allocated bandwidth) are met. * BE connections are admitted. * The amount of allocated bandwidth for a given connection is between b_{min} and b_{max}. * The total amount of allocated bandwidth does not exceed the total available bandwidth. * The thresholds (corresponding to the amount of reserved bandwidth for each service class) are respected.
Queuing model for connection-level performance analysis [32]	Maximize The system revenue <=> Maximize the number of ongoing connections	* The connection blocking probabilities* for UGS, rtPS, nrtPS and BE connections do not exceed the target blocking probabilities.
Efficient and fair Scheduling of Uplink and Downlink in OFDMA Networks [34]	Minimize the unsatisfied demands	* The number of granted slots on a given subchannel do not exceed the number of slots of this subchannel * The amount of bandwidth (slots) allocated per connection do not exceed the whole demand of that connection.

* The blocking probabilities as well as the number of ongoing connections are function of the corresponding threshold.

TABLE 6: Optimization approach: cost function and constraints

where they propose a cross-layer optimization architecture for WiMAX systems. The cross-layer optimizer (CLO) presented in this work, acts as an interface between between MAC and PHY layers to obtain and tune the required and optimum parameters.

The authors in [39] believe that there is no need to introduce the application layer in the cross-layer architecture they are proposing since the application requirements are considered through QoS provisioning at MAC level. Therefore, the proposed CLO is reduced to MAC-PHY cross-layer optimization.

A more technical MAC-PHY cross-layer scheme has been proposed by Liu *et al.* in [35], [36]. The authors in [35], [36] define an AMC design by setting a region boundary defined by signal to noise ratio (SNR) intervals corresponding each to a different transmission mode. The minimum switching threshold of each interval corresponds to the SNR at which the packet error rate (PER) is less or equal to a prescribed PER P_0. The AMC design is not adopted for UGS connections because, according to [35], [36], voice traffic can tolerate “some instantaneous packet loss”. Thus, the number of time slots allocated per frame to UGS connections is fixed. Liu *et al.* define a factor called the normalized channel quality based on the received SNR and a priority function (PRF) is assigned to each non-UGS connection depending on its service class. This PRF depends on:

- the BE class coefficient and the normalized channel quality for BE connections,
- the nrtPS class coefficient, the normalized channel quality, and the rate performance for nrtPS connections,
- the rtPS class coefficient, the normalized channel quality, and the delay requirements for rtPS connections.

The class coefficients are set so that the priority order for the different service classes is rtPS > nrtPS > BE. All the residual time, after scheduling UGS connections, is allocated to the connection having the highest PRF.

The AMC design proposed by Liu *et al.* is quite flexible since it does not depend on any specific traffic or channel model. However, the fact of scheduling only one non-UGS connection per frame might cause a significant delay for real-time applications. This is more likely to happen when the considered PHY is WirelessMAN-OFDM. Indeed, unlike in WirelessMAN-SC PHY where the frame size could take the values: 0.5, 1, or 2 ms, the frame sizes in WirelessMAN-OFDM varies from 2.5 to 20 ms [1], [2] !

Also, the scalability is claimed to be achieved by the proposed scheme since adding new connections would affect connections with low priority prior than those with a high priority. However, this would cause starvation of low priority connections and might even affect high priority ones when the network is overloaded. In order to overcome this shortcoming and guarantee better QoS performance, it would be interesting to combine the proposed scheduling scheme with an efficient CAC algorithm.

4.3.0.2 IP-MAC cross-layer schemes: Unlike Noordin *et al.* in [39] who restricted their cross-layer architecture to PHY and MAC layers, the authors in [37], [38] have focused on a layer 3 (L3) and layer 2 (L2) cross-layer design. They insisted on the importance of an IP and MAC cooperation to provide a better QoS service. The cross-layer framework proposed by Mai *et al.* in [37], [38] includes:

- a mapping between L3 and L2 QoS: where integrated service (IntServ) and differentiated service (DiffServ) classes are mapped to 802.16 MAC service classes as shown in Table 7.
- a simple admission control scheme based on which a new service flow is accepted when the remaining link capacity is more than the new flow required bandwidth.
- a fragment control mechanism that groups fragments of the same IP packet so that they are treated as a whole by L2 (e.g. fragments from the same IP packet are not interleaved in the L2 buffer, they are all removed in the case of congestion)
- a remapping scheme proposed for a better buffer utilization. Indeed, L3 higher priority CL and EF packets may be stored in nrtPS buffers when rtPS buffers are full (this is more likely to happen because of the burstiness of rtPS traffic).

	IP QoS	**MAC 802.16 QoS**
IntServ	Guaranteed Service (GS)	UGS
	Controlled load	rtPS
DiffServ	Expedited Forwarding (EF)	nrtPS
	Assured Forwarding (AF)	
IntServ, DiffServ	Best Effort (BE)	BE

TABLE 7: Mapping rule from IP QoS to MAC 802.16 QoS [37], [38]

4.3.0.3 Application-MAC-PHY cross-layer schemes: The cross-layer optimization mechanism proposed by Triantafyllopoulou *et al.* in [40], [41] takes advantage of the adaptation capabilities existing at both PHY and application layers. They combine the AMC capability of the physical layer and the multi-rate feature of the multimedia applications through a cross-layer optimizer that exists at BS and SS parts. The optimization process consists in collecting an abstraction of the the layer-specific information (such as QoS parameters and channel conditions) and informing the corresponding layers of the required changes. These changes are instructed based on a decision algorithm that decides about the MCS and traffic rate for each SS.

4.4 Connection admission control proposals for WiMAX PMP mode

In order to guarantee QoS in mobile networks, it is important to combine the scheduling policy with an efficient CAC strategy. The main role of a CAC strategy is to decide whether to accept or not new flows while making sure that the available resources would be sufficient for both the ongoing and the incoming connections. In order to take such an important decision, mainly two strategies can be adopted when no resources are available for the new flows. The first one—more flexible—would consist in gracefully degrading existing connections to make room for the new one. The second strategy—more conservative, yet simpler—would maintain the QoS provided for ongoing connections and simply reject the new service flow.

4.4.1 PMP CAC schemes with degradation strategy

This first category of CAC schemes include all the CAC algorithms based on service degradation [46], bandwidth borrowing [47], [48], [49], or bandwidth stealing [50] strategies. The main idea of these policies is to decrease—when necessary and possible—the resources provided to ongoing connections in order to be able to accept a new service flow. As we will see in this section, this strategy could be combined with a threshold-based capacity sharing approach in order to avoid starvation [50], or a guard channel strategy that reserves a dedicated amount of bandwidth for more bandwidth-sensitive flows (like UGS [48], or handover [49] connections).

Service degradation: In [46], service flows (SF) are prioritized according to their respective service type (UGS> (e)rtPS> nrtPS> BE) and among each service type, a priority is assigned to SFs based on their jitter requirements for UGS flows, delay for (e)rtPS flows and traffic priority for both nrtPS and BE flows. If the available bandwidth does not meet the requirements of handover flows, a SF degradation policy is applied. It consists in decreasing the bandwidth assigned to existing SFs whose priority is lower than the handover (HO) SF and whose assigned bandwidth exceeds the minimum reserved bandwidth. SF degradation concerns only handover SFs. A new flow is accepted only if the already available bandwidth guarantees its minimum bandwidth requirement. A two-dimensional continuous Markov model is used to analyze the performance of the proposed scheme. However, many assumptions have been considered: UGS=(e)rtPS and nrtPS=BE. The authors also suppose that all the flow belonging to the same class have the same minimum and maximum requirements which is restrictive.

The proposed scheme is then compared to a threshold-based admission control (TAC) policy [26] in terms of blocking and dropping probabilities and bandwidth utilization. Unlike the TAC algorithm, the AC approach proposed by Ge *et al.* [46] adjusts the grant adaptively to the cell load and does not restrict the SF degradation to a single class of flows when necessary. Thus, the proposed algorithm performs better than the TAC algorithm.

Bandwidth borrowing :

- Bandwidth borrowing in a non-cooperative game

The problem of admission control in IEEE 802.16 networks is formulated by Niyato *et al.* in [47] as a non-cooperative game. The players in this game are the rtPS and nrtPS connections that want to maximize their QoS performance. The payoff of the game is the total utility of the ongoing rtPS and nrtPS connections. The problem consists in finding the equilibrium point between the two types of connections to offer bandwidth for the new connection and meet the QoS requirements of both ongoing and new connection. Based on the solution of the game, a CAC scheme is then proposed to guarantee the QoS requirements of rtPS and nrtPS connections.

- Bandwidth borrowing and stepwise degradation

The CAC scheme, proposed by Wang *et al.* in [48], assigns the highest priority to UGS flows and aims to maximize the bandwidth utilization by bandwidth borrowing and degradation. A predetermined amount of bandwidth U is exclusively reserved for UGS connections. An UGS connection is accepted if there is enough bandwidth to accommodate its requirements otherwise it is rejected. Denote by B the total bandwidth, by b_{ong} the bandwidth set aside for ongoing connections (UGS, rtPS and nrtPS), and by b_{ugs}, b_{rtps} the bandwidth requirement for a new UGS or rtPS connection, respectively. For a new nrtPS connection, b_{nrtps}^{max} and b_{nrtps}^{min} stand for the maximum and minimum bandwidth requirements, respectively. The proposed degradation model is applied when a new rtPS connection is requested and $b_{ong} + b_{rtps} > B - U$ or when the creation of a new nrtPS connection is requested and $b_{ong} + b_{nrtps}^{max} - l_{nrtps}^{n} * \delta \geqq B - U$. where: δ is the amount of degraded bandwidth and l_{nrtps}^{n} is the current degradation level. Note that only nrtPS connections could be degraded to accept more rtPS and nrtPS connections. Thus, the reserved bandwidth for each nrtPS connection is $b_{nrtps}^{max} - l_{nrtps}^{n} * \delta$ which satisfies $b_{nrtps}^{max} - l_{nrtps}^{n} * \delta \geqq b_{nrtps}^{max}$ and the maximum degradation level that can be reached is $(b_{nrtps}^{max} - b_{nrtps}^{min})/\delta$. In this stepwise degradation scheme, the authors assume that all the connections belonging to the same service type (even non-UGS connections) have the same bandwidth requirements and that the bandwidth requested by an rtPS connection is fixed and does not vary between a maximum sustained and a minimum reserved traffic rates. These assumptions simplify the problem but do not take into account the service requirements specified in the standard.

- Proportional bandwidth borrowing and guard channel

In [49], the authors apply the following priority scheme where handover (HO_) connections are prioritized over new (N_) connections: HO_UGS > HO_rtps & HO_ertPS > N_UGS > N_rtPS & N_ertPS > HO_nrtPS > N_nrtPS > HO_BE > N_BE. The reserved bandwidth corresponds to the maximum sustained traffic rate for UGS and to the minimum required rate for polling services. No bandwidth is reserved for BE traffic. This basic algorithm is combined with a guard channel policy and a proportional bandwidth borrowing scheme. Indeed, a guard channel corresponding to $n\%$ of the channel capacity is reserved for handover connections. Thus a new connection is blocked if the available bandwidth is less than $C.n\%$ while a handover connection is blocked only if no bandwidth is available. A proportional bandwidth borrowing scheme is applied when the required bandwidth is not available. The BS borrows from connections having the same or lower priority than the new/HO connection. The connection that occupies more bandwidth lends more to the admitted connection.

Bandwidth stealing : In [50], Jiang *et al.* combine an uplink scheduling algorithm with a CAC policy, both based on a token-bucket approach. In the proposed CAC, each uplink connection is characterized by two parameters: a token rate r_i and a bucket size b_i. rtPS flows, however, have an extra parameter d_i corresponding to their delay requirement. In order to avoid starvation of some classes, the authors define a threshold capacity per service type. Thus, a class using more bandwidth than its dedicated threshold has less chances to use the remaining uplink capacity.

When an SS attempts to establish a new service flow—with parameters r_i, b_i and d_i (for rtPS flows)—with the BS, the proposed CAC algorithm is applied as follows. If the required bandwidth is less than the remaining uplink capacity C_{remain}, the flow is accepted. If not a “bandwidth stealing” strategy is applied. First, if connections belonging to lower classes—than the new one—are using more bandwidth than their respective thresholds, then the new flow is accepted if the sum of this extra C_L and C_{remain} is greater than or equal to its bandwidth requirement. If not, the capacity occupied by connections belonging to the same class of the new one is checked. If it is greater than its threshold, then the new service request is rejected. If not, a bandwidth stealing is attempted from connections belonging to higher classes. This last step is possible only if the capacity of these higher classes exceeds (by $C_U > 0$) their thresholds. If $C_U + C_L + C_{remain}$ is greater than or equal to the new flow bandwidth requirement, then the new flow is accepted. If not, it is rejected. Note that stealing bandwidth from non-real-time classes (BE and nrtPS) amounts to decreasing their capacity, while for real-time classes it consists in degrading the r_i of some of their connections to $c.r_i$ (0<c<1).

4.4.2 PMP CAC schemes without degradation strategy

The hierarchical uplink scheduling algorithm proposed in [24] by Wongthavarawat *et al.* and introduced in Section 4.1.2 was combined with a conservative token-bucket-based admission control module. Indeed, no graceful service degradation of existing connections is foreseen by authors to accept a new flow. Thus, a new connection is accepted only if (1) it will receive QoS guarantees in terms of both bandwidth and delay—for real-time flows—and (2) the QoS of existing connections is maintained.

Unlike most of the works where the admission control decision is only based on bandwidth availability, the CAC algorithm proposed by Chandra *et al.* [51] takes also into account the delay and jitter requirements of the service flows. Because the connections have different QoS requirements, an hyper interval (HI) is defined to test the admissibility of the requests. It represents the interval within which the admission process is performed. The authors however consider the delay and jitter requirements for UGS, rtPS and even nrtPS connections which may cause the blocking of an nrtPS connection for instance just because the jitter requirement—which is not necessary in this case as can be seen in Table 2—cannot be satisfied. Also, Chandra *et al.* include in their scheme a bandwidth estimator agent that is responsible for monitoring the queue length of both rtPS and nrtPS connections and estimating the bandwidth needs based on the instantaneous change in the queue length. Indeed, the authors define a "configurable threshold" BW_{thr} according to which, the bandwidth is requested as in the algorithm shown in Figure 14.

```
if $((minrate \leqq BR)\,\&\&\,(BR \leqq BW_{thr}))$
  then $B_{req} = minrate$
elseif $((BW_{thr} \leqq BR)\,\&\&\,(BR \leqq maxrate))$
  then $B_{req} = BR$
elseif $(maxrate < BR)$
  then $B_{req} = maxrate$
endif
```

Fig. 14: Configurable threshold algorithm [51]

where: BR and B_{req} stand for the bandwidth requirement, and the bandwidth request, respectively.

In [51], the main objective was to ensure QoS guarantee, in terms of bandwidth, delay and jitter. However, only the acceptance ratio was considered to evaluate the performance of the proposed solution.

4.4.3 Other PMP CAC schemes

In this section, we introduce some CAC algorithms that have addressed some of the aspects that have not been (or at least not well) investigated in previous works. The first two works [52], [53] have addressed one of the challenges that we have mentioned in Section 3 i.e. MAC-PHY cross-layer capability, or more specifically the possibility for a SF to change the burst profile (mainly the MCS)—also known as the AMC capability. We have also chosen to introduce the works done by Yang and Lu in [54], [55] because, unlike the other works presented in previous sections, they have proposed a CAC scheme specifically dedicated for real-time video applications.

AMC-induced CAC:: [52] is one of the rare works, addressing CAC in 802.16 networks, that take into account the AMC aspect. Indeed, Kwon *et al.* propose an AMC-induced CAC, for IEEE 802.16 networks, that incorporates the modulation type into the CAC process. The work has then been generalized to AMC networks in [56]. The proposed CAC scheme is based on a Markovian model that considers handoff and new connections as well as connections whose modulation changes. The model however supports only two types of modulations and is built based on the assumption that all the connections have fixed and equal bandwidth requirements which limits its applicability.

CAC for real-time video applications:: Some CAC solutions existing in literature, have been proposed for a specific kind of applications. In [54] and [55] for instance, the authors have taken advantage of the regularity and periodicity of real-time video traffic to propose a CAC process that particularly fits video applications. Indeed the authors have tried to overcome the time-varying bit rate behavior of video traffics by taking advantage of their group of pictures (GOP) structure—identified by a sequence of I, P and B frames. The main idea consists in avoiding the case where I frames—2 to 10 times bigger than B and P frames—of several flows are transmitted too close to each others. Therefore, the authors have defined a pending period during which the CAC module tries to find a proper time to admit the incoming flow. To fix this proper time, a coordination with I frames algorithm is defined to detect and avoid any I-frame superposition—and thus delay violation—between the ongoing flows and the incoming one. A non-I-frame coordination is then applied. This step aims to place the I and non I frames within their delay bounds. If the CAC is able to perform this step, and this before the pending period expires, the flow is admitted otherwise it is rejected. The amount of data corresponding to non-I frames is computed based on an estimation of non-I-frame rate.

In order to maximize the throughput and minimize the difference of delay between admitted flows, the authors have combined their CAC with a scheduling algorithm. Indeed a latest starting time (LST) algorithm is defined and compared to the EDF algorithm used for instance in [24], [25]. The main limit, which is also the advantage, of this solution is that it only addresses a specific kind of application: real-time video.

Table 8 summarizes the different aspects taken into account in the CAC proposals presented in this section. It mainly highlights the criteria (data rate, delay, jitter) based on which the decision, of accepting or rejecting a connection request, has been taken. It also shows whether a degradation and/or a guard channel technique has been adopted by the proposed CAC scheme. Note that we insisted on dedicating a column to AMC even though it has been considered only in [52], [56]. Indeed, we believe that it is a key feature that should not be ignored in the admission control process.

	Data rate	**Delay**	**Jitter**	**Degradation policy**	**Guard channel/ Capacity Thresholds**	**AMC**
[46]	√	—	—	√	—	—
[48]	—	—	—	$\surd^{*}$	$\surd^{**}$	—
[51]	√	√	√	—	—	—
[49]	√	—	—	√	$\surd^{***}$	—
[55], [54] (for video)	√	√	—	—	—	—
[50]	√	√	—	√	√	—
[47]	√	√	—	√	—	—
[52], [56]	√	—	—	—	$\surd^{****}$	√

*** stepwise degradation policy, ** for UGS connections, *** for handover connections**

****** for handover and modulation changing connections**

TABLE 8: CAC in IEEE 802.16 PMP mode: a comparative table

5 Mesh scheduling and CAC proposals

This section presents a possible classification for the scheduling and CAC algorithms for the Mesh mode part of the standard. Figure 15 shows a diagram with the topics used in the classification, the aspects observed are: Operation mode, design level, channel awareness, spectrum reuse, type of traffic and QoS observed.

It is perfectly possible to present more than one characteristic, for example, a proposed scheme that has a centralized approach, with cross layer design, that try to maximize the number of active links and that observe QoS parameters. Actually this is exactly the case of the scheme presented in [57]. However it is important to highlight that the points discussed here are, by no means, an exhaustive list, especially regarding the QoS support aspects. The QoS values listed here are just some of the more commonly found ones. Other classifications can be found in [58] and [59].

- **Operation mode**: The operation mode reflects if the proposed method focuses on the centralized or distributed mode of the mesh part of the IEEE 802.16 standard. In the centralized approach all the scheduling and CAC decisions are made by the Mesh BS. Without a central coordination, distributed approaches are more challenging than centralized ones, since the synchronization problem, in a distributed environment, is considerably harder.
  Both scheduling schemes may coexist, using different messages and configuration slots. Although this is the regular operation mode, explicit in the standard, the work of Cheng *et al.* [59] shows that the avoidance of such division may lead to a better overall network performance.
- **Design Level**: The conventional protocol stack scheme advocates that different protocol layers should be transparent to each other, this intends to make the implementation and operation simple and scalable. Unfortunately, this design approach does not necessarily lead to an optimum solution for wireless networks [4]. The CAC and the scheduling mechanisms are normally agreed to make part of the protocols of the MAC layer. However, some proposals have interfaces to receive information from other network layers and such information may have an impact on the protocol behavior in the MAC layer. Because the unreliability and relatively vulnerability of the wireless links the cross-layer approach may present better results for the schedule and CAC mechanisms.
- **Channel awareness**: The channel awareness aspect is related to how the approach treats and understands the communication channel. Some approaches consider every communication as occurring in one single communication channel, others allow the communication to be divided into different frequencies. The use of multi-channel communication allows more than one communication to occur at the same time, in different frequencies, even among neighbor nodes. This makes the scheduling problem much more interesting, and effective avoiding collisions and increasing the throughput. However, the allocation of frequencies makes the scheduling problem even harder. Other point to observe is that to use multi-frequency, the scheduled channels must be orthogonal, to avoid interference. Considering that, one must be aware that part of the available frequency spectrum is lost.
- **Spectrum reuse**: Some protocols prime for the reuse of frequency spectrum as a mean to increase the network efficiency, others, on the other hand, consider possible just one transmission in the whole network at a time.
- **Type of traffic**: Some protocols make distinction over the kind of traffic they are handling while others do not. The differentiation, normally, targets the possible QoS traffics presented in Section 2.2, as they are the types defined in the standard.
- **QoS aspects observed**: Some scheduling and CAC mechanisms observe QoS aspects to improve the network behavior. The QoS aspect observed may be, for example, in terms of the quality of the flows, e.g. throughput and delay, may be in terms of fairness of medium access for the connections. We consider in this paper, as QoS aspect, the use of other techniques, e.g. interference minimization, also as a QoS aspect. Again,

a proposed method may present more than just one of these aspects and the ones considered here are not an exhaustive list of aspects a protocol could consider.

The next subsections present examples of techniques that fit in each one of these categories. Each literature proposal will be presented in conjunction to the category that best differentiates it, even though it fits in more than one category. Each presented proposed scheme has a short description of its operation mode and an explanation why it fits in that specific category. Table 9 summarizes the classification of the discussed techniques.

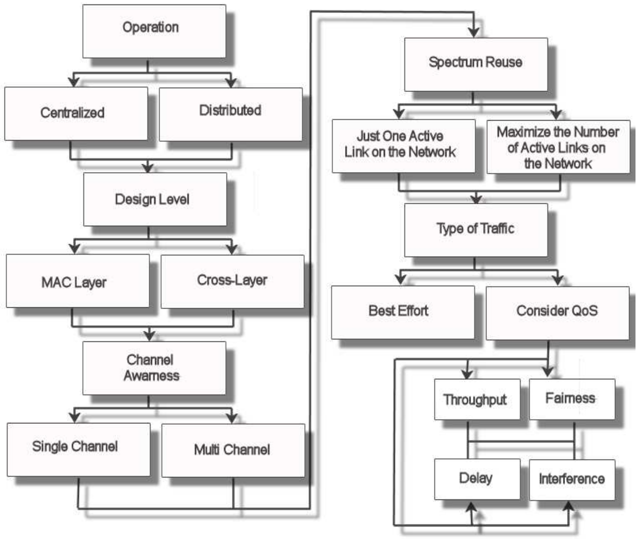


Fig. 15: Proposed classification for WiMAX mesh mode CAC algorithms

5.1 Operation mode-based classification

5.1.1 Centralized proposals

In [60] Kuran *et al.* introduce the Service Adaptive QoS (SAQoS) mechanism, a centralized scheduling and CAC approach. In this method, the BS generates five different node IDs for each SS. Each one of these IDs corresponds to one of the five different kinds of services on the network. Each one of these five virtual nodes request bandwidth individually for the link between it and the next node, observing the rules of the kind of the traffic it represents. Following the standard, the resource requests are incremental. Consider, for example, a topology with three nodes in a row. If the first node needs 100 Kbits, it requests 100 Kbits for the link between it and the next node. If the second node needs, 200 Kbits, it will request grant for 300 Kbits, for the next link, the amount the node needs plus the amount the previous node requested. The scheduling of the granted communications is done through the Fair Adaptive Base Station Scheduler (FABS).

The scheduler is based on the current requests and grants given to each SS. Taking into account a normalization factor, nF, the links are ranked inversely to their granting ratio, gr_i. In other words, links that received the least grants will be in the top of the list.

Using this method, Kuran *et al.* conclude that it is not feasible to have low priority flows beyond the second level in the Mesh mode of IEEE 802.16, since the scheduling will normally favor real-time and multimedia applications. This is an interesting observation, but highly dependent of the kind on traffic on the network and the fairness observed by the scheduler.

A drawback of this approach is that it suffers from spatial bias. It gives more importance to links far from the BS, once these are more likely to present smaller amounts of traffic. However, what normally happens when one uses a communication tree is that the links nearer to the root, in this case the BS, have a higher traffic. This occurs because they carry the traffic of all its children. Indeed, the bandwidth requirement of a link is proportional to the number of child nodes it sponsors. However, following this method, nodes that concentrate the greatest part of the traffic will have the lowest priority. In the extreme case, the network may get disrupted because the leaves will send messages but the sponsor nodes may not be able to forward this traffic to the BS.

5.1.2 Distributed proposals

The first scheduling scheme proposed specifically for IEEE 802.16 mesh mode, that we have knowledge, was introduced by Redana Lott and Capone in [61]. The proposal is a simple adaptation of the PMP basic method to support multihop topologies. The network area is divided into clusters, each cluster has a SS elected to act as a BS for that cluster. This station, called PMP BS, has direct radio contact with the SSs in its area and directed links to the nearby PMP BSs. The idea is quite simple and introduces a new concept, cluster communication. The evaluation shows that for small number of connections the strategy presents better results than the PMP mode.

Unfortunately, too few details are given about both, the techniques and their results. Through the presented information it is hard to reach any conclusion about the technique, but independently of this fact, the proposed method introduces some unique concepts and it is the first initiative in the sense of having some kind of coordination in WiMAX mesh mode networks.

Liu *et al.* presents in [62] a coordinated distributed method for slot allocation based on priorities. After receiving a request the node looks at the resource table to check the slots occupation. The number of allocated minislots represents the utilization of the data subframe in a certain degree. A threshold is considered, varying from 0 and 256. If the network utilization is below the threshold, the network is recognized as in a good condition and all requests are considered with the same priority. If the utilization is above the threshold, indicating network congestion, the algorithm returns failure for low priority resource requests.

Two algorithms are presented, A1 and A2. The difference between them is the number of check points each one evaluates, A1 evaluates just one check point while A2 evaluates two check points. In A2 the second checkpoint is only evaluated if the first checkpoint is below the threshold.

Liu *et al.* approach is simple and elegant, but maybe too simplistic to address completely the problem. In the evaluation *they* do not consider, for example, mobility, drop packets and transmission errors. Another problem regarding the evaluation is that it considers just the two proposed approaches, without considering any other one. Another point to notice is that the check point positions clearly impact the algorithm performance, however, there is no explicit indication of how these points should be set.

Djukic and Valaee also propose, in [63], a distributed link scheduling algorithm for TDMA mesh networks. In the method, each node uses its partial view of the network to solve scheduling conflicts independently. To maximize the concurrent transmissions on the network, the algorithm uses the concept of conflicting graph. In this graph, the edges represents the links on the network and the nodes the conflicts among the links. The conflicting graph is constructed over the communication graph, considering the nodes extended neighborhood. Figure 16 shows an example of topology and the conflicting graph generated by it.

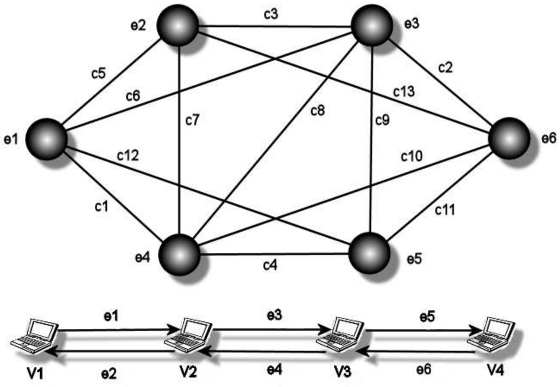


Fig. 16: An example of topology and the conflicting graph generated by it [63]

The method has two different and autonomous procedures. The first procedure is a distributed Bellman-Ford like algorithm running on the conflict graph and the second one is a wave-based procedure that detects the scheduling convergence.

A centralized version for this approach is presented in [64]. In [65] one can find a comparison between the centralized approach, the distributed approach [63] and other two methods [66] and [67]. The algorithm proposed in [68] is considered also in this comparison work but it is not compared with the others. This study shows that both proposals of Djukic and Valaee have consistently better throughput and delay than the others [65]. However, there is a huge penalization for [67] because it allows the links to transmit multiple times in the same time frame. Djukic and Valaee argue that in 802.16 every transmission needs a guard time of three TDMA slots [64], using this the overhead for [67] method increases considerably.

Both proposed schemes, [63] and [64] are interesting and valuable, however, they share the same problem. They consider that all traffic originates or terminates in the BS. In the general case this may not be true, it is perfectly possible to have traffic between two SSs inside the same mesh network. Other valuable observation is that, even though, the algorithm performs well for practical uses, it has a worse case that is two times the number of links in the network. This worst case, mainly for delay purposes, may be a problem for real implementations.

5.1.3 Hybrid (distributed/centralized) proposals

Cheng *et al.* propose in [12] to combine both, centralized and distributed scheduling mechanisms, dynamically modifying the slots allocation to increase the network utilization, Cheng *et al.* suggest to divide the traffic between Internet and intranet ones. The Intranet traffic is the one that goes through the backhaul to outside of the network and the intranet one corresponds to the communication between SSs of the same mesh network. In this division, the centralized scheduler is responsible for the Internet traffic and the distributed scheduler responsible for the intranet traffic.

The IEEE 802.16 standard suggests the use of a partition scheme, where the percentage of minislots, for each kind of traffic, is static and set in the MSH-NCFG messages. This percentage value stays fixed until the next configuration message arrives at the nodes. The problem is that if the demand for one kind of scheduling is greater, or lower, than expected, the network may experience either congestion or waste of bandwidth. The Combined Distributed and Centralized (CDC) scheme, proposes to eliminate the concept of partitions. Figure 17 shows two examples of minislot allocations, one for the standard partition scheme and another one for the CDC scheme.

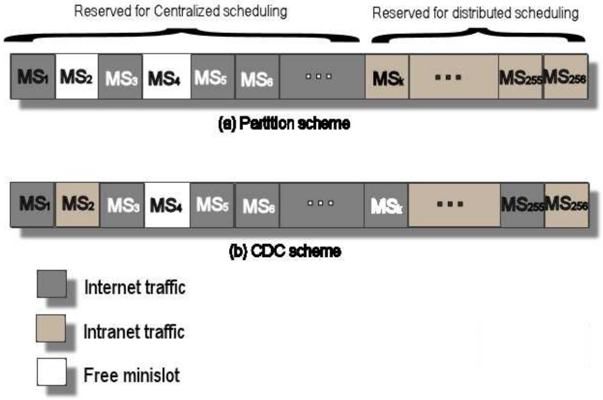

Reserved for Centralized scheduling
Reserved for distributed scheduling
MS1 MS2 MS3 MS4 MS5 MS6 ... MSk ... MS255 MS256
(a) Partition scheme
MS1 MS2 MS3 MS4 MS5 MS6 ... MSk ... MS255 MS256
(b) CDC scheme
Internet traffic
Intranet traffic
Free minislot

Fig. 17: Examples of different minislot allocation schemes [12]

To evaluate the proposed method two different scheduling policies are compared: greedy and round robin. Even though the comparisons between both policies is not fair since the greedy approach takes advantage of spectral reuse, while the round robin scheme does not, the results show that the CDC scheme is better than the partition one. The experiments just present the results for Internet traffic, but for this type of traffic, CDC greatly decreases the ratio of dropped packets in comparison with the partition scheme. It is shown also, that as there is only demand for Internet traffic, this kind of traffic is favored in the slots allocation. More experiments should be done to fully develop the concept and test its validity, but the main idea is quite reasonable and worth being explored.

5.2 Design level-based classification

5.2.1 MAC layer proposals

The centralized scheduling algorithm proposed by Kim and Ganz in [66] tries to maximize the network throughput while reaching fairness in terms of scheduled bandwidth per node. The proposed strategy does not trust, or require, information from any other layer apart from the MAC one. The method is divided into phases, the first one is called Node ordering and the second one is called Link allocation.

The node ordering phase consists of ordering the nodes in accordance to their satisfaction index. The satisfaction index is defined as the ratio between the amount of the allocated bandwidth, during a preconfigured interval time, and the node's total weight. The node's weight is a factor that may be used to reflect the node class or priority and is set during the network initialization. The total weight is the sum of the node weight and the weight of all its children nodes i.e. the nodes to which it provides access.

In the second phase, the link allocation phase, the BS broadcasts the nodes ordering and bandwidth requirements. With this each node determines its own transmission schedule. The method works with two matrices, a schedule matrix and a collision matrix. After inserting a node in the schedule matrix, all nodes in the extended neighborhood, all nodes within 2 or 3 hops, are added to the collision matrix. To avoid collisions, Kim and Ganz present three rules: First, no node may transmit and receive data simultaneously. Second, no neighbor of a sending node may transmit data and third, no neighbor of a receiving node may transmit data.

Even simple, the method reaches efficiency of 94.8%, when compared with the maximum possible throughput, being 5% the maximum fairness variance. Another conclusion of the work is that both phases are needed to reach a high throughput. The main concern about this work is the use of hard fairness. The node is scheduled even if it has no data to transmit. Cao *et al.* show in [6] that such fairness approach undermines the possible network capacity. The efficiency of the network of 94.8% is just possible if all nodes have real demand for bandwidth, assuming this demand to be always nearly the same, what is unlikely to occur in real environments.

5.2.2 Cross-layer proposals

Cross-layer scheduling strategies have been recognized as an effective approach in wireless communication [69]. In [70] Shetiya and Sharma present both a routing and a centralized schedule algorithm for maximizing the network throughput. The scheduling algorithm is obtained through dynamic programming over the optimization of cost functions. In this work Shetiya and Sharma argue that, the IEEE 802.16 standard does not provide any specific routing algorithm, however, to reach an optimal network performance both, scheduling and routing algorithms, must coexist and collaborate. The work [70] not only propose a routing algorithm but also compare a series of different scheduling policies.

The proposed routing algorithm basically creates a tree that maximizes the network stability and minimizes the average work needed to transmit a packet from the SSs to the BS. The routing is fixed over all the frames for each node along the path.

To generate the scheduling algorithms, the proposed scheme introduces a finite horizon dynamic programming framework, which uses the tree to optimize the total reward earned at the end of N slots. The scheduling schemes are compared in terms of system throughput and average queue length versus arrival rate. In accordance to the results presented, at low data rates, up to 1.2 Mbps, all the presented schemes have similar performance. However, when the rates increase, typically around 1.4 Mbps, the differences start to appear. What happens in this context is that, normally, adaptive schemes present better performance than fixed ones. The *Maximum Transmission Scheme* is the one that presents the best overall performance. Although, it has the same throughput of the *Per Slot Maximum Transmission Scheme* and the *Maximum Transmission Scheme*, the *Maximum Transmission Scheme* has a lower queue size.

5.3 Channel awareness-based classification

5.3.1 Single channel proposals

Han *et al.* present in [68] a scheduling algorithm that intends to increase the concurrent transmission and to avoid interferences. The algorithm considers that all nodes use the same frequencies and, because of that, some rules must be followed to avoid interference. The proposed Transmission Tree Scheduling algorithm considers that the transmission tree already exists. Indeed, the tree is considered as an input for the scheduling algorithm. For each link on the tree, if it has a traffic demand, it is labeled as available, otherwise it is labeled as idle. While there are links labeled as available, the link that best fit in the selection criteria is selected and marked as scheduled. All nodes that may cause interference are labeled as interfered and not suitable to be scheduled at this time slot anymore. The process continues while there are links labeled as available.

This work presents a very good explanation about interference in TDMA mode. Let us consider Figure 18, and suppose a communication between A and B is scheduled for the current timeslot, and that A is the transmitter. To avoid interference nodes B, C, E, F, N and P cannot be scheduled to transmit in this time slot. Han *et al.* divide the possible interferences in two types: $Nei\,[B] - \{A\}$ and $Sons\,(Nei\,[A] - \{B\})$. In the first category, neighbors of communication $A - B$, fall nodes B, C and P. On the second category, sons of sponsored nodes of $A - B$, fall E, F and N.

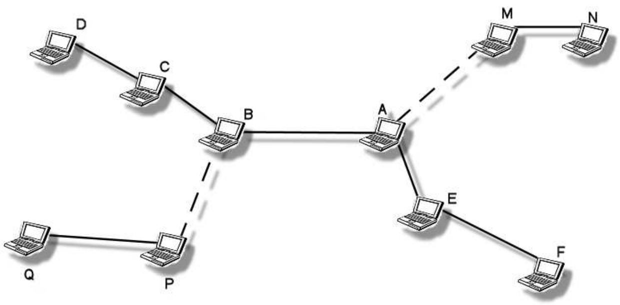


Fig. 18: Tree topology exposing the possible interference in mesh TDMA mode networks [68]

The different scheduling policies evaluated in [68] are: random, min interference, nearest to BS and farthest to BS. In the random police the scheduled link is selected randomly. The min interference selection chooses the link where the transmission will affect the less other nodes. The nearest to BS policy gives preference to the nodes near to the BS and the farthest to BS one, favors the nodes far from the BS.

According to the results, the best policy is the nearest to BS. This makes sense since the nodes nearest to BS will carry the greatest part of the traffic, so they should have some priority in the transmission schedule. Another proof of this need is that, in the random policy, the communication bottle neck was exactly the nodes near the BS. The messages arrive to the nodes near to the BS node, but they are unable to reach the BS because of the scheduler. In the nearest to BS policy the links far from BS do not starve because when the links are empty they are labeled as idle, and this creates opportunities for other links to be scheduled.

5.3.2 Multi channel proposals

In [71] Lee *et al.* present a CAC scheme that intends to guarantee both delay and bandwidth for the nodes in the mesh network. Unlike other methods based in TDMA technology, where every node uses the same channels and concurrent transmissions are allowed only for nodes far apart, this work proposes the use of subchannelization of the frequencies. Doing this, multiple data streams can be sent/received in separate orthogonally divided subchannels by the same, or nearby, nodes at the same time without interference.

The algorithm uses a tree, rooted at the BS, as transmission topology. Each node is labeled, in accordance to its depth in the tree, as even or odd node. The transmissions time slots are also divided into even and odd timeslots. At the even time slot, the even nodes transmits to the odd nodes, and the dual for the odd timeslot. All simultaneous transmissions and receptions must occur into different subchannels. Since all directional links are active half of the time, Lee *et al.* call it as half-idle network.

Each connection presents a reward that expresses how valuable it is to admit such connection on the network. The routing tree construction process is responsible for maximizing the reward of the admitted connections. Lee *et al.* present, through integer linear programming a perfect solution for the problem. However, such solution, even for moderate size networks, is too complex to solve the problem in a feasible time. Because of that, the paper [71] also presents and evaluates four fast heuristics for tree construction. Overall the best approach was the SP-order (Shortest Path order) heuristic. This method iteratively builds a tree attaching the nodes through the links that minimize the resulting maximum node load. The second best approach, called MST, builds a minimum spanning tree using the inverse of the link capacity as edge costs.

Lee *et al.* present also a study of how relax requirements to grant more connections to the network. In this part of the work QoS sensitive connections are granted if and only if their requirements can be met. The other connections are all admitted but with a possible scaled down and stretched end-to-end delay guarantee. The main idea in this case is trying to provide as much bandwidth as possible for each connection, while stretching its delay requirement as small as possible.

5.4 Spectrum reuse-based classification

5.4.1 One active link proposals

The mesh mode proposed by Shetiya and Sharma in [72] considers possible just one transmission at each time on the whole network, although the standard allows spatial reuse. The proposed method intends to provide performance for UDP and TCP traffic on the network. First the paper presents the problem of scheduling UDP and TCP traffic separately, and after that, how to route and schedule both traffics at the same time. Shetiya and Sharma argument that these two kinds of traffic have their own specific traffic patterns. So an optimal routing scheme for one kind of traffic may not be well suitable for the other. Therefore the approach tries to find a scheme that performs well for both cases, even though is not optimal for any one of them.

The chosen route tree is the one that maximizes the network stability. The route does not change unless there is a really good reason to do so, e.g. a broken link. This means that the traffic originated in one node always follows the same path. It is assumed that each node transmits at the maximum allowed power and that if the channel condition on a link changes, the data rate also may change.

The approach adopted to schedule UDP traffic is to minimize the number of dropped packets, once the main worry for this traffic is the delay. For the evaluation all UDP packets, not transmitted during a scheduling frame, are dropped. For TCP traffic there are two main concerns, first, minimizing the queue sizes and second, finding a fair distribution scheme for the exceeding bandwidth. Two different allocation schemes are proposed, one fixed and one adaptive. In the fixed allocation scheme the number of slots attributed to each node is static and dependent on the average data arrival rate and estimated average channel rate. The adaptive allocation scheme works exactly as the fixed one, but it also has a list of good and bad links. Bad links are those ones that either do not have enough data to fulfill its slots or do have poor links. On the adaptive scheme good links have precedence over bad links. To avoid starvation of the nodes under bad links credits are attributed to them when they miss a transmission slot. When the number of credits reaches a limit, the transmission is scheduled even if it occurs through a bad link.

For the joint scheduling Shetiya and Sharma argue that UDP traffic should have precedence over TCP traffic. They base their argument on [73] where it is observed that when giving priority to UDP flows, their delays decrease without affecting the TCP flows throughput.

The experiments show that the joint approach really keep the delay and drop packets for UDP connections low while providing good performance for the greatest part of TCP flows. However, in the experiments, the network bandwidth is over-provisioned. Another important factor, raised by [6], is the lack of spatial reuse which may greatly damage the whole network performance.

5.4.2 Multiple active links proposals

Wei *et al.* [67] present a heuristic centralized scheduling algorithm based on an interference aware route construction approach. The centralized schedule considers both traffic demands and interference conditions to distribute the scheduling grants. Wei *et al.* consider that a good routing approach is required to achieve efficient spectral utilization and high throughput. Keeping this in mind they define a route construction algorithm that minimizes the routes blocking metric.

The blocking value of a node is defined as the number of blocked nodes when a communication of that node occurs, typically it corresponds to all the one hop neighbors. The routing blocking metric is the sum of all blocking values of the nodes that transmits messages on this route. When a new node scans the medium to find a sponsor node, the blocking value is taken into account. New nodes should find sponsors that minimize the interference on the multihop route.

The protocol intends to maximize the concurrent interference to achieve higher spectral utilization and system throughput. For each time slot, the algorithm interactively schedules the node with the higher unallocated traffic demand and that do not interfere with the already allocated traffic. The iterative allocation ends when all the unallocated capacity is fulfilled.

The method is compared to an optimal linear programming solution and to the basic 802.16 mesh scheme. The linear programming solution is optimal and shows the possible upper bound performance. Even being a heuristic, the proposed approach performs very close to the optimal solution with a very low computational cost. However, it is not clear how the other possible interference cases are threaded for this algorithm, e.g. if it is allowed to have a receiver neighbor to a transmitter at the same time slot. Another issue raised in [6] is that this approach does not consider fairness among nodes, despite that it is essential to ensure that subscribers receive acceptable shares of resources from the BS [74].

Taking into account the work of Wei *et al.* [67] Tao *et al.* introduce in [57] an interference aware algorithm to construct a routing tree to improve network performance. The nodes attachments to the network occur in sequential order. Each node, when arriving at the network, selects the sponsor node that minimizes the interference. The interference, for the new node, is defined as the sum of receiving and transmitting interference between the new arrival and its sponsor, plus the interference of the sponsor node itself.

Once a new node is attached, to the network the interference values may change and the new tree may not be the minimum interference one. In this case an adjustment process may be used to optimize the tree. Proceeding in this way, the tree becomes independent from the order the nodes attach to it, with is a problem for [67], for example. The proposed schedule algorithm first gives transmission opportunities for the nodes with the bigger hop-counts and, after that, to the ones with smaller hop-count, once they cause least interference.

The tree adjustment process really improves the overall uplink and downlink throughput, however, the links are considered static and nodes without mobility. The experiments also do not show how this approach behaves in comparison to the original one from Wei *et al.* [67]. It could be interesting to see some comparative evaluation between both to determine the impact of the introduced techniques.

5.5 Type of traffic support-based classification

5.5.1 Best effort support proposals

In [75], Chen *et al.* present a method to schedule only Best Effort traffic. This work does not consider any differentiation among traffic flows. The main objective of the approach to minimize the scheduling period to increase the bandwidth efficiency. The Chen *et al.* base their solution on an access tree over which it are defined the uplink and downlink traffic. Beyond the mathematical model for the scheduling minimization, the paper also proposes a scheduling algorithm to perform the time slot allocation observing spatial reuse.

The proposed method reduces the scheduling frame slots by half when compared to a FIFO approach. The experiments show that the implemented FIFO approach does not benefit from concurrent transmissions. In FIFO mode, only one link is active in most of the scheduling time, serving the node in sequence. The average concurrent rate for the proposed approach is 43% while for the FIFO one is only 12.5%. In contrast to the work presented in [60], this one suggests to provide more slots for the nodes closer to the BS.

5.5.2 All types of services support proposals

The heuristic method proposed in [50], [76] by Jiang *et al.* uses buckets to perform CAC and scheduling with delay guarantee for WiMesh networks. Each one of the different class of traffic has a bucket, with a given number of tokens each, being the number of tokens is set using a Markov Chain. To perform the estimations, a number of parameters are used: the number of tokens in the bucket, the number of packets in the queue, the time interval between two markov chain states, the probability of packets arrival in the interval, the size of the queue and the expected loss rate.

The method avoids flows starvation using a threshold bandwidth value attributed to each class. If a class is using more bandwidth than its threshold value, it will have lower priority in the traffic distribution and it can even have some amount of its bandwidth stolen and shared among other classes.

The proposed scheduler is based on the Earliest Deadline First mechanism proposed in [24]. The scheduler also allocates bandwidth in accordance to the flows priorities, the higher priorities flows are allocated first. The scheduler is conservative in distributing grants, it first gives the less possible amount of rate to each flow and if at the end some resources are still available, they are then redistributed.

The comparison results show that the method is successful in avoiding traffic starvation and distributing the rates among the classes. One of the main reasons attributed to the technique success is the use of the threshold values. The delay achieved by rtPS traffic is also shown to be nearly constant, even with the increase of the number of rtPS connections. However, it is important to notice that in [77] Wang *et al.* call attention to the fact that nrtPS, rtPS and BE have fractal traffic. The connections durations and traffic do not follow exponential distribution, so it is not possible to use Markov Chain to analyze them.

5.5.3 Fairness based proposals

In [6] Cao *et al.* present a new fairness model for centralized scheduling in WiMAX networks. The main objective of this fairness model is to associate the scheduling to the real network traffic demands to increase network capacity. When presenting their fairness model Cao *et al.* introduce two definitions: uplink capacity and pursued fairness.

The uplink capacity in mesh networks, under a scheduling tree T, is defined as:

$$C = \{x : \alpha \in co\left(A_T\right), x \geq 0\}, \tag{1}$$

where x is the bandwidth allocation vector, α is the fraction of time a link need to become active, and $co\left(A_T\right)$ is the convex hull of the activation vector set A_T.

The introduced fairness is based on a fairness profile f, an uplink traffic demand s and a relative bandwidth request R. A bandwidth allocation vector x is fair if:

$$x_i = min\left\{s_i, f_i R\right\}, i \in N. \tag{2}$$

Where N is the set of SSs in the network. The fairness constraint is applied for the nodes where the bandwidth demand can not be met without violating the fairness constraints relative to other nodes.

Through linear programming, the paper [6] presents a reasonably efficient rate allocation algorithm. Having the optimal slot allocation vector, then a greedy coloring algorithm is used to generate the scheduling tree. To evaluate their approach Cao *et al.* compare their method with other three different methods, including hard standard fairness. The approach not only increases the total throughput but also maintains fairness to the nodes having real traffic demands, avoiding spatial bias. In other words the slots share of each SS is independent of their distance from the BS [74].

5.6 Comparison and insights over the presented proposals

Each one of the presented methods has its own objectives and mechanisms. Giving the different objectives of the scheme, any quantitative comparison among the strategies is in principle unfair. For example, some of the presented works just want to test one aspect of the IEEE 802.16 Mesh mode CAC and scheduling problem while others try to go further and really implement the mechanisms in the way standard proposes. Without implementing all proposed methods and comparing them within the same scenarios it is unlikely that any one can affirm without doubt which one is the best one. There are indeed some works, like [65] that have some consistent results comparing the performance of some approaches. However, the comparison summarized in the Table 9, is done in architectural terms and it is based on the taxonomy proposed in Section 5. This comparison does not intend to show which approach is the best or even the most complete. It intends much more to be a summary of the most relevant ideas to guide future works on this field.

As stated previously, no communication is allowed in WiMAX networks, if not previously scheduled. This means that, more than just correct and well designed, the CAC and scheduling mechanisms must also be fast and computational efficient, since all the network communication rely on them. In addition to this, the scheduling problem in multihop networks is proved to be NP-hard [7], [8]. Because of this, even the optimal techniques normally present also an heuristic, not optimal, to solve the problem [63], [67], [71]. In the real world, sub-optimal solutions may be the only way to apply scheduling and CAC techniques to mesh networks.

The fairness is another interesting issue and, probably, the one with the most distinct aspects among the proposed methods. The fairness is in truth an umbrella that accommodates many different definitions. However, it is commonly agreed that some kind of fairness is valuable for the network [74]. A peculiar, although interesting fairness approach, dynamic fairness, is introduced in [6]. The concept of dynamic fairness seems to be more interesting for the link unstable context of mesh networks. Even though, in the general

Proposal	*Operation Mode*	*Design Level*	*Channel Aware*	*Spectrum Reuse*	*Type of Traffic Considered*	*QoS Aspects Observed*
[78]	Distributed	MAC	No	No	No	No
[62]	Distributed	MAC	No	No	Yes	Priority channels
[63]	Distributed	MAC	No	Yes	Yes	Yes
[12]	Dist/Central	MAC	No	Yes/No	No	No
[60]	Centralized	MAC	No	Yes	Yes	5 types of service
[64]	Centralized	MAC	No	Yes	Yes	Yes
[66]	Centralized	MAC	No	Yes	No	No
[68]	Centralized	MAC	No	Yes	No	No
[50]	Centralized	MAC	No	Yes	Yes	Yes, all the classes
[70]	Centralized	Cross-Layer	No	No	No	No
[72]	Centralized	Cross-Layer	No	No	Yes UDP and TCP	Yes
[67]	Centralized	Cross-Layer	No	No	No	Yes
[75]	Centralized	Cross-Layer	No	Yes	No	Yes
[6]	Centralized	Cross-Layer	No	Yes	Yes TCP and UDP	No
[57]	Centralized	Cross-Layer	No	Yes	No	No
[71]	Centralized	Cross-Layer	Yes	Yes	Different rewards for dif. connections	QoS and Non QoS connections

TABLE 9: Mesh scheduling proposed methods comparison, based in the proposed taxonomy

case either one, hard or dynamic fairness, is welcomed. Other simple and efficient ideas related to fairness, like the establishment of threshold for different class of services presented in [50], [57], [76], can also be interesting and even applicable in conjunction to other different techniques.

Many of the proposed approaches also proved that the interference is a real problem that must be treated carefully. The proposed schemes to handle the interference vary in many senses and can use, for example, a conflict graph [63] or a conflict matrix [66]. For TDMA like approaches the techniques can be the constructing better routes [67], [57], [68], [75] or dividing the spectrum [71].

Mainly for the centralized scheduling it is agreed, by many of the proposed methods, that the creation of a scheduling tree is the best approach [6], [57], [60], [63], [68], [70], [71], [72], [75]. If we consider the standard OSI seven layers model [79], the creation of this tree rooted at the mesh BS is routing and, normally, part of the job of the network layer. In this sense, such methods present a cross-layer design. Such kind of scheme normally present really good perspectives and seems to be a good direction for new approaches to follow.

The standard itself [1], [2] defines a series of different types of services, presented here in Section 2.2, to be used by the applications. These services are considered by some approaches [59], [60] in conjunction with their particular characteristics. Some of the approaches, more than just considering differentiation between the different services, also consider during the scheduling and CAC a reward for served connections [71] or nodes [66]. One of the main objectives of the CAC and scheduler in these approaches is to maximize the reward of the network. It is important to notice here that this may really provide better quality to the nodes in the privileged classes, but can be very unfair to other classes. We need to keep in mind that the available amount of resources is always the same. Sometimes to present gains, some techniques may penalize some users. This must be done really carefully to avoid rash unfairness.

The standard states that the grants, even for centralized approaches, should be done hop by hop. Normally the approaches distribute the grants exactly in this way, but some methods go a little further than that. In [60], for example, it is proposed that each node should be represented by *n* different virtual nodes, where *n* is the number of different services. This intends to make easier the manipulation of the scheduling and the grants distribution among the services and nodes.

6 WiMAX manufacturers equipments: main scheduling features

The previous two sections discuss research works in the field of scheduling in WiMAX networks from an academia point of view. This section, however, addresses the problem of scheduling from the vendors and WiMAX Forum point of view.

Established in 2001, the WiMAX Forum is the entity in charge of promoting and certifying wireless broadband equipments based on the IEEE 802.16 and the European telecommunications standards institute (ETSI) HiperMAN standards. In September 2008, the forum had 530 member companies from 51 countries. The first WiMAX Forum Certified products based on IEEE 802.16d, operating in the 3.5 GHz band, were announced in January 2006 and the first certified equipments based on IEEE 802.16e-2005, operating in the 2.3 GHz and 2.5 GHz bands, were announced in the second quarter of 2008. By September 2008, 25 vendors had successfully completed the certification process and 62 products have received the WiMAX Forum certification. At the same time, more than 62 companies were developing WiMAX chipsets and end user devices, and 37 companies developing infrastructure equipments. Their products were used in WiMAX deployments by 407 operators in 133 countries [80].

The certification process has several objectives, among them: enabling interoperability, performance testing, and gradual introduction of new functionalities, providing backward compatibility, making possible the utilization of spectrum bands with common allocations worldwide, enabling economies of scale, accelerating the adoption of the standard, and establishing the WiMAX Forum Certified program as the trusted resource for equipments selection. To receive the certification, a base station, for example, needs to interoperate with a minimum of three subscriber devices from other vendors, and subscriber devices with a minimum of two base stations from other vendors [80].

In IEEE 802.16 standard many points have been left to vendors to differentiate their equipments. Unfortunately, companies do not provide detailed information about the schedulers they implement on their products. Table 10 summarizes the main available information regarding the scheduling of some leading products proposed by Alvarion, Aperto, etc.. As can be seen in Table 10, the resource allocation scheme might be different from one vendor to another. Nevertheless, the WiMAX Forum provides some insightful guidelines for the implementation of the MAC scheduling service. The scheduler must then support some key features to enable the implementation of an efficient broadband data service [81].

- **Fast data scheduler**: the MAC scheduler must efficiently allocate available resources in response to bursty data traffic and time-varying channel conditions.
- **Scheduling for both DL and UL**: the scheduling service is provided for both DL and UL traffics. The UL should also provide information for the efficient allocation of the DL resources.
- **Dynamic resource allocation**: the MAC supports frequency-time resource allocation in both DL and UL on a per-frame basis. Fast and fine granular allocation scheme increases the QoS for data traffic. With the ability to dynamically allocate resources in both DL and UL, the scheduler can provide better QoS for both DL and UL traffics.
- **QoS-oriented:** the MAC scheduler handles data transport on a connection-by-connection basis. Each connection is associated with a single data service with a set of QoS parameters that quantify the aspects of its behavior.
- **Frequency selective scheduling**: the scheduler can operate on different types of sub-channels. The frequency-selective scheduling can enhance the system capacity with a moderate increase in channel quality information (CQI) overhead in the UL [82].

7 Conclusion and directions for future research

This paper presents the state of the art of scheduling and CAC algorithms for IEEE 802.16 networks. This survey is by no means an exhaustive compilation of the works addressing this topic. Yet it describes, classifies, and compares scheduling and CAC proposals for both PMP and Mesh modes. It also summarizes the main challenges and issues that should be considered when designing new scheduling and CAC algorithms.

In the last few years, this research area has been intensively investigated and a lot of progress has been done. It is true that CAC and scheduling in wireless networks are classical problems. However, the comparative study presented in this survey shows that, for WiMAX networks, there is still room for improvement.

If we have a look on the scheduling algorithms proposed in literature for PMP 802.16 networks (Section 4), we would notice that the main challenging problems that arise when trying to develop a CAC and scheduling strategy are:

- to make a trade-off between an efficient solution, that would take into account the QoS requirements of the different applications, and a simple one that would be implementation-friendly and less time consuming.
- to make a compromise between fairness and channel utilization. Indeed giving priority to users having better channel conditions would increase the channel utilization. Nevertheless, it would be unfair to other users.
- to make a choice between an optimized solution that targets a specific kind of applications (like real-time video in [55], [54]) and takes into account its specific needs, and a more general, yet efficient and less complex, scheduling policy that would address heterogeneous types of traffics.
- to take advantage of the adaptive modulation and coding (AMC) capability defined by the standard when proposing a new CAC solution, like it has been proposed in [56].
- to consider the possibility of an adaptive DL/UL bandwidth allocation, as introduced in [18], [20], in order to make an efficient use of the resources and handle unbalanced traffic.
- to investigate more deeply the game theory-based scheduling as an alternative to solve the problem of resource allocation in the context of 802.16 networks. Indeed, despite the efficiency of this approach for wireless networks in general, only a few works like [30], [47] have formulated the scheduling problem as a non-cooperative game.

The WiMAX mesh mode is a good and valuable part of the IEEE 802.16 standard, but it is still a young one.Because of that many aspects of this kind of network are not explored deeply enough, as example of areas that more research would be gladly welcome we could highlight:

- a number of parameters must to be set to reach good protocol performance e.g. holdoff exponent, periodicity

Manufacturer	Product	Main scheduling features
Alvarion	4Motion	Support of DL/UL asymmetric capacity allocation in TDD implementations Real-time scheduling decisions made by the BS based on: * available radio resources, * active SF QoS requirements (e.g. traffic rate and latency), * individual SS service level agreements (SLA), * MCS used by each mobile station, * and connection-per-connection path conditions Advanced scheduling: frame-by-frame capacity allocations to: * support diverse set of SSs, * and meet the committed customer-by-customer SLA
Aperto	PacketMAX (ServiceQ)	Service classes: CBR, committed information rate (CIR), and BE Differentiated scheduling: * CBR: UGS * CIR: rtPS or nrtPS * BE: RR Use of OPNET Modeler's network protocol models QoS-aware scheduling and CAC algorithms
Redline	RedMAX 4C	Predictive scheduling Scheduling classes: UGS, rtPS (for VoIP), nrtPS, and BE TDD/OFDM systems Support of over 250 active SS on a single BS sector
Sequans	S-Cube	Hierarchical QoS Weighted fair queuing (WFQ) Traffic shaping Congestion management Random early detection (RED)

TABLE 10: Main scheduling features supported by WiMAX equipments

of MSH-NCFG messages. Some consistent work have been done analyzing the network performance, but more works exploring these parameters are needed and surely enough would represent a valuable contribution to the field. The holdoff exponent value, for example, strongly affects IEEE 802.16 performance [6] and not many works have explored it.

- the characterization of the traffic distribution in mesh networks is also important, not only for network simulation purposes, but also for designing newer and better algorithms. Some authors, when analyzing and validating their protocols just use Poisson or normal distribution to generate traffic. Also in [77] it is argued that wide-area network traffic is much better modeled using self-similar processes [83]. However, for wireless mesh networks, the traffic distribution and patterns for the different QoS services is still to be studied, at least in a deeper way.
- some works present good results using orthogonal channel allocation for IEEE 802.11 mesh networks [84]. This kind of technique could be even easier when applied to WiMAX networks, but, again, little has been done exploring this field. The frequency reuse is another topic that may be important for Mesh networks, and that has been studied for PMP WiMAX networks [85], but not for the mesh mode.
- a new working group is studying the problem of relay networks, the IEEE 802.16j, that is a problem very near to the mesh networks one. Scheduling and CAC mechanisms for such kind of networks could be an interesting research topic. Apart from that, it would also be interesting to study the mix of both networks, IEEE 802.16 mesh mode and IEEE 802.16j, for example, adding some relay points in the mesh network [77]. This can open new opportunities for scheduling and routing, where new algorithms can take advantage of the relay characteristics to help the network performance.
- other unexplored research area is the use of adaptive power allocation (APA), in scheduling for WiMAX mesh mode, to decrease the interference in the network. Some techniques even consider always node at full power transmission [72]. Some work on this field, using APA and CAC mechanisms have been studied for PMP networks [86], [87], but no work addressed this for WiMAX mesh networks.
- mixing different kinds of networks has also a really important real world appeal. We have many different standards addressing mesh as a valid architectural topology e.g. IEEE 802.16, IEEE 802.20, but so far no work addressed the interconnection of such standards.
- one can always learn with works on other fields, for example, some people have explored hierarchical approaches for CAC for CDMA networks [88]. The

same general idea could be also applied to IEEE 802.16 mesh mode, as well as the cluster based reservation, explored in [89].

- even though mobility is a key aspect for wireless mesh networks, we observed that, so far, no technique has fully considered it. Indeed, there are no guarantees of how the actual methods will behave in a mobility context. New and efficient procedures must be designed to handle handoffs and the constant position changing in the network topology.
- some techniques approach the scheduling and CAC problems using simple heuristics. However, it could be interesting to see how to apply more sophisticated artificial intelligence techniques to solve the scheduling problem, since it is an NP-hard one.
- some techniques propose some schema of reward for connections, which can be used as indicative of revenue, but up to now no one has seriously discussed ways of billing the access to such networks. This is a sensitive subject that even may conflict with some network neutrality aspects [90], but who and how to pay for the access for WiMAX networks and how this will influence the CAC mechanism is not fully comprehended yet.

A good discussion about important emerging trends and future research issues for CAC can be found in [58]. Also Cheng *et al.* present, in [59], a good list of open research issues on CAC mechanisms for wireless networks.

References

[1] IEEE Std 802.16-2004. IEEE Standard for Local and metropolitan area networks- Part 16: Air Interface for Fixed Broadband Wireless Access Systems. 2004.

[2] IEEE Std 802.16e 2005. Ieee standard for local and metropolitan area networks part 16: Air interface for fixed and mobile broadband wireless access systems—amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1. 2005.

[3] S. Sadr, A. Anpalagan, and K. Raahemifar. Radio Resource Allocation Algorithms for the Downlink of Multiuser OFDM Communication Systems. *IEEE Communications Surveys & Tutorials*, 11(3):92–106, 3rd Quarter 2009.

[4] I. F. Akyildiz, X. Wang, and W. Wang. Wireless mesh networks: a survey. In *Computer Networks 2005; 47(4): 445-487*, 2005.

[5] R. Bruno, M. Conti, and E. Gregori. Mesh networks: commodity multihop ad hoc networks. In *IEEE Communications Magazine 43(3): 123-134*. IEEE, 2005.

[6] Min Cao, Vivek Raghunathan, and P. R. Kumar. A Tractable Algorithm for Fair and Efficient Uplink Scheduling of Multi-hop WiMax Mesh Networks. In *Second IEEE Workshop on Wireless Mesh Networks, pp. 101-108 Reston, VA*. IEEE, Sept. 2006.

[7] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu. Impact of Interference on Multi-hop Wireless Network Performance. In *ACM MobiCom, pages 66-80*. ACM Press, 2003.

[8] M. Kodialam and T. Nandagopal. Characterizing the achievable rates in multihop wireless networks. In *ACM MobiCom, pages 868-880*. ACM Press, 2003.

[9] Hua Zhu and Kejie Lu. Performance of IEEE 802.16 Mesh Coordinated Distributed Scheduling Under Realistic Non-Quasi-Interference Channel. In *International Conference on Wireless Networks (ICWN'06), Las Vegas, USA, Jun. 26-29*, 2006.

[10] IEEE Std 802.16a 2003. IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems. 2003.

[11] Yong Sun, Dharma Basgeet, Khurram Rizvi, Zhong Fan, and Paul Strauch. Dynamic Frame Structure for IEEE802.16j Relaying Transmission to Support Efficient Scheduling. In *IEEE C80216j-06_224, Nov., 07*. IEEE, 2006.

[12] S. Cheng, P. Lin, D. Huang, and S. Yang. A study on distributed/centralized scheduling for wireless mesh network. In *2006 International Conference on Communications and Mobile Computing, Vancouver, British Columbia, Canada*. ACM Press, 2006.

[13] R. Jain, D. M. Chiu, and W. R. Hawe. A quantitative measure of fairness and discrimination for resource allocation shared computer systems. In *Digital Equipment Corporation technical report TR-301*, 1984.

[14] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi. Performance Evaluation of the IEEE 802.16 MAC for QoS Support. *IEEE Transactions on Mobile Computing*, 6(1):26–38, Jan. 2007.

[15] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund. Quality of service support in IEEE 802.16 networks. *IEEE Network*, 20(2):50–55, Mar.-Apr. 2006.

[16] Alexander Sayenko, Olli Alanen, Juha Karhula, and Timo Hämäläinen. Ensuring the QoS requirements in 802.16 Scheduling. In *Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems*, pages 108–117, Terromolinos, Spain, 2006. ACM Press New York, NY, USA.

[17] Lin-Fong Chan, Hsi-Lu Chao, and Zi-Tsan Chou. Two-Tier Scheduling Algorithm for Uplink Transmissions in IEEE 802.16 Broadband Wireless Access Systems. In *International Conference on Wireless Communications, Networking and Mobile Computing, 2006. WiCOM 2006*, pages 1–4, Sept. 2006.

[18] Jianfeng Chen, Wenhua Jiao, and Hongxi Wang. A Service Flow management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode. In *IEEE International Conference on Communications (ICC2005)*, Seoul, Korea, 2005.

[19] Naian Liu, Xiaohui Li, Changxing Pei, and Bo Yang. Delay Character of a Novel Architecture for IEEE 802.16 Systems. In *Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005*, pages 293–296, Dec. 2005.

[20] Maode Ma, Jinchang Lu, S.K. Bose, and Boon Chong Ng. A three-tier framework and scheduling to support QoS service in WiMAX. In *6th International Conference on Information, Communications & Signal Processing*, pages 1–5, Dec. 2007.

[21] R. Perumalraja, J.J.J. Roy, and S. Radha. Multimedia Supported Uplink Scheduling for IEEE 802.16d OFDMA Network. In *Annual India Conference, 2006*, pages 1–5, Sept. 2006.

[22] M. Settembre, M. Puleri, S. Garritano, P. Testa, R. Albanese, M. Mancini, and V. Lo Curto. Performance analysis of an efficient packet-based IEEE 802.16 MAC supporting adaptive modulation and coding. In *International Symposium on Computer Networks, 2006*, pages 11–16, Jun. 2006.

[23] J. Sun, Yanling Yao, and Hongfei Zhu. Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems. In *IEEE 63rd Vehicular Technology Conference, 2006. VTC 2006-Spring*, volume 3, pages 1221–1225, 2006.

[24] Kitti Wongthavarawat and Aura Ganz. Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems. *International Journal of Communication Systems*, 16(1):81–96, Feb. 2003.

[25] Kitti Wongthavarawat and Aura Ganz. IEEE 802.16 based last mile broadband wireless military networks with quality of service support. *IEEE Military Communications Conference, 2003. MILCOM 2003*, 2(1):779–784, Oct. 2003.

[26] N. Nasser and H. Hassanein. Prioritized multi-class adaptive framework for multimedia wireless networks. In *IEEE International Conference on Communications, 2004*, volume 7, pages 4295–4300, Jun. 2004.

[27] D.Niyato and E. Hossain. Connection admission control algorithms for OFDM wireless networks. In *IEEE Global Telecommunications Conference, 2005. GLOBECOM '05*, volume 5, page 5 pp, Nov. 2005.

[28] D. Niyato and E. Hossain. Delay-Based Admission Control Using Fuzzy Logic for OFDMA Broadband Wireless Networks. In *IEEE International Conference on Communications*, volume 12, pages 5511–5516, Jun. 2006.

[29] Dusit Niyato and Ekram Hossain. Queue-Aware Uplink Bandwidth Allocation for Polling Services in 802.16 Broadband Wireless Networks. In *IEEE GLOBECOM 2005 proceedings*, 2005.

[30] Dusit Niyato and Ekram Hossain. A Game-Theoretic Approach to Bandwidth Allocation and Admission Control for Polling Services

in IEEE 802.16 Broadband Wireless Networks. In *3rd international conference on Quality of service in heterogeneous wired/wireless networks*, 2006.

[31] Dusit Niyato and Ekram Hossain. Queue-Aware Uplink Bandwidth Allocation and Rate Control for Polling Service in IEEE 802.16 Broadband Wireless Networks. *IEEE TRANSACTIONS ON MOBILE COMPUTING*, 5(6), Jun. 2006.

[32] Dusit Niyato and Ekram Hossain. A Queuing-Theoretic and Optimization-Based Model for Radio Resource Management in IEEE 802.16 Broadband Wireless Networks. *IEEE TRANSACTIONS ON COMPUTERS*, 55(11), Nov. 2006.

[33] D. Niyato and E. Hossain. Joint Bandwidth Allocation and Connection Admission Control for Polling Services in IEEE 802.16 Broadband Wireless Networks. In *IEEE International Conference on Communications (ICC '06)*, volume 12, pages 5540–5545, Jun. 2006.

[34] Vandana Singh and Vinod Sharma. Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks. In *IEEE Wireless Communications and Networking Conference, 2006. WCNC 2006*, volume 2, pages 984–990, Apr. 2006.

[35] Qingwen Liu, Xin Wang, and G.B. Giannakis. Cross-layer scheduler design with QoS support for wireless access networks. In *Second International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, 2005*, page 8 pp., Aug. 2005.

[36] Qingwen Liu, Xin Wang, and G.B. Giannakis. A cross-layer scheduling algorithm with QoS support in wireless networks. *IEEE Transactions on Vehicular Technology*, 55(3):839–847, May 2006.

[37] Yi-Ting Mai, Chun-Chuan Yang, and Yu-Hsuan Lin. Cross-Layer QoS Framework in the IEEE 802.16 Network. In *The 9th International Conference on Advanced Communication Technology*, volume 3, pages 2090–2095, Feb. 2007.

[38] Yi-Ting Mai, Chun-Chuan Yang, and Yu-Hsuan Lin. Design of the Cross-Layer QoS Framework for the IEEE 802.16 PMP Networks. *IEICE Transactions on Communications*, E91-B(5):1360–1369, May 2008.

[39] K. A. Noordin and G. Markarian. Cross-Layer Optimization Architecture for WiMAX Systems. In *The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'07)*, pages 1–4, Sept. 2007.

[40] Dionysia-Katerina Triantafyllopoulou, Nikos Passas, and Alexandros Kaloxylos. A Cross-Layer Optimization Mechanism for Multimedia Traffic over IEEE 802.16 Networks. In *European Wireless 2007 (EW 2007)*, Apr. 2007.

[41] Dionysia Triantafyllopoulou, Nikos Passas, Apostolis K. Salkintzis, and Alexandros Kaloxylos. A heuristic cross-layer mechanism for real-time traffic over IEEE 802.16 networks. *INTERNATIONAL JOURNAL OF NETWORK MANAGEMENT*, 17(5):347–361, Sept. 2007.

[42] D. Stiliadis and A. Varma. Latency-rate servers: a general model for analysis of traffic scheduling algorithms. *IEEE/ACM Transactions on Networking*, 6(5):611–624, Oct. 1998.

[43] H. Fattah and C. Leung. An overview of scheduling algorithms in wireless multimedia networks. *IEEE Wireless Communications*, 9(5):76–83, Oct. 2002.

[44] Howon Lee, Taesoo Kwon, and Dong-Ho Cho. An Enhanced Uplink Scheduling Algorithm Based on Voice Activity for VoIP Services in IEEE 802.16d/e System. *IEEE COMMUNICATIONS LETTERS*, 9(8), Aug. 2005.

[45] Howon Lee, Taesoo Kwon, Dong-Ho Cho, Geunhwi Limt, and Yong Changt. Performance Analysis of Scheduling Algorithms for VoIP Services in IEEE 802.16e Systems. In *IEEE 63rd Vehicular Technology Conference, 2006. VTC 2006-Spring*, volume 3, pages 1231–1235, 2006.

[46] Yin Ge and Geng-Sheng Kuo. An Efficient Admission Control Scheme for Adaptive Multimedia Services in IEEE 802.16e Networks. In *IEEE 64th Vehicular Technology Conference, 2006. VTC-2006 Fall. 2006*, pages 1–5, Sept. 2006.

[47] D. Niyato and E. Hossain. Radio Resource Management Games in Wireless Networks: An Approach to Bandwidth Allocation and Admission Control for Polling Service in IEEE 802.16. *IEEE Wireless Communications*, 14(1), Feb 2007.

[48] H. Wang, W. Li, and D. P. Agrawal. Dynamic admission control and QoS for 802.16 wireless MAN. In *Wireless Telecommunications Symposium, 2005*, pages 60–66, Apr 2005.

[49] Liping Wang, Fuqiang Liu, Yusheng Ji, and Nararat Ruangchaijatupon. Admission Control for Non-preprovisioned Service Flow in Wireless Metropolitan Area Networks. In *Fourth European Conference on Universal Multiservice Networks, 2007. ECUMN '07*, pages 243–249, Feb. 2007.

[50] Chi-Hong Jiang and Tzu-Chieh Tsai. Token bucket based CAC and packet scheduling for IEEE 802.16 broadband wireless access networks. In *3rd IEEE Consumer Communications and Networking Conference, 2006. CCNC 2006*, volume 1, pages 183–87, Jan. 2006.

[51] S. Chandra and A. Sahoo. An Efficient Call Admission Control for IEEE 802.16 Networks. In *Proceedings of the 15th IEEE LAN/MAN Workshop, LANMAN 2007*, pages 188–193, Jun. 2007.

[52] E. Kwon, J. Lee, K. Jung, and S. Ryu. A Performance Model for Admission Control in IEEE 802.16. In *3rd International Conferences on Wireless/Wired Internet Communications (WWIC 2005)*, pages 159–168, Xanthi , Greece, May 2005.

[53] J.Y. Lee and K.B. Kim. Statistical Admission Control for Mobile WiMAX Systems. In *IEEE Wireless Communications and Networking Conference. WCNC 2008*, Apr. 2008.

[54] O. Yang and J. Lu. A New Scheduling and CAC Scheme for Real-Time Video Application in Fixed Wireless Networks. In *3rd IEEE Consumer Communications and Networking Conference, 2006. CCNC 2006*, volume 1, pages 303–307, Jan 2006.

[55] O. Yang and J. Lu. Call Admission Control and Scheduling Schemes with QoS Support for Real-time Video Applications in IEEE 802.16 Networks. *JOURNAL OF MULTIMEDIA (JMM)*, 1(2):21–29, May 2006.

[56] Jaiyong Lee Eunhyun Kwon, Hun-je Yeon and Kyunghun Jung. Markov Model for Admission Control in the Wireless AMC Networks. *IEICE Transactions on Communications*, E89-B(8):2230–2233, 2006.

[57] Jian Tao, Fuqiang Liu, Zhihui Zeng, and Zhangxi Lin. Throughput enhancement in WiMAX mesh networks using concurrent transmission. In *Wireless Communications, Networking and Mobile Computing 2005, v.2, 871-874*, Sept. 2005.

[58] M.H. Ahmed. Call admission control in wireless networks: a comprehensive survey. In *Communications Surveys & Tutorials, IEEE, First Qtr. 7(1):49-68*. IEEE, 2005.

[59] H. T. Cheng, H. Jiang, and W. Zhuang. Distributed medium access control for wireless mesh networks. In *Wireless Communications and Mobile Computing, V.6, I.6*, Sept. 2006.

[60] M. S. Kuran, B. Yilmaz, F. Alagoz, and T. Tugcu. Quality of Service in Mesh Mode IEEE 802.16 Networks. In *SOFTCOM 2006, Split, Croatia*, Sept. 2006.

[61] S. Redana, M. Lott, and A. Capone. Performance evaluation of point-to-multi-point (PMP) and mesh air-interface in IEEE standard 802.16a. In *Vehicular Technology Conference, VTC2004-Fall*, Sept. 2004.

[62] Fuqiang Liu, Zhihui Zeng, Jian Tao, Qing Li, and Zhangxi Lin. Achieving QoS for IEEE 802.16 in Mesh Mode. In *8th International Conference on Computer Science and Informatics, Salt Lake City, Utah, USA*, 2005.

[63] P. Djukic and S. Valaee. Distributed link scheduling for TDMA mesh networks. In *ICC 2007, Glasgow, Scotland*, Jun. 2007.

[64] P. Djukic and S. Valaee. Link scheduling for minimum delay in spatial re-use TDMA. In *26th Annual IEEE Conference on Computer Communications, Anchorage, Alaska, USA*. IEEE, May 2007.

[65] Petar Djukic. *Scheduling Algorithms for TDMA Wireless Multihop Networks*. PhD thesis, University of Toronto, http://www.wirlab.utoronto.ca/ref/text/DjukicPHDThesis.pdf, 2008.

[66] D. Kim and A. Ganz. Fair and efficient multihop scheduling algorithm for IEEE 802.16 BWA systems. In *Broadband Networks, 2005*, volume 2, pages 833–839, Oct. 2005.

[67] H-Y. Wei, S. Ganguly, R. Izmailov, and Z. J. Haas. Interference-Aware IEEE 802.16 WiMax Mesh Networks. In *61st IEEE Vehicular Technology Conference (VTC 2005 Spring), Stockholm, Sweden*. IEEE, 2005.

[68] Bo Han, Fung Po Tso, Lidong Ling, and Weijia Jia. Performance Evaluation of Scheduling in IEEE 802.16 Based Wireless Mesh Networks. In *2006 IEEE International Conference Mobile Adhoc and Sensor Systems (MASS), 789-794*. IEEE, Oct. 2006.

[69] Mehrdad Shariat, Atta U. Quddus, Seyed Ali Ghorashi, and Rahim Tafazolli. Scheduling as an Important Cross-Layer Operation for Emerging Broadband Wireless Systems. *IEEE Communications Surveys & Tutorials*, 11(2):74–86, 2nd Quarter 2009.

[70] Harish Shetiya and V Sharma. Algorithms for Routing and centralized Scheduling in IEEE 802.16 Mesh Networks. In *IEEE Wireless*

Communications and Networking Conference, Las Vegas, NV, USA. IEEE, Apr. 2006.

[71] S. Lee, G. Narlikar, M. Pal, G. Wilfong, and L. Zhang. Admission Control for Multihop Wireless Backhaul Networks with QoS Support. In *IEEE Wireless Communications and Networking Conference*. IEEE, 2006.

[72] Harish Shetiya and Vinod Sharma. Algorithms for routing and centralized scheduling to provide QoS in IEEE 802.16 mesh networks. In *1st ACM workshop on Wireless multimedia networking and performance modeling, Montreal, Quebec, Canada, 140 - 149*. ACM Press, 2005.

[73] A. Gupta. A unified approach for analyzing persistent, non-persistent and ON-OFF TCP sessions with RED control and exogenous trafic. In *M.Sc. Thesis, Dept. of ECE, IISc, Bangalore*, 2002.

[74] V. Gambiroza, B. Sadeghi, and E. Knightly. End-to-end performance and fairness in multihop wireless backhaul networks. In *MobiCom 04, pages 287-301, New York, NY*. ACM SIGMOBILE, ACM Press, Sept.-Oct. 2004.

[75] Jianfeng Chen, Caixia Chi, and Qian Guo. A Bandwidth Allocation Model with High Concurrence Rate in IEEE802.16 Mesh Mode. In *2005 Asia-Pacific Conference on Communications, Perth, Western Australia*, Oct. 2005.

[76] Tzu-Chieh Tsai, Chi-Hong Jiang, and Chuang-Yin Wang. CAC and Packet Scheduling Using Token Bucket for IEEE 802.16 Networks. In *Journal of Communications, V1, N2*. Academy Publisher, May 2006.

[77] Haitang Wang, Bing He, and D. P. Agrawal. Admission control and bandwidth allocation above packet level for IEEE 802.16 wireless MAN. In *12th International Conference on Parallel and Distributed Systems, 2006. ICPADS 2006*, volume 1, page 6 pp, Jul. 2006.

[78] Simone Redana and Matthias Lott. Performance Analysis of IEEE 802.16a in Mesh Operation Mode. In *IST SUMMIT 2004, Lyon, France, Jun. 2004*, Jun. 2004.

[79] Andrew S. Tanenbaum. *Computer Networks*. Prentice Hall, 4 edition edition, 2002.

[80] WiMAX Forum. The WiMAX Forum Certified$^{\text{TM}}$ program Driving the adoption of interoperable wireless broadband worldwide. WiMAX Forum White paper, Sep. 2008.

[81] WiMAX Forum. Mobile WiMAX Part I: A Technical Overview and Performance Evaluation. WiMAX Forum, Mar., 2006.

[82] R. Love K. Stewart R. Ratasuk R. Bachu Y. Sun Q. Zhao F. Wang, A. Ghosh. IEEE 802.16e System Performance-Analysis and Simulation Results. In *PIMRC*. IEEE, Sept. 2005.

[83] W. E. Leland. On the self-similar nature of ethernet traffic (extended version). In *IEEE Communications Magazine, 40:1-15*. IEEE, Jun. 2002.

[84] Murali Kodialam and Thyaga Nandagopal. Characterizing Achievable Rates in Multi-Hop Wireless Mesh Networks With Orthogonal Channels. In *IEEE/ACM Transactions on Networking, vol. 13, no. 4*. IEEE/ACM, Aug. 2005.

[85] Carsten Ball, E. Humburg, K. Ivanov, and F. Treml. Comparison of IEEE802.16 WiMax Scenarios with Fixed and Mobile Subscribers in Tight Reuse. In *IST Summit Dresden*, 2005.

[86] Bo Rong, Yi Qian, and Hsiao-Hwa Chen. Adaptive power allocation and call admission control in multiservice WiMAX access networks. *IEEE Wireless Communications*, 14(1):14–19, Feb. 2007.

[87] Bo Rong, Yi Qian, and Kejie Lu. Revenue and Fairness Guaranteed Downlink Adaptive Power Allocation in WiMAX Access Networks. In *16th IST Mobile and Wireless Communications Summit*, Jul. 2007.

[88] Hsiao-Hwa Chen and Wee-Teck Tea. Hierarchy Schedule Sensing Protocol for CDMA Wireless Networks Performance Study under Multipath, Multiuser Interference, and Collision-Capture Effect. In *IEEE Transactions on Mobile Computing, vol. 4, no. 2*. IEEE, Mar./Apr. 2005.

[89] Lin CR and Gerla M. Adaptive clustering for mobile wireless networks. In *IEEE Journal on Selected Areas in Communications, 15(7):1265-1275*. IEEE, 1997.

[90] Tim Wu. Network Neutrality, Broadband Discrimination. In *Journal of Telecommunications and High Technology Law, Vol. 2, p. 141*. Academy Publisher, 2003.

Appendix A: Abbreviations and acronyms

2TSA	two-tier scheduling algorithm
AF	assured forwarding
AMC	adaptive modulation and coding
APA	adaptive power allocation
BE	best effort
BS	base station
BWA	broadband wireless access
BWN	broadband wireless network
CAC	connection admission control
CBR	constant bit rate
CIR	committed information rate
CDMA	code division multiple access
CDC	combined distributed and centralized
CID	connection identifier
CL	controlled load
CQI	channel quality information
CRC	cyclic redundancy check
CTMC	continuous time markov chain
DCD	downlink channel descriptor
DFPQ	deficit fair priority queuing
DiffServ	differentiated services
DIUC	downlink interval usage code
DL	downlink
DLFP	downlink frame prefix
DRR	deficit round robin
DSA	dynamic service addition
DSC	dynamic service change
DSD	dynamic service deletion
EDD	earliest due date
EDF	earliest deadline first
EF	expedited forwarding
ertPS	extended real-time polling service
ETSI	european telecommunications standards institute
FDD	frequency division duplex or duplexing
FIFO	first in first out
FQ	fair queuing
FTP	file transfer protocol
GOP	group of pictures
GPC	grant per connection
GPSS	grant per subscriber station
GS	guaranteed service
HO	handover
HTTP	hypertext transfer protocol
IE	information element
IP	Internet protocol
IEEE	institute of electrical and electronics engineers
IntServ	integrated services
IUC	interval usage code
ISP	Internet service provider
L2	layer 2
L3	layer 3
LOS	line-of-sight
LR	latency-rate
LST	latest starting time
MAC	media access control
MCS	modulation and coding scheme
MMFS	max-min fair sharing
MPEG	moving picture experts group
MSH-CSCH	mesh centralized schedule

MSH-DSCH	mesh centralized schedule configuration
MSH-NCFG	mesh network configuration
MSH-NENT	mesh network entry
MST	minimum spanning tree
NLOS	non-line-of-sight
nrtPS	non-real-time polling service
OFDM	orthogonal frequency division multiplexing
OFDMA	orthogonal frequency division multiple access
OSI	open systems interconnection
PDRR	pre-scale dynamic resource reservation
PDU	protocol data unit
PF	proportional fair
PHS	payload header suppression
PHY	physical layer
PMP	point-to-multipoint
PQLW	priority-based queue length weighted
QoS	quality of service
RED	random early detection
RF	radio frequency
RR	round-robin
RRM	radio resources management
RTG	receive/transmit transition gap
rtPS	real-time polling service
SAQoS	service adaptive quality of service
SCFQ	self-clocked fair queuing
SD	silence detector
SF	service flow
SFID	service flow identifier
SLA	service level agreement
SMTP	simple mail transfer protocol
SPLF	shortest packet length first
SP-order	shortest path order
SS	subscriber station
TAC	threshold-based admission control
TCP	transmission control protocol
TDD	time division duplex or duplexing
TDM	time division multiplexing
TDMA	time division multiple access
TTG	transmit/receive transition gap
UCD	uplink channel descriptor
UDP	user datagram protocol
UGS	unsolicited grant service
UIUC	uplink interval usage code
UL	uplink
VAD	voice activity detection
VoIP	voice over IP (Internet protocol)
WFQ	weighted fair queuing
WiMAX	worldwide interoperability for microwave access
WiMesh	wireless mesh
WirelessMAN	wireless metropolitan area networks
WMN	wireless mesh networks
WRR	weighted round-robin

Ikbal Chammakhi Msadaa received her Computer Science Engineering diploma in 2005 and her Master's degree in 2006, both from ENSI (Ecole Nationale des Sciences de l'Informatique), Tunisia. In January 2007, she joined the Mobile Communications Department of EURECOM, Sophia-Antipolis France to prepare a Ph.D. under the supervision of Pr. Fethi FILALI. Her research work is focused on QoS and mobility management in 802.16 networks.

Daniel Câmara holds a M.Sc. in Computer Science from the Computer Science Department of the Federal University of Minas Gerais, Brazil and a B.Sc. in Computer Science from the Department of Informatics of the Federal University of Paraná, also in Brazil. Currently he is a Ph.D. candidate at TELECOM ParisTech working at EURECOM Sophia-Antipolis, France. His research interests include routing for mobile networks, call admission control and topology management protocols for wireless mesh networks.

Fethi Filali received his Computer Science Engineering and DEA degrees from the National College of Computer Science (ENSI) in 1998 and 1999, respectively. At the end of 1999, he joined the Planète research team at INRIA (The French National Institute for Research in Computer Science and Control) in Sophia-Antipolis to prepare a Ph.D. in Computer Science which he has defended on November 2002. During 2003, he was an ATER (Attaché Temporaire d'Enseignement et de Recherche) at the Université of Nice Sophia-Antipolis (UNSA) and he joined on September 2003 the Mobile Communications department of EURECOM in Sophia-Antipolis as an Assistant Professor. He is/was involved in several French-funded (Dipcast, Constellation, Rhodos, Cosinus, Airnet, WiNEM) and IST FP6/7 (Widens, Newcom, Daidalos, E2R, Multinet, Unite, Chorist, iTetris, Newcom++) projects. In the context of some of these projects, he designed and developed an open, flexible, and efficient architecture for the support of heterogeneous radio technologies. This architecture was integrated in EURECOM's wireless software-radio platform. His current research interests include the design, development, and performance evaluation of communication protocols and systems for: Ubiquitous networking, Communications for Intelligent Transportation Systems (in particular Vehicle to Vehicle and Vehicle to Infrastructure communications), WIMAX wireless networks, sensor and actuator networks (SANETs), and wireless mesh networks (WMNs). He served as a technical reviewer of several international conferences and journals. Additionally, he is a member of IEEE and IEEE Communications Society. In April 2008, he was awarded the "Habilitation à Diriger des Recherches" (HDR) from the University of Nice Sophia-Antipolis for his research on wireless networking.