2009

Quality of service in wireless local and metropolitan area networks

Mohamed K. Watfa
University of Wollongong in Dubai, mwatfa@uow.edu.au

Haidar Safa

Publication Details

Watfa, M. & Safa, H. 2009, Quality of service in wireless local and metropolitan area networks, in J. Ma, L. Yang & Y. Zhang (eds), Unlicensed Mobile Access Technology: Protocols, Architecture, Security, Standards and Applications, Auerbach Publications, Boca Raton, Florida, 163-185.

Chapter 9

Quality of Service in Wireless Local and Metropolitan Area Networks

Haidar Safa and Mohamed K. Watfa

CONTENTS

Wireless technology has shown tremendous growth and acceptance as a solution for both wireless local area networks and wireless metropolitan area networks. Ā e use of multimedia applications over IP with quality-of-service (QoS) support is now a reality in corporate networks and is rapidly expanding to the wireless networks. In this chapter, the state-of-the-art in supporting the QoS concepts in the IEEE 802.11-based wireless local area networks and the IEEE 802.16-based wireless metropolitan area networks is presented. Āe chapter is divided into two parts. Ā e first part starts by describing the IEEE 802.11 standard that supports only best effort (BE) services before examining the new IEEE 802.11e that is introduced to support sophisticated services that guarantee QoS attributes such as bandwidth, delay, and jitter. Āe second part explores the QoS in the wireless metropolitan area networks as introduced in IEEE 802.16 standard and its IEEE 802.16e amendment. Open research issues pertaining to realizing QoS in these networks are identified and some of the solutions that are proposed to address these challenges are also presented.

9.1 Quality of Service in Wireless Local Area Networks

In this section we present the architectures, basic elements, and the QoS of wireless local area networks as introduced in IEEE 802.11 and IEEE 802.11e standards [1,2]. In this context, we describe two access mechanisms, the distributed coordination function (DCF) and the point coordination function (PCF), of the IEEE 802.11 medium access control (MAC) layer and the hybrid coordination function (HCF) access mechanism introduced in the IEEE 802.11e standard.

9.1.1 IEEE 802.11 Wireless Network Architectures

Āe IEEE 802.11 standard defines two basic architectures for wireless local area networks: Infrastructure-Based Architecture and Ad Hoc Architecture. In the Infrastructure-Based Architecture, the wireless network consists of access points (AP) and a set of mobile stations. Āe mobile stations and the AP that are within the same radio coverage form a basic service set (BSS) as shown in Figure 9.1a. If two stations in the same BSS want to communicate, then the communications flow from the source station to the AP and then from the AP to destination station. Ā e BSS is the basic building block of IEEE 802.11 LAN. Several BSSs can be connected via a distribution system to form a single network called extended service set. Āe Ad Hoc Architecture contains no APs as shown in Figure 9.1b. In this architecture, mobile stations using the same frequency and situated in the transmission range of each other may form an independent BSS (IBSS) and communicate directly.

9.1.2 IEEE 802.11 MAC Layer

Āe IEEE 802.11 MAC layer defines two basic access mechanisms: the mandatory contention-based DCF, which offers an asynchronous data service, and the optional contention-free PCF that is built on top of the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA)-based

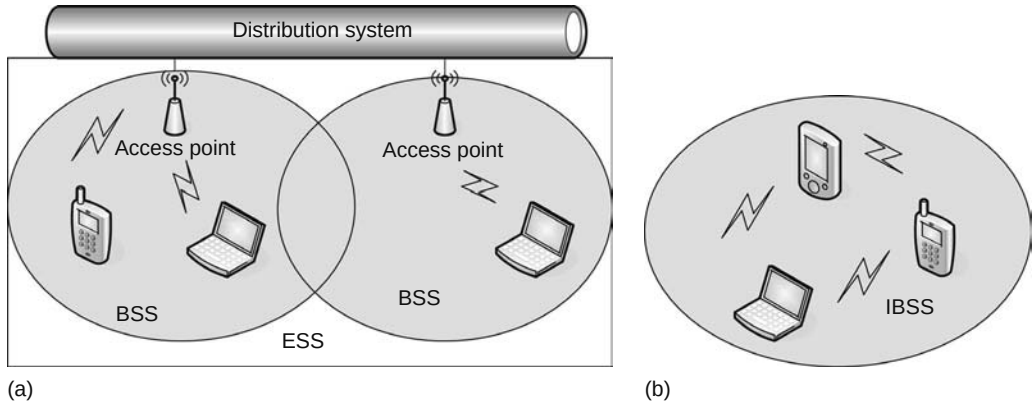

Figure 9.1 (a) Architecture of 802.11 infrastructure-based network, (b) Architecture of 802.11 ad hoc network.

DCF, as shown in Figure 9.2, to offer both asynchronous and time-bounded services. Ā e DCF is based on the well-known CSMA/CA MAC access algorithm [1].

9.1.2.1 Carrier Sense Multiple Access with Collision Avoidance

Ā e CSMA/CA Protocol is designed to reduce the collision probability between multiple stations accessing a shared medium [1]. A station wants to transmit senses if the medium is idle for a specific period. If it is, it starts transmitting. Otherwise, the station waits till the medium becomes idle again then resumes its operation as explained later.

Āe highest probability of a collision exists when the medium becomes idle following a busy medium because multiple stations could have been waiting for the medium to become idle again. Āi s situation necessitates a random backoff procedure to resolve medium contention conflicts. Another type of collision may occur when two stations hidden from each other want to communicate with a third station. Āe two stations may sense an idle medium then transmit a frame that may cause a collision at the receiver. To avoid such collision, the CSMA defines two reservation information packets that announce the impending use of the medium prior to transmitting the actual data frame. Āese packets are the Request-to-Send (RTS) and the Clear-to-Send (CTS). Āe RTS and CTS frames contain a duration field that defines the period during which the medium is

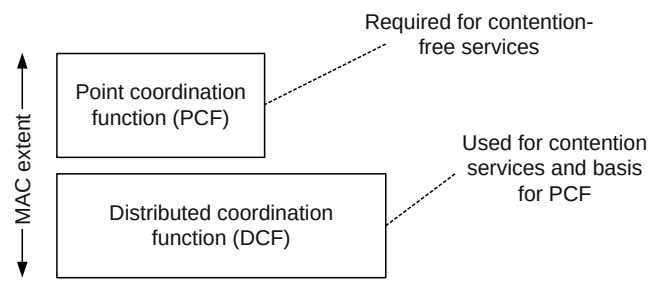

Figure 9.2 Access mechanisms of MAC layer.

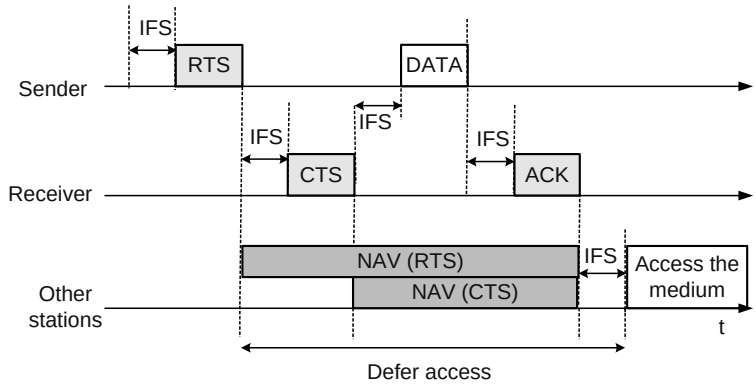

Figure 9.3 CSMA/CA.

to be reserved to transmit the actual data frame and the returning acknowledgment (ACK) frame. All stations within the reception range of either the station that transmits the RTS or the station that transmits the CTS learn of the medium reservation and set their network allocation vector (NAV) accordingly as shown in Figure 9.3. Āus, the NAV maintains a prediction of future traffic on the medium based on the duration information that is announced in the RTS/CTS frames.

In the example shown in Figure 9.3, a station that wants to transmit senses an idle channel for a period (inter-frame space [IFS] is explained in the next subsection) then transmits an RTS packet to the destination. All the nondestination neighboring stations that receive the transmission set their NAV to the duration announced in the RTS. Āe destination station replies to the RTS originator by transmitting a CTS packet after sensing an idle channel for a shorter period. After this transmission, the neighbors of the CTS sender will know about the time needed to complete the frame transmission and set their NAV accordingly. Āen the data is transmitted and acked. All stations that have a NAV value larger than zero must refrain from using the medium. We may think about the NAV as a counter, which counts down to zero at a uniform rate. A NAV value of zero is an indication of an idle medium.

9.1.2.2 Inter-Frame Space

Āe minimum time interval between frames is called the IFS. A station determines that the medium is idle through the use of the carrier-sense function for the interval specified. Different IFSs are defined in IEEE 802.11 to provide priority levels for access to the wireless medium as shown in Figure 9.4; they are listed in order, from the shortest to the longest: short inter-frame space (SIFS), PCF inter-frame space (PIFS), and DCF inter-frame space (DIFS) [1].

Āe SIFS is used when a station has seized the medium and needs to keep it for the duration of the completion of the frame exchange sequence. Using the smallest gap between transmissions within the frame exchange sequence prevents other stations, which are required to wait for a longer gap for an idle medium, from attempting to use the medium, thus giving priority to complete the frame exchange sequence that is in progress. SIFS are mostly used for an ACK frame, a CTS frame, the second or subsequent fragment burst, and by a station responding to any polling by the PCF.

Āe PIFS is used only by stations operating under the PCF to gain priority access to the medium at the start of the contention-free period (CFP).

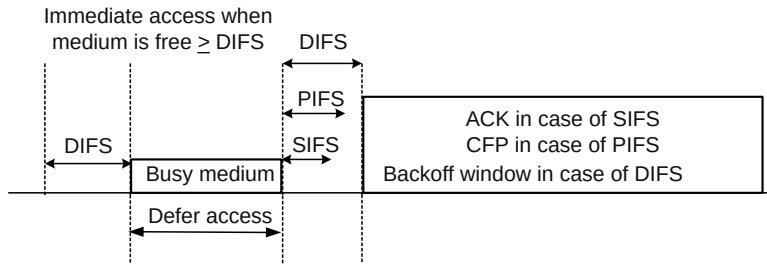

Figure 9.4 IFS types.

Āe DIFS is used by stations operating under the DCF to transmit data and management frames. A station using the DCF is allowed to transmit if its carrier-sense mechanism determines that the medium is idle for a DIFS period and its backoff time has expired.

9.1.2.3 Distributed Coordination Function

Āe DCF is the fundamental access method of the IEEE 802.11 MAC layer. It is implemented in all stations, for use within both infrastructure and ad hoc architectures [1]. Ā e DCF access technique employs a contention window (CW)-based channel access function and uses the CSMA/CA Access Protocol to avoid collision in the the event of two or more stations attempting to transmit simultaneously. Under DCF, a station that intends to transmit must sense an idle medium for a DIFS period then selects a backoff timer (time slot) within a backoff window. Ā e backoff timer is decreased only when the medium is idle; it is frozen when another station is transmitting, as shown in Figure 9.5. Each time the medium becomes idle, the station waits for a DIFS period then starts continuously decrementing the backoff timer. As soon as the backoff timer expires, the station is authorized to access the medium and transmit. Ā e backoff timer is derived from a uniform distribution over the interval [0, CW], where CW, the contention window size, is a value between $[CW_{min}, CW_{max}]$.

Āe set of CW values shall be sequentially ascending integer powers of 2, minus 1. At the very first transmission attempt, CW value is equal to the initial backoff window size CW_{min}. For every unsuccessful transmission, the value of CW + 1 is doubled until CW_{max} is reached. After transmitting a frame, the station expects to receive an ACK from the destination station following SIFS time period. If the acknowledgment is not received, the sender assumes that the transmitted frame was collided, so it schedules a retransmission and enters the backoff process again. After every successful transmission, the CW is reset to CW_{min}.

Ā e DCF employs a discrete time backoff scale. Ā e time that immediately follows the DIFS is slotted and the station is permitted to transmit only at the beginning of each slot. Ā e length of the slot is set equal to the time needed at any station to detect the transmission of a packet from any other station. More precisely, a slot time is defined in the standard as aCCATime + aRxTxTurnaroundTime + aAirPropagationTime + aMACProcessingDelay where aCCATime is the minimum time the clear channel assessment mechanism has available to assess the medium within every time slot to determine whether the medium is busy or idle; aRxTxTurnaroundTime is the maximum time the physical layer requires to change from receiving to transmitting the start of the first symbol; aAirPropagationTime is the anticipated time it takes a transmitted signal to go

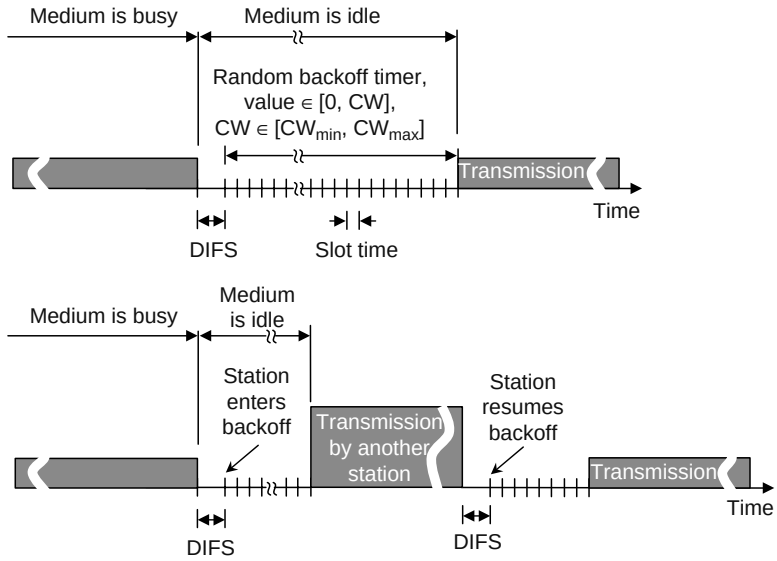

Figure 9.5 IEEE 802.11 DCF channel access.

from the transmitting station to the receiving station; aMACProcessingDelay is the nominal time that the MAC layer uses to process a frame and prepare a response to the frame.

9.1.2.4 Point Coordination Function

Ā e PCF provides a time-bounded service and is especially utilized for asynchronous data, voice, and mixed applications (voice, data, video) [1]. It is a polling-based contention-free MAC access mechanism, used in a wireless local area network that operates in an infrastructure mode where APs are used as point coordinators (PC). Ā e PCF is built on top of the CSMA/CA-based DCF access mechanism. It controls the frame transfers during a CFP. Āe CFP should alternate with the contention period (CP) in which the DCF controls the frame transfers as shown in Figure 9.6 [1]. Āus, PC allows contention and contention-free mechanisms to coexist. Each CFP should begin with a Beacon frame. Ā e CFPs occur at a defined repetition rate that is synchronized with the beacon interval. Ā e contention-free repetition rate (CFPRate) is defined as a number of delivery traffic indication message (DTIM) intervals where DTIM interval itself is a number of beacon intervals.

At the nominal beginning of each CFP, the PC senses the medium and gains control of it by waiting a PIFS period between transmissions. When the medium is determined to be idle for one PIFS period, the PC transmits a Beacon frame, which specifies the maximum time needed starting from the transmission of this beacon to the end of this CFP. All stations in the BSS set their NAVs to the duration value of the CFP. Ā is prevents contention by preventing transmissions by other stations. Ā e PC transmits a contention-free end frame at the end of each CFP. Stations that receive this frame reset their NAVs.

After the initial Beacon frame, the PC waits for a SIFS period, and then transmits a Data frame, a Polling frame, a contention-free end frame, or a combination of these frames. Āe Contention-Free Transfer Protocol is based on a polling scheme controlled by a PC operating at the AP of the

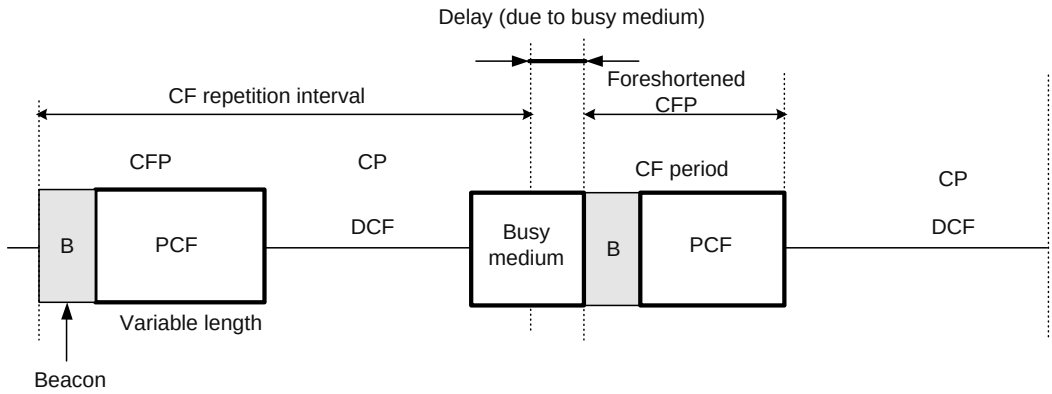

Figure 9.6 IEEE 802.11 PCF channel access.

BSS. During a CFP, the PC maintains a list of registered stations and polls them accordingly. A station can start transmitting only after it is polled. Āe size of each Data frame is bounded by the maximum MAC frame size. Stations receiving directed, error-free frames from the PC are expected to acknowledge the frame after a SIFS period. With PCF, stations are allowed to transmit even if the frame transmission cannot finish before the start of the next CF repetition interval. Āe duration of the beacon to be sent defers the transmission of data frames during the next CFP, as shown in Figure 9.6.

9.1.3 QoS in IEEE 802.11e MAC Layer

Āe new IEEE 802.11e standard provides an HCF through the services of DCF as shown in Figure 9.7 [2]. Ā e HCF combines and enhances aspects of the access methods to provide QoS-stations (QSTAs) with prioritized and parameterized QoS access to the wireless medium, while continuing to support non-QoS stations for BE transfer. Ā e HCF is compatible with both DCF and PCF. It defines two medium access mechanisms: a contention-based channel access referred to as enhanced distributed channel access (EDCA), and controlled channel access referred to as HCF controlled channel access (HCCA).

9.1.3.1 Enhanced Distributed Channel Access

Āe IEEE 802.11 DCF access mechanism can support only the BE services [3,4]. In DCF mode, all the stations compete for the resources and channel with the same priorities. Ā ere is no differentiation mechanism to guarantee bandwidth, packet delay, and jitter for high-priority traffic or multimedia flows [5]. Āe EDCA is proposed in IEEE 802.11e to support prioritized QoS services [2]. It provides differentiated, distributed access to the wireless medium for QoS stations using eight different user priorities (UPs) that are shown in Table 9.1. Ā e user priority is a value assigned to a data packet in the layers above the MAC layer to indicate how the packet is to be handled. At the MAC layer, EDCA introduces four different first-in first-out queues, called access categories (ACs), and multiple independent backoff entities that are shown in Figure 9.8 [2]. Ā e eight traffic priorities are mapped into four queues (ACs) as shown in Table 9.1. Ā us, four backoff entities exist in every 802.11e station. Each AC queue works as an independent DCF station and uses its own

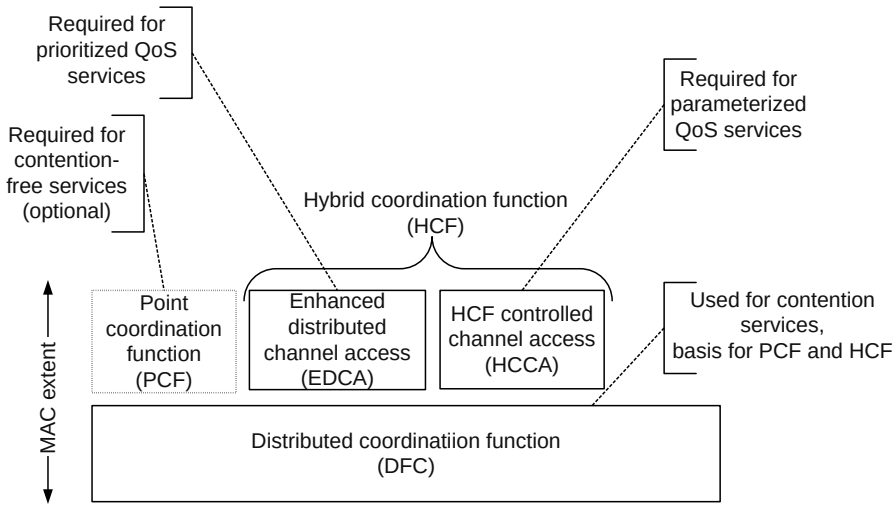

Figure 9.7 MAC architecture in 802.11e.

contention parameters such as CW_{min}, CW_{max}, and arbitrary inter-frame space (AIFS), as shown in Figure 9.8. Ā e AIFS is introduced in EDCA in place of DIFS in DCF. Each AIFS has arbitration length that is computed as follows: AIFS[AC] = SIFS + AIFSN[AC] × slot_time where AIFSN[AC] is called the AIFS number.

Similar to a DCF station, each AC starts a backoff timer after detecting an idle channel for a time interval equal to an AFIS length. Āe backoff value is chosen to be a random number between $1, CW + 1$, where CW is initially set to CW_{min} and increases whenever collision occurs up to CW_{max}, CW increases in accordance with the following equation [5]:

$$CW_{new}[AC] = (CW_{current}[AC] + 1) \times 2 - 1$$

Table 9.1 User Traffic Priorities Mapped to ACs

User Priority from Lowest to Highest	*Designation*	*Access Category*
1	BK (Background)	AC_BK
2	BK (Background)	AC_BK
0	BE (Best-effort)	AC_BE
3	EE (Video/excellent-effort)	AC_BE
4	CL (Video/controlled load)	AC_VI
5	VI (Video)	AC_VI
6	VO (Voice)	AC_VO
7	NC (Network control)	AC_VO

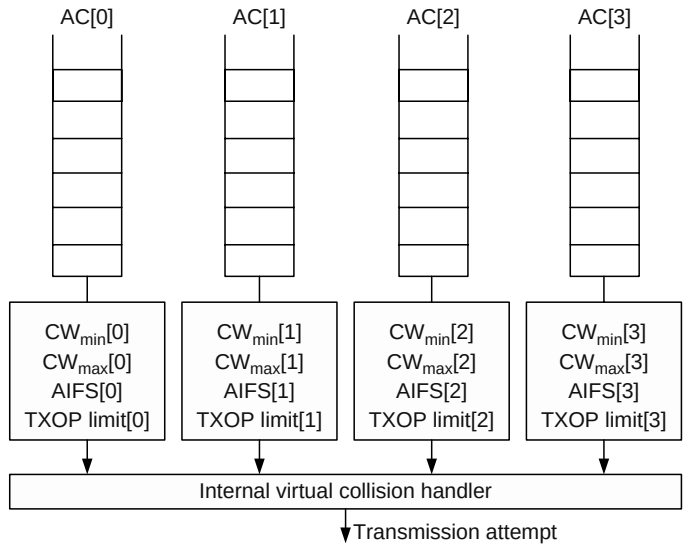

Figure 9.8 AC queues.

Whenever a station seizes the channel, it can transmit for a transmission opportunity time interval (TXOP). A TXOP is defined by its starting time and duration. Ā e duration of TXOP is limited by a parameter referred to as TXOPlimit. In case of successful transmission, the CW value of the AC queue is reset to CW_{min}.

Additionally, the purpose of using different contention parameters for different queues is to give the low-priority traffic the longer waiting time than a high-priority trafic. Āus, the high-priority traffic has values for AIFS, CW_{max}, and CW_{min} smaller than those of the low-priority traffic as shown in Table 9.2. Consequently, the higher priority traffic will enter the CP and access the wireless medium earlier than the lower priority traffic.

Note that the backoff timers of different ACs in one QoS station are randomly generated and may reach zero simultaneously. Ā is can cause an internal collision. In such a case, a virtual scheduler inside every QoS station, as shown in Figure 9.8, allows only the highest-priority AC to transmit frames [4].

Table 9.2 User Traffic Priorities Mapped to ACs

Access Category	*AC VO*	*AC VI*	*AC BE*	*AC BO*
AIFS	2	2	3	7
CW_{min}	7	15	31	31
CW_{max}	15	31	1023	1023
TXOPLimit	3.264	6.016	0	0

9.1.3.2 HCF Controlled Channel Access

Āe PCF mode of the IEEE 802.11 standard has some major problems that lead to poor QoS performance [3,4,6]. Indeed, the PCF defines only a single-class round-robin scheduling algorithm, which cannot handle the various QoS requirements of different types of traffic. In PCF, stations are allowed to transmit even if the frame transmission cannot finish before the start of the next CF repetition interval. Ā is delays the following data frames. A polled station is allowed to send a frame of any length between 0 and 2304 bytes, which may introduce a variable transmission time.

Āe IEEE 802.11e HCCA was proposed as the contention-free part of HCF to overcome the limitations of the PCF [2]. Unlike PCF, HCCA stations are not allowed to transmit packets if the frame transmission cannot finish before the next beacon starts [2]. In addition, HCCA uses a $TXOP_{Limit}$ parameter to bound the transmission time of polled QoS stations. HCCA provides parameterized QoS support. It uses a QoS-aware centralized coordinator, called a hybrid coordinator (HC) that is collocated with the QoS AP and has a higher priority of access to the wireless medium. Ā is allows HCCA to initiate frame exchange sequences and allocate transmission opportunities (TXOPs) to itself and other QoS stations.

Āe most important enhancement of the HCCA is the ability to provide a limited-duration CAP for contention-free transfer of QoS data during the CP. When the HC needs to access the wireless medium to start a CFP or a CAP in the CP, it should sense an idle wireless medium for one PIFS period. After an idle medium for PIFS period, the HC transmits the first permitted frame, with the duration value set to cover CFP or the CAP. Āe first permitted frame in a CFP is the Beacon frame, as shown in Figure 9.9. Ā e HC can start a CAP during the CP by sending a poll or data frame after sensing an idle medium for a PIFS period. Because the PIFS length is shorter than the AIFS length (used by EDCA), the HC is able to interrupt the contention operation and generate a CAPs at almost any moment (with at most one packet length delay). Ā e HC can start several CAPs after detecting a medium being idle for a PIFS period. To leave enough space for EDCA, the maximum duration for HCCA is limited by a $T_{CAPLimit}$ parameter. After the TXOP, the HC may reclaim the channel if the channel remains idle for a duration of PIFS. A CAP ends when the HC does not reclaim the channel after the end of a transmission opportunity.

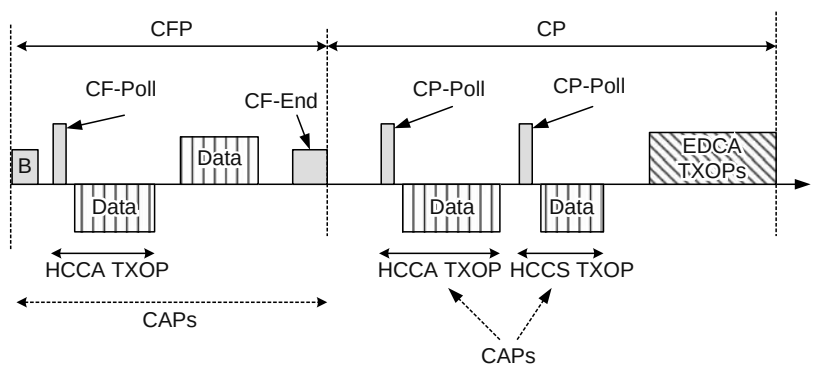

Figure 9.9 Controlled access phases in HCCA.

9.1.4 Current Challenges and Enhancements

Āe two access mechanisms of the HCF IEEE 802.11e have several drawbacks. Ā e main drawback of the EDCA mechanism is that the values of CW_{min}, CW_{max}, and backoff time of each queue are static and do not take into account wireless channel conditions. Āe static reset method has been proved to be ineffective in maximizing channel utilization whenever the demand for channel access (e.g., traffic load) increases [6–8]. Indeed, the probability of collisions in a busy wireless channel is high and will likely cause CW to approach CW_{max} almost every time before the station succeeds in transmitting its data. After a successful transmission, and the next time the station wants to transmit a frame, CW will start at CW_{min} and will subsequently increase (double) at each unsuccessful step until the station succeeds in transmitting its packet. As a result, a station may have to try several times before it succeeds and hence, significant time is wasted due to resetting CW to CW_{min} and consequently, the channel becomes underused.

Āe HCCA of the 802.11e standard does not specify the scheduling discipline that determines when the controlled access phase are generated and leaves it to system developers to devise such a scheme [9]. In addition, the HCCA allocates transmission opportunities (TXOPs) to itself and to other QoS stations using the reported mean transmission rates. Āe HCCA scheduler allocates a fixed TXOP to each QoS station based on its mean rate requirements. When the transmitted flow is of variable bit rate (VBR) and its rate is larger than the mean transmission rate, the packets will be queued causing a delay increase. If the peak transmission rate of VBR applications are reported to the HC and used to calculate the TXOPs, the TXOPs will be large enough for delivering packets. However, the channel will be underutilized when a smaller number of VBR flows are admitted and the gaps between the peak and mean rates are considerable [5].

Several techniques have been proposed to enhance the performance of IEEE 802.11e by adapting CW to the network state [7,9–12]. In Ref. [12], a scheme is proposed for the purpose of improving the IEEE 802.11e performance under different load rates and increasing the service differentiation in EDCA-based networks. Āe scheme uses a dynamic procedure to change the CW value after a collision or a successful transmission by resetting CW to adaptive values that are different from the CW_{min}, taking into account their current sizes and the average collision rate. Furthermore, the scheme suggests changing the mechanism of doubling the CW when a collision occurs. In Ref. [7], an approach was proposed to replace the EDCA technique and is based on adapting the AIFS in response to network conditions. Ā e rationale behind adapting the AIFS is to reduce the waiting time for the high-priority applications and increasing it for the low-priority ones. When the network is congested, the AIFS of the high-priority traffic is decreased while the AIFS of the BE traffic is increased. Conversely, if the network is in normal conditions, high-priority AIFS is increased while BE AIFS is decreased. It follows that this technique favors high-priority streams when the network is overloaded and tries to serve all traffic when the network is in normal conditions. A link adaptation strategy that provides differentiation not only at the MAC layer but also at the PHY layer was proposed in Ref. [10]. Ā is strategy exploits the positive ACK procedure to evaluate the quality of the link. If the transmitter does not receive an ACK frame it concludes that the last transmission was failed. For each active link, a transmitter maintains two counters, a success counter and a failure counter. If a frame is successfully transmitted, then the success counter is incremented by one and the failure counter is reset to zero. If the transmission fails then the failure counter is incremented by one and the success counter is reset to zero. Ā ese two counters are used to determine whether the quality of the link is good or not. Āe transmission rate is adjusted with respect to the link quality. A low-complexity adaptive EDCA algorithm that adapts the CW to channel conditions and adjusts it depending on the network utilization and performance

was proposed in Ref. [11]. Ā e proposed technique outperforms IEEE 802.11e and is comparable to the other enhancement schemes while maintaining relatively low-complexity requirements and providing a faster adaptation to the network state. A new access scheduling framework designed to improve the HCCA access mechanism was proposed in Ref. [9]. Ā is framework is capable of providing per-session QoS guarantees for interactive voice and video applications over WLAN. It provides guaranteed services to flows that make reservation with the WLAN AP by means of the available MAC signaling methods, while at the same time, allowing the normal contention-based access to take place using the remaining capacity of the channel. Āi s approach is different from the existing polling mechanisms in which long alternating contention-free and CPs are generated, resulting in uncontrolled delay bounds and an inefficient operation.

9.2 Quality of Service in Wireless Metropolitan Area Networks

Āe WiMAX technology based on the IEEE 802.16 standard plays a key role in fixed broadband wireless metropolitan area networks [13]. It has proven to be a cost-effective wireless alternative to cabled access networks (i.e., fiber optic links, digital subscriber line [DSL]). Āi s section gives an overview of the WiMAX networks and its QoS requirements as presented in the IEEE 802.16 standard and its IEEE802.16e amendment [13,14].

9.2.1 WiMAX Network: Entities, Architecture, and Operation Modes

To give an understanding of how QoS can be achieved in WiMAX networks, we first present the architecture of these networks, the main entities, and the operation modes. We then describe the data and scheduling services supported by the MAC layer and how resources are allocated.

9.2.1.1 Mesh Mode versus Point-to-Multipoint Mode

Āe basic architecture of WiMAX consists of two fixed stations: base station (BS) and subscriber station (SS). Ā e BS is a central equipment set providing connectivity, management, and control of several SSs situated at varying distances. An SS can represent a building equipped with a conventional wireless or wired local area network. Āe IEEE 802.11 standard defines two operation modes: an optional mesh mode and a mandatory point-to-multipoint (PMP) mode [13]. Ā e mesh mode supports direct communications between SSs without the need for a BS. Access coordination is distributed among the SSs. In PMP mode, a controlling BS connects multiple SSs to various public networks as shown in Figure 9.10.

9.2.1.1.1 PMP Mode

In PMP, the communication path is bidirectional, downlink (from the BS to the SS) and uplink (from the SS to the BS). Ā e uplink and downlink data transmissions are duplexed using frequency division duplex (FDD) or time division duplex (TDD). In FDD, the uplink and downlink subframes occurs simultaneously on separate frequencies while in TDD they occur at different times but usually share the same frequency as shown in Figure 9.11. To schedule the uplink and downlink grants to meet the negotiated QoS requirements, the BS starts the downlink subframe with a downlink map (DL-MAP) followed by an uplink map (UL-MAP). Ā e downlink MAP contains the

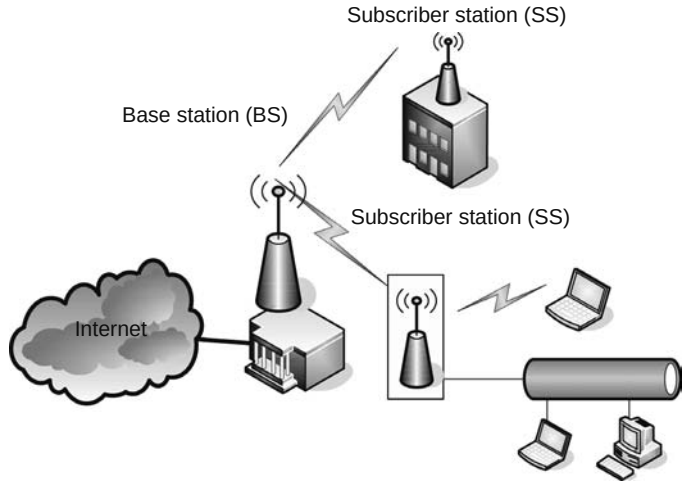

Figure 9.10 WiMAX in PMP mode.

timetable for downlink grants in the forthcoming downlink subframe. Downlink grants directed to SSs with the same downlink interval usage code (DIUC) are advertised in the DL-MAP as a single burst. Āe uplink MAP tells each SS about the boundaries of its allocated bandwidth within

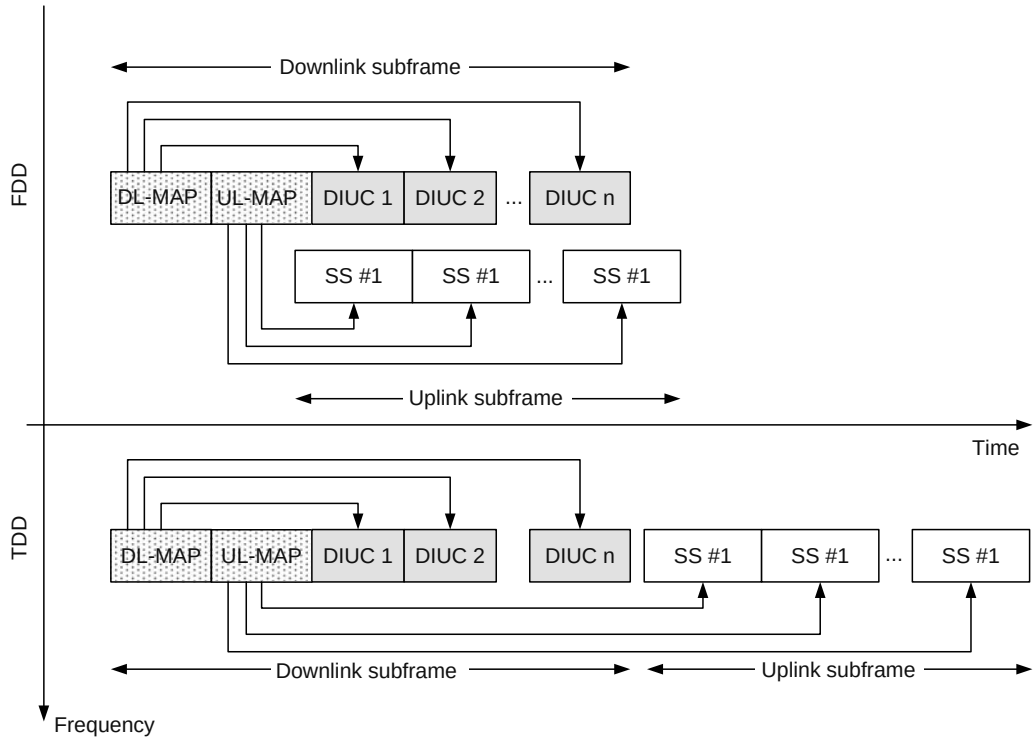

Figure 9.11 Frame structure with FDD and TDD.

the current uplink subframe. Ā e SSs transmit in their assigned allocations using the burst profile specified in the uplink MAP entry granting them bandwidth. Āe downlink is generally broadcast. In cases where the DL-MAP does not explicitly indicate that a portion of the downlink subframe is for a specific SS, all SSs capable of listening to that portion of the downlink subframe shall listen. However, SSs check the received subframe and retain only the parts addressed to them. In addition to messages that are individually addressed, messages may also be sent on multicast connections (control messages and video distribution are examples of multicast applications) as well as broadcast to all stations. SSs share the uplink to the BS on a demand basis. Depending on the class of service utilized, the SS may be issued continuing rights to transmit, or the right to transmit may be granted by the BS after receipt of a request from the user.

9.2.1.1.2 Mesh Mode

In mesh mode, the traffic can be routed through other SSs and can occur directly between SSs. A system that has a direct connection to backhaul services outside the mesh network is termed a mesh BS. All the other systems of a mesh network are termed mesh SS. Within mesh context, the uplink and downlink are defined as a traffic in the direction of the mesh BS and traffic away from the mesh BS, respectively. Ā e other three important terms of mesh systems are neighbor, neighborhood, and extended neighborhood. Ā e stations with which a node has direct links are called neighbors. Neighbors of a node form a neighborhood. Nodes, neighbors are considered to be one hop away from the node. An extended neighborhood contains, additionally, all the neighbors of the neighborhood. Using distributed scheduling, all the nodes including the mesh BS coordinate their transmissions in their two-hop neighborhood to ensure that the resulting transmissions do not cause collisions with the data and control traffic scheduled by any other node. Using centralized scheduling, resources are granted in a more centralized manner. Āe mesh BS gathers resource requests from all the mesh SSs within a certain hop range. It then determines the amount of granted resources for each link in the network and communicates these grants to all the mesh SSs within the hop range. All the communications are in the context of a link, which is established between two nodes. One link is used for all the data transmissions between the two nodes.

Āe rest of this chapter focuses on PMP mode because it is anticipated that providers will use it to connect customers to the Internet [15].

9.2.1.2 Addressing and Connections

Each SS has a 48-bit universal MAC address that uniquely defines its air interface and serves mainly as an equipment identifier [14]. Ā is address is used during the initial ranging process to establish the appropriate connections for an SS and is also used in the authentication process between the BS and the SS.

Āe MAC layer of the IEEE 802.16 is a connection-oriented layer. All data services are in the context of a connection. A connection is defined as a unidirectional mapping between the MAC peers of the BS and the SS. Āe MAC defines two kinds of connections: management connections that are used for the purpose of transporting management messages or standard-based messages and transport connections that are used to transport user data. Connections are identified by a 16-bit connection identifier (CID). At the SS initialization, two pairs of management connections (uplink and downlink) are established between the SS and the BS and a third pair may be optionally generated. Āese three pairs of management connections reflect the fact that there are inherently three different levels of QoS for managing traffic between an SS and the BS. Ā ese connections

are (1) the basic connection that is used by the BS MAC and the SS MAC to exchange short, time-urgent MAC management messages; (2) the primary management connection that is used by the BS MAC and the SS MAC to exchange longer, more delay-tolerant MAC management messages such as those used for authentication and connection setup; (3) the optional secondary management connection that is used by the BS and the SS to transfer delay tolerant, standard-based (Dynamic Host Configuration Protocol [DHCP], Trivial File Transfer Protocol [TFTP], SNMP, etc.) messages. In addition, the IEEE 802.16 standard defines another two management connections: the broadcast connection that is used by the BS to send MAC management messages on a downlink to all SSs and the initial ranging connection that is used by the SS and the BS during the initial ranging process. Āe initial ranging connection is identified by a well-known constant value within the protocol because an SS has no addressing information available until the initial ranging process is complete.

9.2.1.3 Data and Scheduling Services

Scheduling services represent the data handling mechanisms supported by the MAC scheduler for data transport on a connection. Each connection is associated with a single scheduling service. A scheduling service is determined by a set of QoS parameters that quantify aspects of its behavior. Āe se parameters are managed using dynamic service messages that allow the BS and the SS to add, modify, or delete the characteristics of a service flow. Ā e key QoS parameters are [13]

- Traffic priority: specifies the priority assigned to a service flow. Given two service flows identical in all QoS parameters besides priority, the higher priority service flow should be given lower delay and higher buffering preference.
- Maximum sustained traffic: defines the peak information rate of the service and is expressed in bits per second. Data units deemed to exceed the maximum sustained traffic rate may be delayed or dropped.
- Maximum traffic burst: describes the maximum continuous burst that the system should accommodate for the service.
- Minimum reserved traffic rate: specifies the minimum amount of data to be transported on behalf of the service flow when averaged over time. It is expressed in bits per second. Ā e specified rate is only honored when sufficient data is available for scheduling. Āe BS and the SS are able to transport traffic up to its minimum reserved traffic rate. If less than the minimum reserved traffic rate is available for a service flow, the BS and the SS may reallocate the excess reserved bandwidth for other purposes.
- Tolerated jitter: defines the maximum delay variation (jitter) for the connection.
- Maximum latency: specifies the maximum latency between the reception of a packet by the BS or SS on its network interface and the forwarding of the packet to its radio frequency Interface.

Well-known scheduling services can be implemented by specifying a specific set of QoS parameters. Four scheduling services are supported [14]:

1. Scheduling service to support real-time data streams consisting of fixed-size data packets issued at periodic intervals, such as Voice-over-IP (VoIP) without silence suppression. Ā e key QoS parameters are the maximum sustained traffic rate, the maximum latency, and the tolerated jitter.

2. Scheduling service to support real-time data streams consisting of variable-sized data packets that are issued at periodic intervals, such as moving pictures experts group (MPEG) video. Āe key QoS service flow parameters for this scheduling service are the minimum reserved traffic rate, the maximum sustained traffic rate, and the maximum latency.
3. Scheduling service to support delay-tolerant data streams consisting of variable-sized data packets for which a minimum data rate is required, such as FTP. Āe key QoS service flow parameters for this scheduling service are the minimum reserved traffic rate, the maximum sustained traffic rate, and the traffic priority.
4. Scheduling service to support data streams for which no minimum service level is required and therefore may be handled on a space-available basis. Āe key QoS service flow parameters for this scheduling service are the maximum sustained traffic rate and the traffic priority.

9.2.1.3.1 Uplink Request/Grant Scheduling

Āe uplink request/grant scheduling is performed by the BS with the intent of providing each subordinate SS with bandwidth for uplink transmissions or opportunities to request bandwidth. By specifying a scheduling type and its associated QoS parameters, the BS scheduler can anticipate the throughput and latency needs of the uplink traffic and provide polls or grants at the appropriate time. Ā ere are five uplink scheduling algorithms [14]:

1. Unsolicited grant service (UGS) scheduling algorithm is designed to support real-time uplink service flows that transport fixed-size data packets on a periodic basis such as VoIP. Ā e BS provides data grant burst to the SS at periodic intervals based upon the maximum sustained traffic rate of the service flow. Ā e size of these grants should be sufficient to hold the fixed-length data associated with the service flow but may be larger at the discretion of the BS scheduler. Ā e grant size and period are negotiated in the session initialization process. Ā is eliminates the overhead and latency of SS bandwidth requests (BW-REQ) and assures that grants are available to meet the flow's real-time needs.
2. Real-time polling service (rtPS) scheduling algorithm is designed to support real-time uplink service flows that transport variable size data packets on a periodic basis such as MPEG video. Ā e service offers real-time, periodic, unicast BW-REQ opportunities, which allow the SS to specify the size of the desired grant. Āus , this service requires more request overhead than UGS, but supports variable grant sizes for optimum data transport efficiency. Āe BS provides unicast request opportunities. For this service to work correctly, the SS is prohibited from using any contention request opportunities for that connection. Āe BS may issue unicast request opportunities as prescribed by this service even if prior requests are currently unfulfilled. Āi s results in the SS using only unicast request opportunities and data transmission opportunities to obtain uplink transmission opportunities.
3. Extended rtPS (ertPS) scheduling algorithm is designed to support real-time service flows that generate variable size data packets on a periodic basis such as VoIP. Āi s scheduling mechanism, recently proposed in IEEE 802.16e, builds on the efficiency of both UGS and rtPS. Āe BS provides unicast grants in an unsolicited manner like in UGS, thus saving the latency of a BW-REQ. However, whereas UGS allocations are fixed in size, ertPS allocations are dynamic. Ā e SS requests the bandwidth using extended piggyback request (PBR) bits of the grant management subheader. Āe BS provides periodic uplink allocations according to the requested size until the SS requests a different size of bandwidth. When the SS data rate increases, SS requests the bandwidth using BR (BW-REQ) bits of the bandwidth request

header. Ā e BS assigns uplink bandwidth according to the requested size periodically. Ā e BS does not change the size of uplink allocations until receiving another bandwidth change request from the SS.

4. Non-real-time polling (nrtPS) uplink scheduling is designed to support delay-tolerant data streams consisting of variable-sized data packets for which a minimum data rate is required such as FTP. Āe nrtPS offers unicast polls on a regular basis, which assures that the uplink service flow receives request opportunities even during network congestion. Ā e BS typically polls nrtPS CIDs on an interval on the order of one second or less. Āe BS provides timely unicast request opportunities. For this service to work correctly, the SS is allowed to use contention request opportunities as well as unicast request opportunities and data transmission opportunities.
5. BE scheduling is designed to support data streams for which no minimum service level is required. Ā e intent of the BE grant scheduling type is to provide an efficient service for BE-traffic in the uplink. For this service to work correctly, the SS is allowed to use contention request opportunities as well as unicast request opportunities and data transmission opportunities.

9.2.1.4 Bandwidth Allocation and Request Mechanisms

In IEEE 802.16, all packets from the application layer in the SS are classified by the connection classifier based on the CID and are forwarded to the appropriate queue as shown in Figure 9.12. According to the incoming traffic service flow, the SS sends BW-REQ messages that report the current queue size of each SS connection to the BS uplink bandwidth allocation scheduler, which controls all the uplink packet transmissions. Āe BS schedules the requests in the different service flow queues according to their QoS requirements and generates a MAP message to the SS scheduler. Ā e MAP message contains the information element (IE) parameter, which includes the time slotsin which the SS can transmit during the uplink subframe. Āe three management connec-

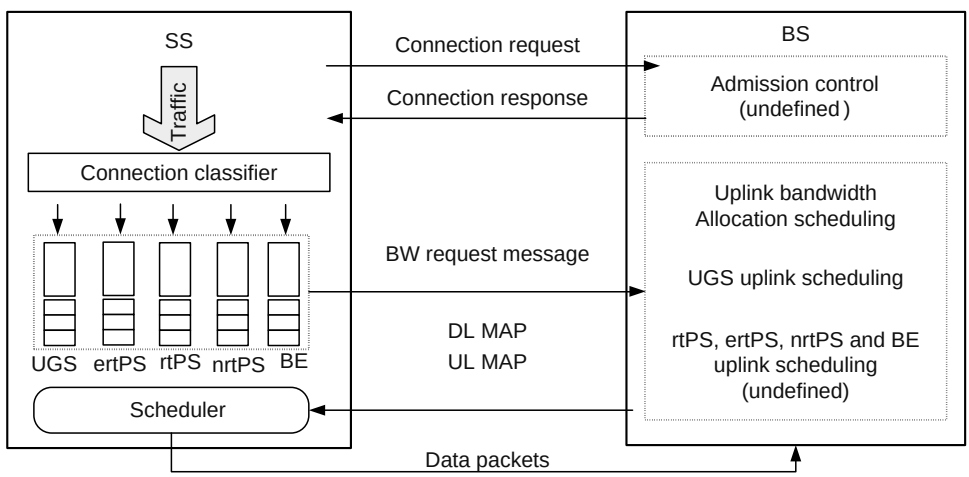

Figure 9.12 QoS architecture of the IEEE 802.16.

tions that are assigned to the SS during the initialization process are used for sending and receiving BW-REQ messages and other control messages. Āe se connections allow differentiated levels of QoS to be applied to the different connections carrying MAC management traffic. Increasing (or decreasing) bandwidth requirements is necessary for all services except the incompressible constant bit rate UGS connections. Ā ere are numerous methods by which the SS can get the BW-REQ message to the BS.

Requests: Requests refer to the mechanism that SSs use to indicate to the BS that they need uplink bandwidth allocations. A request may come as a standalone BW-REQ header or may come as a PBR. BW-REQs are made in terms of the number of bytes needed to carry the MAC header and the payload. Āe BW-REQ message may be transmitted during any uplink allocation, except during the initial ranging interval. SS bandwidth requests reference individual connections. Ā ey may be incremental or aggregate. When the BS receives an incremental BW-REQ, it adds the quantity of the requested bandwidth to its current perception of the connection bandwidth needs. When the BS receives an aggregate BW-REQ, it replaces its perception of the connection bandwidth needs with the quantity of the requested bandwidth. Capability of incremental BW-REQs is optional for the SS and mandatory for the BS. Capability of aggregate BW-REQs is mandatory for SS and BS.

Grants: Ā e BS grants bandwidth resources to the SS, not to individual CIDs. When the SS receives a grant shorter than expected (scheduler decision, request message lost, etc.), no explicit reason is given. On the basis of the latest information received from the BS and the status of the request, the SS may decide to perform backoff and request again or to discard the transmission.

Polling: Polling is the process by which the BS allocates bandwidth to the SSs for making BW-REQs. Āese allocations may be to individual SSs (unicast polling) or to groups of SSs (multicast polling). Āe allocations are not in the form of an explicit message, but are contained as a series of IEs within the uplink MAP. When an SS is polled, no explicit message is transmitted to poll the SS. Rather, the SS is allocated in the uplink MAP, bandwidth sufficient to respond with a BW-REQ. If insufficient bandwidth is available to individually poll many inactive SSs, some SSs may be polled in multicast groups or a broadcast poll may be issued. Certain CIDs are reserved for multicast groups and for broadcast messages. An SS belonging to the polled group may request bandwidth during any request interval allocated to that CID in the UL-MAP. To reduce the likelihood of collision with multicast and broadcast polling, only SSs needing bandwidth reply; they should apply the contention resolution algorithm to select the slot in which the initial BW-REQ is to be transmitted.

Note that polling is done on SS basis. Bandwidth is always requested on a CID basis and is allocated on an SS basis.

9.2.2 IEEE 802.16 QoS Architecture

Āe IEEE 802.16 standard defines several QoS related mechanisms: (1) service flow QoS scheduling; (2) dynamic service establishment; and (3) two-phase activation model. Ā ese concepts are used to support QoS for both uplink and downlink traffic through the SS and the BS. Āe principal mechanism for providing QoS is to associate packets traversing the MAC interface into a service flow as identified by the transport connection.

9.2.2.1 Service Flow QoS Scheduling

A service flow is a MAC transport service that provides unidirectional transport of packets either to uplink packets transmitted by the SS or to downlink packets transmitted by the BS. A service flow is partially characterized by several attributes that include details of how the SS requests uplink bandwidth allocations and the expected behavior of the BS uplink scheduler. Some of these attributes are

- Service flow ID (SFID) that is assigned to each existing service flow to serve as the principal identifier for the service flow between a BS and an SS.
- Connection ID (CID) of the transport connection exists only when the service flow is admitted or active. Āe relationship between the SFID and the transport CID is unique. An SFID is associated with only one transport CID, and a transport CID is associated with only one SFID.
- Admitted QoS parameters that define QoS parameters for which the BS (and possibly the SS) are reserving resources.
- Active QoS parameters that specify a set of QoS parameters defining the service actually being provided to the service flow. Only an active service flow may forward packets.

It is useful to think of three types of service flows:

1. Provisioned: A service flow may be provisioned but not immediately activated (sometimes called deferred).
2. Admitted: Āi s type of service flow has resources reserved by the BS for its admitted QoS parameters, but these parameters are not active. Admitted service flows may have been provisioned or may have been signaled by some other mechanisms.
3. Active: Ā is type of service flow has resources committed by the BS for its active QoS parameters.

Service flows are first admitted, then activated. An authorization module in the BS approves or rejects a request regarding a service flow. Āe authorization module can activate a service flow immediately or defer activation to a later time. Every change to the service flow QoS parameters should be approved by the authorization module. Ā is includes every Dynamic Service message to activate, update, or delete an existing service flow.

9.2.2.2 Dynamic Service Flow Establishment

Creation of service flows may be initiated by either the SS (optional capability) or the BS (mandatory capability). In the SS-initiated protocol, an SS wishing to create either an uplink or downlink service flow sends a request to the BS using a Dynamic Service Addition Request (DSA-REQ) message (Figure 9.13a). Āe BS checks the integrity of the message and, if the message is intact, sends a Dynamic Service Received (DSX-RVD) response message to the SS. Āe BS checks the SSs authorization for the requested service and whether the QoS requirements can be supported, generating an appropriate response using a DSA Response (DSA-RSP) message indicating acceptance or rejection. Āe SS concludes the transaction with a DSA Acknowledgment (DSA-ACK) message.

In the BS-initiated protocol, a BS wishing to establish either an uplink or a downlink dynamic service flow with an SS checks first the authorization of the destination SS for the requested service

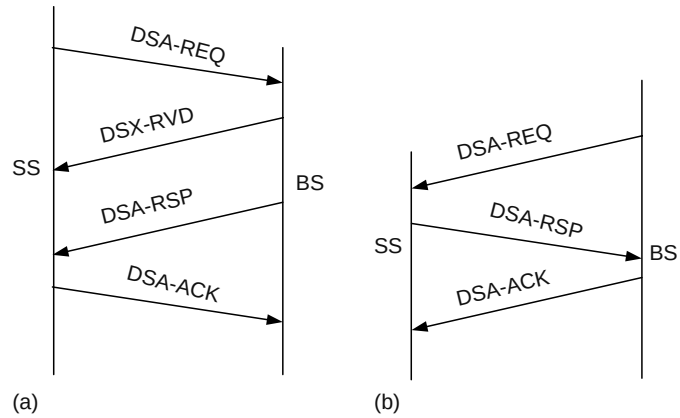

Figure 9.13 (a) SS-initiated protocol; (b) BS-initiated protocol.

flow and determines whether the QoS requirements can be supported. If the service can be supported, the BS generates a new SFID with the required class of service and informs the SS using a DSA-REQ message (see Figure 9.13b). If the SS checks that it can support the service, it responds using a DSA-RSP message. Āe transaction completes with the BS sending a DSA-ACK message.

In addition to the methods for creating service flows, protocols are defined for modifying and deleting service flows. Āe Dynamic Service Change (DSC) set of messages is used to modify the flow parameters associated with a service flow. Specifically, DSC can modify the service flow specification. Implementation of the DSC initiated by BS is mandatory while it is optional by SS.

A single DSC message exchange can modify the parameters of either one downlink service flow or one uplink service flow. Ā e BS controls the uplink scheduling, the downlink scheduling, and the downlink transmit behavior. Āe BS always changes scheduling on receipt of a DSC-REQ (SS-initiated transaction) or DSC-RSP (BS-initiated transaction). Āe timing of scheduling changes is independent of direction and whether it is an increase or decrease in bandwidth. Ā e change in the downlink transmit behavior is always coincident with the change in downlink scheduling as BS controls both.

Āe SS controls the uplink transmit behavior. Ā e timing of SS transmit behavior changes is a function of which device initiated the transaction and whether the change is an increase or decrease in bandwidth. If an uplink service flows bandwidth is being reduced, the SS reduces its payload bandwidth first and then the BS reduces the bandwidth scheduled for the service flow. If an uplink service flows bandwidth is being increased, the BS increases the bandwidth scheduled for the service flow first and then the SS increases its payload bandwidth.

Any service flow can be deleted with the the Dynamic Service Delete (DSD) messages. When a service flow is deleted, all resources associated with it are released.

9.2.2.3 Activation Model

Āe IEEE 802.16 standard supports a two-phase activation model that is often utilized in telephony applications. In the two-phase activation model, the resources for a call are first admitted, and then once the end-to-end negotiation is completed the resources are activated. Āe two-phase model serves the following purposes:

- Conserving network resources until a complete end-to-end connection has been established.
- Performing policy checks and admission control on resources as quickly as possible, and in particular, before informing the far end of a connection request.
- Preventing several potential theft-of-service scenarios.

9.2.3 Undefined QoS Requirements, Challenges, and Enhancements

Ār ee components are necessary to manage QoS in the IEEE 802.16 standard. Ā ese are (1) admission control, which determines whether a new request for a connection can be granted or not according to the remaining free bandwidth; (2) scheduling, which determines which packet will be served first to guarantee QoS requirements; and (3) buffer management, which controls buffer size and decides which packets to drop. IEEE 802.16 defines the signaling mechanism for information exchange between the BS and the SS such as connection setup, BW-REQ, and MAP messages. It defines also the UGS uplink scheduling to support real-time data streams consisting of fixed-size data packets issued at periodic intervals. However, IEEE 802.16 does not define the uplink rtPS and ertPS scheduling to support real-time uplink service flows that transport variable size data packets on a periodic basis (see Figure 9.12). IEEE 802.16 does not define also the nrtPS and BE uplink scheduling. In addition, the admission control in the BS is left undefined as well.

In IEEE 802.16, service data units deemed to exceed the maximum sustained traffic rate may be delayed or dropped. However, the standard does not define or recommend any algorithm for measuring whether a flow exceeds its maximum sustained traffic rate.

Āe QoS mechanisms of the IEEE 802.11-based networks are difficult to apply on the IEEE 802.16 networks due to the difference in the nature of these two technologies. Indeed, the IEEE 802.11 MAC is a connectionless and a contention-based technology in which the MAC uses acknowledgments and timeouts, which may cause overhead and delays. However, the IEEE 802.16 is a connection-oriented protocol that uses service flows. In IEEE 802.16, overhead and delays between users are eliminated because of its grant-based nature that does not require the use of acknowledgment and timers like SIF, PIFS, DIFS, and AIFS of the IEEE 802.11. Āi s allows a better QoS handling. In addition, IEEE 802.11 has a fixed channel size while the channel size is changeable in IEEE 802.16.

Several architectures were recently proposed [16–22] to support QoS in WiMAX. Most of these proposals aim to complete the missing parts of the IEEE 802.16 QoS architecture. In Ref. [16], a two-layer scheduling structure of the bandwidth allocation was proposed to support all types of service flows. In the first layer, the deficit fair priority queue (DFPQ) was used to distribute total bandwidth among flow services in different queues. Six queues were defined according to their direction (uplink or downlink) and service classes. In the second layer scheduling, packets within each of the six queues will be served according to a certain scheduling algorithm. For rtPS connections, packets with earliest deadline will be scheduled first. Āe information module determines the packets' deadline that is calculated by its arrival time and maximum latency. For nrtPS connections, packets are scheduled based on their weight, which is defined as ratio between a connection's nrtPS minimum reserved traffic rate and the total sum of the minimum reserved traffic rate of all nrtPS connections [23]. For BE connections, the remaining bandwidth is allocated to each BE connection using round robin [24]. Because UGS will be allocated fixed bandwidth in transmission, their bandwidths will be directly cut before each scheduling. In Ref. [17], a preemption-based variation of the scheduling algorithm presented in Ref. [16] was proposed. Āe proposed scheme focuses on giving rtPS service flow packets more chances to meet their deadline and decrease their

delay to better guarantee the QoS requirements of this class. Indeed, in addition to checking if the available bandwidth is enough for granting requests, information related to the rtPS service flows that are admitted are tracked by maintaining a table that is used to approximate the expected delay of each rtPS connection. Āese delays are used later in the scheduling algorithm.

9.3 Summary

Āi s chapter presented the QoS concepts in wireless local area networks (i.e, WiFi/IEEE 802.11) and wireless metropolitan area networks (i.e., WiMAX/IEEE 802.16). It described the BE services IEEE 802.11 standard and the IEEE 802.11e standard that was introduced to support sophisticated services that guarantee QoS attributes such as bandwidth, delay, and jitter. In this context, the architectures, basic elements, the DCF MAC access mechanism, the PCF access mechanism, and the HCF access mechanism were described. Ā e chapter also examined the QoS in the wireless metropolitan area networks as introduced in the IEEE 802.16 standard and its IEEE 802.16e amendment. In this context, the WiMAX Architecture, operation modes, scheduling services and algorithms, resource allocation, and QoS requirements were explored. Ā roughout the chapter, the QoS aspects of the wireless local and metropolitan area networks were addressed. Current challenges and drawbacks of the IEEE 802.11 and IEEE 802.16 were highlighted and some proposed enhancements were surveyed.

REFERENCES

1. IEEE Std 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, 1999 edition
2. IEEE Std 802.11e-2005 IEEE Standard for Information Technology—Telecommunications and Information Exchange between Systems—LAN/MAN Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements.
3. S. Mangold, S. Choi, P. May, G. Hiertz, O. Klein, and B. Walke, Analysis of IEEE 802.11e for QoS support in wireless LAN, *IEEE Wireless Communications*, December 2003.
4. Q. Ni, L. Romdhani, and T. Turletti, A survey of QoS enhancements for IEEE 802.11 wireless LAN, *Wireless Communications and Mobile Computing*, 4, 2004.
5. D. Gu and J. Zhang, QoS enhancement in IEEE 802.11 wireless area networks, IEEE *Communications Magazine*, 41(6), June 2003.
6. Q. Ni, Performance analysis and enhancements for IEEE 802.11e wireless networks, *IEEE Network*, 19(4), 21–27, July–August 2005.
7. A. Ksentini, M. Naimi, A. Nafss, and M. Gueroui, Adaptive service differentiation for QoS provisioning in IEEE 802.11 wireless ad hoc networks, Proceedings of the 1st ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks, Venezia, Italy, 2004.
8. J. Naoum-Sawaya, B. Ghaddar, S. Khawam, H. Safa, H. Artail, and Z. Dawy, Adaptive approach for QoS support in IEEE 802.11e wireless LAN, IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, 2005 (WiMob'2005), Montreal, Canada, August 22–24, 2005.
9. Y. Fallah and H. Alnuweiri, Hybrid polling and contention access scheduling in IEEE 802.11e WLANs, *Journal of Parallel and Distributed Computing*, 67, 2007.
10. M. Bandinelli, F. Chifi, R. Fantacci, D. Tarchi, and G. Vannuccini, A link adaptation strategy for QoS support in IEEE 802.11e-based WLANs, IEEE Wireless Communications and Networking Conference, New Orleans, Louisiana March 2005.

11. H. Artail, H. Safa, J. Naoum-Sawaya, B. Ghaddar, and S. Khawam, A simple recursive scheme for adjusting the contention window size in IEEE 802.11e wireless ad hoc networks, *Computer Communications*, 29(18) 18, November 2006.
12. L. Romdhani, Q. Ni, and T. Turletti, Adaptive EDCF: Enhanced service differentiation for IEEE 802.11 wireless ad-hoc networks, IEEE Wireless Communications and Networking Conference (WCNC'03), New Orleans, Louisiana, 2003.
13. IEEE Std 802.16-2004, IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.
14. IEEE Std 802.16e-2005, IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, 2006.
15. A. Ghosh, D. Walter, J. Andrews, and R. Chen, Broadband wireless access with WiMax/802.16: Current performance benchmarks and future potential, *IEEE Communications Magazine*, February 2005.
16. J. Chen, W. Jiao, and H. Wang, A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode, Proceedings IEEE International Conference on Communications (ICC 2005), Seoul, Korea, May 2005.
17. H. Safa, H. Artail, M. Karam, R. Soudah, and S. Khayat A new scheduling architecture for IEEE 802.16 wireless metropolitan area network, Proceedings of the 5th ACS/IEEE International Conference on Computer Systems and Applications, AICCSA '2007, Amman, Jordan, May 2007.
18. H. Alavi, M. Mojdeh, and N. Yazdani, A QoS architecture for IEEE 802.16 standards, Proceedings of IEEE Asia Pacific Conference on Communications, Perth, Australia, October 2005.
19. J. Chen, W. Jiao, and Q. Guo, An integrated QoS control architecture for IEEE 802.16 broadband wireless access systems, Proceedings of IEEE Global Telecommunications Conference (GLOBECOM'05), St. Louis, Missouri, 2005.
20. D.-H Cho, J.-H Song, M.-S Kim, and K.-J Han An architecture for efficient QoS support in the IEEE 802.16 broadband wireless access network, Proceedings of 4th International Conference on Networking, April 2005.
21. G. Chu, D. Wang, and S. Mei, A QoS architecture for the MAC protocol of the IEEE 802.16 BWA system, Proceedings of IEEE Conference on Communications, Circuits, and Systems, St. Petersburg, Russia, 2002.
22. Y. Shang and S. Cheng, An enhanced packet scheduling algorithm for QoS support in IEEE 802.16 wireless network, Proceedings of 3rd International Conference on Networking and Mobile Computing, August 2005.
23. A. Demers, S. Keshav, and S. Shenker Analysis and simulation of a fair queuing algorithm, ACMSIGCOMM19, Austin, Texas, 1989.
24. E. L. Hahne and R. G. Gallager, Round robin scheduling for fair flow control in data communication networks, International Conference on Communications, June 1986.