802.11 Arbitration

White Paper

September 2009
Version 1.00

Author:
Marcus Burton, CWNE #78
CWNP, Inc.
marcus.burton@cwnp.com

Technical Reviewer:
GT Hill, CWNE #21
gt@gthill.com

Certified Wireless Network Professional®

Table of Contents

Executive Summary

This whitepaper is intended to provide a thorough exploration of the channel access methods used in IEEE 802.11 WLANs. While the 802.11 standard specifies multiple subsets of functionality, not all of them are currently used in today's WLANs; therefore, this paper will address, in detail, only those mechanisms currently used in DCF and HCF implementations. Because the topic of 802.11 arbitration is multifaceted and often very confusing, this paper will attempt to explain these mechanisms in detail, dissecting the many components included in 802.11 channel access. Some details have been intentionally excluded due to the overwhelming depth of this material. After examining the details, this paper will also provide a brief discussion of the applications and significance of understanding these functions as a WLAN professional.

Approach / Intent

The intended audience for this paper is any WLAN professional who seeks to understand the methods by which medium access is obtained in an 802.11 WLAN. As written for the CWNP Program, CWNA candidates will benefit from this understanding, as will those who have already become CWNA certified. Professionals desiring a detailed understanding of WLAN analysis will also benefit from this paper. In the absence of the 802.11 specification, this document can serve as a reference tool for professionals who must know the processes and parameters related to management of the wireless medium.

Introduction to 802.11 Channel Access

Several mechanisms are used as a part of 802.11 channel access to minimize the likelihood of frame collisions by multiple STAs attempting to access the transmission medium simultaneously. This function is necessary because the RF medium is half-duplex, meaning only one device can successfully transmit at a given time. The half-duplex constraint applies to devices that share a relative physical area and RF frequency.

In order to demonstrate the need for channel access mechanisms, a simple illustration may be helpful. People use something similar to uphold social propriety and to maintain effective conversations within a group. For example, imagine that your company has a meeting in which a manager, Noah, asks the group "does anybody have an update to share?" Each member of the group may look around before speaking, but perhaps one person, Mali, will immediately begin speaking and share an update. While Mali speaks, everyone else obeys the rules and waits patiently, listening. When Mali is finished, someone else will begin speaking. But, before the next person begins speaking, everyone quietly obeys the rules by following protocol. For example, it may be assumed that Noah will thank the first person for sharing, and then ask if anyone else would like to share. Let's assume that Noah thanks Mali and poses the question again. "Thanks Mali, anyone else?" Now, the floor is open again for someone else to begin speaking. Obeying the rules, everyone glances around the room, and perhaps two people start speaking at the same time. They can't both continue speaking or no one will understand either comment. Thus, they both stop, wait a random amount of time, and hopefully only one person begins speaking again. In some cases, neither of the two original speakers will take the floor, but someone else will begin speaking while the other two are deferring. Many variations of this process can and do occur, but as individuals follow the rules of social propriety, group conversation is more efficient and more effective.

This illustration demonstrates the need for a set of rules when there can only be one transmitter—or speaker—at a time. WLANs function in this manner. In order to minimize collisions, each station heeds the protocol to ensure that operation continues as efficiently as possible. Collisions often occur, but the processes used by the 802.11 protocol are in place to minimize the likelihood of collisions and define the appropriate response in the event that a collision is inferred[1]. As in our example, wireless transmitters will not know that a collision has occurred. The scenario above more closely reflects the approach of Ethernet devices that can physically detect a collision. However, the point remains that protocols that minimize collisions are necessary.

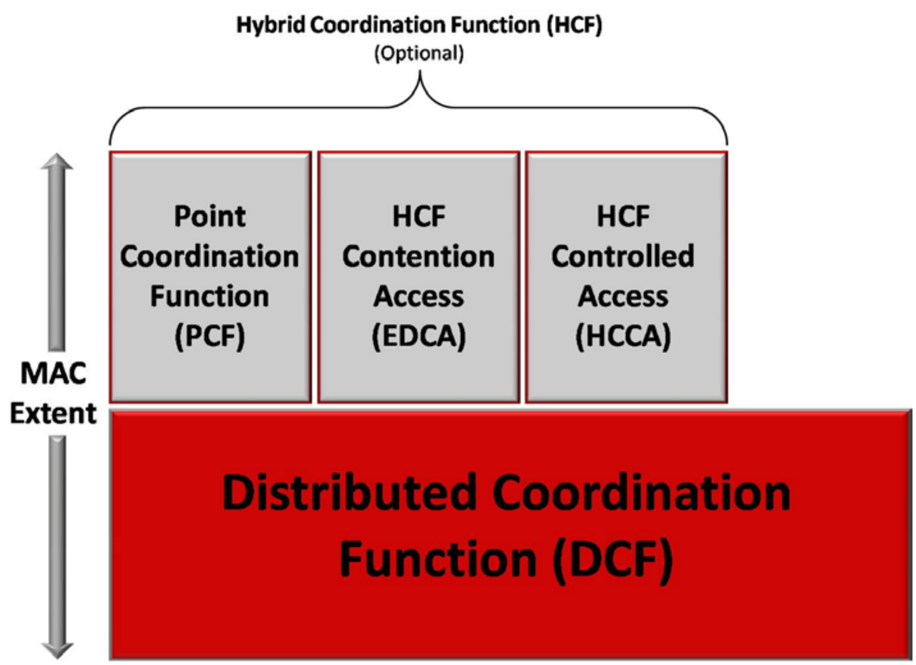
However, just as social expectations vary depending on the situation, WLAN access protocols can be affected by the features supported within a network. As in our example, it may be the case that a group of peers are conducting a meeting, in which case each member has an equal chance to speak; a separate meeting may include executives, managers, and laborers, where executives have a higher priority for speaking opportunities than both managers and laborers. These two different types of meetings and protocols could be compared, respectively, with the two most common coordination access methods used in WLANs: Distributed Coordination Function (DCF) and Enhanced Distributed Coordination Access (EDCA). The many, and often complex, elements of 802.11 access coordination are initially a little intimidating, but a detailed understanding of these mechanisms will help network designers, consultants, and other WLAN professionals to perform their jobs more effectively. As latency-sensitive applications begin to play a predominant role in WLANs, QoS becomes increasingly important. Similarly, airtime fairness mechanisms that help to distribute the use of the wireless medium more effectively are becoming more popular, and necessary. These are just two examples, but there are many more reasons why a thorough understanding of 802.11 access methods is important.

To that end, the following section will briefly explore the architecture of 802.11 access coordination, including the different access modes, most notably, DCF and EDCA. Following that, the specific components used to coordinate access in 802.11 networks will be examined in great detail. Finally, we will discuss the real-world uses for this knowledge and attempt to bring the monster to life.

802.11 MAC Channel Access Architecture

The 802.11-2007 standard renders a MAC channel access architecture to delineate the role of each channel access method used at the MAC layer[2]. The architecture demonstrates the foundational role of Distributed Coordination Function (DCF) and the optional role of Hybrid Coordination Function (HCF) and Point Coordination Function (PCF). As yet, neither PCF nor HCF Controlled Channel Access (HCCA) has been implemented in 802.11 networks, so they will be addressed cursorily.

FIGURE 1 – 802.11 MAC Architecture



Distributed Coordination Function (DCF)

Using the foundational DCF access method, the same coordination function logic is active in every station (STA) in a basic service set (BSS) whenever the network is in operation[3]. Stated differently, each station within a DCF follows the same channel access rules. This method is contention-based, which means that each device "competes" with one another to gain access to the wireless medium. After a transmission opportunity is obtained and observed, the contention process begins again. As the original 802.11 network access method, DCF is the most simple channel access method; however, being the first access method, it lacks support for quality of service (QoS). In order to maintain support for non-QoS devices in QoS-enabled networks, support for DCF is required for all 802.11 networks.

Point Coordination Function (PCF)

As an optional access method that may be used in addition to DCF, PCF is a contention-free access method. In contrast with DCF's contention-based access methods, PCF provides polling intervals to allow uncontended transmission opportunities for participating client devices. In this approach, the AP of a BSS acts as a point coordinator (PC), initiating contention-free periods in which prioritized medium access is granted to clients, one at a time. Despite the support of this mechanism in the 802.11 standard, IEEE left some details unaddressed, such as scaling PCF in an enterprise network, which left PCF as an immature access method. Vendors never saw the need to meet this development need, and as a result, PCF has gone unused in 802.11 WLANs.

Hybrid Coordination Function (HCF)

As an optional access method that may be used in addition to DCF, HCF was introduced to support QoS. HCF assimilated elements of both DCF and PCF mechanisms, creating a contention-based HCF method, called **EDCA**, and a contention-free HCF method, called **HCCA**. EDCA inaugurated a means of prioritizing contention-based wireless medium (WM) access by classifying 802.11 traffic types by User Priorities (UP) and Access Categories (AC). There are a total of 8 UPs, which map to 4 ACs, as shown in Figure 2. EDCA is used by stations that support QoS in a QoS BSS to provide prioritized WM access, but HCF is not used in non-QoS BSSs.

FIGURE 2 – EDCA User Priority Table

Priority	User Priority (UP)	802.11 Access Category (AC)	802.11 AC Designation	802.1D Designation
Lowest Priority	1	AC_BK	Background	BK
	2	AC_BK	Background	--
	0	AC_BE	Best Effort	BE
	3	AC_BE	Best Effort	EE
	4	AC_VI	Video	CL
	5	AC_VI	Video	VI
Highest Priority	6	AC_VO	Voice	VO
	7	AC_VO	Voice	NC

***Note** — UP values are identical to 802.1D UPs

The second, optional, access method offered by HCF is HCCA. HCCA was designed to provide parameter-based contention-free transmission opportunities (TXOPs) within a QoS BSS. Similarly to PCF, HCCA employs the AP as a QoS-aware centralized Hybrid Coordinator (HC), which initiates contention-free limited-duration controlled access to the WM for QoS-enabled STAs. For better or worse, HCCA has suffered a similar fate as PCF and has not been implemented by WLAN vendors.

Summary

- **DCF** is the fundamental, required contention-based access service for all networks
- **PCF** is an optional contention-free service, used for non-QoS STAs
- **HCF** Contention Access (**EDCA**) is required for prioritized contention-based QoS services
- **HCF** Controlled Access (**HCCA**) is required for parameterized contention-free QoS services

At the time of this writing, no vendor has implemented PCF or HCCA.

802.11 Channel Access Mechanisms

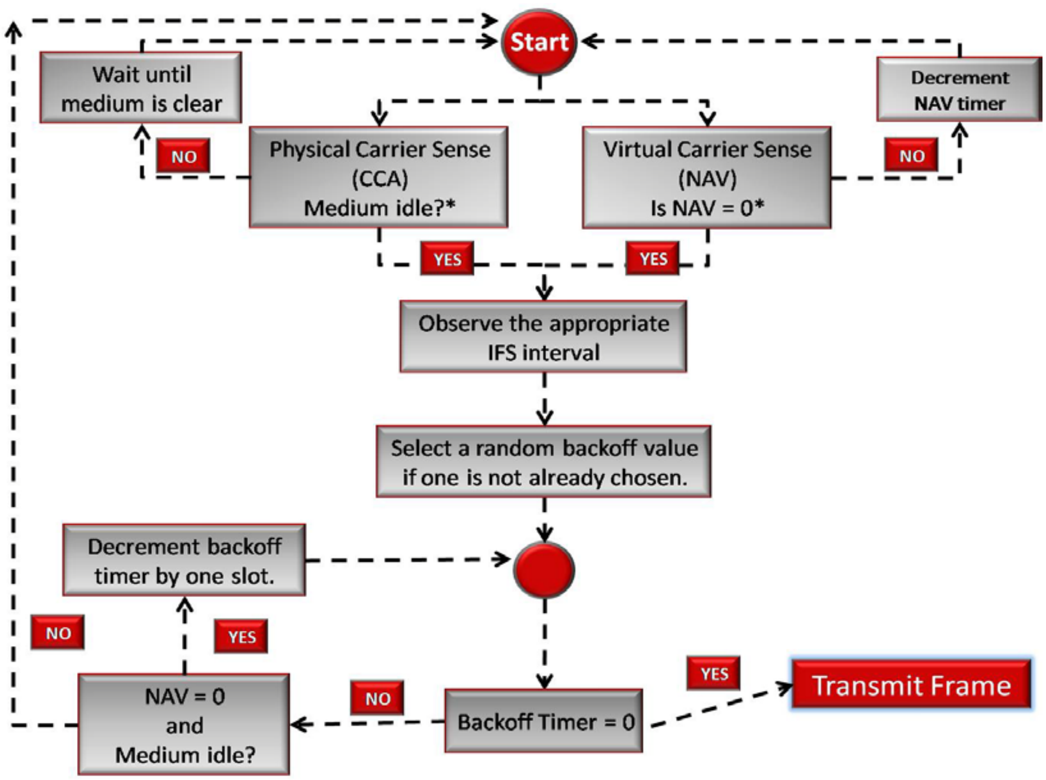
The 802.11 implementation of CSMA/CA includes several mechanisms that make coordinated WM access possible when many STAs are vying for access. Both contention-based access methods described previously (i.e. DCF and EDCA) employ similar mechanisms to moderate channel access and to minimize collisions. Some notable differences exist and we will elaborate on these where necessary. As a preparation for the details of the 802.11 arbitration mechanisms, a simplified outline and flowchart may be helpful.

Outline

1. STAs use a physical carrier sense (Clear Channel Assessment—CCA) to determine if the WM is busy.
2. STAs use virtual carrier sense (Network Allocation Vector—NAV) to detect if the WM is busy. When the virtual timer (NAV) reaches zero, STAs may proceed.
3. If conditions 1 and 2 are met, STAs wait the necessary IFS interval, as prescribed by the protocol.
4. If conditions 1 and 2 are met through the duration of condition 3, STAs generate a random backoff number in accordance with the range of allowed values.
5. STAs begin decrementing the backoff timer by one for every slot time duration that the WM is idle.
6. After decrementing the backoff value to zero, with an idle medium, a STA may transmit the allotted frame exchange, in accordance with the parameters of the obtained transmission opportunity.
7. If another STA transmits before Step 6 is completed, STAs observe steps 1, 2, 3, and 5 until the backoff timer is equal to zero.

8. After a successful transmission, repeat as needed.

FIGURE 3 – Arbitration Flowchart



* **Note** – These steps are performed continually throughout the flowchart. If either of these conditions change, the process returns to "Start"

Figure 3 provides a graphical perspective of the steps described previously. Despite the figure's simplicity, it should help readers to organize the flow of 802.11 arbitration. A detailed analysis of these functions follows.

Carrier Sense

The carrier sense protocol used by 802.11 WLANs is the well-known Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). CSMA/CA is used to minimize the probability of collisions on the network by determining the status of the wireless medium. For STAs to cooperate effectively on a half-duplex channel, each STA must be able to determine when the medium is clear, and when another device is actively transmitting. Both physical and virtual carrier sense mechanisms are defined by the 802.11 standard. As logic dictates, the physical carrier sense uses the physical radio interface to sample the wireless medium to detect transmissions. Virtual carrier sense refers to the use of Duration values in the MAC header of a frame and NAV timers to virtually determine if another STA is transmitting. Used in concert, these two mechanisms determine the status of the medium, whether idle or busy.

Physical Carrier Sense

The physical carrier sense mechanism defined by IEEE is known as clear channel assessment (CCA). CCA is the physical measurement taken by a radio interface to determine if the wireless medium is currently in use. The physical layer, which can be divided into two sublayers—the physical medium dependent (PMD, lower sublayer) and the physical layer convergence procedure (PLCP, upper sublayer)—performs this task and communicates this information to the MAC. According to the CCA mode in use, the PMD issues service primitives (protocol speak for "instructions") to the PLCP sublayer indicating whether the wireless medium is in use. The PLCP sublayer then communicates with the MAC layer to indicate a busy or idle medium, which prevents the MAC from attempting to forward a frame for transmission. Each PHY within the 802.11-2007 standard dictates the specific operations and signal thresholds used to carry out the CCA mechanism.

CCA can be divided into two separate processes: *Energy detection (ED)* and *carrier sense (CS)*. ED functionality is based upon raw RF energy. When an energy level detected in the channel crosses a certain threshold for a certain period of time, the "busy medium" indication will be triggered. On the other hand, CS—more precisely, "preamble detection"—monitors and detects 802.11 preambles, which are used to trigger the CCA mechanism and indicate a busy medium.

In the legacy 802.11 PHY specifications—particularly DSSS and HR/DSSS—CCA operations, including signal thresholds, were defined in great detail. However, most chipset manufacturers saw these details as overly restrictive and narrow, seeing more incentive to modify these parameters than to simply follow the specification. To improve the usefulness and reliability of CCA, chipset vendors have deviated from the 802.11 specifications, which has made this section of the 802.11 standard a poor representation of real-world implementations. Many infrastructure vendors consider these parameters in the chipsets of their APs, optimizing CCA behavior according to their specific architecture and desired functionality. More recent, and higher rate, PHY amendments—OFDM, ERP, and HT—have included fewer restrictions for CCA functionality, allowing chipset vendors to follow the specification more closely, while maintaining implementation flexibility for specific applications.

As an example of the CCA details for a PHY, let's look at 802.11a. According to 802.11-2007, Clause 17 (OFDM), CCA indicates a busy medium when the start of a valid OFDM transmission is received at a level equal to or greater than the minimum modulation and coding rate sensitivity (-82 dBm)—this is an example of CS. If the preamble of the frame is missed, the CCA holds a busy medium for any signal 20 dB above the minimum modulation and coding rate sensitivity, which equates to -62 dBm—this is an example of ED. Clause 17's CCA sensitivity is much higher (meaning, the signal can be lower power) when the frame header is received and processed correctly. The sensitivity threshold for RF noise, including frames whose header was not processed, as well as other non-802.11 RF sources, is much lower, at -62 dBm. Thus, it takes a much stronger signal to indicate a busy medium when ED is used instead of CS.

Virtual Carrier Sense

Due to the importance of proper interactivity on a WLAN channel, a virtual carrier sense mechanism is defined in addition to the physical carrier sense. Similar to the CS mechanism —"preamble detection"—defined previously, virtual carrier sense uses information found in 802.11 frames to predict the status of the wireless medium. This is performed by means of the network allocation vector (NAV), which is a timer that is set using the duration values in the MAC header of a frame. Each frame contains a duration value, indicating the time required for a station to complete the conversation. All STAs use these duration values to set their NAV, and then they count down the NAV timer, waiting for the medium to become available.

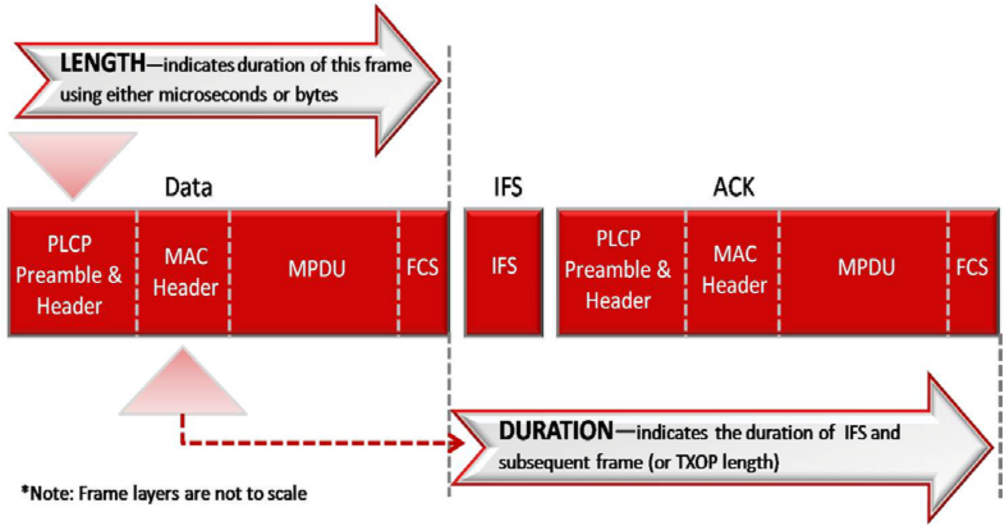
In order to accomplish this stage in 802.11 arbitration, all STAs attempt to process all frames—and a minimum of the first in a frame exchange—on their channel. The first frame in a frame exchange is significant because it can, and sometimes must, be used to determine how long a given transmission opportunity will occupy the wireless medium. It is important to understand the frame contents used in this process, so we will examine them in detail.

Each 802.11 frame begins with a PLCP (upper sub-layer of OSI layer 1) preamble that contains a sync field, allowing receivers to synchronize with the transmitter before the frame contents are transmitted. Following the synchronization bit sequence, a start frame delimiter (SFD) triggers the beginning of the PLCP header of a frame. Among other things, the PLCP header includes a LENGTH field, which indicates the length of that frame[4]. The LENGTH field allows STAs to know when a frame will end, even if they can't receive the entire frame due to data rate limitations. This is shown in Figure 4.

As described in the previous paragraph, the LENGTH field informs STAs when a frame will end. Virtual carrier sense uses a similar process to determine when a set of frames (transmission opportunity) will end, as follows. The PLCP header of a frame is immediately followed by the MAC header. The MAC header of each frame contains a Duration field, which indicates the amount of time necessary to complete the entire frame exchange, or the entire TXOP duration. In DCF, a transmission opportunity only allows for the transmission of one frame, thus the Duration value represents the required IFS interval and the acknowledgement frame (ACK), if one is required. The exception to this rule is for networks in which RTS/CTS or CTS-to-self protection is enabled. In this case, the transmission opportunity allows for the use of these frames. In HCF, several frames may be transmitted within a transmission opportunity. Thus, the Duration value refers to the TXOP duration. In either case, non-transmitting STAs must remain idle while the medium is reserved.

When STAs read the Duration value in a frame, they set their NAV timer accordingly and count down this duration. The duration value in the MAC header indicates the time required to complete the transmission opportunity after the current—the frame in which the Duration value resides—frame if completed. Thus, the duration value does not cover the frame in which it resides, as Figure 4 demonstrates. Figure 4 shows the LENGTH and DURATION fields for a data exchange followed by an acknowledgement frame (ACK). For simplicity, Figure 4 shows only these fields for the data frame, and not for the ACK, though they are also present in the ACK. If a STA is counting down its NAV and it receives another frame with a longer duration (would increase its NAV), the STA increases its NAV accordingly. Conversely, when a STA receives a frame with a shorter duration value (would decrease its NAV), the STA ignores this value and continues to observe the longer NAV duration.

FIGURE 4 – Length and Duration fields

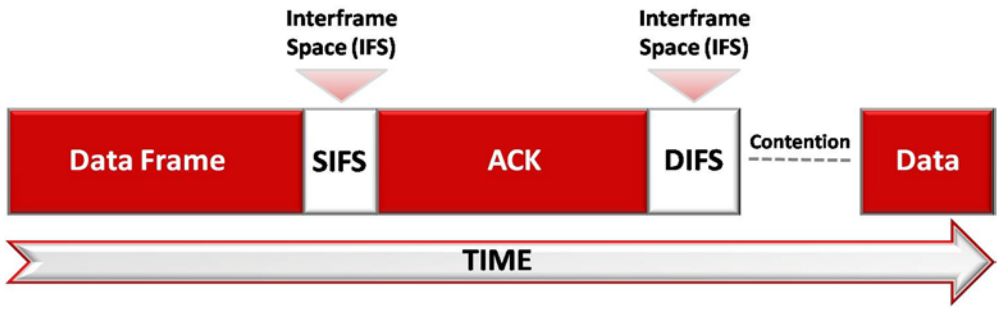


Because newer PHY technologies provide higher data rates, more complex modulation techniques, and other performance-enhancing features, transmissions using these more complex methods must be “protected” from legacy stations. In networks where mixed PHY technologies are supported, protection mechanisms are enabled to satisfy the requirements of frame processing and adherence to the common channel access protocol. Frames used as a protection mechanism (often an RTS/CTS exchange or CTS-to-Self) are transmitted at a common rate understood by all PHYs in the network. Legacy PHY STAs read the Duration value in the protection frame(s), set their NAV timer—we will discuss this function later in the paper—to this value, and then count down the NAV timer before proceeding with channel access mechanisms.

Interframe Spacing (IFS)

After each frame transmission, 802.11 protocols require an idle period on the medium, called an interframe space (IFS). The length of the IFS is dependent upon a number of factors, such as the previous frame type, the following frame type, the coordination function in use, the access category of the following frame (in a QoS BSS), as well as the PHY type. The purpose of an IFS is both to provide a buffer between frames to avoid interference as well as to add control and to prioritize frame transmissions. Each IFS “is the time from the end of the last symbol of the previous frame to the beginning of the first symbol of the preamble of the subsequent frame as seen at the air interface[5].” In other words, the IFS interval is observed beginning with the completion of the previous frame. The length of each IFS interval, excluding AIFS, is fixed for each PHY.

FIGURE 5 – Interframe Space Concept



Short Interframe Space (SIFS)

SIFS are used within all of the different coordination functions. For 802.11-2007, SIFS is the shortest of the IFSs and is used prior to ACK and CTS frames as well as the second or subsequent MPDUs of a fragment burst. However, with 802.11n, a shorter IFS (RIFS) was introduced. The IEEE explains the use of SIFS accordingly:

> "SIFS shall be used when STAs have seized the medium and need to keep it for the duration of the frame exchange sequence to be performed. Using the smallest gap between transmissions within the frame exchange sequence prevents other STAs, which are required to wait for the medium to be idle for a longer gap, from attempting to use the medium, thus giving priority to completion of the frame exchange sequence in progress."

In other words, SIFS is used as a priority interframe space once a frame exchange sequence has begun. It allows the participants of a frame exchange sequence to complete their conversation uninterrupted. This is true when multiple frames are transmitted within a TXOP (as with frame bursting) and it is also true when a single frame is transmitted (as with typical data-ack exchanges).

PCF Interframe Space (PIFS)

PIFS are used by STAs during the contention-free period (CFP) in PCF mode. Because PCF has not been implemented in 802.11 devices, you will not see PIFS used for this purpose. However, PIFS may be used as a priority access mechanism for Channel Switch Announcement frames, as used to meet DFS requirements. In order to gain priority over other STAs during contention, the AP can transmit a Channel Switch Announcement frame after observing a PIFS.

DCF Interframe Space (DIFS)

When a STA desires to transmit a data frame (MPDU) or management frame (MMPDU) for the first time within a DCF network, the duration of a DIFS must be observed after the previous frame's completion. The duration of a DIFS is longer than both the SIFS and PIFS.

Arbitration Interframe Space (AIFS)

"The AIFS shall be used by QoS STAs to transmit all data frames (MPDUs), all management frames (MMPDUs), and the following control frames: PS-Poll, RTS, CTS (when not transmitted as a response to the RTS), BlockAckReq, and BlockAck (when not transmitted as a response to the BlockAckReq)[6]." With EDCA, the basic contention logic is the same as with non-QoS networks, but in order to facilitate QoS, there are some notable differences. While DCF can designate a single DIFS value for each PHY, EDCA establishes unique AIFS durations for access categories (AC). For this reason, an AIFS is typically notated as an AIFS[AC]. QoS STA's TXOPs are obtained for a specific access category, so delineation between ACs must be made.

For improved control of QoS mechanisms, AIFS values are user-configurable. By default, QoS APs announce an EDCA parameter set in the Beacon frame that notifies stations in the BSS about QoS values. By changing these values in the AP configuration, the AP will broadcast a different set of parameters to the BSS.

FIGURE 6 – Default AIFSN in Beacon Frames

Figure 6 shows a Beacon frame that includes the AIFSN value for this BSS. These values can be found for each AC in the EDCA Parameter Set Element within a Beacon frame. In Figure 6, a blue highlight shows the Best Effort AC and the red highlight shows that the AIFSN is currently set to 3, the default setting. The AIFSN value determines the duration of the AIFS[AC], as we will discuss in a later section, titled *Calculating an Interframe Space*.

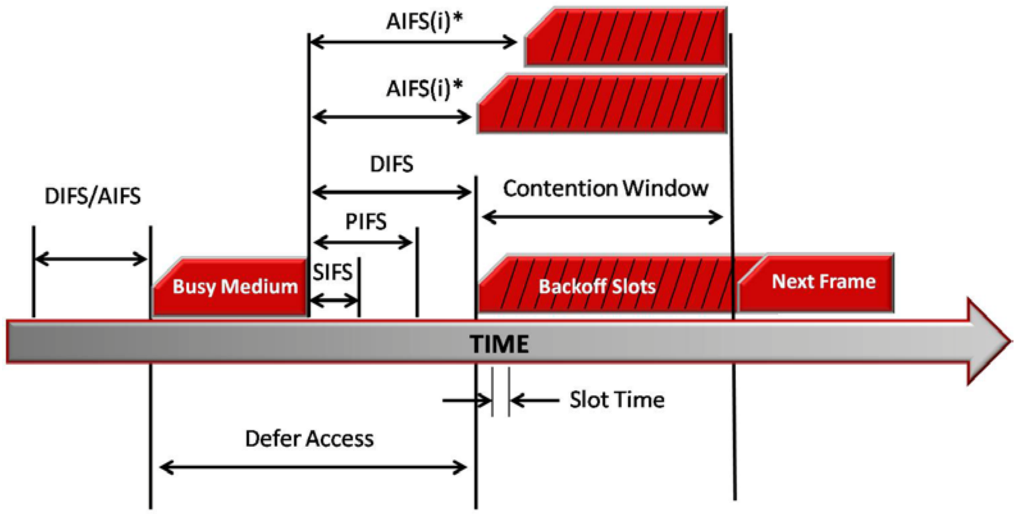
Extended Interframe Space (EIFS)

The EIFS value is used by STAs that have received a frame that contained errors. By using this longer IFS, the transmitting station will have enough time to recognize that the frame was not received properly before the receiving station commences transmission. If, during the EIFS duration, the STA receives a frame correctly (regardless of intended recipient), it will resume using DIFS or AIFS, as appropriate.

Reduced Interframe Space (RIFS)

RIFS were introduced with 802.11n to improve efficiency for transmissions to the same receiver in which a SIFS-separated response is not required, such as a transmission burst.

FIGURE 7 –Interframe Space Comparison



The graphic demonstrates the relationship between the different IFS intervals. You will notice that the initial frame ("Busy Medium") transmission is preceded by a DIFS or AIFS. Though not shown, a contention window value would have been selected, and that value would have determined the number of slot times the STA would wait before transmitting the frame. The section of the graphic that is most important is the middle of the graphic, which shows the relative relationship of the IFS lengths. SIFS are the shortest IFS (excluding 802.11n's RIFS) and PIFS are second shortest, while DIFS and AIFS take up the caboose. One small note to point out about this graphic is that it also demonstrates that when SIFS and PIFS are used, the contention window and random backoff process is eliminated. SIFS, PIFS, and RIFS are used to provide priority access for a given type of frame, which eliminates the need for added contention.

Calculating an Interframe Space

The 802.11-2007 specification provides the information necessary for us to calculate the durations for each IFS. As noted previously, SIFS, PIFS, and DIFS are fixed values for each PHY, while AIFS will vary in accordance with the AC in use. EIFS are fixed per PHY in DCF networks, but vary when used with EDCA. The formulas and components used for SIFS, PIFS, DIFS, EIFS, and AIFS calculations are as follows:

aSIFSTime = aRxRFDelay + aRxPLCPDelay + aMACProcessingDelay + aRxTxTurnaroundTime

aSlotTime = aCCATime + aRxTxTurnaroundTime + aAirPropagationTime + aMACProcessingDelay.

While these two formulas look intimidating, the "aSIFSTime" is the same as a SIFS, measured in microseconds (µs). Similarly, the "aSlotTime" is the same as a slot time, which we will discuss in a later section of the paper. Both of these values are provided for each PHY in the 802.11 specification.

PIFS = aSIFSTime + aSlotTime

DIFS = aSIFSTime + 2 × aSlotTime

Given that the SIFS and slot time values are provided for us in the standard, these calculations are pretty simple. Figure 8 shows these values, and the math to produce them should be intuitive enough.

FIGURE 8 – IFS Calculations

PHY	SIFS*	Slot Time*	PIFS	DIFS
HR/DSSS (802.11b)	10 µs	20 µs	30 µs	50 µs
ERP (802.11g)	10 µs	Long = 20 µs Short = 9 µs	Long = 30 µs Short = 19 µs	Long = 50 µs Short = 28 µs
OFDM (802.11a)	16 µs	9 µs	25 µs	34 µs
HT (802.11n)	10 µs – 2.4 GHz 16 µs – 5 GHz	Long = 20 µs – 2.4 GHz Short = 9 µs – 2.4 GHz 9 µs – 5 GHz	Long = 20 µs – 2.4 GHz Short = 9 µs – 2.4 GHz 25 µs – 5 GHz	Long =50 µs – 2.4 GHz Short = 28 µs – 2.4 GHz 34 µs – 5 GHz

*** Both the SIFS and Slot Time values are provided in each PHY amendment/clause.**
--Note: 802.11n specifies a RIFS interval = 2 µs

EIFS (DCF) = aSIFSTime + DIFS + ACKTxTime

In this formula, the "ACKTxTime" is the amount of time it takes to transmit an ACK frame at the lowest mandatory rate in the BSS.

EIFS (EDCA) = aSIFSTime + AIFS[AC] + ACKTxTime

The EIFS (EDCA) formula mirrors the same for DCF, but replaces the DIFS with the appropriate AIFS[AC].

In the previous AIFS introduction, we mentioned the relationship between an AIFSN and an AIFS. An AIFSN is a number (AIFS Number) value that is user-configurable and determines the brevity (or length) of an AIFS interval. AIFSN values are set for each access category, giving the AIFS[AC] a shorter or longer duration, in accordance with the desired priority. This is demonstrated by the AIFS[AC] formula:

AIFS[AC] = AIFSN[AC] × aSlotTime + aSIFSTime

The default values for each AIFSN[AC] are shown in Figure 9 along with the resulting AIFS[AC] value for each PHY. By modifying the AIFSN[AC], administrators can manipulate the AIFS[AC] duration.

FIGURE 9 – Default AIFS Parameter Set

AC	AIFSN	802.11b AIFS[AC]	802.11g AIFS[AC]	802.11a AIFS[AC]	802.11n 2.4 GHz AIFS[AC]	802.11n 5 GHz AIFS[AC]
AC_BK	7	150 µs	Long = 150 µs Short = 73 µs	79 µs	Long = 150 µs Short = 73 µs	79 µs
AC_BE	3	70 µs	Long = 70 Short = 37 µs	43 µs	Long = 70 µs Short = 37 µs	43 µs
AC_VI	2	50 µs	Long = 50 µs Short = 28 µs	34 µs	Long = 50 µs Short = 28 µs	34 µs
AC_VO	2	50 µs	Long = 50 µs Short = 28 µs	34 µs	Long = 50 µs Short = 28 µs	34 µs

In comparison with Figure 8, which shows the SIFS, PIFS, and DIFS values, you can see in Figure 9 that AIFS are always longer than either a SIFS or a PIFS for a given PHY. This ensures that SIFS- and PIFS-separated frame exchanges take priority.

Slot Times

As we have discussed in the previous section, the slot time parameter is a fixed value for each PHY, regardless of the type of network access method in use. The slot time value is a part of both the IFS interval calculation as well as the random backoff process, which is described in the next section. Only two slot times are used in modern 802.11 networks, specifically 9 µs and 20 µs. More details about slot times and their role in 802.11 contention is included in the next section, but this brief introduction should provide the foundation necessary to proceed.

Contention Window / Random Backoff

After a STA has observed an idle wireless medium with carrier sense (CS) for the appropriate IFS interval (DIFS, EIFS, or AIFS[7]), an additional deferral period must be observed before transmission of an MPDU or MMPDU. This deferral period is known as a *random backoff period* and is selected at random by the STA from a window of possible values called a *contention window* (CW). We will look at the details of this mechanism in the following paragraphs, but before looking at the intricate details of the CW and the random backoff process, it is important to understand its purpose. As we discussed earlier, IFS exist to provide priority access to certain types of transmissions in certain situations. Interframe spaces help to control and streamline the flow of conversation between STAs so that interruptions do not occur. The random backoff process is the most important mechanism used with 802.11 CSMA/CA to prevent collisions. When the 802.11 specification refers to a contention period, random backoff is the primary mechanism for contention.

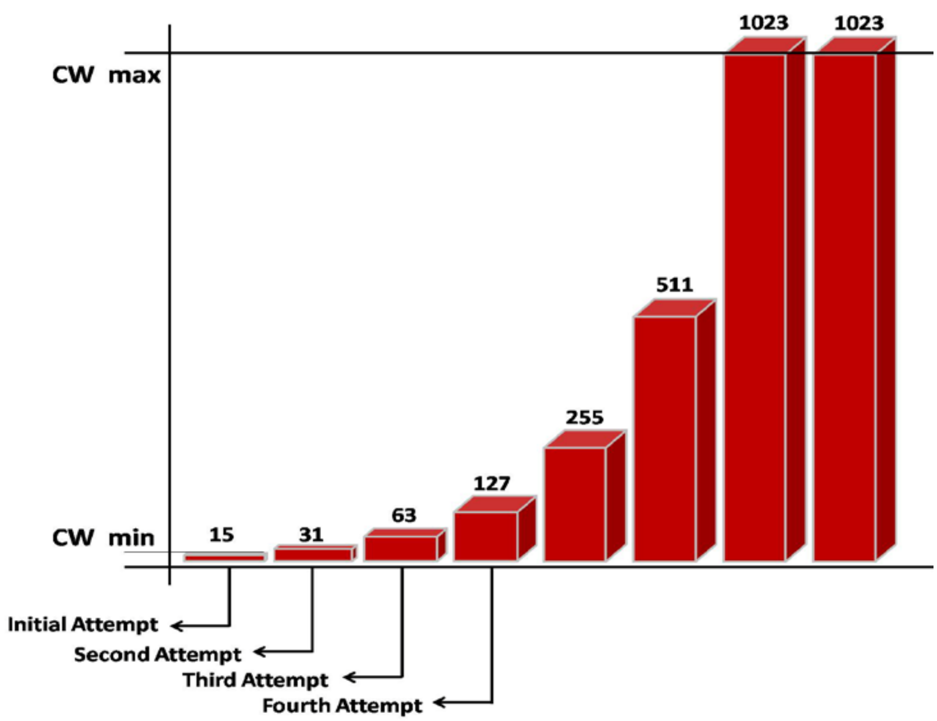
To contend for medium access after the IFS, each station selects a backoff value randomly so that the statistical likelihood of a collision on the WM is decreased. If the STA already has a non-zero value for the random backoff timer, then a new backoff value is not selected. To elaborate, when the backoff timer reaches zero on a STA, the STA may transmit a frame. After completing a TXOP, the backoff timer is reset and a new one must be selected. For all STAs that did not transmit a frame, after waiting for the WM to be available and then waiting the appropriate IFS interval, they may proceed decrementing the backoff timer until it reaches zero or another STA transmits a frame.

The backoff timer is decremented by one for every slot time that the WM is idle, as determined by the CS function. So, if the selected backoff value is seven, then the WM must be idle for the duration of seven slot times before the station can transmit a frame.

The possible backoff values for a STA (with DCF) or a specific AC (with EDCA) are determined by the CW. Each STA (or AC) has an *aCWmin* and *aCWmax* parameter, which designates the lowest and highest possible CW. For an initial attempt at a frame transmission, the CW is the aCWmin value. An example may help to clarify. In a DCF network, an 802.11a STA has a fixed aCWmin of 15[8]. Thus, for the first attempt at a frame transmission, the STA randomly selects a backoff value between 0 and 15; the selected backoff value is then decremented for each slot time that the WM is free. It is important to realize that the minimum value in a CW is always zero. While the maximum value in the CW varies, the minimum value is constant. This means that at any time, any STA can randomly select the shortest backoff timer and transmit a frame. However, as the maximum value in the CW increases, the statistical likelihood of randomly choosing a higher value also increases. This means that the CW does not always guarantee that one STA will take priority, but it can provide a high degree of statistical predictability to this effect.

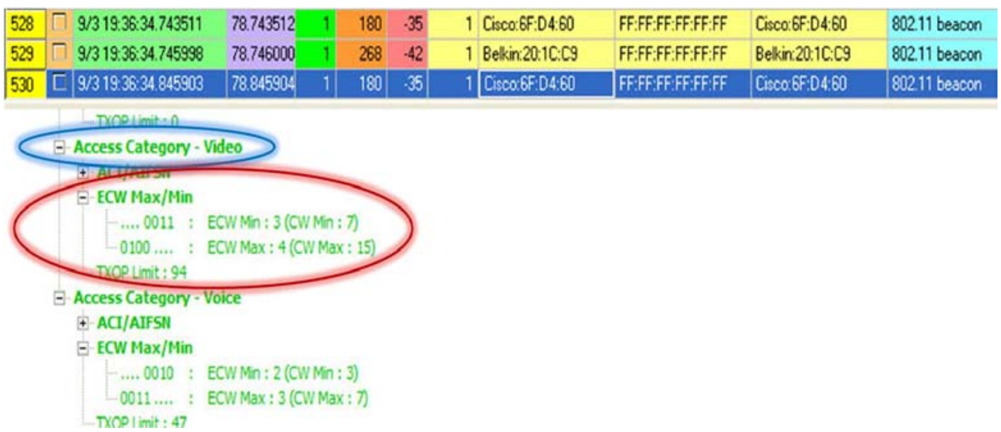
For each attempted retransmission, the CW is increased exponentially until the CW reaches the aCWmax. In this case, the CW remains at the aCWmax until a successful frame sequence is performed or the number of attempted retransmissions reaches the max retry count. For an example of this process, see Figure 10.

FIGURE 10 – Incrementing the CW until it reaches aCWmax



The aCWmin and aCWmax values are fixed for each PHY in DCF networks, and similar to other EDCA parameters, are user-configurable in EDCA networks. These parameters are shown in Figures 11, 12, 13, and 14. Figure 11 shows a QoS-supported Beacon frame that includes the EDCA Parameter Set, including the aCWmin and aCWmax for each AC. Administrators can manipulate these values on WLAN controllers and access points to provide greater priority for specific ACs. Some vendors even have proprietary solutions in which these values can be manipulated on a per-station basis. This provides a high level of predictability for wireless medium access.

FIGURE 11 –Default CWmin and CWmax Values in Beacon Frame



The fixed DCF aCWmin and aCWmax values are shown in Figure 12 and the default EDCA aCWmin and aCWmax values are shown in Figure 13. Comparing the values shown in Figure 11, you will notice that both the video and voice ACs in this Beacon frame represent the default CW parameters for the EDCA function.

FIGURE 12 – Fixed aCWmin and aCWmax Values for DCF

PHY	aCWmin	CWmax
802.11b	31	1023
802.11g	aCWmin(0) = 31 aCWmin(1) = 15	1023
802.11a	15	1023
802.11n	15	1023

You may notice from Figure 12 that 802.11g supports two different aCWmin values. As you might guess, this is done as an effort to mirror 802.11b values. For 802.11g, the variable aCWmin is set to aCWmin(0) if the *characteristic rate set* includes only rates in the set 1, 2, 5.5, 11; otherwise, aCWmin is set to aCWmin(1). The *characteristic rate set* is equal to the AP's supported rate set when the STA is associated with an AP. At all other times, it is equal to the STA's mandatory rate set[9].

FIGURE 13 – Default EDCA aCWmin and aCWmax Values

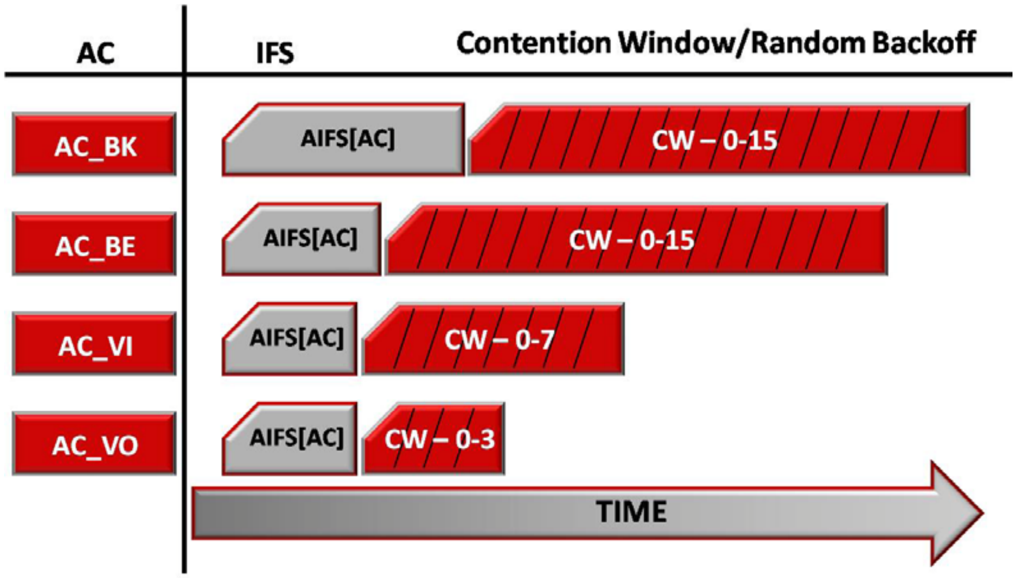
AC	AIFSN	CWmin	CWmax
AC_BK	7	aCWmin	aCWmax
AC_BE	3	aCWmin	aCWmax
AC_VI	2	(aCWmin+1)/ 2-1	aCWmin
AC_VO	2	(aCWmin+1)/ 4-1	(aCWmin+1)/ 2-1

By comparing the values in Figures 12 and 13, you can determine that the CW for the voice and video ACs are exponentially less than the CW for best effort and background ACs, which use the same CW values as DCF STAs. The aCWmin and aCWmax values referred to in Figure 13 reflect the fixed PHY (aCWmin and aCWmax) values shown in Figure 12. To reiterate, DCF uses fixed values based on each PHY. The EDCA chart shows variable values that are derived from the CW values for each PHY.

As further clarification, let's break down the formula used for the CWmin of AC_VO: (aCWmin+1)/4-1. As in all math formulas, the parenthetical steps are performed first, and division/multiplication is performed before addition/subtraction. So, we start with the aCWmin (let's use the values for OFDM), which is 15 (refer to Figure 11). Inserting this value into the formula, we get (15+1)/4-1. After doing the math, we can determine that the CWmin for the voice AC on an OFDM STA is 3. Let's do the same for the CWmax of AC_VO, again with OFDM values: (aCWmin+1)/2-1. (15+1)/2-1. Thus, the CWmax for the voice AC on an OFDM STA is 7. When you compare these numbers with the same numbers for OFDM in a DCF, the statistical priority of AC_VO really stands out. For DCF, OFDM has an aCWmin of 15 and an aCWmax of 1023. Compare that with the AC_VO queue, which we've established as having an aCWmin of 3 and an aCWmax of 7. The statistical likelihood that AC_VO will transmit first is substantial.

Understanding EDCA parameters is a daunting task. For the sake of clarity and conceptual understanding, Figure 14 shows a high-level comparison of the AIFS and CWs for each AC, using the default EDCA Parameter Set as well as the aCWmin value.

FIGURE 14 – Conceptual comparison of AIFS and CW for each AC



*** Note** – This graphic reflects default EDCA Parameter Set values and contention windows using aCWmin

Understanding the inner workings of the contention window and the selection of a random backoff timer is no picnic at the beach. However, spending the time to put the puzzle together can be amazingly helpful for understanding some of the ways in which wireless medium access can be managed.

Why Does 802.11 Arbitration Matter?

If you've read this entire paper in one sitting, you probably need a bathroom break. If you had to take it in several steps, at least we've confirmed that you're human. 802.11 arbitration is a complex topic with many confusing nooks and crannies. Despite the immensity of details covered in this paper, many details have still been left out. No less, the details included in the paper were carefully selected and pruned according to their applicability to administrators and wireless network engineers. Some details are included because they're necessary to simply to completely understand 802.11 arbitration. Other details are included because they're absolutely essential for engineers who are attempting to optimize a wireless network. The following section of the paper is provided to reinforce the value of understanding 802.11 arbitration and hopefully to bring the monster to life by applying the details to real-world uses.

Network Design

It is common to think of WLANs as a marriage between radio frequency (RF) technology and wired networking. As network designers approach capacity and coverage planning for wireless, they must consider the impact that 802.11 contention will have on the efficiency of a network's operation. As we've discussed in this paper, 802.11 STAs compete with one another for access to the half-duplex, shared wireless medium. When one STA in an area transmits, others do not. For this reason, arbitration processes are a primary consideration when AP locations, channel plans, power output settings, and antenna coverage patterns are selected. While network designers can't take the time to consider each client station's impact on 802.11 arbitration, the concepts can be applied generally. For example, when there is a high number of high-priority devices (such as VoWiFi phones), designers should consider minimizing the load on each AP by decreasing the coverage area for each AP. When low latency devices like phones are in use, they observe shorter contention periods, which means they attempt to access the network more frequently during an active call. For this reason, a fewer number of devices can typically be supported by a single AP. When random backoff values range between 0 and 3, the statistical likelihood of two devices choosing the same number increases. This may lead to a significant number of collisions when client density is too high. For this reason, vendors typically have a recommend maximum client density for each AP. These factors should all be included in proper network design, and 802.11 arbitration sets this foundation.

QoS

As the quality of RF handling technologies continues to increase, a greater number of user devices and applications are being supported on the wireless network. For networks that are attempting to support multiple applications like latency- and delay-sensitive voice and video along with data, QoS is essential. When wireless is the primary end-user network access method in an enterprise and user densities and network utilization continue to climb, the efficiency and effectiveness of 802.11 contention protocols can be put to the test. For this reason, assessing and proving QoS configuration, such as EDCA Parameter Sets, is critical. To properly manage these settings and to know when changes may be helpful, a thorough understanding of the functions of 802.11 arbitration is necessary. As administrators recognize a need for greater priority of certain device types or user groups, understanding the parameters that affect these needs can make the difference between having an effective and reliable multi-service WLAN and having to look for a new job. In most networks, the default parameters designated by the IEEE should be suitable, but there may be times for adjustments.

Because QoS operation is fundamentally dependent upon 802.11 arbitration, QoS is the single most important reason to understand how STAs contend for medium access. While Ethernet has its own well-known prioritization and forwarding processes, 802.11 processes are less well known. This is a fairly cursory explanation of the significance of 802.11 arbitration and QoS, but its importance cannot be understated.

Airtime Fairness

Airtime fairness is a relatively new feature in WLANs that is designed to maximize throughput in networks where there is a significant disparity in data rate between the STAs. The problem is that slower STAs require more airtime than faster STAs to transmit the same amount of data. This problem is exacerbated by 802.11n. By intelligently managing network access, this hurdle can be overcome. To date, the implementation of this solution is always proprietary. Similar methods are employed by different vendors to achieve airtime fairness, but each WLAN architecture introduces unique nuances and capabilities. Understanding the way that 802.11 arbitration can be managed by a given solution requires an intimate understanding of 802.11 arbitration fundamentals. Administrators are often required to identify and provide recommended solutions to the problems faced by highly utilized WLANs. Having an in-depth knowledge of 802.11 protocols allows administrators to understand the way in which their solution addresses these problems.

Airtime fairness is a perfect example of this requirement, because airtime fairness may take advantage of the EDCA Parameter Set values to manage wireless medium access. Some vendors can control both downlink and uplink communications using proprietary solutions. Some vendors can even control these parameters on a per STA basis. Undoubtedly, these solutions are extremely sophisticated and the details of their operation are not generally public knowledge. Further, an understanding of the algorithms and backend processes of these technologies is not necessary in most cases, but knowing how performance is managed in a network *is* extremely important.

Summary

As you've discovered from this paper, 802.11 arbitration is both frustratingly complex and, unfortunately, essential to learn. Managing a wireless network, especially one that depends on QoS, requires an intimate knowledge of 802.11 protocols. There are many different parts and pieces to keep track of, but having a general understanding is helpful. While this paper is intended to be thorough and useful as a teaching aid, it is also designed as a reference for those readers who either don't have access, or don't want access, to the 802.11 specification. The graphics and tables in this paper are intended to be a resource for the administration of QoS in a WLAN.

If this paper was overwhelming and you don't anticipate remembering every detail, refer back to the flowchart towards the beginning of the paper. This should help to organize the steps and processes according to their use. The details will come with time and review. Another helpful way of cementing this knowledge is to review a packet trace that shows some of this information in QoS AP Beacon frames.

Recap

The wireless medium used by WLAN devices is half-duplex, meaning that it is shared. In the same way that people must follow social expectations to participate in conversations, so wireless STAs must follow the 802.11 protocols to maximize the efficiency and effectiveness of wireless communications.

Carrier Sense (CS) mechanisms are used by STAs to determine if the wireless medium is currently in use. CS is often divided into two separate processes: physical carrier sense and virtual carrier sense. Clear Channel Assessment (CCA) is the physical carrier sense, used for detecting physical noise on the RF medium, and Network Allocation Vectors (NAV) are used as the virtual carrier sense, where Duration values in MAC headers are used to virtually predict if the WM is busy.

When CS indicates an idle medium, STAs defer for the appropriate IFS interval before either contending for WM access or transmitting a frame. IFS are dictated by the type of frame being transmitted, the PHY in use by the STA, and the type of network access method in use.

After deferring the appropriate IFS interval, backoff values are selected at random from a range of values called the Contention Window (CW). After selecting a random backoff value, this value is decremented by one for every slot time in which the WM remains idle. When this value reaches zero, a TXOP is obtained and a frame be transmitted.

As a dynamic process with physically disparate devices, each STA within a network is typically at a different stage in the process at any given time.

[1] In WLANs, collisions are never known with certainty. They can only be inferred based on a failure to receive an ACK, or the appropriate response frame. CSMA with Collision Detection is used in Ethernet networks, where the physical medium can be sensed and a collision detected. Wireless transmitters cannot detect a collision.

[2] 802.11-2007, Section 9.1

[3] 802.11-2007, Section 3.38

[4] For DSSS, HR/DSSS, and ERP frames, "the PLCP LENGTH field shall be an unsigned 16-bit integer that indicates the number of microseconds…required to transmit the MPDU." The MPDU includes all of the contents of a frame from layer 2 through 7. For OFDM, "The PLCP LENGTH field shall be an unsigned 12-bit integer that indicates the number of octets in the PSDU that the MAC is currently requesting the PHY to transmit." "In addition, the CCA mechanism can be augmented by predicting the duration of the packet from the contents of the RATE and LENGTH fields, even if the data rate is not supported by the STA." In other words, 2.4 GHz devices use the LENGTH field of the PLCP header to indicate the length of time, in microseconds, it will take to transmit the frame. Conversely, OFDM uses the LENGTH field to indicate the number of bytes that are present in the frame. By combining the information in the LENGTH and RATE fields, STAs can determine the length of time necessary to complete the OFDM frame, even if they do not support the rate at which the remainder of a frame will be transmitted.

[5] 802.11-2007, Section 9.2.3.1

[6] 802.11-2007, Section 9.2.3.4

[7] Backoff timers are not used with SIFS- and PIFS-separated transmissions.

[8] It may be helpful to think of the CW values as a formula: 2^x-1. In other words, the value of 2 with a given exponent, minus 1. For HR/DSSS the exponent is 5 and for ERP and OFDM, the exponent is 4, yielding aCWmin values of 31 and 15, respectively.

[9] 802.11-2007, Section 9.2.12