Quality of Service Support in IEEE 802.16 Networks

Claudio Cicconetti, Luciano Lenzini, and Enzo Mingozzi, University of Pisa
Carl Eklund, Nokia Research Center

Abstract

During the last few years, users all over the world have become more and more accustomed to the availability of broadband access. This has boosted the use of a wide variety both of established and recent multimedia applications. However, there are cases where it is too expensive for network providers to serve a community of users. This is typically the case in rural and suburban areas, where there is slow deployment (or no deployment at all) of traditional wired technologies for broadband access (e.g., cable modems, xDSL). In those cases, the most promising opportunity rests with Broadband Wireless Access technologies, such as the IEEE 802.16, also known as WiMAX. One of the features of the MAC layer of 802.16 is that it is designed to differentiate service among traffic categories with different multimedia requirements. This article focuses on mechanisms that are available in an 802.16 system to support quality of service (QoS) and whose effectiveness is evaluated through simulation.

It is common knowledge that over the last decade there has been a major boost in communication networks. In fact, the development of high-performance backbone networks was immediately followed by the rapid dissemination of broadband wired access technologies, such as leased lines based on fiber-optic links, cable modems using coaxial systems, and digital subscriber line (xDSL) access networks. This gave users a whole new class of services that exploit the increasing number of available network resources. Many new services are based on multimedia applications, such as voice over IP (VoIP), video conferencing, video on demand (VoD), massive online gaming, and peer-to-peer. Unlike traditional TCP/IP services, multimedia applications usually require strict network guarantees such as reserved bandwidth or bounded delays.

The broadband access phenomenon has been investigated by the International Telecommunication Union (ITU), which reported in [1] that Broadband Wireless Access (BWA), although still in the early stage of its growth, is one of the most promising solutions for broadband access. Standards for BWA are being developed within IEEE Project 802, working group 16, also referred to as 802.16 [2]. To promote 802.16-compliant technologies, the Worldwide Interoperability for Microwave Access (WiMAX) Forum was founded, with more than 300 member companies. According to the WiMAX forum, 802.16 technology is attractive in a wide variety of environments, including high-speed Internet access, WiFi hotspot backhaul, cellular backhaul, public safety services, and private networks. However, in [3], it is envisaged that the first 802.16-compliant products to be deployed will very likely be aimed at providing last-mile Internet access for residential users and small and medium-sized enterprises (SMEs). Specifically, 802.16 technology will address the market segment of high-speed Internet access for the residential customers market, especially in those cases where broadband services based on DSL or cable are not available, such as rural areas or developing countries. Instead, for the SME market, 802.16 will provide a cost-effective alternative to existing solutions based on very expensive leased-line services.

The challenge for BWA networks is in providing quality of service (QoS) simultaneously to services with very different characteristics. QoS support in wireless networks is a much more difficult task than in wired networks, mainly because the characteristics of a wireless link are highly variable and unpredictable, both on a time-dependent basis and a location-dependent basis. To cope with such issues, QoS in wireless networks is usually managed at the medium access control (MAC) layer. Despite the fact that the launch of 802.16 products has already been announced on the market by several manufacturers, the research literature still lacks a sufficient number of studies that specifically address the analysis of the 802.16 MAC protocol. In [4] the author performed a hybrid analytical-simulative analysis of the effect on the system performance of several MAC mechanisms, including the fragmentation of service data units (SDUs) and the padding of OFDM symbols. The performance with the time-division duplex (TDD) mode was partially analyzed in [5]. Finally, in [6] the authors analyzed the performance of expected WiMAX compatible systems, from a physical-layer perspective.

In this article, we review and analyze the mechanisms for supporting QoS at the IEEE 802.16 MAC layer. We then analyze by simulation the performance of IEEE 802.16 in two application scenarios, which consist of providing last-mile Internet access for residential and SME subscribers, respectively. Our analysis is aimed at showing the effectiveness of the 802.16 MAC protocol in providing differentiated services to applications with different QoS requirements, such as VoIP and Web.

Results reported in this article were obtained in the framework of a project financed by Nokia to the Dipartimento di Ingegneria dell'Informazione *of the University of Pisa*

0890-8044/06/$20.00 © 2006 IEEE

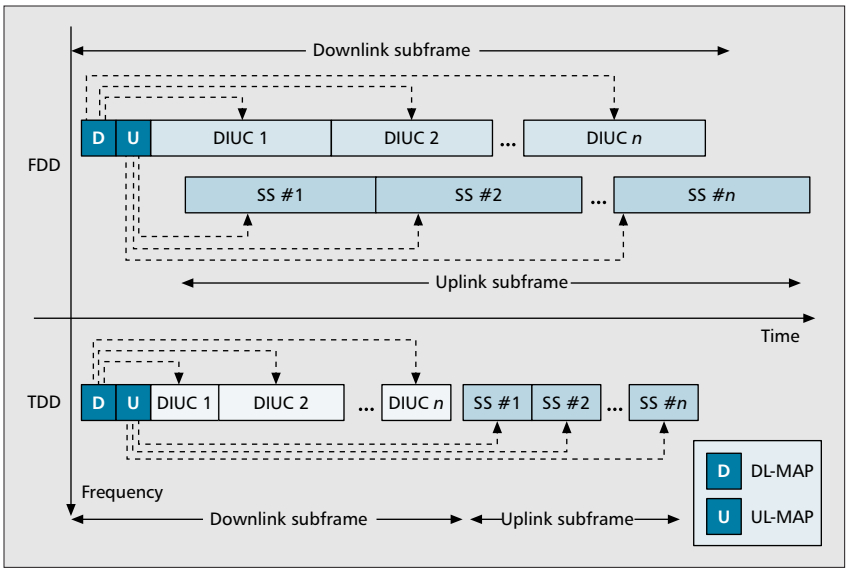

Figure 1. *Frame structure with FDD and TDD.*

QoS Support in IEEE 802.16

This section reviews the basic MAC mechanisms defined by 802.16 which are specifically related to QoS provisioning. For further details related to the 802.16 standard, see [7].

The 802.16 standard specifies two modes for sharing the wireless medium: point-to-multipoint (PMP) and mesh (optional). With PMP, the BS serves a set of SSs within the same antenna sector in a broadcast manner, with all SSs receiving the same transmission from the BS. Transmissions from SSs are directed to and centrally coordinated by the BS. On the other hand, in mesh mode, traffic can be routed through other SSs and can occur directly among SSs. Access coordination is distributed among the SSs. The PMP operational mode fits a typical fixed BWA scenario, where multiple service subscribers are served by one centralized service provider so that they can access external networks (e.g., the Internet) or services (e.g., Digital Video Broadcasting — DVB). In this study we focus on the PMP mode alone.

In PMP mode, uplink (from SS to BS) and downlink (from BS to SS) data transmissions occur in separate time frames. In the downlink subframe, the BS transmits a burst of MAC protocol data units (PDUs). Since the transmission is broadcast, all SSs listen to the data transmitted by the BS. However, an SS is only required to process PDUs that are addressed to itself or that are explicitly intended for all the SSs. In the uplink subframe, on the other hand, any SS transmits a burst of MAC PDUs to the BS in a time-division multiple access (TDMA) manner. Based on measurements at the physical layer, any SS adapts over time the interval usage code (IUC) in use, that is, modulation, rate, and forward error correction (FEC) scheme, for both downlink (downlink IUC, DIUC) and uplink (uplink IUC, UIUC) transmissions. Downlink and uplink subframes are duplexed using one of the following techniques, as shown in Fig. 1: Frequency-division duplex (FDD) is where downlink and uplink subframes occur simultaneously on separate frequencies, and time-division duplex (TDD) is where downlink and uplink subframes occur at different times and usually share the same frequency. SSs can be either full duplex (i.e., they can transmit and receive simultaneously) or half-duplex (i.e., they can transmit and receive at nonoverlapping time intervals).

The MAC protocol is connection-oriented: all data communications, for both transport and control, are in the context of a unidirectional connection. At the start of each frame, the BS schedules the uplink and downlink grants in order to meet the negotiated QoS requirements. Each SS learns the boundaries of its allocation within the current uplink subframe by decoding the UL-MAP message. On the other hand, the DL-MAP message contains the timetable of the downlink grants in the forthcoming downlink subframe. More specifically, downlink grants directed to SSs with the same DIUC are advertised by the DL-MAP as a single burst. Both maps are transmitted by the BS at the beginning of each downlink subframe, as shown in Fig. 1, for both FDD and TDD modes.

Figure 2 shows the blueprint of the functional entities for QoS support, which logically reside within the MAC layer of the BS and SSs. Each downlink connection has a packet queue (or *queue*, for short) at the BS (represented with solid lines). In accordance with the set of QoS parameters and the status of the queues, the *BS downlink scheduler* selects from the downlink queues, on a frame basis, the next service data units (SDUs) to be transmitted to SSs. On the other hand, uplink connection queues (represented in Fig. 2 by solid lines) reside at SSs.

Since the BS controls the access to the medium in the uplink direction, bandwidth is granted to SSs on demand. For this purpose, a number of different bandwidth-request mechanisms have been specified. With *unsolicited granting*, a fixed amount of bandwidth on a periodic basis is requested during the setup phase of an uplink connection. After that phase, bandwidth is never explicitly requested. A *unicast poll* consists of allocating to a polled uplink connection the bandwidth needed to transmit a bandwidth request. If the polled connection has no data awaiting transmission (*backlog*, for short), or if it has already requested bandwidth for all of its backlog, it will not reply to the unicast poll, which is thus wasted. Instead, *broadcast polls* are issued by the BS to all uplink connections. The main drawback in this mechanism is that a collision occurs whenever two or more uplink connections send a bandwidth request by responding to the same poll, in which case a truncated binary exponential backoff algorithm is employed. Finally, bandwidth requests can be *piggybacked* on a PDU. However, this mechanism is effective only if the connection has some backlog for which bandwidth reservation has already been issued.

Bandwidth requests are used on the BS for estimating the residual backlog of uplink connections. In fact, based on the amount of bandwidth requested (and granted) so far, the *BS uplink scheduler* estimates the residual backlog at each uplink connection (represented in Fig. 2 as a virtual queue by dashed lines), and allocates future uplink grants according to the respective set of QoS parameters and the virtual status of the queues. However, although bandwidth requests are per connection, the BS nevertheless grants uplink capacity to each SS as a whole. Thus, when an SS receives an uplink grant, it cannot deduce from the grant which of its connections it was intended for by the BS. Consequently, an *SS scheduler* must also be implemented within each SS MAC, in order to redistribute the granted capacity to all of its own connections (Fig. 2).

The 802.16 document clearly states that the definition of both the BS (uplink and downlink) and the SS scheduling algorithms is out of the scope of the standard, and is thus left up to the manufacturers [2, p. 139]. However, based on the above mentioned functions and mechanisms, the 802.16 MAC specifies four different *scheduling services* in order to meet the QoS requirements of multimedia applications: unsolicited grant service (UGS), real-time polling service (rtPS), non-real-time

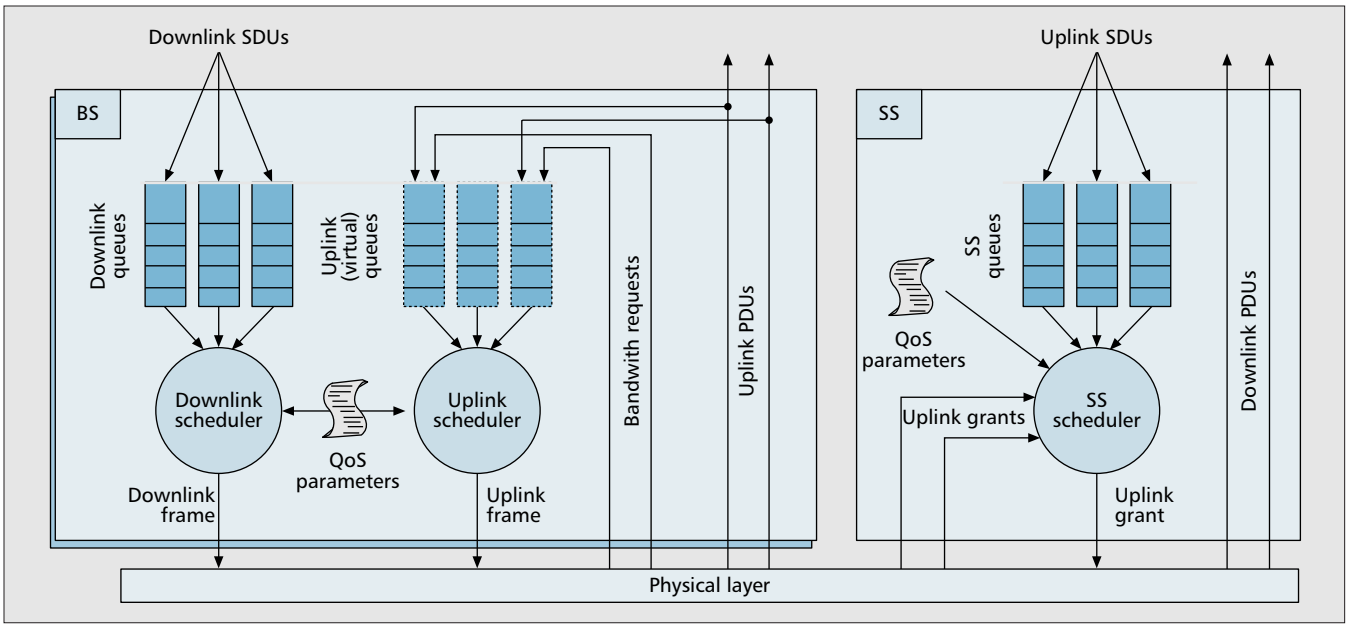

Figure 2. *QoS functions within the BS and SSs.*

polling service (nrtPS), and best effort (BE). Each scheduling service is characterized by a mandatory set of QoS parameters, which is tailored to best describe the guarantees required by the applications that the scheduling service is designed for. Furthermore, for uplink connections, it also specifies which mechanisms to use in order to request bandwidth.

UGS is designed to support real-time applications (with strict delay requirements) that generate fixed-size data packets at periodic intervals, such as T1/E1 and VoIP without silence suppression. The guaranteed service is defined so as to closely follow the packet arrival pattern (i.e., grants occurring on a periodic basis), with the base period equal to the unsolicited grant interval and the offset upper bounded by the tolerated jitter. Uplink grants are granted by the BS regardless of the current estimation of backlog; hence, UGS connections use the unsolicited granting bandwidth-request mechanism (i.e., UGS connections never request bandwidth). The grant size is computed by the BS based on the minimum reserved traffic rate, which is defined as the minimum amount of data transported on the connection when averaged over time.

rtPS is designed to support real-time applications (with less stringent delay requirements) that generate variable-size data packets at periodic intervals, such as Moving Pictures Expert Group (MPEG) video and VoIP with silence suppression. The key QoS parameters for rtPS connections are the minimum reserved traffic rate, which has the same meaning as with UGS, and the maximum latency, which upper bounds the waiting time of a packet at the MAC layer. Since the size of arriving packets with rtPS is not fixed, as it is with UGS-tailored applications, rtPS connections are required to notify the BS of their current bandwidth requirements. The BS periodically grants unicast polls to rtPS connections. The polling period may be explicitly specified as an optional QoS parameter, namely, the unsolicited polling interval. If it is not, then the BS is free to use any optimized polling scheme, so that the maximum latency requirement is met.

Unlike UGS and rtPS scheduling services, nrtPS and BE are designed for applications that do not have any specific delay requirement. The main difference between the two is that nrtPS connections are reserved a minimum amount of bandwidth (by means of the minimum reserved traffic rate parameter), which can boost performance of bandwidth-intensive applications, such as File Transfer Protocol (FTP). Both nrtPS and BE uplink connections request bandwidth by either responding to broadcast polls from the BS or piggybacking a bandwidth request on an outgoing PDU. Additionally, the BS grants unicast polls to nrtPS connections at a time-scale of one second or less.

Performance Evaluation

In this section we assess the performance of 802.16 in two of the most promising application scenarios envisaged by the WiMAX forum [3]. They consist in providing last-mile Internet access for residential and SME subscribers. In this case, the use of 2–11 GHz frequency bands is essential so that non-line-of-sight operations are allowed, thus avoiding the need to mount rooftop antennas. The most promising air interface for this environment is therefore the WirelessMAN-OFDM,[1] with a typical channel bandwidth of 7 MHz, operating in FDD duplexing mode. We assume that all SSs have full-duplex capabilities, and that the frame duration is 10 ms.

As mentioned above, neither the BS nor SS schedulers are specified by the standard. Therefore, it was necessary to specify both these schedulers for the purpose of carrying out the system evaluation. Since a minimum reserved rate is the basic QoS parameter negotiated by a connection within a scheduling service, the class of latency-rate [8] scheduling algorithms is particularly suited for implementing the schedulers in the 802.16 MAC. Specifically, within this class, we selected deficit round robin (DRR) as the downlink scheduler to be implemented at the BS [9], since it combines the ability of providing fair queuing in the presence of variable length packets with the simplicity of implementation. In fact, it can exhibit O(1) complexity, provided that specific allocation constraints are met. In particular, DRR requires a minimum rate to be reserved for each packet flow being scheduled. Therefore, although not required by the 802.16 standard, BE connections should also be guaranteed a minimum rate. This fact can be exploited in order to both avoid BE traffic starvation in overloaded scenarios, and let BE traffic take advantage of the excess bandwidth which is not reserved for the other scheduling services. On the

[1] *WirelessMAN-SCa development has been discontinued, while WirelessMAN-OFDMA mainly addresses nomadic and mobile applications.*

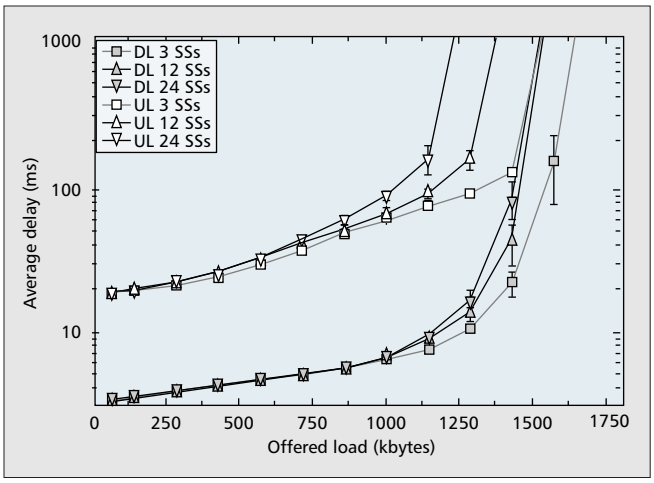

Figure 3. *Residential: average delay vs. offered load.*

other hand, DRR assumes that the size of the head-of-line packet is known at each packet queue; thus, it cannot be used by the BS to schedule transmissions in the uplink direction. In fact, with regard to the uplink direction, the BS is only able to estimate the overall amount of backlog of each connection, but not the size of each backlogged packet. Therefore, we selected weighted round robin (WRR) [10] as the uplink scheduler in our 802.16 simulator. Like DRR, WRR belongs to the class of rate-latency scheduling algorithms. Finally, we decided to implement DRR as the SS scheduler, because the SS knows the sizes of the head-of-line packets of its queues.

Although accurately modeling channel conditions is a key aspect of simulation for network provisioning and resource management, this study only focuses on the functions and mechanisms available at the MAC layer to provide QoS; thus, we assumed ideal channel conditions, that is, no packet corruption due to the wireless channel. In addition, we simulated a system in which the set of active connections does not change. The simulation scenarios consisted in several SSs located at various distances from the BS. In realistic conditions, the nearer the SS is to the BS, the more robust the physical profile. We assumed that the SSs employ the following modulations: QPSK 3/4, 16-QAM 3/4, and 64-QAM 3/4, which are evenly partitioned among SSs.

The metrics used for assessing the performance of 802.16 are the average packet-transfer *delay* and the *delay variation*. Packet-transfer delay (delay, for short) is defined as the time between the arrival of the packet at the MAC transmit buffer of the source node (SS/BS) and the time that this packet is completely delivered to the upper protocol layer of the destination node (BS/SS). The delay variation is defined as the difference between the maximum packet-transfer delay and the packet transmission delay, that is, the time it takes for a packet of minimum length to be transmitted over the air from the source to the destination.

The simulations were carried out by means of a prototypical simulator of the IEEE 802.16 protocol. The simulator was event-driven and was developed using the C++ programming language. Specifically, the MAC layers of the SSs and the BS were implemented, including all functions for uplink/downlink data transmission. In all the simulation runs, we estimated the 95 percent confidence interval of each performance measure (a confidence interval is not drawn whenever it is negligible).

Residential Scenario

The *Residential* scenario consists of a BS providing Internet access to its subscribers, by means of a variable number of BE connections evenly distributed among the SSs. We considered three cases, which differ in terms of the number of SSs served by the BS (3, 12, and 24 SSs). Internet traffic is modeled as a Web traffic source [11]. Packet sizes are drawn from a Pareto distribution with cutoff (shape factor = 1.1, mode = 4.5 KB, cutoff threshold = 2 MB), while packet interarrival times are distributed exponentially (average = 5 s), which yields an average load of 25 KB/s.

Figure 3 shows the average delay when the number of connections, and hence the offered load, increases. Since the BS knows the current status of downlink queues, as soon as a downlink packet is enqueued at the BS, it is immediately eligible for transmission to its intended SS. As long as the system is underloaded (i.e., the offered load is lower than 1250 KB/s), the connection queues are almost always empty. Thus, the average delay of downlink packets is almost constant. However, the average delay increases sharply as soon as the system starts to get overloaded (i.e., the offered load is higher than 1250 KB/s), because the BS is not able to fully serve the backlog of downlink connections before new packets are enqueued. Figure 3 also highlights that the average delay of uplink traffic is higher than that of downlink traffic. In fact, any SS has to request bandwidth to the BS, in order to receive an uplink grant to transmit its backlog. More specifically, if the incoming packet finds a nonempty queue, then the SS piggybacks the bandwidth request on the first outgoing SDU from that queue. Otherwise, the SS requests bandwidth by responding to a broadcast poll from the BS. However, in our 802.16 simulator the BS reserves for broadcast polls the portion of the uplink subframe that has not been granted for data transmission. Therefore, the average time needed by an SS to request bandwidth by using contention increases with the offered load. This justifies the fact that the uplink curves increase earlier than the downlink curves.

From Fig. 3 we can also derive information about the maximum achievable throughput. In fact, for each curve, there is an offered load so that the average delay increases almost asymptotically. In this case, connection queues are almost always full; hence, it is not possible to obtain a higher system throughput by increasing the offered load further. Note that the maximum achievable throughput decreases when the number of SSs increases, for both downlink and uplink curves. In the uplink case, this is because the higher the number of different SSs granted in a frame, the higher the number of physical preambles transmitted[2] and this reduces (from the payload standpoint) the amount of available uplink capacity. On the other hand, in the downlink case, having a higher number of SSs entails a higher MAC overhead due to the transmission of uplink and downlink MAPs, and this consumes downlink capacity.

SME Scenario

We now consider the *SME* scenario, which involves a BS providing several enterprise customer premises with three different types of services: VoIP, videoconference, and data. We assume that each SS has four VoIP sources multiplexed into an rtPS connection, two videoconference sources multiplexed into an rtPS connection, and a BE connection loaded with data traffic at 25 KB/s. The number of SSs increases from 6 to 48. We model VoIP traffic as an ON/OFF source [12]: during the ON periods, packets 66 bytes in length are generated at fixed intervals of 20 ms, so that a GSM adaptive multirate encoder at 3.3 KB/s is simulated [13]. The OFF periods, on the other hand, correspond to the intervals during which no voice activity is detected, and thus packets are not generated. The duration of the ON and OFF periods is distributed exponentially

[2] *Recall that any burst of uplink PDUs must be prepended a physical preamble for synchronizing the BS with the SS's transmission.*

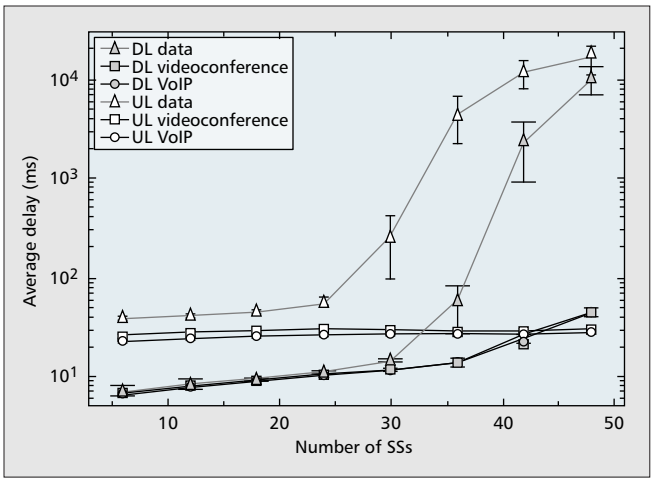

Figure 4. *SME: average delay vs. number of SSs.*

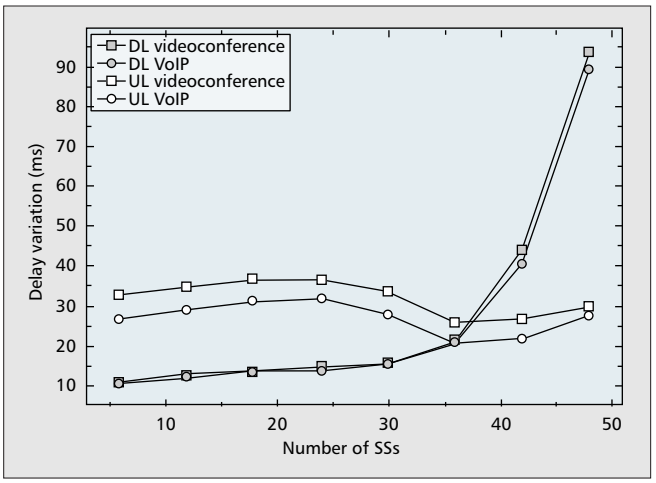

Figure 5. *SME: delay variation vs. number of SSs.*

(average duration of ON periods = 1.34 s, average duration of OFF periods = 1.67 s). On the other hand, a videoconference source consists of a VoIP and a video source. VoIP is modeled as described above, employing a G.723.1 codec at 2 KB/s instead of the GSM codec. The packet size of the video source traffic [14] is drawn from a lognormal distribution (average = 4.9 bytes, standard deviation = 0.75 bytes), and the packet interarrival time is drawn from a normal distribution (average = 33 ms, standard deviation = 10 ms). We assume that the BS grants a unicast poll to each VoIP and videoconference connection every 20 ms. Finally, for data traffic we use the same Web traffic source model as in the previous scenario.

Figure 4 shows the average delay of each type of traffic, when the number of SSs increases from 6 to 48. As can be seen, the downlink curves increase smoothly when the system is underloaded (i.e., the number of SSs is smaller than or equal to 30). Under this condition, all downlink queues are almost always empty. Thus, when a packet is received by the BS, the downlink scheduler will likely serve it in the subsequent downlink subframe. Therefore, there is no service differentiation between connections with data and multimedia traffic. However, when the system becomes overloaded (i.e., the number of SSs becomes greater than 30), the average delay of data traffic increases much more sharply than that of multimedia traffic. This is due to the way in which capacity has been provisioned to the different connections. Specifically, scheduling algorithms have been configured so that rtPS connections have a reserved rate equal to the mean rate of VoIP and videoconference applications, respectively. That is, the reserved rate for VoIP connections is 13.2 KB/s, which accounts for the four VoIP sources multiplexed into the same connection. On the other hand, the reserved rate for videoconference connections is 16 KB/s, which accounts for the two videoconference sources multiplexed into the same connection. Finally, BE connections are reserved a rate of 10 B/s. Note that the rate guaranteed to BE connections is negligible with respect to the rate guaranteed to rtPS connections, and this justifies the different performance of the BE and rtPS connections, respectively.

With regard to uplink connections, the average delay of rtPS connections is almost constant when the number of SSs increases. This behavior can be justified as follows. The BS grants a unicast poll to each rtPS connection every 20 ms. Consequently, the time between the arrival of an rtPS SDU and the transmission of the corresponding bandwidth request does not depend on the number of SSs, and is bounded to 20 ms. Since this time interval is the main component in the delay of uplink rtPS connections, the average delay is almost constant when the number of SSs increases. On the other hand, SSs have to request bandwidth for BE connections on a contention basis. Thus, when the system is underloaded (i.e., the number of SSs is smaller than or equal to 24), the average delay of uplink BE connections is almost equal to the average time that those connections need to request bandwidth. On the other hand, as soon as the system gets overloaded (i.e., the number of SSs is greater than 24) the average delay of BE connections increases remarkably, whereas the average delay of rtPS connections remains low, even with high traffic rates. Again this is because, as mentioned above, the reserved rate of BE connections is negligible with respect to the one guaranteed to rtPS connections.

Since VoIP and videoconference are interactive multimedia applications, a relevant performance index is also the 99th percentile of the delay variation. Figure 5 shows the 99th percentile of the delay variation of VoIP and videoconference traffic, when the number of SSs increases from 6 to 48. We firstly analyze the performance of downlink traffic. In this case, the curves of VoIP and videoconference connections almost coincide, even though those applications generate traffic at different rates. This is because both VoIP and videoconference connections have a reserved rate equal to the mean rate of their respective applications. When the number of SSs increases from 6 to 36, the delay variation increases smoothly from 10 to 20 ms. Under these conditions the system is underloaded; thus, the BS is able to serve almost all VoIP and videoconference packets with a maximum delay of 20 ms, which is the packet interarrival time of VoIP packets. From Fig. 4, we note that the data traffic also experiences a relatively small average delay. On the other hand, when the number of SSs increases further, the BS downlink scheduler is not able, on average, to schedule each VoIP packet before the next one is generated from the same application. Hence, the delay variation of multimedia traffic increases sharply.

We will now comment on the performance of uplink multimedia traffic. As also highlighted by Fig. 5, unlike downlink traffic, the delay variation of VoIP traffic is always smaller than that of videoconference traffic. This is because the traffic generated by video source applications is much more variable than the traffic generated by VoIP applications, in terms of interarrival times and packet sizes. Thus, the difference between the *virtual* backlog computed by the BS and the *real* backlog buffered at the respective SS connection queue in the VoIP case is lower, on average, than in the videoconference case (this is confirmed by the results discussed at the end of this section). Consequently, videoconference traffic experiences a slightly higher delay variation than VoIP traffic. The delay variation of uplink traffic is higher than that of downlink traffic when the system is underloaded, whereas it is lower when the system is overloaded. Moreover, note that the delay variation decreases when the system gets overloaded (i.e., the number of SSs

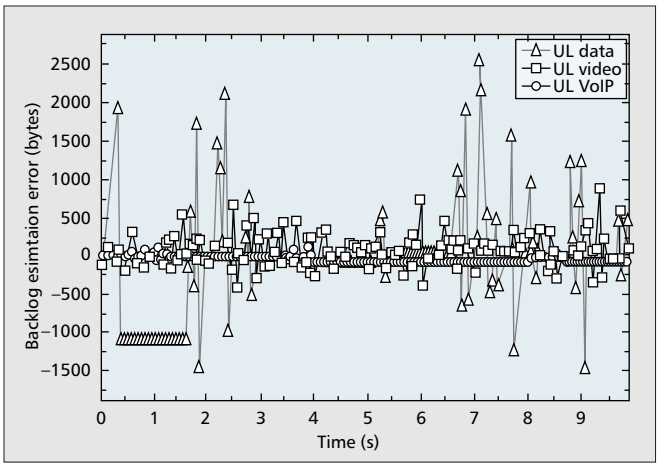

Figure 6. *SME: backlog estimation error versus time.*

ranges between 24 and 36). This result can be justified as follows. When the system is underloaded, the BS almost always issues an uplink grant immediately after receiving a bandwidth request. When the next packet is generated by the application, the SS has to wait until the next unicast poll from the BS before it can request bandwidth. On the other hand, when the system becomes overloaded, a new packet is generated by an application before the BS issues a grant to serve a previous packet. SSs are thus able to piggyback the bandwidth request for the just enqueued packet on the first outgoing packet from the same connection, for which a bandwidth reservation has been already issued. In this way, a connection anticipates the unicast poll from the BS. However, when the number of SSs is greater than 36, the above phenomenon cannot compensate further for the offered load increase; hence, the delay-variation curve starts increasing again with the number of SSs.

Before concluding our analysis, we would like to briefly examine the bandwidth-request mechanisms used by rtPS and BE connections. To this aim, Fig. 6 shows a time diagram of the backlog estimation error of the BS for each type of traffic, with 48 SSs. We define the backlog estimation error as the difference between the BS's estimate of the backlog of a connection (as acquired via bandwidth requests), and the actual backlog of that connection on the SS. Recall that the BS has to estimate the backlog of uplink queues, which reside at the SSs. As can be seen, the curve that experiences the highest variability is that of data traffic, which is served using the BE scheduling service. This is because the time needed for an SS to request bandwidth for a BE connection is unpredictable. In fact, the time depends on the backoff procedures running on all the SSs that have at least one BE connection with a pending contention-based bandwidth request. On the other hand, the time for requesting bandwidth for an rtPS connection is bounded by the unicast polling time (20 ms, in this case).

Conclusions

In this article we have assessed, via simulation, the performance of an IEEE 802.16 system under two traffic scenarios. The first one (residential scenario) dealt with data (non-QoS) traffic only, and was thus managed by the BE scheduling service. Our results have shown that the average delay of the uplink traffic is higher than that of the downlink traffic. Furthermore, the former increases more sharply than the latter with the offered load. This behavior can be explained by means of both the bandwidth-request mechanism and the overhead introduced by physical preambles. In the second scenario (SME scenario), on the other hand, we have shown the service differentiation, in terms of delay, between data (served via BE) and multimedia traffic (served via rtPS). This is achieved because scheduling in 802.16 is controlled by the BS in both the downlink and uplink directions. Therefore, it is possible to employ scheduling algorithms that have been proposed for wired environments, which are able to provide QoS guarantees. In our simulations, we have evaluated the DRR and WRR scheduling algorithms as possible candidates for algorithms to be implemented in a production system. Moreover, we have shown that requesting bandwidth using unicast polls yielded a better estimation of the connection requirements at the BS, as compared to requesting bandwidth on a contention basis by responding to broadcast polls.

References

[1] International Telecommunication Union, "ITU Internet Reports: Birth of Broadband," Sept. 2003.
[2] IEEE 802.16-2004, "IEEE standard for Local and Metropolitan Area Networks — Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Oct. 2004.
[3] WiMAX forum, "Business Case Models for Fixed Broadband Wireless Access based on WiMAX Technology and the 802.16 Standard," Oct. 2004.
[4] C. Hoymann, "Analysis and Performance Evaluation of the OFDM-based Metropolitan Area Network IEEE 802.16," *Comp. Net.*, vol. 49, no. 3, Oct. 2005, pp. 341–63.
[5] O. Gusak, N. Oliver, and K. Sohraby, "Performance Evaluation of the 802.16 Medium Access Control layer," *Proc. ISCIS 2004*, Kemer-Antalya, Turkey, Oct. 27-29, 2004, pp. 228–37.
[6] A. Ghosh *et al.*, "Broadband Wireless Access with WiMax/802.16: Current Performance Benchmarks and Future Potential," *IEEE Comm. Mag.*, vol. 43, no. 2, Feb. 2005, pp. 129–36.
[7] C. Eklund *et al.*, "IEEE standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access," *IEEE Comm. Mag.*, vol. 40, no. 6, June 2002, pp. 98–107.
[8] D. Stiliadis and A. Varma, "Latency-Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms," *IEEE/ACM Trans. Net.*, vol. 6, Oct. 1998, pp. 675–89.
[9] M. Shreedhar and G. Varghese, "Efficient Fair Queuing using Deficit Round Robin," *IEEE Trans. Net.*, vol. 4, no. 3, June 1996, pp. 375–85.
[10] M. Katevenis, S. Sidiropoulos, and C. Courcoubetis, "Weighted Round-Robin Cell Multiplexing in a General-Purpose ATM Switch Chip," *IEEE JSAC*, vol. 9, no. 8, Oct. 1991, pp. 1265–79.
[11] Motorola, "Evaluation Methods for High Speed Downlink Packet Access (HSDPA)," TSG-R1 document, TSGR#14(00)0909, 2000.
[12] P.T. Brady, "A Model for Generating On-off Speech Patterns in Two-way Conversation," *Bell Syst. Tech. J.*, vol. 48, Sept. 1969, pp. 2445–72.
[13] Cisco Press, "Traffic Analysis for Voice over IP," Nov. 2001.
[14] D. P. Heyman, A. Tabatabai, and T. V. Lakshman, "Statistical Analysis and Simulation Study of Video Teleconference Traffic in ATM Networks," *IEEE Trans. Circuits Sys. Video Tech.*, vol. 2, no. 1, Mar. 1992, pp. 49–59.

Biographies

CLAUDIO CICCONETTI (claudio.cicconetti@iet.unipi.it) was graduated in computer systems engineering from the University of Pisa, Italy, in October 2003. He is currently pursuing his Ph.D. degree at the same university. His research interests include quality of service in wireless networks, medium access protocols for mobile computing, and mesh networks. He is involved in the EuQoS (End-to-end Quality of Service support over heterogeneous networks) project, which participates in the EU Information Society Technologies (IST) Program.

CARL EKLUND (carl.eklund@nokia.com) is a Principal Engineer with Nokia Research Center, Helsinki, Finland. He chaired the MAC Task Group that developed the IEEE 802.16 medium access control protocol and served as a lead MAC editor of IEEE Standard 802.16-2001. He received his M.Sc. in engineering physics from Helsinki University of Technology in 1996. He currently is heading the program for the research and standardization activities related to WiMAX and IEEE 802.16 in Nokia.

LUCIANO LENZINI (l.lenzini@iet.unipi.it) holds a degree in physics from the University of Pisa, Italy. In 1994 he joined the Department of Information Engineering of the University of Pisa as a Full Professor. His current research interests include the design and performance evaluation of MAC protocols for wireless networks and the QoS provision in integrated and differentiated services networks.

ENZO MINGOZZI (e.mingozzi@iet.unipi.it) has been an associate professor at the Faculty of Engineering of the University of Pisa, Italy since January 2005. He received Laurea (cum laude) and Ph.D. degrees in Computer Systems Engineering in 1995 and 2000, respectively, from the University of Pisa. His current research interests focus on the design and analysis of MAC protocols for wireless networks, and QoS provisioning and service integration in computer networks.