Quality of Service for wireless networks

How is QoS for wireless networks implemented?

Analytical Network Project
Master education in System and Network Engineering
University of Amsterdam, the Netherlands

ing. E. Hos
hossie@chello.nl

ing. N. Visser
nivi@chello.nl

Version 1.1
29 June 2004

Abstract

In this paper, an overview of the Quality of Service (QoS) mechanisms in the upcoming IEEE 802.11e and IEEE 802.16 standards for wireless networking is given. First, Ethernet and the current 802.11 standard and their concepts are discussed, focussing on the Medium Access Control (MAC) layer of the OSI model. The new 802.11e standard is evaluated and the new mechanisms in the MAC layer for QoS support, Enhanced Distributed Coordination Function (EDCF) and Hybrid Coordination Function (HCF), are introduced. The 802.16 standard is discussed on both the physical and MAC layer also focussing on QoS. The concept of grants, introduced in the DOCSIS standard and used by 802.16 is discussed. Finally, the way the upcoming standards will fit in with an existing example of a low-latency/low-bandwidth environment, gaming, is evaluated.

Contents

List of figures

List of tables

1. Introduction

With online/networked gaming and wireless Internet growing [8], the need for some kind of quality guarantee on the medium grows. In order for online games to run at acceptable speeds and with a minimum amount of lag, a network with the lowest possible latency is needed.

One way to guarantee the quality on a network speaking in terms of traffic prioritisation and bandwidth allocation is a so-called quality of service. In wired networking there are a number of standards implementing quality of service. In the currently often used IEEE 802.11 standard for wireless LAN, quality of service is not implemented. However a taskforce at the IEEE is developing the 802.11e standard which will add quality of service functionalities to 802.11. Another standard in development is IEEE 802.16 wireless MAN. This is an all new wireless networking standard which, by default comes with quality of service implemented.

The purpose of this report is to give an overview of the upcoming quality of service standards for wireless networking and to reflect on their impact on low-latency/low-bandwidth network environments.

Chapter 2 describes the set-up of this research project and the research question. In the following chapter the current wired and wireless standards for quality of service are described as well as the problems with these standards. Chapter 5 describes the newly proposed quality of service for wireless networking standards. Finally chapter 6 describes the effects of the newly proposed standards on an example low-latency/low-bandwidth environment, gaming. The report is completed by a conclusion.

2. Research set-up

This chapter describes the set-up of this research project. First is explained what quality of service is and elaborated on low-latency/low-bandwidth requirements. Finally, the main research question for this project is stated.

2.1. *What is quality of service?*

Quality of service (QoS) is the ability for a network element (e.g. a router, a host or an access point) to provide some level of quality assurance regarding consistent network data transmission and delivery. Depending on the application there can be different approaches to QoS:

- Resource reservation; this approach focuses on the proportioning of network resources depending on the quality of service request of an application. Also the bandwidth management policy applies.
- Prioritisation; network traffic is given a qualification and network resources are applied depending on the priority management policy.

Using one approach does not exclude using the other. Often, a QoS implementation will use a combination of both approaches to offer the needed services.

2.2. *Low-latency/low-bandwidth*

Latency, or delay, is the time it takes for a packet to get from one point to another across a network. In gaming situations, latency is referred to as lag. In an ideal network setup, it is assumed that it is possible for a packet to get delivered instantly. Anything less than instantly is called latency. Latency can be caused by the following reasons [23]:

- Propagation, this is the time it takes for a packet to travel from one point to another at the speed of light.
- Transmission, the medium itself can cause latency, the amount of latency caused depends on the medium used (optical fibre, coaxial fibre, etc.). The size of the packet can also cause delays in round trip times. The larger the packet the higher the delay.
- Processing, when a packet travels through a network it encounters all sorts of devices, like bridges, routers and hosts. Some of these devices open and examine the packet and change information (routers change hop counts for example). The examination and changing of this information takes time and thus causes delay.
- Utilisation, this is the amount of users (or bandwidth for that matter) that is in use at a given time.

Jitter is the variation in time between packets arriving. This can have several reasons like, network congestion, timing drift or route changes. Jitter is especially disruptive to real-time streaming media like audio, video or voice over IP

2.3. *Research question*

The main research question for this report is:

How do existing and proposed quality of service standards for wireless networks cope with low-latency/low-bandwidth requirements?

3. Quality of Service implementations

This chapter describes the inner workings of the 802.11 WLAN standards which are widely used and accepted nowadays.

3.1. 802.3 Ethernet

The IEEE 802.3 Ethernet standard forms the basis for the 802.11 WLAN standards. It introduces the collision sense multiple access/collision detection model which is also used in 802.11. Most QoS implementations focus on the MAC layer of the OSI model. This layer also takes an important place in the Ethernet standard.

3.1.1. Layered network model

The IEEE standards used for standardising network interconnections are organised according to the Open Systems Interconnection (OSI) reference model (*Figure 1* below displays the OSI model). The OSI model was developed in 1978 by the International Standardisation Organisation (ISO) to create a standard reference model for network communication between two end stations. Its seven layers describe the different networking functions used in network interconnections. Layers are defined by the functions they perform and the services they provide to its adjacent upper layer.

Each layer in the OSI model only communicates to its neighbouring layers, other layers are completely invisible. Data sent to another end station via the network passes all the layers, from layer 7 to layer 1, over the network and then at the other end station from layer 1 back to layer 7.

The IEEE introduced a few sub layers in both the physical and data link layer. Layer 2 is divided in the Logic Link Control (LLC) sub layer and another, implementation specific, sub layer. For the IEEE 802.3 standard [19] [7], the Ethernet protocol, this layer is the Media Access Control (MAC) sub layer. Layer 1 is also divided in two implementation specific sub layers, for Ethernet these are the physical signalling sub layer and the media specification sub layer.

	OSI layers	Ethernet layers
layer 7	Application	Application
layer 6	Presentation	Presentation
layer 5	Session	Session
layer 4	Transport	Transport
layer 3	Network	Network
layer 2	Data link	LLC
		MAC
layer 1	Physical	Physical signalling
		Media specification

Figure 1 - OSI model and added Ethernet sub layers

The LLC sub layer is implementation independent and concerned with managing traffic (flow- and error control) over the network. It also shields the upper layers from the connection method since the upper layers do not need to know what kind of media they're connected with. The MAC sub layer is concerned with sharing the physical connection to the network with other end stations.

3.1.2. CSMA/CD

With Ethernet, before sending its data, all stations first access the network to see if the communication channel is available and ready for transmission. This is called Carrier Sense (CS). When no stations are sending data, the communication channel is available and all stations have the same priority in starting a data transmission. This is called Multiple Access (MA). When the station detects a busy signal it does not try to send its data. After a short, random, interval time the station tries again. When, due to propagation delays for example, two data transmissions collide on the medium, a collision signal is generated which informs all stations about the collision occurrence. After a certain backoff time, the stations try to resend their data. The backoff time is a random number of slot times between 0 and 2^(n-1), where n is an integer representing the amount of retransmission attempts. The maximum amount of slot times to wait before a retransmission is 1023, which is reached at the tenth attempt. If after 16 retransmission attempts collisions still occur, the station gives up and the packet is dropped. This mechanism is called

Collision Detection (CD). The method for providing fair access to a shared channel is thus called Carrier Sense Multiple Access with Collision Detection (CSMA/CD).

3.2. Quality of service in wired networks

Most of current QoS standards are focussed on wired networks and much of the ideas introduced in these standards are also present in the proposed wireless QoS standards [28].

3.2.1. 802.1d QoS in MAC

In an attempt to create a standard for Quality of Service for LAN, the IEEE developed the 802.1d standard [15], which defines the protocol architecture for bridges and Layer 2 switches. The 802.1d protocol works on the MAC layer and deals with connecting LANs with equal MAC protocols as well as LANs with different MAC protocols through bridges and switches. In addition to passing frames form the incoming port to the outgoing port, bridges can also pass parameters from software controlling the incoming port to software controlling the outgoing port. Two of these parameters introduced in 802.1d are user_priority and access_priority which handle frame priorities.

The layer 2 device will derive the values of these introduced parameters from the priority fields in the incoming MAC frames. MAC protocols like IEEE 802.3 (Ethernet) and IEEE 802.11 (wireless LAN) don't have priority fields; see *Figure 2* which shows an Ethernet frame. Ethernet and wireless LAN therefore do not support the priority features of 802.1d. If the outgoing frame is Ethernet or wireless LAN, the priority field of the incoming 802.1d frame is ignored. If the incoming frame is Ethernet or wireless LAN, the priority field of the outgoing 802.1d frame is set to the default user priority.

preamble
destination address
source address
MAC length/type
data
FCS

Figure 2 - Ethernet IEEE 802.3 frame

Frames of other MAC protocols like Token Ring (IEEE 802.5 [20]) and FDDI (ISO 9314) do contain priority fields. *Figure 3* shows a Token Ring frame, consisting of various fields. The access control field controls several tasks including the frame priority.

starting delimiter
access control
frame control
destination address
source address
data
FCS
ending delimiter
frame status

Figure 3 - Token Ring IEEE 802.5 frame

The *access control* field contains 3 bits available for prioritisation, which corresponds to 8 priority levels. The *priority* field implies that a token is prioritised and indicates when a station may use the token. The access control field also contains 3 bits available for reservation. Stations that send high priority frames can request that the next token will be issued at the requested priority.

User priority in 802.1d is directly derived from the priority field in a MAC frame and is used to create output streams with different priorities. Using the priority field in a MAC frame, this value is communicated to other layer 2 devices and end stations. The access priority is used for prioritising access to shared media like Token Ring. A third concept introduced with 802.1d is called traffic classes. When various MAC frame queues are waiting to be transmitted, these queues can be classified in different traffic classes to determine the relative priority of these queues

3.2.2. 802.1q VLAN tagging

The IEEE 802.1Q standard [18], defines the architecture, protocols and mappings of bridges and switches to provide interoperability and consistent management of virtual LANs. A virtual LAN (VLAN) is a logical group of users, servers and other resources that appear to be on the same LAN segment, but can physically be on various segments, divided by VLAN aware switches (i.e. same broadcast domain, different collision domain).
To identify different VLANs connected tot a VLAN aware switch and to provide extra features, the VLAN tag, also called Q-tag, is added to the Ethernet frame. *Figure 4* below shows the VLAN tag (in grey) added to the Ethernet frame, in between the *source address* field and the *MAC length/type* field.

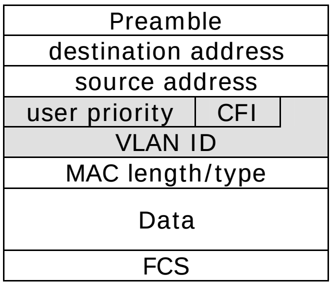

Figure 4 - Ethernet IEEE 802.3 frame with IEEE 802.1Q VLAN tag

The so-called Q-tag consists of three different fields. The *user priority* field is a 3 bit field, which is defined by the 802.1p standard. It describes the frame priority elaborated in chapter 3.2.3. The CFI (Canonical Format Indicator) field is a one bit field which indicates whether the MAC address is in canonical or non canonical format. The VLAN ID field is a 12 bit field defining the identification number of the different virtual LANs.

IEEE 802.1Q-compliant devices recognise the Q-tag and can use the information in the tag to forward the frame to the specified virtual LAN. Non 802.1Q-compliant devices don't understand the tag and will drop the frames.

3.2.3. 802.1p Priority bits in VLAN tag

The IEEE 802.1p standard [16][17], has two primary goals. First it expedites traffic capabilities by defining the User Priority field from the Q-tag. It also specifies a filtering service to support the dynamic use of group MAC addresses.

Using the three priority bits in the Q-tag, 802.1p provides 8 levels of priority. VLAN aware switches with multiple output queues use the value of the priority field of the frames to group the incoming data packets in different traffic classes. It's up to the network administrator to decide how the traffic classes are allocated, although IEEE has made some recommendations, described below.

Priority level 7, the highest priority, is preferably used for network critical traffic like RIP or OSPF table updates. Video and voice streams, which are delay-sensitive data transmissions, are recommended to get assigned to level 5 and 6. Levels 1 to 4 are best assigned to applications like streaming media and business-critical traffic. The lowest priority level, level 0, is the default. Level 0 is automatically assigned when no priority value is specified.

With switches only taking care of the order in which incoming frames are forwarded, 802.1p can be defined as best-effort QoS (Quality of Service). Frames are prioritised by the sender and classified by the switch and sent to the destination, no bandwidth reservation mechanisms are implemented.

3.3. 802.11 Wireless LAN

Just as the Ethernet standard (described in chapter 3.1) the key to the 802.11 specification [9] is the MAC sub layer. Like Ethernet, the 802.11 standard uses a form of CSMA to provide access to the (wireless) medium. However, unlike Ethernet it does not use Collision Detection (CD) but rather uses Collision Avoidance (CA). Using CA saves valuable transmission capacity. The main difference between 802.11 Wireless LAN and 802.3 Ethernet is the underlying medium (wired vs. wireless).

3.3.1. Wireless LAN network composition

In this document it is assumed for a wireless LAN to consist of at least two parts, an access point and a host station.

3.3.2. MAC access modes

In 802.11 access to the wireless medium is provided by so-called coordination functions. The 802.11 standard defines two different coordination functions:

- Distributed Coordination Function (DCF), this is much like the CSMA/CA mechanism. It first checks if the medium is free before transmitting. To avoid collisions (hence CA) each station uses a random backoff after each frame. In this method the first-come-first-serve principle is used, so the first transmitter seizes the channel. More on DCF is described in chapter 3.3.3.
- Point Coordination Function (PCF), provides time bounded services to allow stations to have priority access to the medium. PCF uses special stations called point coordinators (PC) to ensure the medium is available without contention. More about PCF and the way it implements some sort of QoS is described in chapter 3.3.4.

3.3.3. Distributed Coordination Function

DCF uses a carrier-sensing mechanism in order to determine if the medium is available. In 802.11 two carrier-sensing mechanism manage this problem, physical carrier-sensing and virtual carrier-sensing [1].

Physical carrier-sensing is provided by the physical layer (OSI layer 1) and depends on the medium and modulation used. Building physical layer carrier-sensing is difficult for wireless networks and not very reliable because of the existence of hidden nodes (nodes that are invisible to other nodes).

Virtual carrier-sensing is implemented in the 802.11 standard by using the so-called Network Allocation Vector (NAV). Most 802.11 frames contain a duration field. This field can be used to reserve the medium for a fixed period of time. The NAV is a counter that indicates the amount of time the medium will be reserved. Some stations set the NAV to the time they expect to use the medium while other count down form the NAV to zero. Whenever the NAV is nonzero it indicates the medium is busy.

If the medium is available, as detected by one of the carrier-sensing methods, the station can begin transmitting. As with Ethernet, 802.11 uses spaces between two frames, called inter-frame spaces. The 802.11 standard has four different inter-frame spaces all of which are used to determine access to the medium. Varying inter-frame spaces allow different types of traffic to have different priority levels allowing high-priority traffic to grab the channel before any lower priority traffic does. The four inter-frame spaces (in ascending order) as implemented by 802.11 are:

- Short Inter-frame Space (SIFS), is used for highest priority transmissions such as control or acknowledge frames.
- PCF Inter-frame Space (PIFS) is used by PCF during contention free operation.
- DCF Inter-frame Space (DIFS) is used by DCF and to be considered the medium idle time for contention based services.
- Extended Inter-frame Space (EIFS) has no fixed length and is used only when an error in frame transmission occurs.

The inter-frame spaces overlap so if one station grabs the medium during the PIFS a station using DIFS cannot grab the medium. By using both SIFS and NAV, stations can reserve the medium for as long as necessary.

After frame transmission has completed and the inter-frame space has elapsed, a station may transmit congestion-based data. Especially for this purpose a so-called Contention Window (CW) follows the DIFS inter-frame space and serves a backoff period. When having entered the CW the station waits before accessing the medium again. Each time a transmission fails, the backoff window is increased. Different physical layers use different CW sizes. CW sizes are always a power of two decremented by one. So if a re-transmission occurs, the CW size is raised to the next greatest power of two. If, during the backoff (i.e. before the backoff reaches zero), the medium is sensed busy the station will have to wait for the medium being idle for DIFS again. The CW is divided into a number of slots. The length of each slot is dependent of the medium. Media with higher transfer speeds will have shorter slot times.

3.3.4. Point Coordination Function – Limited QoS support

The PCF can be used if contention free delivery is required. However, PCF is an optional part of the 802.11 specification and thus will not be found on all products. The PCF operation is divided in two periods, a Contention Free Period (CFP) and a Contention Period (CP). During the CP the DCF based service runs while during the CFP the PCF runs. As stated, PCF uses point coordinators to restrict access to the medium. Stations on the network can only transmit data when they are allowed to do so by the point coordinator (a special function implemented in the access points).

At the beginning of a CFP the point coordinator transmits a Beacon Frame (BF). This BF consists of several parameters: synchronisation of local timers, protocol parameters and the maximum duration of the contention free period. Access points generate beacon frames at regular intervals. This makes sure every station knows when to receive a BF, the time between the transmissions of two beacon frames is called the Target Beacon Transition Time (TBTT). Whenever any station receives the BF it sets it's NAV to the maximum duration in order to block any DCF traffic to the wireless medium. To make even more sure DCF traffic cannot reach the medium, all PCF transmissions are separated only by the SIFS and PIFS inter-frame spaces. These inter-frame spaces are both shorter than the DIFS inter-frame space so any DCF based station cannot grab the medium.

The access point keeps a polling list. After an access point has seized control of the medium it polls any associated stations on this polling list for data transmissions. If, during a CFP, a station wants to transmit, it can only do so if the access point invites it to send data with a polling frame (CF-Poll). Every CF-Poll sent by the access point allows for one frame of data to be send.

Generally, all transmissions during the CFP are separated by the SIFS. In order to make sure the PC retains control of the medium, it sends to the next station on its polling list after it has waited for one PIFS. By using the PIFS the access point makes sure it retains control of the medium. The CFP is ended by the CF-End frame sent by the point coordinator. *Figure 5* below shows the operation of the point coordination function.

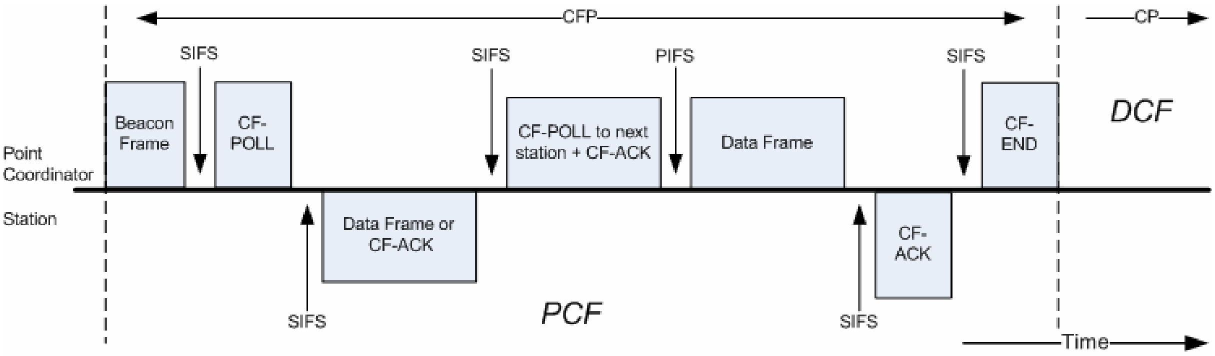

Figure 5 - PCF Operation

3.3.5. 802.11 Timings

The various inter-frame spaces have different timings which are listed in *Table 1* below. The following formulas are used to calculate the different times:

- PIFS = SIFS + Slot-time
- DIFS = SIFS + (2 x Slot-time)

	SIFS (µs)	**Slot-time (µs)**	**PIFS (µs)**	**DIFS (µs)**
802.11a [9]	16	9	25	34
802.11b [10]	10	20	30	50
802.11g [12]	10	9/20	19/30	28/50

Table 1 - 802.11 Timings

4. QoS for wireless networks challenges

The quality of service support in the 802.11 standard has some limitations which make it far from an ideal quality of service implementation. Besides this, some general problems arise when implementing quality of service in wireless networks. This chapter describes these limitations.

4.1. Technical challenges

The implementation of quality of service in wireless networks imposes several new technical challenges:

- Wireless channel quality typically differs for different users, and randomly changes with time.
- Wireless bandwidth is a scarce resource which, if at all possible, should be used as efficiently as possible.
- In contrary to wired networks, wireless networks typically suffer from higher interference and higher error rates.
- Because of the mobile nature of wireless networks, it’s difficult to allocate the available resources properly.

A number of these challenges can be overcome by using the right physical layer (802.11a/b/g physical layers all differ, same goes for 802.16). What also helps is using non-overlapping channels for different access points (the 802.11 standard defines 14 channels which happen to overlap, leaving only three truly free channels). Avoid RF interference to make sure other devices do not interfere with the 802.11 stations. Placing sufficient access points will allow for better coverage and will limit the number of users per access point and thereby increase the bandwidth available per user.

4.2. 802.11 PCF limitations

While the point coordination function provides a rudimentary form of quality of service support, it also has its limitations [4]:

- Beacon delays; the contention period has no fixed length, the minimum length of the CP is the time required to transmit and acknowledge one frame. There is no maximum length however. It is possible for the CP to run longer than planned due to traffic still being transmitted. While the target beacon transition times are fixed the CP will run past the TBTT. The CP running past the TBTT results in the beacon frame being transmitted later than planned and thus shortens the following contention free period (i.e. the BF is delayed). Besides delaying the BF it also delays the transmission of time bound frames during the CFP. This severely hampers the QoS as it introduces a possible unpredictable delay for each CFP.
- Unknown transmission durations of the polled stations; as stated, a station polled by the point coordinator is allowed to transmit a single frame. However there is no specified time a station can take to transmit the frame (perhaps it needs to be encoded or fragmented). This will affect the QoS for other stations that are polled during the rest of the CFP.
- Lack of knowledge of the offered traffic at the stations; the access point has no knowledge of the traffic offered at the polled stations. This can result in a situation in which the PC polls every station using a round robin scheduling algorithm and waists a lot of time doing this. Thus any station having time critical traffic will have to wait until the PC has finished polling the other stations. This will affect QoS parameters for these traffic categories.
- The PCF is an optional part of the 802.11 specification and is therefore not widely available, as most manufacturers of wireless equipment chose not to implement the PCF. One of the few access points implementing the PCF is an access point manufactured by Actiontec [5].

5. Proposed standards

This chapter focuses on the newly proposed wireless standards: 802.11e and 802.16. The way these standards handle quality of service is discussed

5.1. 802.11e – MAC Enhancements

The proposed 802.11e standard [11][4][24][26] introduces a new access method which is called the Hybrid Coordination Function (HCF). The HCF has two basic methods, the enhanced DCF (EDCF) and the HCF controlled access method.

5.1.1. Hybrid Coordination Function

The HCF supports three different QoS levels. *Table 2* below shows these levels together with their scheduling policy:

QoS Levels	**Channel access mechanism**	**Scheduling policy**
Level 0	DCF, PCF	None
Level 1	HCF (EDCF only)	Prioritised
Level 2	HCF (EDCF and HCF controlled channel access)	Prioritised
Level 3	HCF (EDCF and HCF controlled channel access)	Parameterised

Table 2 - QoS Levels in HCF

An access point that implements the new 802.11e standard called a QAP (Quality of Service Access Point) while a station doing so is called QSTA (Quality of Service Station). The HCF is executed by a so called Hybrid Coordinator (HC) (implemented in a QAP) which can be seen as a centralised controller in one QoS Basic Service Set (QBSS). The HC performs its duties during both the contention and contention free periods. Essential to the workings of the 802.11e standard is the TXOP (Transmission Opportunities). A TXOP is an interval of time in which a station has the right to start transmissions on the wireless medium. TXOP is distributed by the HC using the point coordination function in order to provide Controlled Access Periods (CAP). During a CAP QoS data can be transferred. In order to provide the predefined delivery, priority, service rate and delay/jitter rates the HC may initiate a CCI (Controlled Contention Interval). During the CCI contention only occurs only between QSTA's because they need to request new TXOP's. Transmissions during the CAP are protected by using a virtual carrier sensing mechanism (NAV in this case, see chapter 3.3.3).

5.1.2. Enhanced Distribution Coordination Function

Just like its predecessor, DCF, EDCF is a contention based channel access method. EDCF provides differentiated service, distributed access to the medium for up to eight delivery priority levels. The maximum number of priority queues a QSTA should implement is eight, one for each priority level. A prioritised output queue is called a Traffic Category (TC). In EDCF the contention window is used to assign priority to each TC. For example, assigning a short CW to a high priority TC makes sure that, in most case high priority traffic is transmitted before low priority traffic. As with DCF, EDCF can use various inter-frame spaces. However, instead of using DIFS a newly defined inter-frame space is used: AIFS (Arbitration Inter-frame Space). The length of an AIFS is a DIFS plus a number (possibly zero) of time slots.

Within a QSTA there are running multiple backoff instances, each responsible for its own data delivery. Additionally, each backoff instance is parameterised with TC-specific parameters. Each running instance within a QSTA behaves just like a virtual station. Just like a station, it waits for a TXOP and also starts a backoff after it has detected the channel is idle. The backoff time is calculated as follows:

Backoff_time[TCi] = Rand(1,CW[TCi]) x aSlotTime

The backoff time is a random value between 1 and the value of the contention window times the slot time. As with DCF, the bakcoff will have to wait for AIFS if the medium is sensed busy before the backoff counter reaches zero.

Whenever the backoff counter of two or more parallel TCs in one station reaches zero at the same time, the highest priority TC is granted TXOP. In this case, the lower priority TC will act just as if a collision on the medium occurred. However, EDCF can only resolve the possibility of internal collisions so there still is a possibility of a collision on the external medium (the wireless medium itself). To enhance performance and achieve better medium utilisation, the Contention Free Burst (CFB) can be used in 802.11e. However, using this is completely optional.

The CFB may be used by both the QAP and the QSTA. If a QSTA or QAP has time remaining in its granted TXOP it may use this time to transmit any data it has left after waiting for one SIFS as long as the total access time does not exceed a certain limit, TxOpLimit. If a collision occurs during the CFB, packet bursting is terminated. CFB may increase delay jitter so TxOpLimit should be chosen carefully and just be enough to transmit the largest frame.

5.1.3. HCF controlled channel access

HCF controlled channel access uses a hybrid coordinator to manage the allocation of wireless medium bandwidth. The HC is often implemented in the QAP of a QBSS. In order to be able to initiate TXOPs to the wireless stations the HC needs to have a higher priority than the wireless stations in the QBSS. The 802.11e standard takes care of this problem by making sure the HC can access the medium after PIFS which is shorter than both DIFS and AIFS granting it priority over both DCF and EDCF based traffic. The HC can perform its tasks during both the CP and the CFP in order to meet the QoS requirements of different traffic categories.

Within HCF controlled channel access some special periods are defined, CAPs. A CAP is an interval within a CP in which short bursts of frames can be transmitted using policy based controlled access mechanisms. When, within a CP, the frames are not transmitted using the CAPs, the EDCF rules are used. The CAPs can be started at any time the medium remains idle for at least one PIFS.

HCF controlled channel access also offers the option to handle periodic traffic. This is implemented by so called Contention Controlled (CC) periods. During which stations can request the allocation of polled TXOPs. CC is used by the HC to create lists of which stations need to be polled, at which time and for which duration. This way, HCF controlled channel access can provide more guaranteed services than EDCF.

5.2. 802.16 – WirelessMAN

Even though the demand for broadband connections is continually growing [13], in both corporate as well as residential environments, not all users have access to a broadband connection. There are several reasons for the lack of nationwide broadband cover. People might think it’s too expensive or they may be in an area where there’s no DSL or cable infrastructure. The IEEE introduced the 802.16 standard [14], which defines the wireless MAN (metropolitan area network) air interface specification, also known as WirelessMAN (which is a registered trademark from the IEEE). This standard can help crossing the “last mile” to the homes and offices of users that do not have wired access to a broadband connection and it could be the beginning of fixed Broadband Wireless Access (BWA) offering fibre-optic-like speeds.

5.2.1. 802.16 standard

The 802.16 standard addresses frequencies in the range from 10 to 66 GHz [2]. It requires a line of sight between the sending and the receiving station, because no functions are implemented for correcting odd incoming signals, caused by the reflection of these signals on buildings or other large objects. Chapter 5.2.4 describes the 802.16a standard which addresses a lower frequency range, not requiring a line of sight, at the cost of offering lower data rates.

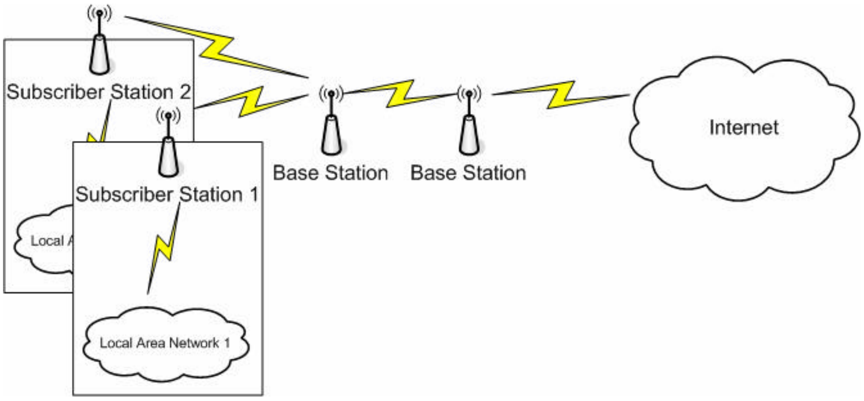

Figure 6 - 802.16 setup

Like wireless LAN, WirelessMAN uses an access point (also called a base station) and an end station with antenna (also called a subscriber station). The base station is connected via a high speed backbone to the Internet and can provide up to 60 users with T1 connection speeds (up to 1.5 MBps) and hundreds of users with DSL-like speeds. WirelessMAN is not a point-to-point architecture like wireless LAN, but a point-to-multipoint architecture. One base station serves multiple subscriber stations. A subscriber station typically serves a building; either commercial or residential (see Figure 6 above).

Because of the line-of-sight requirement IEEE opted for single carrier modulation which they called WirelessMAN-SC [21]. Because of the point-to-multipoint architecture, the base station transmits a TDM (time division multiplexing) signal. The TDM signal serially allocates different time slots to each subscriber station. The subscriber stations transmit by TDMA (time division multiple access). Two duplexing variants are defined. With time-division duplexing (TDD) the base station and the subscriber station both use the same channel, but do not transmit at the same time, resulting in a half-duplex operation mode. With frequency-division duplexing (FDD) the base station and the subscriber station both use different channels and can transmit at the same time, resulting in a full-duplex operation mode. A cheap variant was added later, half-duplex FDD.

5.2.2. 802.16 MAC Layer

For 802.16, the IEEE subdivided the Media Access Control sub layer in a common part sub layer and a service specific convergence sub layer. The service specific convergence sub layer takes care of mapping services to and from 802.16 MAC connections. For ATM services the ATM convergence sub layer was defined. For mapping packet services like IPv4, IPv6, Ethernet and VLAN the packet convergence sub layer was defined. The service specific convergence sub layer checks the incoming service data units (SDU's) and classifies them to the correct MAC connection. It also preserves or enables QoS and bandwidth allocation.

The MAC layer of 802.16 is connection-oriented. All services are mapped to a connection, even those that are actually connectionless. Mapping services to a connection helps when applying QoS options and reserving bandwidth. Each connection is identified with a 16 bit connection identifier (CID), the 48 bit MAC addresses of the subscriber stations are only used as equipment identifier.

Subscriber stations that enter the network get three unidirectional uplink and three unidirectional downlink management connections. These management connections represent three different QoS requirements used by different management levels. The connections are unidirectional, providing the ability to use different QoS parameters for the uplink and downlink. The first connection, the basic connection, is used for transmitting time critical MAC and radio link control (RLC) messages. The primary management connection is for transmitting longer, more delay-tolerant messages like authentication and connection setup messages. The secondary management connection is for transmitting management messages like DHCP, TFTP and SNMP. For transmitting normal data, the subscriber stations are assigned transport connections

The data units sent between the base station and the subscriber stations are called MAC protocol data units (PDU). A PDU consists of a header with a fixed length and a payload with a variable length. *Figure 7* below shows a generic MAC PDU header. There are two types of headers; the generic header and the bandwidth request header, differentiated by the *header type* (HT) field. Generic MAC PDU's contain MAC management messages or convergence sub layer data, bandwidth request header don't contain any payload.

header type
encryption control
type
reserved
CRC indicator
encryption key sequence
reserved
length
connection identifier
header check sequence

Figure 7 - WirelessMAN IEEE 802.16 generic MAC header

5.2.3. QoS in 802.16

The 802.16 standard has a Quality of Service mechanism [22] which includes concepts like service flow QoS scheduling, dynamic service establishment and the two-phase activation model. Each uplink connection is mapped to a single scheduling service, while a scheduling service can be associated with multiple connections. A scheduling service represents the data handling mechanisms supported by the MAC scheduler and consists of a set of rules the base station scheduler uses for bandwidth allocation and for the request-grant protocol between the base station and the subscriber station. When a subscriber station first connects to the base station the specifications of the scheduling service are negotiated. The scheduling services in the DOCSIS standard [6] (developed by CableLabs and approved by the International Telecommunication Union (ITU) in March 1998) form the basis for the scheduling services in 802.16.

Service flow QoS scheduling is the principal mechanism for providing QoS in 802.16. A service flow is defined by a QoS Parameter Set like latency, jitter and bandwidth reservations. For each service flow, a different type of QoS can be assigned. The base station and the subscriber station both take care of providing the QoS defined by the service flow.

The scheduling services support four services: unsolicited grant service (UGS), Real-time polling service (rtPS), non-real-time polling service (nrtPS), and best effort (BE). A grant is an opportunity for the subscriber station to begin a data transmission. The UGS is used when the base station is allowing the subscriber station to transmit without the subscriber station having requested bandwidth. This is used for subscriber stations which transmit fixed data packets at regular intervals, like voice over IP. The used use QoS service flow parameters for this scheduling service are Time Base, Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Maximum Latency, SDU Size, Tolerated Jitter, and Request/Transmission Policy for upload service flows.

The rtPS is used for data streams with variable-sized data packets sent at fixed intervals, like MPEG video. This service uses Time Base, Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Maximum Latency, Tolerated Jitter, and Request/Transmission Policy for upload service flows as QoS parameters. Delay-tolerant data streams with variable-sized data packets which require a minimum data rate, like FTP, use the nrtPS. NrtPS uses Time Base, Minimum Reserved Traffic Rate, Maximum Sustained Traffic Rate, Traffic Priority, and Request/Transmission Policy for upload service flows as QoS parameters. When no minimum service level is required, the best effort service is used which uses Time Base, Maximum Sustained Traffic Rate, Traffic Priority and Request/Transmission Policy as QoS parameters.

The scheduling services also implement bandwidth allocation [22] [3]. A subscriber station can request bandwidth allocation in three ways.

- Use the “contention request opportunities” interval when he’s polled by the base station (multicast or broadcast poll).
- Send a standalone MAC message called “BW request” (bandwidth request) in an already granted slot.
- Piggyback a BW request message on a data packet.

The base station has two ways to allocate bandwidth. It can be allocated to a single connection between a base station and a subscriber station, this mode is called grant per connection (GPC). It can also allocate bandwidth to a subscriber station (i.e. to all connections of the subscriber station), this mode is called grant per subscriber station (GPSS). Subscriber stations and the base station can negotiate about the bandwidth allocation and some subscriber station can take care of their own allocations. With this architecture various types of services, like connectionless IP traffic and connection-oriented ATM traffic, are supported at the same time.

5.2.4. 802.16a

The main differences between 802.16 and 802.16a lie on the physical layer. Where the high frequencies of 802.16 require a line of sight between the base station and the end station, 802.16a uses frequencies ranging from 2 to 11 GHz. It addresses both licensed an unlicensed frequency bands.

When residential implementations were expected, the need for non-line-of-sight (NLOS) arose. Rooftops may be to low and other buildings or trees may obstruct the line of sight. Antennas mounted on top of buildings are more expensive in hardware and installation then indoor antennas. The 802.16a standard has functions implemented for correcting incoming reflected signals.

802.16a has three air interface specifications. WirelessMAN-SC2 uses a single-carrier modulation format. WirelessMAN-OFDM uses 256-carrier orthogonal frequency-division multiplexing and is compulsory for the license free frequency bands. Last there is WirelessMAN-OFDMA which uses 2,048-carrier orthogonal frequency-division multiple access, supporting multiple access by addressing part of the available carriers for each receiver.

6. QoS in low-latency/low-bandwidth

This chapter describes how the 802.11e and 802.16 standards fit in an example low-latency/low-bandwidth networking environment introduced below. An answer to the research question (see chapter 2.3) focussed on one specific low-latency/low-bandwidth environment is given.

6.1. Example low-latency/low-bandwidth environment

A good example of a low-latency/low-bandwidth environment is online gaming. With online gaming one can think of games like first person shooters, real-time strategy games or racing games. More classic examples are online chess, backgammon or card games. A feature these games have in common is that only game updates (for example unit/chess piece position, player position or player speed) are sent over the network. Most of the time, the packets with game updates will be small (often less than 20 bytes [27]), only containing those updates necessary for that moment. When no user updates occur, status updates are sent on a regular interval [25].

Of the game genres mentioned above, first person shooters require players to quickly respond to gaming situations and to be able to interact fast to actions of other players (for example one player is shooting another player who wants to dodge and shoot back). Any amount of lag will be experienced as annoying and prevents the player from correctly responding to a gaming situation. To emphasise, in fast action games like a first person shooter, small data packets are sent on irregular intervals.

The prioritisation of traffic will allow packets to traverse the network with a minimum amount of lag, offering maximum playability. To increase speed even further, most games use UDP as transport protocol. Besides offering one great advantage over TCP (speed) it also has some disadvantages; no error checking and no 100% guarantee all data is sent.

6.2. 802.11e

In 802.11e the HCF solves the problems with the beacon delays (as described in chapter 4.2) by introducing the controlled access periods. By using the CAP mechanism, the hybrid coordinator can ensure for a period of contention free access (i.e. no collisions). During this CAP, any station has a period of time (of size TXOP) in which it can send data. The lengths of the TXOP and CAP periods as well as the point in time when they occur are fixed. When data is send at regular intervals (e.g. with streaming media) this is very useful because client stations will know when they can send with low latency values.

In the gaming example above (chapter 6.1) TXOP/CAP is less useful. Because of the nature of gaming, packets are typically not sent on regular intervals, not making full use of the advantages the TXOP/CAP mechanism offers. If games are played over a TXOP/CAP enabled network, the use of the contention free burst will allow a station to transmit more than one data packet within the boundaries of its TXOP. Using this, the transmission will regain some of the speed lost as a result of using TXOP/CAP.

EDCF offers the opportunity to use up to eight different priority levels. Traffic using a prioritised queue will have priority over non-prioritised traffic allowing for lower latencies. This is useful in gaming environments as these environments profit from low-latency values.

HCF controlled channel access is used to manage the allocation of wireless bandwidth. The allocation of bandwidth is less useful in a gaming environment while typically only small data packets are sent. In other applications, like streaming media, bandwidth allocation will be more useful while these environments require a continuous data stream as well as low latency values.

All of the quality of service mechanisms implemented in 802.11(e) (EDCF, PCF and HCF) offer the opportunity, in one way or another, to send data in a contention free period. While most games use UDP as a means of data transport this is especially useful. Because of the connectionless nature of the UDP protocol, no acknowledges upon receipt are used and no retransmission functionality is implemented. Therefore, a contention free period is useful because this mechanism can guarantee no collisions occur, increasing the chance of data delivery.

One thing to keep in mind is that, no matter what quality of service is used, the wireless medium is much more sensitive to jamming/link failure than a wired medium. Much of the service quality guarantees are thus only valid in case of a reliable wireless link.

6.3. 802.16

From the four scheduling mechanisms 802.16 offers (UGS, rtPS, nrtPS and BE) only best effort is useful (however, the usefulness of BE is determined only by the fact that the other mechanisms are not useful at all).

- UGS can provide QoS for fixed-sized data packets sent at regular intervals, making it not useful for games but useful for other applications like voice over IP.
- rtPS provides QoS for variable-sized data packets sent at regular intervals, making it useful for applications like MPEG video.
- nrtPS is focussed on delay tolerant variable-sized data streams, making it not useful for a fast action game like a first person shooter. It could be used for games like chess however.
- BE is used when no minimum service level is required but provides traffic rating and prioritisation making it useful for use in gaming environments.

The 802.16 standard is designed for crossing the “last mile” to users who don’t have direct wired access to a broadband connection, like in remote, rural areas. 802.16 WirelessMAN can be used as a backbone to these areas, being able to span distances up to 50 km. A WirelessMAN base station can thus be considered as a point of access to the backbone and to the other subscriber stations. When two subscriber stations, connected to the same base station, maintain a networked gaming session, the base station can use bandwidth allocation between the two subscriber stations. A guaranteed data stream will be available for the data transmission between the two subscriber stations, minimising lag.

Bandwidth can be allocated in two ways, GPC and GPSS. For a single online gamer per subscriber station GPC can be used, not consuming all available bandwidth. In case of a LAN-party (a gathering of multiple online gamers) with users connected to multiple subscriber stations, GPSS can be used. This provides a guaranteed data stream for all the connections of the subscriber stations.

7. Conclusion and Future Work

7.1. Conclusion

This report introduced the upcoming IEEE 802.11e and IEEE 802.16 standards both of which provide some form of quality of service for wireless networks. The standards were researched theoretically and their most distinctive properties regarding QoS were provided. Finally a look on how the proposed standards fit in an example low-latency/low-bandwidth environment, gaming, was given.

Both proposed standards for quality of service in wireless networking do not provide specific mechanisms for use in low-latency/low-bandwidth networks. 802.11e provides priority mechanisms useful in low-latency environments while the 802.16 standard offers multiple mechanisms for bandwidth allocation. Unfortunately these are less useful in low-latency/low-bandwidth environments.

802.16 in contrary to 802.11 will be less suitable for use in an office environment because of its line-of-sight nature. This problem is solved with an extension for 802.16, 802.16a which will allow for non-line-of-sight transmissions. The fact that the 802.16 standard is designed to be used over large distances (up to 50 km) and provides high bandwidth makes it more suitable to be used as a wireless backbone.

Concluding, the mechanisms 802.11e offers will be the most useful in the example low-latency/low-bandwidth environment previously mentioned. 802.16 will have most of its uses as a wireless backbone.

7.2. Future Work

This report provided an abstract theoretical analysis of the IEEE 802.11e and IEEE 802.16 standards. It is not yet finished however; a number of recommendations for future work are listed below.

- Simulation. In this document, a theoretical analysis is given. A computer simulation of the two standards will give a more thorough look into the performance of the two standards.
- Real-life testing. Because both standards are not yet implemented in any product no real tests could have been performed.
- Other environments. Reflecting on both standards in another environment than gaming. Streaming media or voice over IP for example.
- Other wireless standards. Other wireless standards like IEEE 802.15.3 (wireless PAN) could be researched to provide a more thorough look into the subject.

8. References

[1] 802.11 Wireless Networks – The Definitive Guide; Matthew S. Gast; O'Reilly, first edition April 2002; ISBN: 0-596-00183-5

[2] 802.16: A Future Option for Wireless MANs; Geier J.; July 17, 2003

[3] 802.16: A look under the hood; Cohen B., Deutsch D.; Wifiplanet 26 August 2003

[4] A survey of QoS Techniques and Enhancements for IEEE 802.11 Wireless LANs; Gannoune L., Robert S., Rodellar D.; Swisscom Innovations

[5] Actiontec 802.11 wireless access point; http://www.actiontec.com/NETH/products/broadband/80211bwap/80211bwap_overview.html

[6] Docsis 2.0 specification; Cable Labs; 7 March 2004

[7] Ethernet - The Definitive Guide; Charles E. Spurgeon; O'Reilly, first edition February 2000; ISBN: 1-56592-660-9

[8] Face of the Web 2003 Press Release; Ipsos-insight may 11 2004; http://www.ipsos-na.com/news/pdf/media/mr040511-1revis.pdf

[9] IEEE 802.11, 1999 Edition (ISO/IEC 8802-11: 1999) IEEE Standards for Information Technology -- Telecommunications and Information Exchange between Systems -- Local and Metropolitan Area Network -- Specific Requirements -- Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications

[10] IEEE 802.11b-1999 Supplement to 802.11-1999,Wireless LAN MAC and PHY specifications: Higher speed Physical Layer (PHY) extension in the 2.4 GHz band

[11] IEEE 802.11e Wireless LAN for Quality of Service; Mangold S., Choi S., May P., Klein O., Hiertz G., Stibor L.; ComNets RWTH Aachen Univ. of Technology, Philips Research USA, Philips Research Germany

[12] IEEE 802.11g-2003 IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications—Amendment 4: Further Higher-Speed Physical Layer Extension in the 2.4 GHz Band

[13] IEEE 802.16 WirelessMAN Specification Accelerates Wireless Broadband Access; Johnston D.J., LaBrecque M.; Intel

[14] IEEE 802.16-2001; IEEE Standard for Local and Metropolitan Area Networks--Part 16 -- Air Interface for Fixed Broadband Wireless Access Systems

[15] IEEE 802.1D, 1998 Edition (ISO/IEC 15802-3:1998), IEEE Standard for Information technology--Telecommunications and information exchange between systems--IEEE standard for local and metropolitan area networks--Common specifications--Media access control (MAC) Bridges

[16] IEEE 802.1p; http://www.nwfusion.com/details/475.html

[17] IEEE 802.1p; http://www.xilinx.com/esp/networks_telecom/optical/net_tech/ieee8021p.htm

[18] IEEE 802.1Q, 2003 Edition, IEEE Standards for Local and metropolitan area networks—Virtual Bridged Local Area Networks

[19] IEEE 802.3-2002 IEEE Standard for Information technology--Telecommunications and information exchange between systems--Local and metropolitan area networks--Specific requirements--Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications

[20] IEEE 802.5, 1998 Edition (ISO/IEC 8802-5:1998) IEEE Standard for Information technology--Telecommunications and information exchange between systems--Local and metropolitan area networks--Specific requirements--Part 5: Token Ring Access Method and Physical Layer Specification

[21] IEEE Advances Wireless MAN Standard; Olavsrud T.; January 31, 2003

[22] IEEE Standard 802.16: A Technical Overview of the WirelessMAN™ Air Interface for Broadband Wireless Access; Eklund C.; Nokia Research Center Marks R.B.; National Institute of Standards and Technology Stanwood K.L., Wang S.; Ensemble Communications Inc. IEEE Communications Magazine - June 2002

[23] Latency; http://www.whatis.com; keyword latency

[24] Making QoS a reality over WLAN; Godfrey T.; PRISM Wireless Product Group Conexant Systems Inc. 30 April 2004

[25] Networked Java Game Specifcation; 26 May 2004; http://icawww1.epfl.ch/sc250_2004/project/protocol.pdf

[26] Performance Evaluation of IEEE 802.11e; Grilo A., Nunes M.; INESC/IST, Lisboa Portugal

[27] Quake II Network protocol specifications; Ferguson T.; 3 December 1998 http://www.csse.monash.edu.au/~timf/bottim/q2network-0.03.sgml
[28] Quality of Service for Local Area Networks; http://www.informit.com/articles/article.asp?p=25315
[29] Routing and Switching – Time of Convergence?; Rita Pužmanová; Addison-Wesley, first edition 2002; ISBN: 0-201-39861-3
[30] Scheduling Services for 802.16; Yanover V.; IEEE 802.16 Broadband Wireless Access Working Group 15 April 2004

9. List of abbreviations

AIFS	Arbitration Inter-Frame Space
ATM	Asynchronous Transfer Mode
BE	Best Effort
BF	Beacon Frame
BWA	Broadband Wireless Access
CA	Collision Avoidance
CAP	Controlled Access Period
CC	Controlled Contention
CCI	Controlled Contention Interval
CFB	Contention Free Burst
CF-End	Contention Free – End
CFI	Canonical Format Indicator
CFP	Contention Free Period
CF-Poll	Contention Free – Poll
CID	Connection Identifier
CP	Contention Period
CSMA/CA	Carrier Sense Multiple Access/Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access/Collision Detection
CW	Contention Window
DCF	Distributed Coordination Function
DHCP	Dynamic Host Configuration Protocol
DIFS	DCF Inter-Frame Space
DOCSIS	Data Over Cable Service Interface Specifications
DSL	Digital Subscriber Line
EDCF	Enhanced DCF
EIFS	Extended Inter-Frame Space
FCS	Frame Check Sequence
FDD	Frequency Division Duplexing
FDDI	Fibre Distributed Data Interface
GPC	Grant Per Connection
GPSS	Grant Per Subscriber Stations
HC	Hybrid Coordinator
HCF	Hybrid Coordination Function
HT	Header Type
IEEE	Institute of Electrical and Electronics Engineers
IPv4/IPv6	Internet Protocol version 4/Internet Protocol version 6
ISO	International Organisation for Standardisation
ITU	International Telecommunication Union
LAN	Local Area Network
LLC	Logical Link Control
MAC	Media Access Control
MAN	Metropolitan Area Network
MPEG	Moving Picture Expert Group
NAV	Network Allocation Vector
NLOS	Non Line Of Sight
nrtPS	Non Real Time Polling Service
OFDMA	Orthogonal Frequency Division Multiple Access
OSI	Open Systems Interconnection
OSPF	Open Shortest Path First
PAN	Personal Area Network
PC	Point Coordinator
PCF	Point Coordination Function
PDU	Protocol Data Unit
PIFS	PCF Inter Frame Space
QAP	Quality of Service Access Point
QBSS	Quality of Service Basic Service Set
QoS	Quality of Service
QSTA	Quality of Service Station
RIP	Routing Information Protocol
RLC	Radio Link Control
rtPS	Real Time Polling Service
SDU	Service Data Unit
SIFS	Short Inter Frame Space
SNMP	Simple Network Management Protocol
TBTT	Target Beacon Transmission Time
TC	Traffic Category
TCP	Transport Control Protocol
TDD	Time Division Duplexing

TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
TFTP	Trivial File Transfer Protocol
TXOP	Transmission Opportunity
UDP	User Datagram Protocol
UGS	Unsolicited Grant Service
VLAN	Virtual Local Area Network
WLAN	Wireless Local Area Network