CHAPTER 5

Cisco Unified Wireless QoS

This chapter describes quality-of-service (QoS) in the context of WLAN implementations. This chapter describes WLAN QoS in general, but does not provide in-depth coverage on topics such as security, segmentation, and voice over WLAN (VoWLAN), although these topics have a QoS component. This chapter also provides information on the features of the Cisco Centralized WLAN Architecture.

This chapter is intended for those who are tasked with designing and implementing enterprise WLAN deployments using the Cisco Unified Wireless technology.

QoS Overview

QoS refers to the capability of a network to provide differentiated service to selected network traffic over various network technologies. QoS technologies provide the following benefits:

- Provide building blocks for business multimedia and voice applications used in campus, WAN, and service provider networks
- Allow network managers to establish service-level agreements (SLAs) with network users
- Enable network resources to be shared more efficiently and expedite the handling of mission-critical applications
- Manage time-sensitive multimedia and voice application traffic to ensure that this traffic receives higher priority, greater bandwidth, and less delay than best-effort data traffic

With QoS, bandwidth can be managed more efficiently across LANs, including WLANs and WANs. QoS provides enhanced and reliable network service by doing the following:

- Supporting dedicated bandwidth for critical users and applications
- Controlling jitter and latency (required by real-time traffic)
- Managing and minimizing network congestion
- Shaping network traffic to smooth the traffic flow
- Setting network traffic priorities

Wireless QoS Deployment Schemes

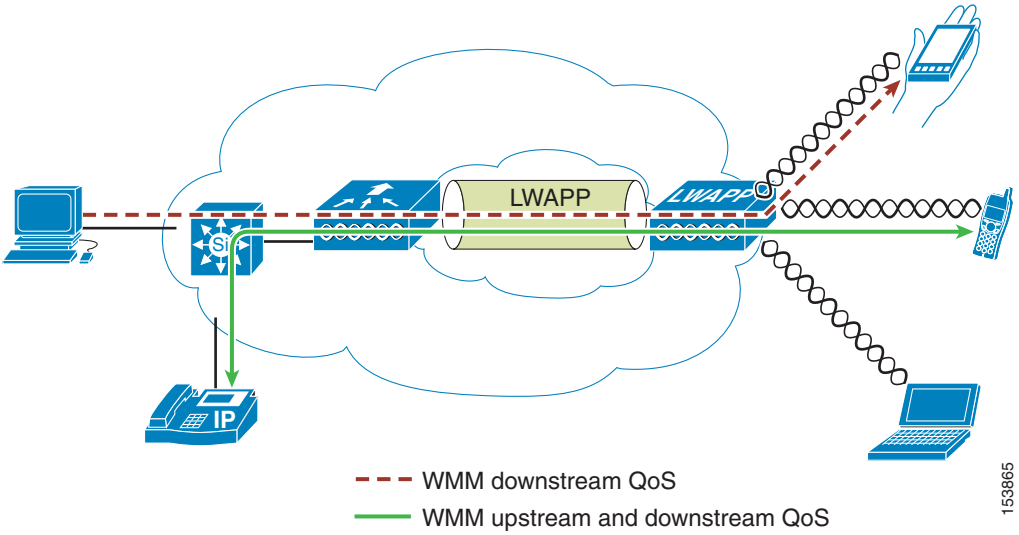
In the past, WLANs were mainly used to transport low-bandwidth, data-application traffic. Currently, with the expansion of WLANs into vertical (such as retail, finance, and education) and enterprise environments, WLANs are used to transport high-bandwidth data applications, in conjunction with time-sensitive multimedia applications. This requirement led to the necessity for wireless QoS.

Several vendors, including Cisco, support proprietary wireless QoS schemes for voice applications. To speed up the rate of QoS adoption and to support multi-vendor time-sensitive applications, a unified approach to wireless QoS is necessary. The IEEE 802.11e working group within the IEEE 802.11 standards committee has completed the standard definition, but adoption of the 802.11e standard is in its early stages, and as with many standards there are many optional components. Just as occurred with 802.11 security in 802.11i, industry groups such as the Wi-Fi Alliance and industry leaders such as Cisco are defining the key requirements in WLAN QoS through their WMM and Cisco Compatible Extensions programs, ensuring the delivery of key features and interoperation through their certification programs.

Cisco Unified Wireless products support Wi-Fi MultiMedia (WMM), a QoS system based on IEEE 802.11e that has been published by the Wi-Fi Alliance, and WMM Power Save, as well as Admission Control.

Figure 5-1 shows a sample deployment of wireless QoS based on Cisco Unified Wireless technology features.

Figure 5-1 QoS Deployment Example



QoS Parameters

QoS is defined as the measure of performance for a transmission system that reflects its transmission quality and service availability. Service availability is a crucial element of QoS. Before QoS can be successfully implemented, the network infrastructure must be highly available. The network transmission quality is determined by latency, jitter, and loss, as shown in Table 5-1.

Table 5-1 QoS Parameters

Transmission Quality	Description
Latency	Latency (or delay) is the amount of time it takes for a packet to reach the receiving endpoint after being transmitted from the sending endpoint. This time period is called the end-to-end delay and can be divided into two areas: • Fixed network delay—Includes encoding and decoding time (for voice and video), and the finite amount of time required for the electrical or optical pulses to traverse the media en route to their destination. • Variable network delay—Generally refers to network conditions, such as queuing and congestion, that can affect the overall time required for transit.
Jitter	Jitter (or delay-variance) is the difference in the end-to-end latency between packets. For example, if one packet requires 100 ms to traverse the network from the source endpoint to the destination endpoint, and the next packet requires 125 ms to make the same trip, the jitter is calculated as 25 ms.
Loss	Loss (or packet loss) is a comparative measure of packets successfully transmitted and received to the total number that were transmitted. Loss is expressed as the percentage of packets that were dropped.

Upstream and Downstream QoS

Figure 5-2 illustrates the definition of *radio upstream* and *radio downstream* QoS.

Figure 5-2 Upstream and Downstream QoS

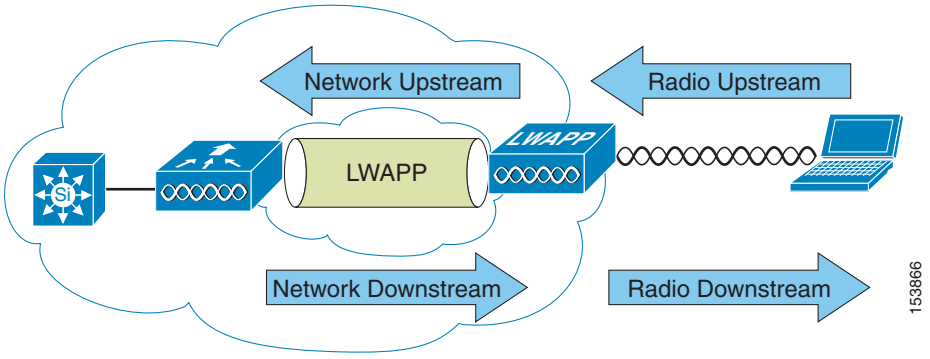


Figure 5-2 shows the following:

- *Radio downstream* QoS—Traffic leaving the AP and traveling to the WLAN clients. Radio downstream QoS is the primary focus of this chapter, because this is still the most common deployment. The radio client upstream QoS depends on the client implementation.
- *Radio upstream* QoS—Traffic leaving the WLAN clients and traveling to the AP. WMM provides upstream QoS for WLAN clients supporting WMM.

- *Network downstream*—Traffic leaving the WLC traveling to the AP. QoS can be applied at this point to prioritize and rate-limit traffic to the AP. Configuration of Ethernet downstream QoS is not covered in this chapter.
- *Network upstream*—Traffic leaving the AP, traveling to the WLC. The AP classifies traffic from the AP to the upstream network according to the traffic classification rules of the AP.

QoS and Network Performance

The application of QoS features might not be easily detected on a lightly loaded network. If latency, jitter, and loss are noticeable when the media is lightly loaded, it indicates either a system fault, poor network design, or that the latency, jitter, and loss requirements of the application are not a good match for the network. QoS features start to be applied to application performance as the load on the network increases. QoS works to keep latency, jitter, and loss for selected traffic types within acceptable boundaries. When providing only radio downstream QoS from the AP, radio upstream client traffic is treated as best-effort. A client must compete with other clients for upstream transmission as well as competing with best-effort transmission from the AP. Under certain load conditions, a client can experience upstream congestion, and the performance of QoS-sensitive applications might be unacceptable despite the QoS features on the AP. Ideally, upstream and downstream QoS can be operated either by using WMM on both the AP and WLAN client, or by using WMM and a client proprietary implementation.

Note Even without WMM support on the WLAN client, the Cisco Unified Wireless solution is able to provide network prioritization in both network upstream and network downstream situations.

Note WLAN client support for WMM does not mean that the client traffic automatically benefits from WMM. The applications looking for the benefits of WMM assign an appropriate priority classification to their traffic, and the operating system needs to pass that classification to the WLAN interface. In purpose-built devices, such as VoWLAN handsets, this is done as part of the design. However, if implementing on a general purpose platform such as a PC, application traffic classification and OS support must be implemented before the WMM features can be used to good effect.

802.11 DCF

Data frames in 802.11 are sent using the Distributed Coordination Function (DCF), which is composed of the following two main components:

- Interframe spaces (SIFS, PIFS, and DIFS).
- Random backoff (contention window) DCF is used in 802.11 networks to manage access to the RF medium.

A baseline understanding of DCF is necessary to deploy 802.11e-based enhanced distributed channel access (EDCA). For more information on DCF, see the IEEE 802.11 specification at the following URL: http://ieeexplore.ieee.org/xpl/standardstoc.jsp?isnumber=14251&isYear=1997.

Interframe Spaces

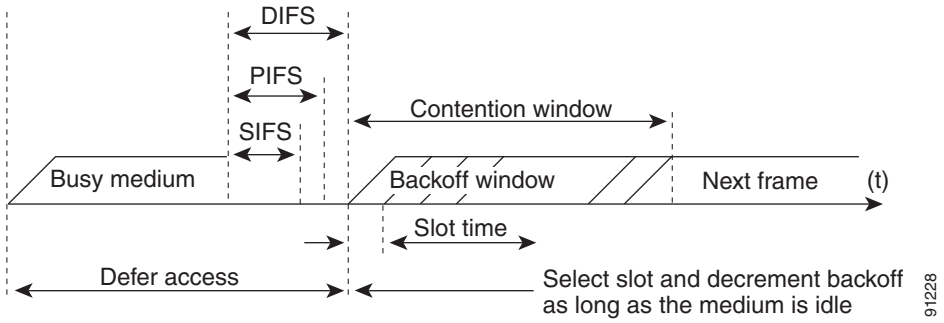
802.11 currently defines three interframe spaces (IFS), as shown in Figure 5-3:

- Short interframe space (SIFS)—10 µs
- PCF interframe space (PIFS)—SIFS + 1 x slot time = 30 µs
- DCF interframe space (DIFS)—50 µs SIFS + 2 x slot time = 50 µs

Note The base timing used in this interframe space example are for 802.11b; the timing in 802.11g and 802.11a are different, but the principles applied are the same.

The interframe spaces (SIFS, PIFS, and DIFS) allow 802.11 to control which traffic gets first access to the channel after carrier sense declares the channel to be free. Generally, 802.11 management frames and frames not expecting contention (a frame that is part of a sequence of frames) use SIFS, and data frames use DIFS.

Figure 5-3 Interframe Spaces



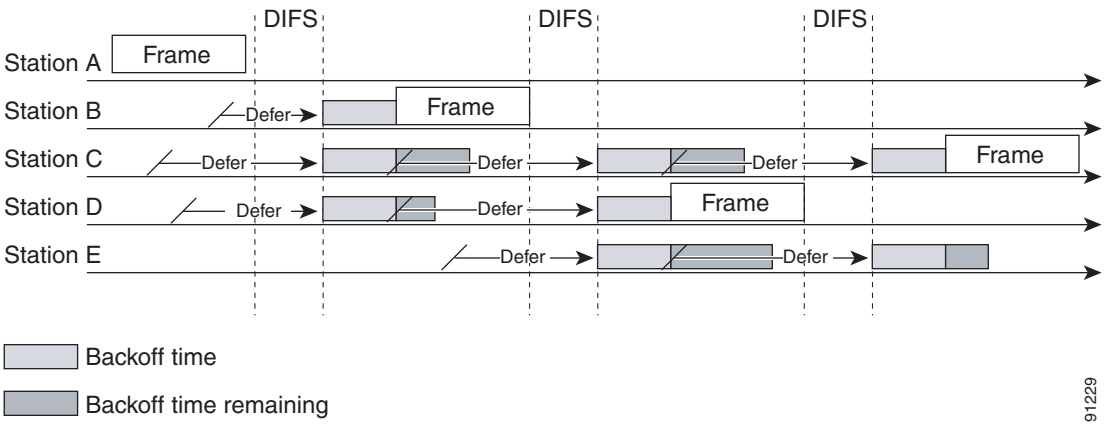
Random Backoff

When a data frame using DCF is ready to be sent, it goes through the following steps:

1. Generates a random backoff number between 0 and a minimum contention window (CWmin).
2. Waits until the channel is free for a DIFS interval.
3. If the channel is still free, begins to decrement the random backoff number, for every slot time (20 µs) that the channel remains free.
4. If the channel becomes busy, such as another station getting to 0 before your station, the decrement stops and steps 2 through 4 are repeated.
5. If the channel remains free until the random backoff number reaches 0, the frame can be sent.

Figure 5-4 shows a simplified example of how the DCF process works. In this simplified DCF process, no acknowledgements are shown and no fragmentation occurs.

Figure 5-4 Distributed Coordination Function Example



The DCF steps illustrated in Figure 5-4 are as follows:

1. Station A successfully sends a frame; three other stations also want to send frames but must defer to Station A traffic.
2. After Station A completes the transmission, all the stations must still defer to the DIFS. When the DIFS is complete, stations waiting to send a frame can begin to decrement the backoff counter, once every slot time, and can send their frame.
3. The backoff counter of Station B reaches zero before Stations C and D, and therefore Station B begins transmitting its frame.
4. When Station C and D detect that Station B is transmitting, they must stop decrementing the backoff counters and defer until the frame is transmitted and a DIFS has passed.
5. During the time that Station B is transmitting a frame, Station E receives a frame to transmit, but because Station B is sending a frame, it must defer in the same manner as Stations C and D.
6. When Station B completes transmission and the DIFS has passed, stations with frames to send begin to decrement the backoff counters. In this case, the Station D backoff counter reaches zero first and it begins transmission of its frame.
7. The process continues as traffic arrives on different stations.

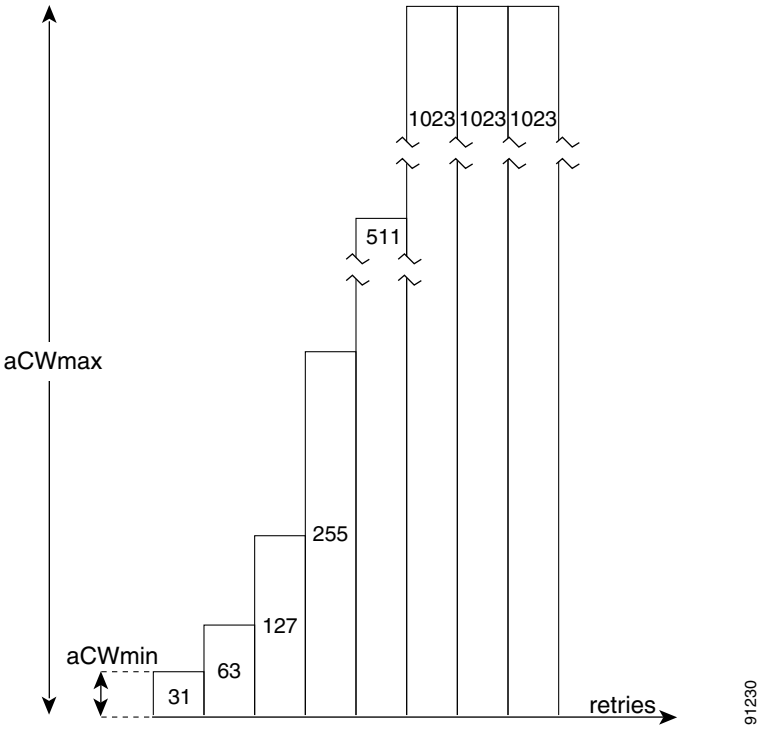
CWmin, CWmax, and Retries

DCF uses a contention window (CW) to control the size of the random backoff. The contention window is defined by two parameters:

- aCWmin
- aCWmax

The random number used in the random backoff is initially a number between 0 and aCWmin. If the initial random backoff expires without successfully sending the frame, the station or AP increments the retry counter, and doubles the value random backoff window size. This doubling in size continues until the size equals aCWmax. The retries continue until the maximum retries or time-to-live (TTL) is reached. This process of doubling the backoff window is often referred to as a *binary exponential backoff*, and is illustrated in Figure 5-5 where the aCWmin if 2^5-1, and increases to 2^6-1, on the next backoff level, up to the aCWmax value of 2^{10}-1.

Figure 5-5 Growth in Random Backoff Range with Retries



Note These values are for 802.11b, and values can be different for different physical layer implementations.

Wi-Fi Multimedia

This section describes three WMM implementations:

- WMM access
- WMM power save
- Access control

WMM Access

WMM is a Wi-Fi Alliance certification of support for a set of features from an 802.11e draft. This certification is for both clients and APs, and certifies the operation of WMM. WMM is primarily the implementation of the EDCA component of 802.11e. Additional Wi-Fi certifications are planned to address other components of the 802.11e.

WMM Classification

WMM uses the 802.1P classification scheme developed by the IEEE (which is now a part of the 802.1D specification).

This classification scheme has eight priorities, which WMM maps to four access categories: AC_BK, AC_BE, AC_VI, and AC_VO. These access categories map to the four queues required by a WMM device, as shown in Table 5-2.

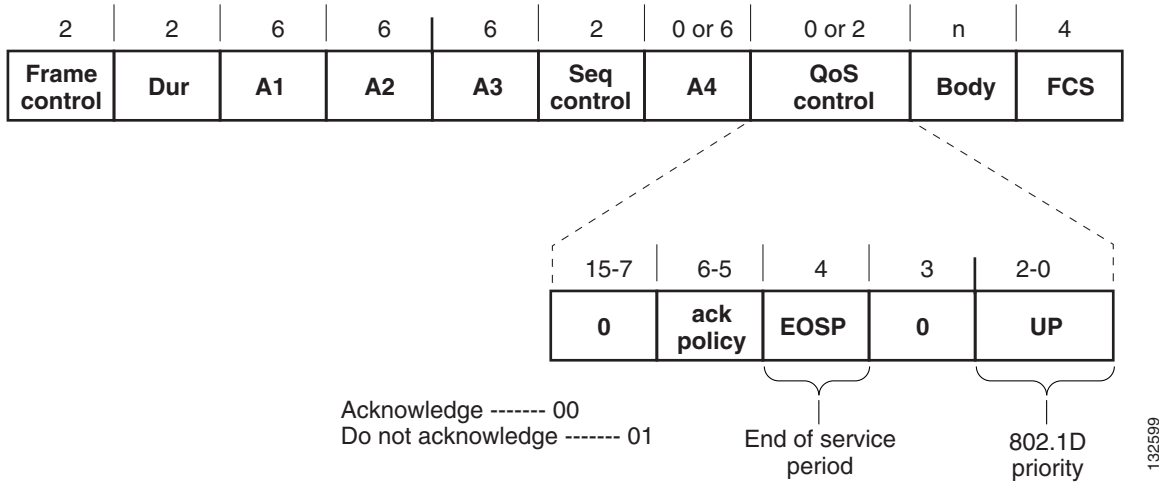
Table 5-2 Table 2 802.1P and WMM Classification

Priority	802.1P Priority	802.1P Designation	Access Category	WMM Designation
Lowest	1	BK	AC_BK	Background
	2	-		
	0	BE	AC_BE	Best Effort
	3	EE		
	4	CL	AC_VI	Video
	5	VI		
	6	VO	AC_VO	Voice
Highest	7	NC		

Figure 5-6 shows the WMM data frame format. Note that even though WMM maps the eight 802.1P classifications to four access categories, the 802.11D classification is sent in the frame.

Note The WMM and IEEE 802.11e classifications are different from the classifications recommended and used in the Cisco network, which are based on IETF recommendations. The primary difference in classification is the changing of voice and video traffic to 5 and 4, respectively. This allows the 6 classification to be used for Layer 3 network control. To be compliant with both standards, the Cisco Unified Wireless solution performs a conversion between the various classification standards when the traffic crosses the wireless-wired boundary.

Figure 5-6 WMM Frame Format



WMM Queues

Figure 5-7 shows the queuing performed on a WMM client or AP. There are four separate queues, one for each of the access categories. Each of these queues contends for the wireless channel in a similar manner to the DCF mechanism described previously, with each of the queues using different interframe space, CWmin, and CWmax values. If more than one frame from different access categories collide internally, the frame with the higher priority is sent, and the lower priority frame adjusts its backoff parameters as though it had collided with a frame external to the queuing mechanism. This system is called enhanced distributed channel access (EDCA).

Figure 5-7 WMM Queues

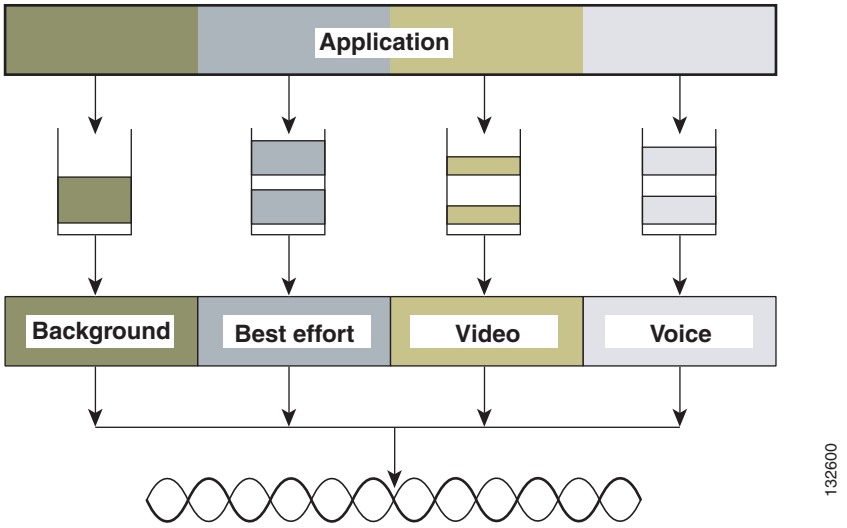
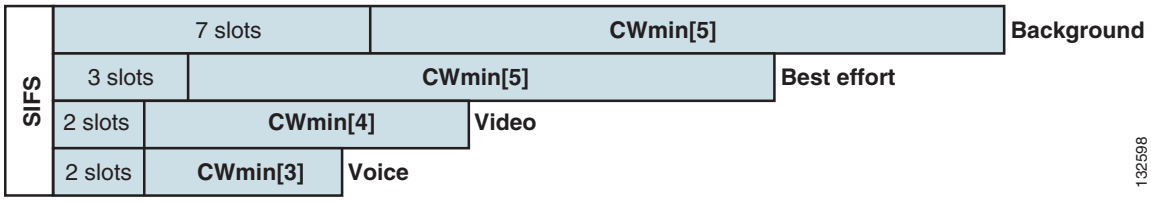


Figure 5-8 shows the principle behind EDCF, where different interframe spacing and CWmin and CwMax values (for clarity CwMax is not shown) are applied per traffic classification. Different traffic types can wait different interface spaces before counting down their random backoff, and the CW value used to generate the random backoff number also depends on the traffic classification. For example, the CWmin[3] for Voice is 2^3-1, and CWmin[5] for Best effort traffic is 2^5-1. High priority traffic has a small interframe space and a small CWmin value, giving a short random backoff, whereas best-effort traffic has a longer interframe space and large CWmin value that on average gives a large random backoff number.

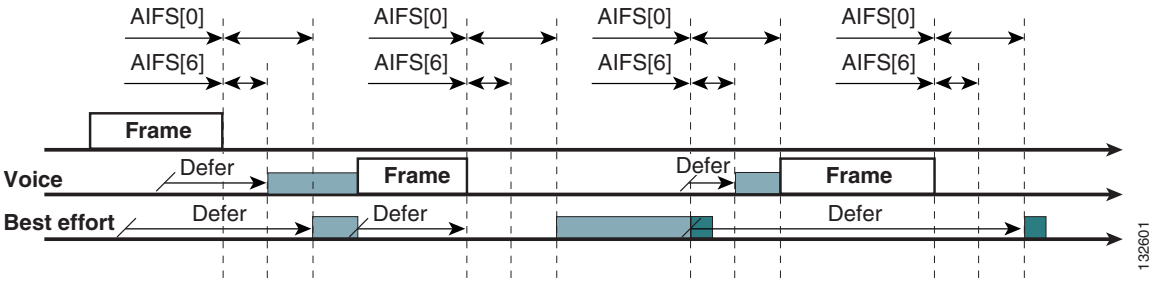
Figure 5-8 Access Category Timing



EDCA

The EDCA process is illustrated in Figure 5-9.

Figure 5-9 EDCA Example



The EDCA process follows this sequence:

1. While Station X is transmitting its frame, three other stations determine that they must send a frame. Each station defers because a frame was already being transmitted, and each station generates a random backoff.
2. Because the Voice station has a traffic classification of voice, it has an arbitrated interframe space (AIFS) of 2, and uses an initial CWmin of 3, and therefore must defer the countdown of its random backoff for 2 slot times, and has a short random backoff value.
3. Best-effort has an AIFS of 3 and a longer random backoff time, because its CWmin value is 5.
4. Voice has the shortest random backoff time, and therefore starts transmitting first. When Voice starts transmitting, all other stations defer.
5. After the Voice station finishes transmitting, all stations wait their AIFS, then begin to decrement the random backoff counters again.
6. Best-effort then completes decrementing its random backoff counter and begins transmission. All other stations defer. This can happen even though there might be a voice station waiting to transmit. This shows that best-effort traffic is not starved by voice traffic because the random backoff decrementing process eventually brings the best-effort backoff down to similar sizes as high priority traffic, and that the random process might, on occasion, generate a small random backoff number for best-effort traffic.
7. The process continues as other traffic enters the system. The access category settings shown in Table 5-3 and Table 5-4 are, by default, the same for an 802.11a radio, and are based on formulas defined in WMM.

Note Table 5-3 refers to the parameter settings on a client, which are slightly different from the settings for an AP. The AP has a larger AIFSN for voice and video ACs.

Table 5-3 WMM Client Parameters

AC	CWmin	CWmax	AIFSN	TXOP Limit (802.11b	TXOP Limit (802.11a/g)
AC_BK	aCWmin	aCWmax	7	0	0
AC_BE	aCWmin	4*(aCQmin+1)-1	3	0	0

Table 5-3 WMM Client Parameters

AC	CWmin	CWmax	AIFSN	TXOP Limit (802.11b	TXOP Limit (802.11a/g)
AC_VI	(aCWmin+1)/2-1	aCWmin	1	6.016 ms	3.008 ms
AC_VO	(aCWmin+1)/4-1	(aCWmin+1)/2-1	1	3.264 ms	1.504 ms

Table 5-4 WMM AP Parameters

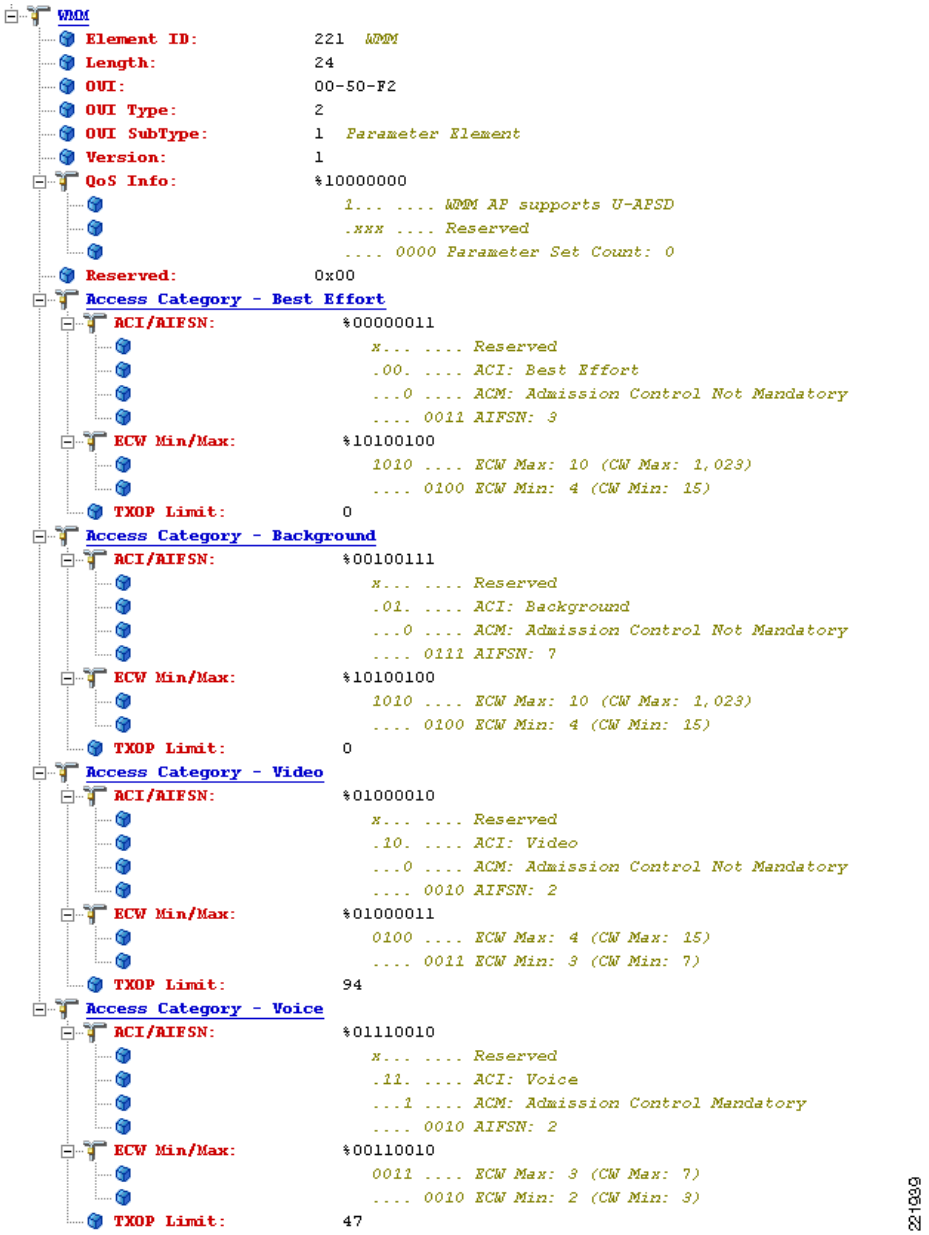
Access Category	CWmin	CWmax	AIFSN	TXOP Limit (802.11b	TXOP Limit (802.11a/g)
AC_BK	aCWmin	aCWmax	7	0	0
AC_BE	aCWmin	4*(aCQmin+1)-1	3	0	0
AC_VI	(aCWmin+1)/2-1	aCWmin	2	6.016 ms	3.008 ms
AC_VO	(aCWmin+1)/4-1	(aCWmin+1)/2-1	2	3.264 ms	1.504 ms

The overall impact of the different AIFS, CWmin, and CWmax values is difficult to illustrate in timing diagrams because their impact is more statistical in nature. It is easier to compare the AIFS and the size of the random backoff windows, as shown in Figure 5-8.

When comparing voice and background frames as examples, these traffic categories have CWmin values of 2^3-1 (7) and 2^5-1 (31), and AIFS of 2 and 7, respectively. This an average delay of 5 (2+7/1) slot times before sending a voice frame, and an average of 22 slot (7+31/2) times for background frame. Therefore, voice frames are statistically much more likely to be sent before background frames.

Figure 5-10 shows the WMM information in a probe response. Apart from the WMM access category information contained in this element, the client also learns which WMM categories require admission control. As can be seen in this example, the Voice AC has admission control set to mandatory. This requires the client to send the request to the AP, and have the request accepted, before it can use this AC. Admission control is discussed in more detail later in this chapter.

Figure 5-10 Probe Response WMM Element Information



U-APSD

Unscheduled automatic power-save delivery (U-APSD) is a feature that has two key benefits:

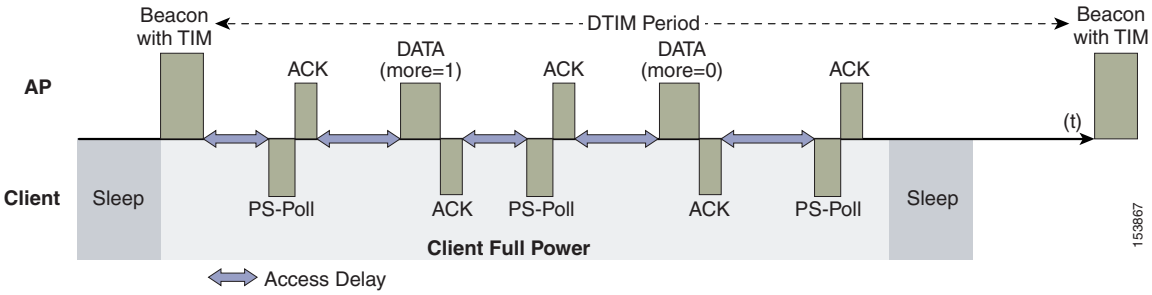
- The primary benefit of U-APSD is that it allows the voice client to synchronize the transmission and reception of voice frames with the AP, thereby allowing the client to go into power-save mode between the transmission/reception of each voice frame tuple. The WLAN client frame transmission in the access categories supporting U-APSD triggers the AP to send any data frames queued for that WLAN client in that access category. A U-APSD client remains listening to the AP until it receives a frame from the AP with an end-of-service period (EOSP) bit set. This tells the client that it can now go back into its power-save mode. This triggering mechanism is considered a more efficient

use of client power than the regular listening for beacons method, at a period controlled by the delivery traffic indication message (DTIM) interval, because the latency and jitter requirements of voice are such that a WVoIP client would either not be in power-save mode during a call, resulting in reduced talk times, or would use a short DTIM interval, resulting in reduced standby times. The use of U-APSD allows the use of long DTIM intervals to maximize standby time without sacrificing call quality. The U-APSD feature can be applied individually across access categories, allowing U-APSD can be applied to the voice ACs in the AP, but the other ACs still use the standard power save feature.

- The secondary benefit of this feature is increased call capacity. The coupling of transmission buffered data frames from the AP with the triggering data frame from the WLAN client allows the frames from the AP to be sent without the accompanying interframe spacing and random backoff, thereby reducing the contention experience by call.

Figure 5-11 shows a sample frame exchange for the standard 802.11 power save delivery process.

Figure 5-11 Standard Client Power-Save

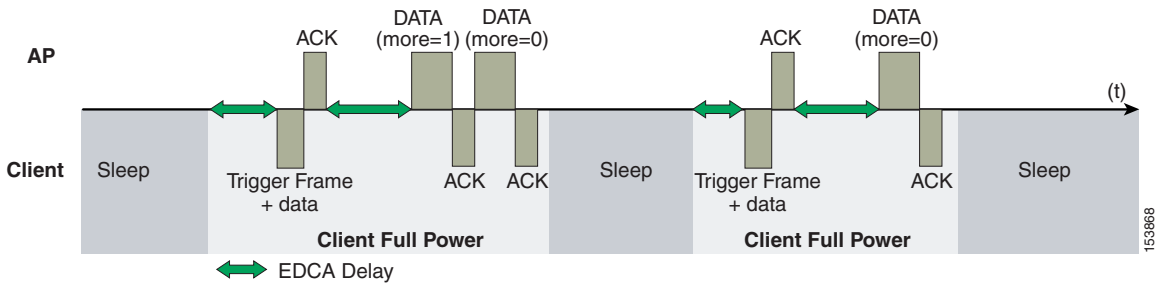


The client in power-save mode first detects that there is data waiting for it at the AP via the presence of the TIM in the AP beacon. The client must power-save poll (PS-Poll) the AP to retrieve that data. If the data sent to the client requires more than one frame to be sent, the AP indicates this in the sent data frame. This process requires the client to continue sending power-save polls to the AP until all the buffered data is retrieved by the client.

This presents two major problems. The first is that it is quite inefficient, requiring the PS-polls, as well as the normal data exchange, to go through the standard access delays associated with DCF. The second issue, being more critical to voice traffic, is that retrieving the buffered data is dependent on the DTIM, which is a multiple of the beacon interval. Standard beacon intervals are 100 ms, and the DTIM interval can be integer multiples of this. This introduces a level of jitter that is generally unacceptable for voice calls, and voice handsets switch from power-save mode to full transmit and receive operation when a voice call is in progress. This gives acceptable voice quality but reduces battery life. The Cisco Unified Wireless IP Phone 7921G addresses this issue by providing a PS-Poll feature that allows the 7921G to generate PS-Poll requests without waiting for a beacon TIM. This allows the 7921G to poll for frames when it has sent a frame, and then go back to power-save mode. This feature does not provide the same efficiency as U-APSD, but improves battery life for 7921Gs on WLANs without U-APSD.

Figure 5-12 shows an example of traffic flows with U-APSD. In this case, the trigger for retrieving traffic is the client sending traffic to the AP. The AP, when acknowledging the frame, tells the client that data is queued for it, and that it should stay on. The AP then sends data to the client typically as a TXOP burst where only the first frame has the EDCF access delay. All subsequent frames are then sent directly after the acknowledgment frame. In a VoWLAN implementation there is only likely to be one frame queued at the AP, and VoWLAN client would be able to go into sleep mode after receiving that frame from the AP.

Figure 5-12 U-APSD



This approach overcomes both the disadvantages of the previous scheme in that it is much more efficient. The timing of the polling is controlled via the client traffic, which in the case of voice is symmetric, so if the client is sending a frame every 20 ms, it would be expecting to receive a frame every 20 ms as well. This would introduce a maximum jitter of 20 ms, rather than an n * 100 ms jitter.

TSpec Admission Control

Traffic Specification (TSpec) allows an 802.11e client to signal its traffic requirements to the AP. In the 802.11e MAC definition, two mechanisms provide prioritized access. These are the contention-based EDCA option and the controlled access option provided by the transmit opportunity (TXOP). When describing TSpec features where a client can specify its traffic characteristics, it is easy to assume that this would automatically result in the use of the controlled access mechanism, and have the client granted a specific TXOP to match the TSpec request. However, this does not have to be the case; a TSpec request can be used to control the use of the various access categories (ACs) in EDCA. Before a client can send traffic of a certain priority type, it must have requested to do so via the TSpec mechanism. For example, a WLAN client device wanting to use the voice AC must first make a request for use of that AC. Whether or not AC use is controlled by TSpec requests is configurable with voice and video ACs controlled by TSpec requests, and best-effort and background ACs can be open for use without a TSpec request. The use of EDCA ACs, rather than the 802.11e Hybrid Coordinated Channel Access (HCCA), to meet TSpec requests is possible in many cases because the traffic parameters are sufficiently simple to allow them to be met by allocating capacity, rather than creating a specific TXOP to meet the application requirements.

Note Unlike the 7921G, which does have support for TSpec, the Cisco 7920 WVoIP handset does not support TSpec admission control.

Add Traffic Stream

The Add Traffic Stream (ADDTS) function is how a WLAN client performs an admission request to an AP. Signalling its TSpec request to the AP, an admission request is in one of two forms:

- ADDTS action frame—This happens when a phone call is originated or terminated by a client associated to the AP. The ADDTS contains TSpec and might contain a traffic stream rate set (TSRS) IE (Cisco Compatible Extensions v4 clients).
- Association and re-association message—The association message might contain one or more TSpecs and one TSRS IE if the STA wants to establish the traffic stream as part of the association. The re-association message might contain one or more TSpecs and one TSRS IE if an STA roams to another AP.

The ADDTS contains the TSpec element that describes the traffic request. See Figure 5-13 and Figure 5-14 for examples of an ADDTS request and response between a Cisco 7921 WLAN handset and a Cisco AP. Apart from key data describing the traffic requirements, such as data rates and frame sizes, the TSpec element also tells the AP the minimum physical rate that the client device will use. This allows the calculation of how much time that station can potentially consume in sending and receiving in this TSpec, and therefore allowing the AP to calculate whether it has the resources to meet the TSpec. TSpec admission control is used by the WLAN client (target clients are VoIP handsets) when a call is initiated and during a roam request. During a roam, the TSpec request is appended to the re-association request.

Figure 5-13 ADDTS Request Decode

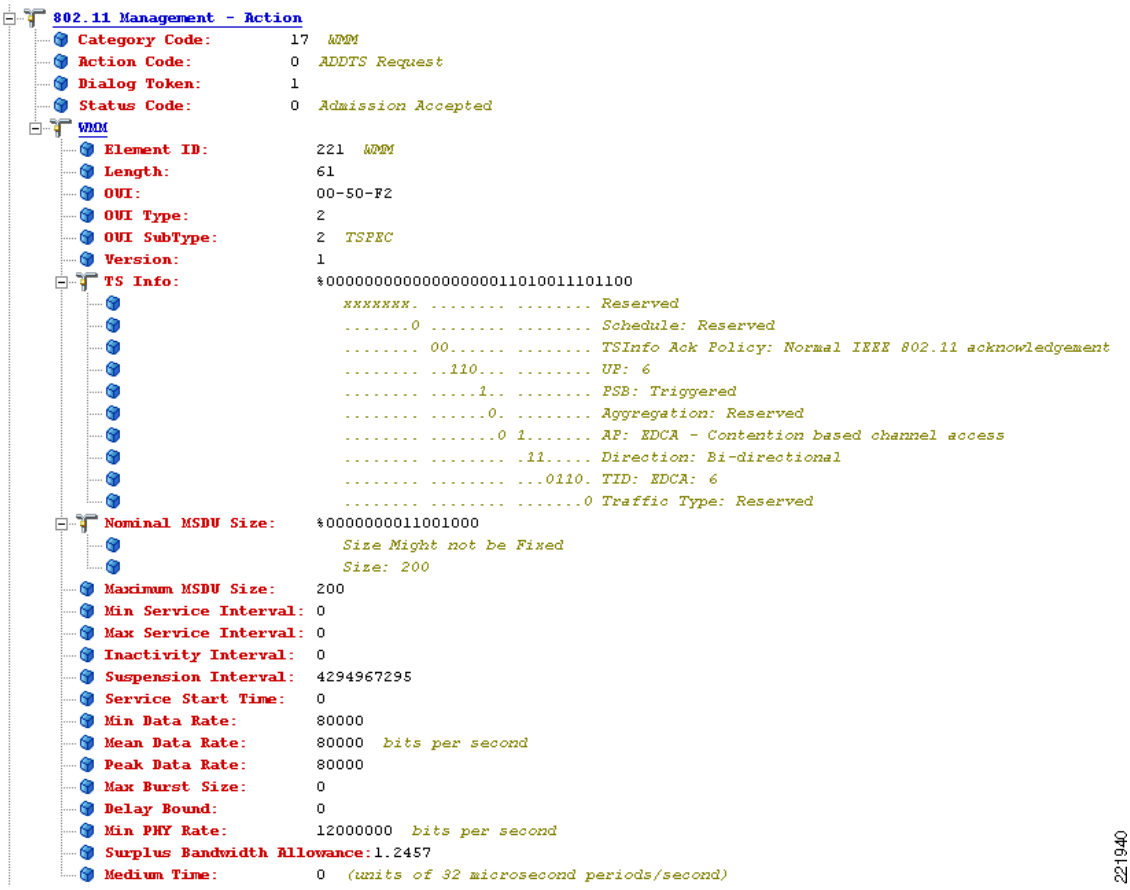
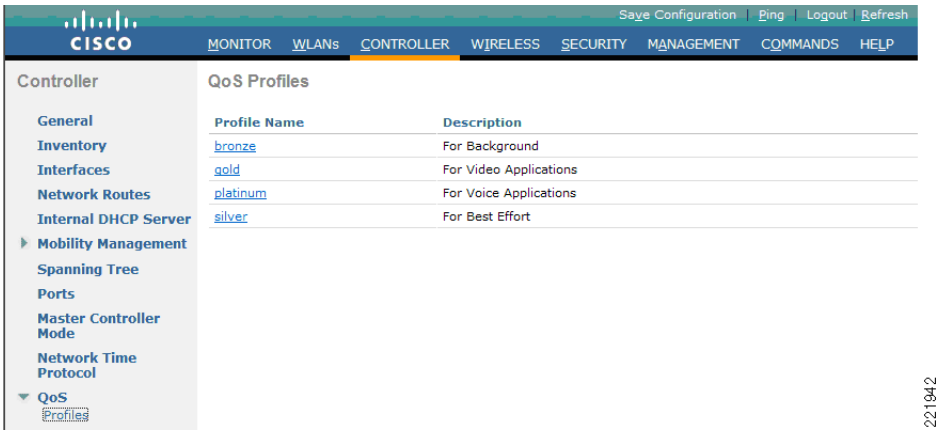


Figure 5-14 ADDTS Response Decode

QoS Advanced Features for WLAN Infrastructure

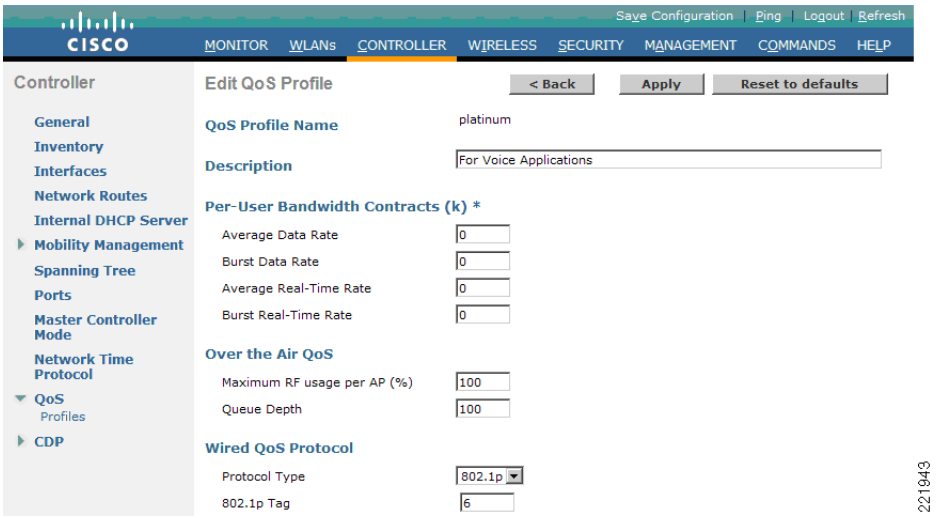
The Cisco Centralized WLAN Architecture has multiple QoS features, in addition to WMM support. Primary among these are the QoS profiles in the WLC. Four QoS profiles can be configured: platinum, gold, silver, and bronze, as shown in Figure 5-15.

Figure 5-15 QoS Profile Options



Each of the profiles shown in Figure 5-16 allows the configuration of bandwidth contracts, RF usage control, and the maximum 802.1P classification allowed.

Figure 5-16 QoS Profile Settings



It is generally recommended that the Per-User Bandwidth Contracts settings be left at their default values, and that the 802.11 WMM features be used to provide differentiated services.

For WLANs using a given profile, the 802.1P classification in that profile controls two important behaviors:

- Determines what class of service (CoS) value is used for packets initiated from the WLC.

  The CoS value set in the profile is used to mark the CoS of all LWAPP packets for WLAN using that profile. So a WLAN with a platinum QoS profile, and the 802.1P mark of 6, will have its LWAPP packets from the ap-manager interface of the controller marked with CoS of 5. The controller adjusts the CoS to be compliant with Cisco QoS baseline recommendations. The reason why it is important to maintain the IEEE CoS marking in the configuration is covered in the next point. If the network is set to trust CoS rather a DSCP at the network connection to the WLC, the CoS value determines the DSCP of the LWAPP packets received by the AP, and eventually the WMM classification and queuing for WLAN traffic, because the WLAN WMM classification of a frame is derived from the DSCP value of the LWAPP packet carrying that frame.

- Determines the maximum CoS value that can be used by clients connected to that WLAN.

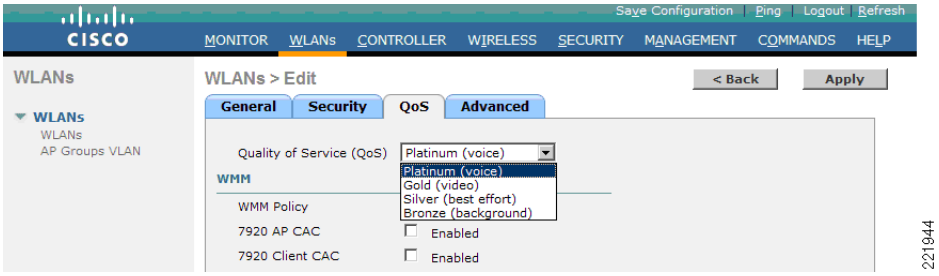
  The 802.1P classification sets the maximum CoS value that is admitted on a WLAN with that profile.

WMM voice traffic arrives with a CoS of 6 at the AP, and the AP automatically performs a CoS-to-DSCP mapping for this traffic based on a CoS of 6. If the CoS value in the WLC configuration is set to a value less than 6, this changed value is used by the WLAN QoS profile at the AP to set the maximum CoS marking used and therefore which WMM AC to use.

The key point is that with the Unified Wireless Network, you should always think in terms of IEEE 802.11e classifications, and allow the Unified Wireless Network Solution to take responsibility for converting between IEEE classification and the Cisco QoS baseline.

The WLAN can be configured with various default QoS profiles, as shown in Figure 5-17. Each of the profiles (platinum, gold, silver, and bronze) are annotated with their typical use. In addition, clients can be assigned a QoS profile based on their identity, through AAA. For a typical enterprise, WLAN deployment parameters, such as per-user bandwidth contracts and over-the-air QoS, should be left at their default values, and standard QoS tools, such as WMM and wired QoS, should be used to provide optimum QoS to clients.

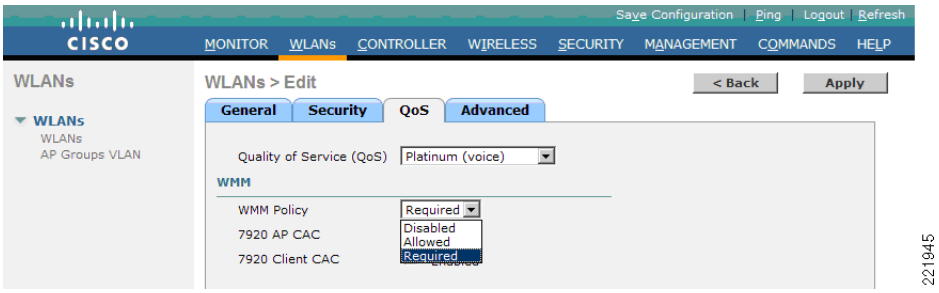
Figure 5-17 WLAN QoS Profile



In addition to the QoS profiles, the WMM policy per WLAN can also be controlled, as shown in Figure 5-18. The three WMM options are as follows:

- Disabled—The WLAN does not advertise WMM capabilities, or allow WMM negotiations,
- Allowed—The WLAN does allow WMM and non-WMM clients
- Required—Only WMM-enabled clients can be associated with this WLAN.

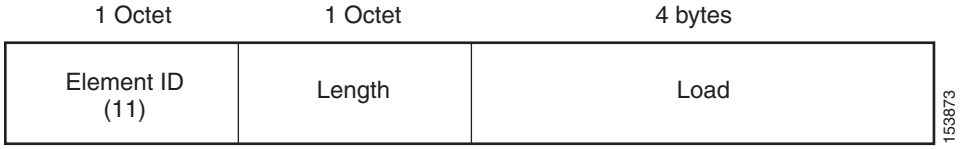
Figure 5-18 WLAN WMM Policy



IP Phones

Figure 5-19 shows the basic QoS Basis Service Set (QBSS) information element (IE) advertised by a Cisco AP. The Load field indicates the portion of available bandwidth currently used to transport data on that AP.

Figure 5-19 QBSS Information Element



There are actually three QBSS IEs that need to be supported in certain situations:

- Old QBSS (Draft 6 (pre-standard))
- New QBSS (Draft 13 802.11e (standard))
- New distributed CAC load IE (a Cisco IE)

The QBSS used depends on the WMM and 7920 settings on the WLAN.

7920 phone support, shown in Figure 5-18, is a component of the WLC WLAN configuration that enables the AP to include the appropriate QBSS element in its beacons. WLAN clients with QoS requirements, such as the 7920 and 7921G, use these advertised QoS parameters to determine the best AP with which to associate.

The WLC provides 7920 support through the client call admission control (CAC) limit, or AP CAC limit. These features provide the following:

- Client CAC limit—The 7920 uses a call admission control setting that is set on the client. This supports legacy 7920 code-pre 2.01.
- AP CAC limit—The 7920 uses CAC settings learned from WLAN advertisement.

The various combinations of WMM, client CAC limit, and AP CAC limit result in different QBSS IEs being sent:

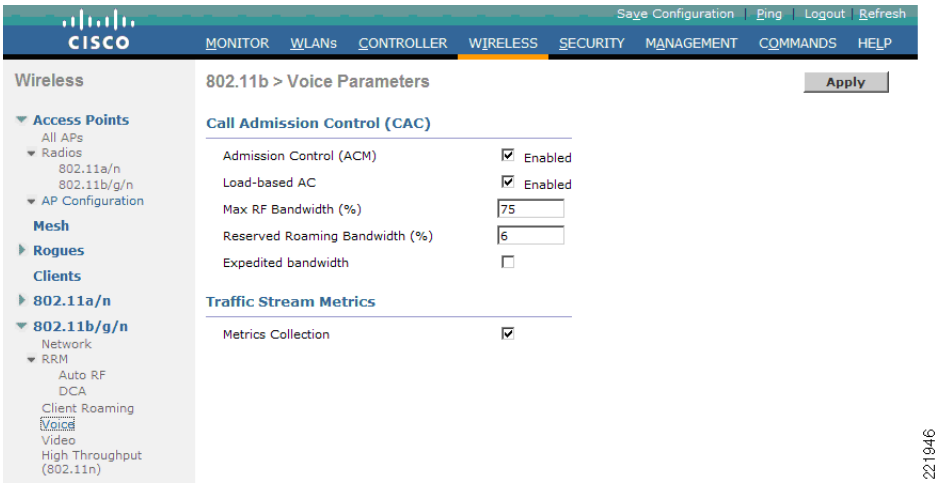
- If only WMM is enabled, IE number 2 (802.11e standard) QBSS Load IE is sent out in the beacons and probe responses.
- If 7920 client CAC limit is to be supported, IE number 1 (the pre-standard QBSS IE) is sent out in the beacons and probe responses on the bg radios.
- If 7920 AP CAC limit is to be supported, the number 3 QBSS IE is sent in the beacons and probe responses for bg radios.

Note The various QBSS IEs use the same ID, and therefore the three QBSSs are mutually exclusive. For example, the beacons and probe responses can contain only one QBSS IE.

Setting the Admission Control Parameters

Figure 5-20 shows a sample configuration screen for setting the voice parameters on the controller.

Figure 5-20 Voice Parameter Setting



The admission control parameters consist of the maximum RF Bandwidth that a radio can be using and still accept the initiation of a VoWLAN call through a normal ADDTS request.

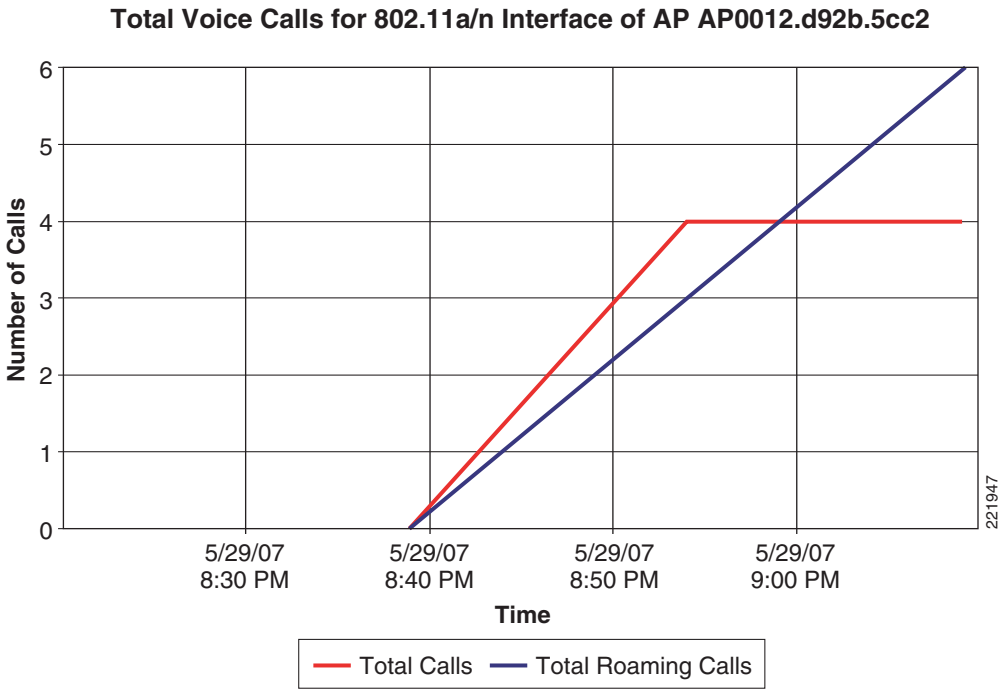
The reserved roaming bandwidth is how much capacity has been set aside to be able to respond to ADDTS requests during association or re-association, and which are VoWLAN clients with calls in progress that are trying to roam to that AP.

To enable admission control based upon these parameters to, use the Admission Control (ACM) checkbox. This enables admission control, based upon the APs capacity, but does not take into account the possible channel loading impact of other APs in the area. To include this “channel load” in capacity calculations, check the Load-Based AC checkbox as well as the Admission Control (ACM) checkbox.

The Metrics Collection option determines whether data is collected on voice or video calls for use by the WCS.

Figure 5-21 shows an example of one of the voice statistics reports available on the WCS, which shows the calls established on the radio of one AP, and the number of calls that roamed to that AP. This report and other voice statistics can be scheduled or ad-hoc, and either graphically displayed or posted as a data file.

Figure 5-21 Voice Statistics from WCS



Note Call admission control is performed only for voice and video QoS profiles.

Impact of TSpec Admission Control

The purpose of TSpec admission control is not to deny clients access to the WLAN; it is to protect the high priority resources. Therefore, a client that has not used TSpec admission control does not have its traffic blocked; it simply has its traffic re-classified if it tries to send (which it should not do if the client is transmitting WMM-compliant traffic in a protected AC).

Table 5-5 and Table 5-6 describe the impact on classification if access control is enabled and depending on whether a traffic stream has been established.

Table 5-5 Upstream Traffic

	Traffic Stream Established	No Traffic Stream
No admission control	No change in behavior; the packets go into the network as they do today-UP is limited to max= WLAN QoS setting.	No change in behavior; the packets go into the network as they do today-UP is limited to max= WLAN QoS setting.
Admission control	No change in behavior; the packets go into the network as they do today-UP is limited to max= WLAN QoS setting.	Packets are remarked to BE (both CoS and DSCP) before they enter the network for WMM clients. For non-WMM clients, packets are sent with WLAN QoS.

Table 5-6 Downstream Traffic

	Traffic Stream Established	No Traffic Stream
No admission control	No change	No change
Admission control	No change	Remark UP to BE for WMM client. For non-WMM clients, use WLAN QoS.

802.11e, 802.1P, and DSCP Mapping

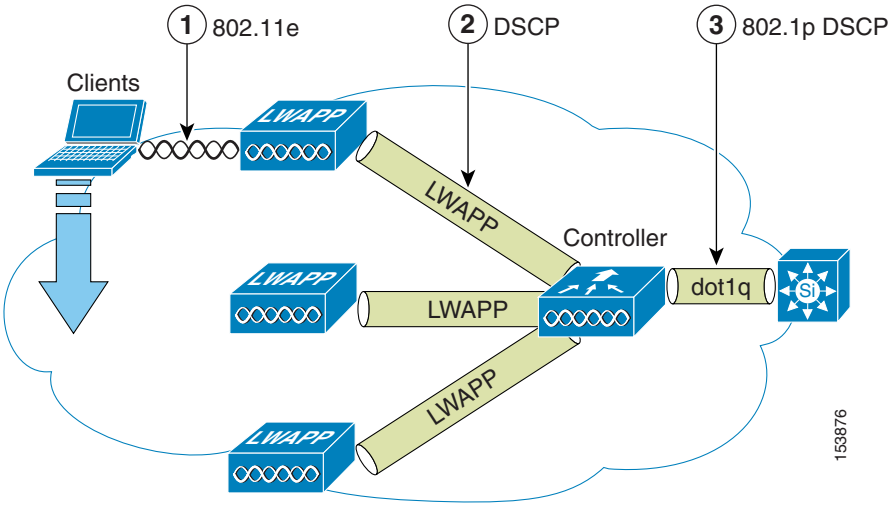
WLAN data in a Unified Wireless network is tunneled via LWAPP (IP UDP packets). To maintain the QoS classification that has been applied to WLAN frames, a process of mapping classifications to and from DSCP to CoS is required.

For example, when WMM classified traffic is sent by a WLAN client, it has an 802.1P classification in its frame. The AP needs to translate this classification into a DSCP value for the LWAPP packet carrying the frame to ensure that the packet is treated with the appropriate priority on its way to the WLC. A similar process needs to occur on the WLC for LWAPP packets going to the AP.

A mechanism to classify traffic from non-WMM clients is also required, so that their LWAPP packets can also be given an appropriate DSCP classification by the AP and the WLC.

Figure 5-22 shows the various classification mechanisms in the LWAPP WLAN network.

Figure 5-22 WMM and 802.1P Relationship



The multiple classification mechanisms and client capabilities require multiple strategies:

- LWAPP control frames require prioritization, and LWAPP control frames are marked with a DSCP classification of CS6.
- WMM-enabled clients have the classification of their frames mapped to a corresponding DSCP classification for LWAPP packets to the WLC. This mapping follows the standard IEEE CoS-to-DSCP mapping, with the exception of the changes necessary for QoS baseline compliance. This DSCP value is translated at the WLC to a CoS value on 802.1Q frames leaving the WLC interfaces.

- Non-WMM clients have the DSCP of their LWAPP tunnel set to match the default QoS profile for that WLAN. For example, the QoS profile for a WLAN supporting 7920 phones would be set to platinum, resulting in a DSCP classification of EF for data frames packets from that AP WLAN.
- LWAPP data packets from the WLC have a DSCP classification that is determined by the DSCP of the wired data packets sent to the WLC. The 80211.e classification used when sending frames from the AP to a WMM client is determined by the AP table converting DSCP to WMM classifications.

Note The WMM classification used for traffic from the AP to the WLAN client is based on the DSCP value of the LWAPP packet, and not the DSCP value of the contained IP packet. Therefore, it is critical that an end-to-end QoS system is in place.

QoS Baseline Priority Mapping

The LWAPP AP and WLC perform QoS baseline conversion, so that WMM values as shown in Table 5-7 are mapped to the appropriate QoS baseline DSCP values, rather than the IEEE values.

Table 5-7 Access Point QoS Translation Values

AVVID 802.1 UP-Based Traffic Type	AVVID IP DSCP	AVVID 802.1p UP	IEEE 802.11e UP
Network control	-	7	-
Inter-network control (LWAPP control, 802,11 management	48	6	7
Voice	46 (EF)	5	6
Video	34 (AF41)	4	5
Voice Control	26 (AF31)	3	4
Background (gold)	18 (AF21)	2	2
Background (gold)	20 (AF22)	2	2
Background (gold)	22 (AF23)	2	2
Background (silver)	10 (AF11)	1	1
Background (silver)	12 (AF12)	1	1
Background (silver)	14 (AF13)	1	1
Best Effort	0 (BE)	0	0, 3
Background	2	0	1
Background	4	0	1
Background	6	0	1

Deploying QoS Features on LWAPP-based APs

When deploying WLAN QoS on the APs, consider the following:

- The wired LWAPP AP interface does read or write Layer 2 CoS (802.1P) information, the WLC and the APs depend on Layer 3 classification (DSCP) information to communicate WLAN client traffic classification. This DSCP value may be subject to modification by intermediate routers, and therefore the Layer 2 classification received by the destination might not reflect the Layer 2 classification marked by the source of the LWAPP traffic.
- The APs no longer use NULL VLAN ID. As a consequence, L2 LWAPP does not effectively support QoS because the AP does not send the 802.1P/Q tags, and in L2 LWAPP there is no outer DSCP on which to fall back.
- APs do not re-classify frames; they prioritize based on CoS value or WLAN profile.
- APs carry out EDCF-like queuing on the radio egress port only.
- APs do FIFO queueing only on the Ethernet egress port.

WAN QoS and the H-REAP

For WLANs that have data traffic forwarded to the WLC, the behavior is same as non-hybrid remote edge access point (H-REAP) APs. For locally-switched WLANs with WMM traffic, the AP marks the dot1p value in the dot1q VLAN tag for upstream traffic. This occurs only on tagged VLANs; that is, not native VLANs.

For downstream traffic, the H-REAP uses the incoming dot1q tag from the Ethernet side and uses this to queue and mark the WMM values on the radio of the locally-switched VLAN.

The WLAN QoS profile is applied both for upstream and downstream packets. For downstream, if an 802.1P value that is higher than the default WLAN value is received, the default WLAN value is used. For upstream, if the client sends an WMM value that is higher than the default WLAN value, the default WLAN value is used. For non-WMM traffic, there is no CoS marking on the client frames from the AP.

Note Bug CSCsi78368 currently impacts the CoS Marking of traffic from the WLC and the CoS marked on frames sent by the WLC represents the value set by the QoS profile and not the WMM CoS marked by the client.

Guidelines for Deploying Wireless QoS

The same rules for deploying QoS in a wired network apply to deploying QoS in a wireless network. The first and most important guideline in QoS deployment is to know your traffic. Know your protocols, the sensitivity to delay of your application, and traffic bandwidth. QoS does not create additional bandwidth; it simply gives more control over where the bandwidth is allocated.

Throughput

An important consideration when deploying 802.11 QoS is to understand the offered traffic, not only in terms of bit rate, but also in terms of frame size, because 802.11 throughput is sensitive to the frame size of the offered traffic.

Table 5-8 shows the impact that frame size has on throughput: as packet size decreases, so does throughput. For example, if an application offering traffic at a rate of 3 Mbps is deployed on an 11 Mbps 802.11b network, but uses an average frame size of 300 bytes, no QoS setting on the AP allows the

application to achieve its throughput requirements. This is because 802.11b cannot support the required throughput for that throughput and frame size combination. The same amount of offered traffic, having a frame size of 1500 bytes, does not have this issue.

Table 5-8 Throughput Compared to Frame Size

	300	600	900	1200	1500	Frame Size (bytes)
11g–54 Mbps	11.4	19.2	24.6	28.4	31.4	Throughput Mbps
11b–11 Mbps	2,2	3.6	4.7	5.4	6	Throughput Mbps

QoS Example LAN Switch Configuration

AP Switch Configuration

The QoS configuration of the AP switch is relatively trivial because the switch must trust the DSCP of the LWAPP packets that are passed to it from the AP. There is no CoS marking on the LWAPP frames coming from the AP. The following is an example of this configuration. Note that this configuration addresses only the classification, and that queueing commands may be added, depending on local QoS policy.

```
interface GigabitEthernet1/0/1
 switchport access vlan 100
 switchport mode access
 mls qos trust dscp
 spanning-tree portfast
end
```

In trusting the AP DSCP values, the access switch is simply trusting the policy set for that AP by the WLC. The maximum DSCP value assigned to client traffic is based on the QoS policy applied to the WLANs on that AP.

WLC Switch Configuration

The QoS classification decision at the WLC-connected switch is a bit more complicated than at the AP-connected switch, because the choice can be to either trust the DSCP or the CoS of traffic coming from the WLC. In this decision there are a number of points to consider:

- Traffic leaving the WLC can be either upstream (to the WLC or network) or downstream (the AP and WLAN client). The downstream traffic is LWAPP encapsulated, and the upstream traffic is from AP and WLAN clients, either LWAPP encapsulated or decapsulated WLAN client traffic, leaving the WLC.
- DSCP values of LWAPP packets are controlled by the QoS policies on the WLC; the DSCP values set on the WLAN client traffic encapsulated by the LWAPP tunnel header has not been altered from those set by the WLAN client.
- CoS values of frames leaving the WLC are set by the WLC QoS policies, regardless of whether they are upstream, downstream, encapsulated, or decapsulated.

The following example chooses to trust the CoS of settings of the WLC, because this allows a central location for the management of WLAN QoS, rather than having to manage the WLC configuration and an additional policy at the WLC switch connection. Other customers wishing to have a more precise degree of control may wish to implement QoS classification policies on the WLAN-client VLANs.

```
interface GigabitEthernet1/0/13
```

```
 switchport trunk encapsulation dot1q
 switchport trunk allowed vlan 11-13,60,61
 switchport mode trunk
 mls qos trust cos
end
```

Traffic Shaping, Over the Air QoS, and WMM Clients

Traffic shaping and over-the-air QoS are useful tools in the absence of WLAN WMM features, but they do not address the prioritization of 802.11 traffic directly. For WLANs that support WMM clients or 7920 handsets, the WLAN QoS mechanisms of these clients should be relied on; no traffic shaping or over-the-air QoS should be applied to these WLANs.

WLAN Voice and the Cisco 7921G and 7920

The Cisco 7921G and the Cisco 7920 are Cisco VoWLAN handsets. Their use is one of the most common reasons for deploying QoS on a WLAN.

For more information on each of these handsets, see the following:

- Cisco Unified Wireless IP Phone 7921G Version 1.0(2)—http://www.cisco.com/en/US/products/hw/phones/ps379/products_data_sheet0900aecd805e315d.html
- Cisco Unified Wireless IP Phone 7920 Version 3.0—http://www.cisco.com/en/US/products/hw/phones/ps379/products_data_sheet09186a00801739bb.html

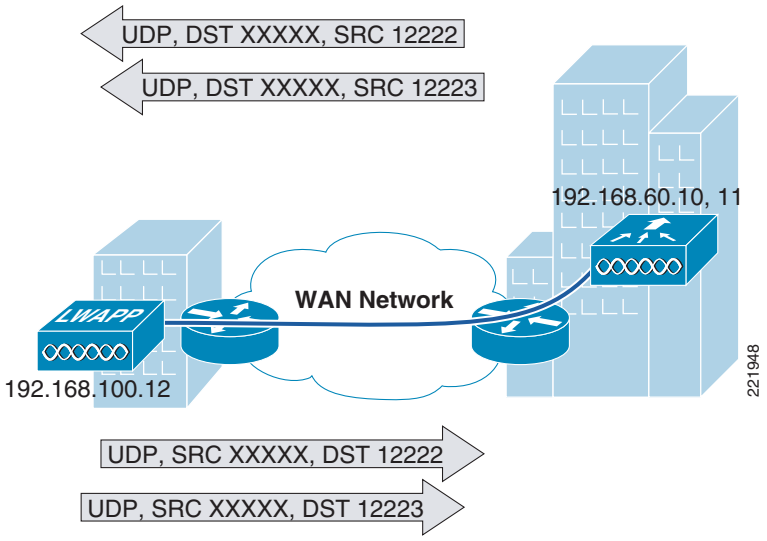
Deploying VoWLAN infrastructure involves more than simply providing QoS on WLAN. A voice WLAN needs to consider site survey coverage requirements, user behavior, roaming requirements, and admission control. These are covered in the following guides:

- Design Principles for Voice Over WLAN—http://www.cisco.com/en/US/solutions/collateral/ns340/ns394/ns348/net_implementation_white_paper0900aecd804f1a46.html
- Cisco Wireless IP Phone 7920 Design and Deployment Guide—http://www.cisco.com/en/US/docs/voice_ip_comm/cuipph/7920/5_0/english/design/guide/7920ddg.html

LWAPP over WAN Connections

This section describes QoS strategies when LWAPP APs are deployed across WAN links, as illustrated in Figure 5-23.

Figure 5-23 LWAPP Traffic Across the WAN



LWAPP Traffic Classification

LWAPP APs can be generally separated into the following two types:

- LWAPP control traffic—Identified by UDP port 12223
- LWAPP 802.11 traffic—Identified by UPD port 12222

LWAPP Control Traffic

LWAPP control traffic can be generally divided into the following two additional types:

- Initialization traffic—Generated when an LWAPP AP is booted and joins an LWAPP system. For example, the traffic generated by controller discovery, AP configuration, and AP firmware updates.

Note LWAPP image packets from the controller are marked best-effort, but their acknowledgement is marked CS6. Note that there is no windowing of the protocol, and each additional packet is sent only after an acknowledgement. This type of handshaking minimizes the impact of downloading files over a WAN

- Background traffic—Generated by an LWAPP AP when it is an operating member of a WLAN network. For example, LWAPP heartbeat, RRM, rogue AP measurements. Background LWAPP control traffic is marked CS6.

Figure 5-24 and Figure 5-25 show examples of the initial LWAPP control messages. Figure 5-26, Figure 5-27, and Figure 5-28 show examples of background LWAPP control messages.

A full list of initial LWAPP control messages includes the following:

- LWAPP discovery messages
- LWAPP join messages
- LWAPP config messages

- Initial LWAPP RRM messages

Although AP image download is also discussed in this section, it is not typically part of an AP initialization, and would only occur during firmware changes.

Figure 5-24 LWAPP Discovery Message

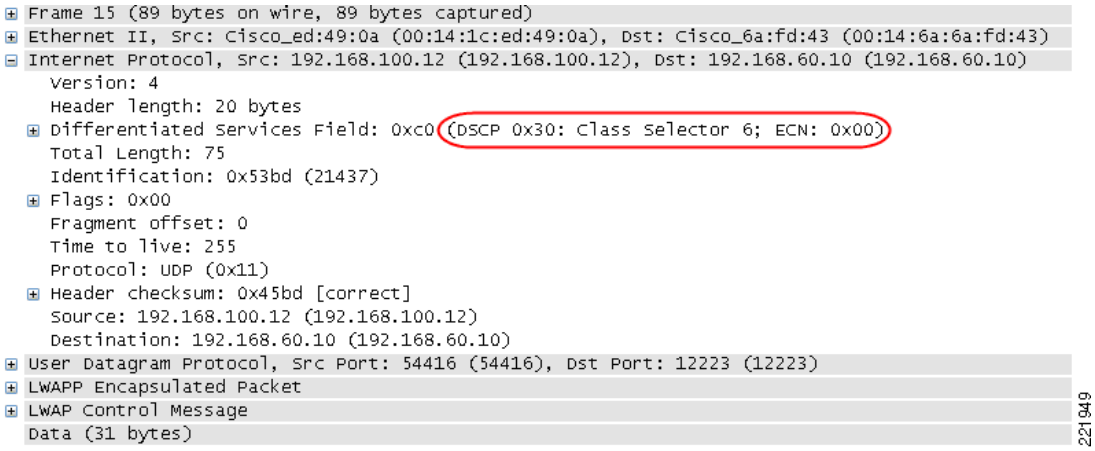


Figure 5-25 LWAPP Image Response

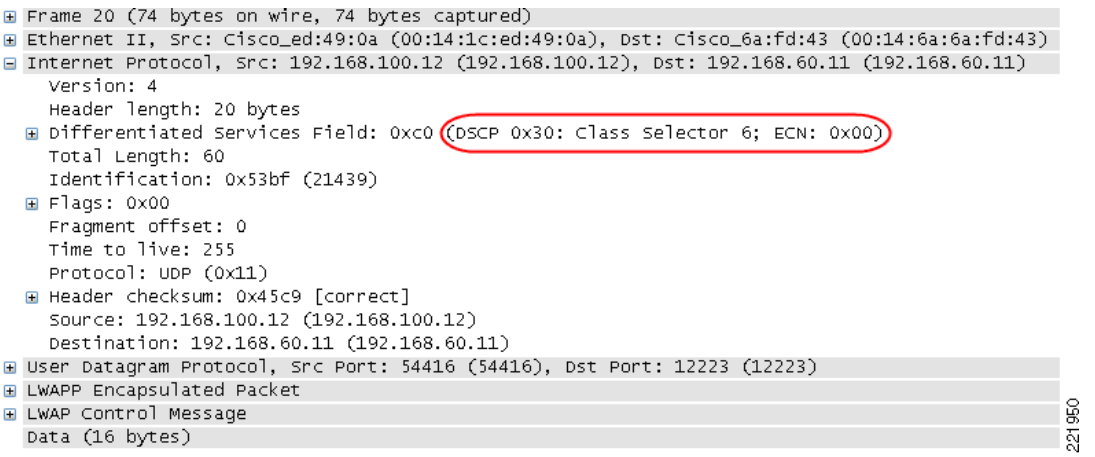


Figure 5-26 LWAPP Heartbeat Messages

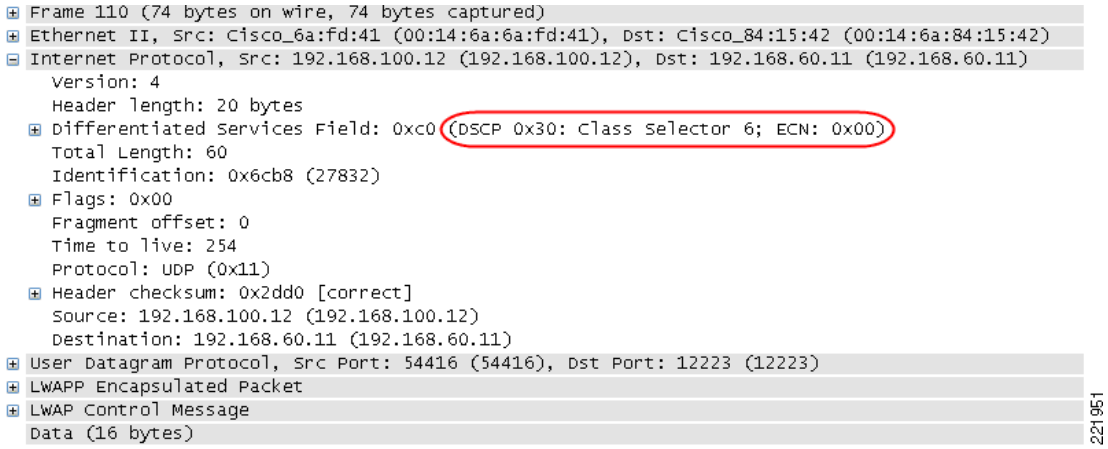


Figure 5-27 LWAPP Statistics

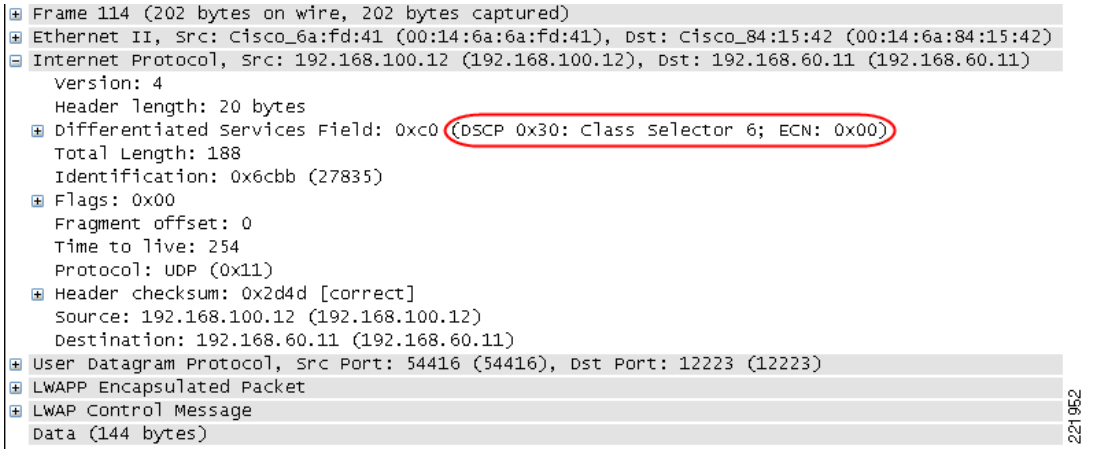
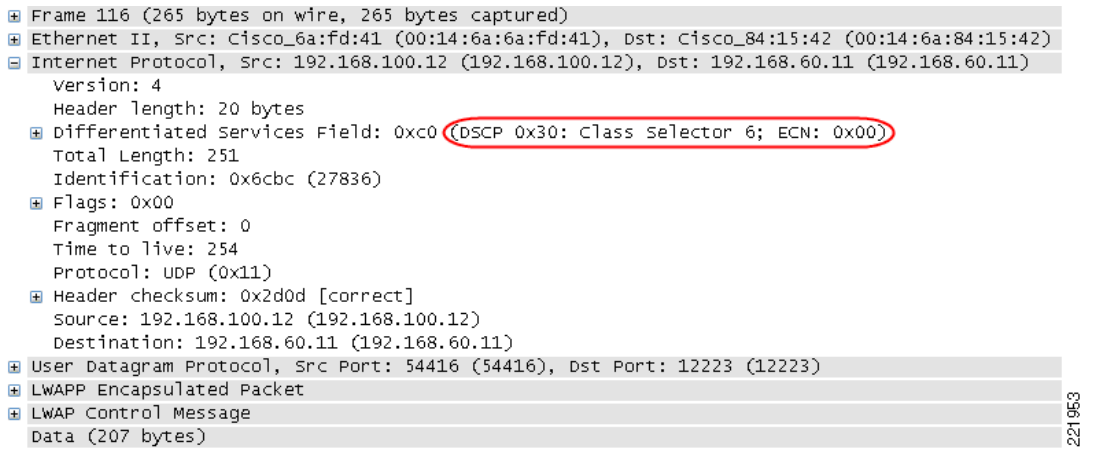


Figure 5-28 LWAPP RRM



LWAPP 802.11 Traffic

LWAPP 802.11 traffic can be divided generally into the following two additional types:

- 802.11 management frames—802.11 management frames such as probe requests, and association requests and responses are classified automatically with a DSCP of CS6.
- 801.11 data frames—Client data and 802.1X data from the client is classified according to the WLAN QoS settings, but packets containing 802.1X frames from the WLC are marked CS4. 802.11 data traffic classification depends on the QoS policies applied in the WLAN configuration, and is not automatic. The default classification for WLAN data traffic is best-effort.

Classification Considerations

The DSCP classification of used LWAPP control traffic is CS6, which is an IP routing class, and is intended for IP routing protocols such as Border Gateway Protocol (BGP), Open Shortest Path First (OSPF), Enhanced Interior Gateway Routing Protocol (EIGRP), and so on.

The current LWAPP DSCP classification represents a classification that, although optimal for the WLAN system, may not align with the QoS policies and needs of each customer.

In particular, a customer may wish to minimize the amount of CS6-classified traffic generated by the WLAN network. They may wish to stop CS6 traffic generated by client activity such as probe requests. The simplest mechanism to do this would be to reclassify the LWAPP 802.11 CS6 traffic to a different DSCP. The fact that the LWAPP UDP port used is different from that used by LWAPP data, and the default DSCP marking allow for remarking this traffic without resorting to deep packet inspection.

In addition, a customer may wish to ensure that LWAPP initialization traffic does not impact routing traffic. The simplest mechanism for ensuring this is to mark LWAPP control traffic that is in excess of the background rate with a lower priority.

LWAPP Traffic Volumes

Cisco testing has found that the average background traffic per AP is approximately 305 bits/sec.

Calculating the average initial traffic per AP is more difficult, because the average time taken for an AP to go from rebooted to operational is a function of the WAN speed, as well as that of the WLC and AP. In reality, the difference is minimal. While on a lab test network, the best of initial traffic might average 2614 bit/sec over 18 seconds. With a WAN link with a 100 ms RTT, the average is 2318 bits/sec over 20.3 seconds.

Example Router Configurations

This section contains router configuration examples to be used as guides when addressing CS6 remarking or LWAPP control traffic load.

This example uses LWAPP APs on the 192.168.101.0/24 subnet, and two WLCs with ap-managers at 192.168.60.11, and 192.168.62.11.

Remarking Client Generated CS6 Packets

The following shows a sample configuration for remarking LWAPP data packets marked as CS6 to a more appropriate value of CS3. This moves the traffic to a more suitable classification, at the level of call control, rather than at the level of network control.

```
class-map match-all LWAPPDATACS6
 match access-group 110
 match  dscp cs6
!
policy-map LWAPPDATACS6
 class LWAPPDATACS6
  set dscp cs3
!
interface FastEthernet0
 ip address 192.168.203.1 255.255.255.252
service-policy input LWAPPDATACS6
!
access-list 110 remark LWAPP Data
access-list 110 permit udp 192.168.101.0 0.0.0.255 host 192.168.60.11 eq 12222
access-list 110 permit udp 192.168.101.0 0.0.0.255 host 192.168.62.11 eq 12222
access-list 111 remark LWAPP Control
access-list 111 permit udp 192.168.101.0 0.0.0.255 host 192.168.60.11 eq 12223
access-list 111 permit udp 192.168.101.0 0.0.0.255 host 192.168.62.11 eq 12223
```

Changing the DSCP of LWAPP Control Traffic above a predefined rate

The following shows an example of rate limiting the LWAPP control traffic from the WAN site to minimize the impact of the CS6-marked control traffic on routing traffic. Note that the rate limit configuration does not drop non-conforming traffic, but simply reclassifies that traffic.

Note Note that this is an example, and not a recommendation. Under normal circumstances, and following the design guidelines for deploying APs over a WAN connection, it is unlikely that LWAPP control traffic would impact the WAN routing protocol connection.

```
interface Serial0
 ip address 192.168.202.2 255.255.255.252
 rate-limit output access-group 111 8000 3000 6000 conform-action transmit exceed-action
set-dscp-transmit 26
access-list 111 remark LWAPP Control
access-list 111 permit udp 192.168.101.0 0.0.0.255 host 192.168.60.11 eq 12223
access-list 111 permit udp 192.168.101.0 0.0.0.255 host 192.168.62.11 eq 12223
!
```

For more information on WLAN QoS and 802.11e, refer to the IEEE 802.11 Handbook, A designers companion (second edition), Bob O'Hara and Al Petrick.