White Paper

Enhancing Cisco High Density Experience with Cisco Air Time Fairness

Cisco has enhanced our High Density Experience solution with a new set of features. This white paper explains a key new feature that is one of many enhancements to HDX.

Introduction

The primary issue that Cisco® Air Time Fairness for High Density Experience (hereafter referred to as Air Time Fairness) addresses is the inability of other solutions to permit wireless LAN (WLAN) administrators to predictably define what fairness means for over-the-air client traffic consumption. Thus, the advantage of Air Time Fairness is an improved, simple mechanism for allocating traffic consumption by groups of client devices that need differing levels of service.

Air Time Fairness for High Density Experience allows a network administrator to group devices of a defined category and enable some groups to receive traffic from the WLAN more often than other groups. Therefore, some groups become entitled to more airtime than other groups. Our solution:

1. Allocates Wi-Fi air time for groups of user and device categories
2. Allows the network administrator to define airtime fairness (not the network)
3. Provides a simple mechanism for allocating airtime
4. Dynamically adapts to changing conditions in the WLAN
5. Provides improved fulfillment of service-level agreements (SLAs)
6. Augments (does not replace) standards-based Wi-Fi quality-of-service (QoS) mechanisms

This is particularly advantageous when there is a need to provide different services to different types of users based on predefined SLAs (which is another way of saying entitlement). Air Time Fairness simplifies management of WLAN traffic consumption for heterogeneous types of users. See the use cases in the next section of this document for more information.

Our solution can enable network administrators to define what fairness means within their environments with regard to the amount of on-air time per groups of clients. Correspondingly, the amount of traffic that can be consumed is also controlled.

Air Time Fairness Use Cases

Public Hotspots (Stadiums, Airports, Convention Centers, and Other Venues)

In these areas, a public network shares a WLAN between two or more service providers and services are offered locally by the venue. Subscribers to each service provider can be grouped and each group can be allocated a percentage of airtime. The venue staff itself is also a group that can also be allocated a percentage of airtime. This percentage can be fixed and statically allocated if needed. Other groups cannot acquire the venue airtime.

Education

In this case, a university shares a WLAN between students, faculty, and guests. The guest network can be further partitioned by the service provider. Each group can be assigned a percentage of airtime.

Enterprise, Hospitality, Retail

In these areas, the venue shares a WLAN between employees and guests. The service provider can further partition the guest network. Guests could be sub-grouped by tier of service type, with each subgroup assigned a percentage of airtime. For example, a paid group can be entitled to more airtime than a free group.

Time-Shared Managed Hotspot

In this case, the business entity managing the hotspot, whether a service provider or an enterprise, can allocate and subsequently lease airtime to other business entities.

Cisco Air Time Fairness - Basic Concepts and Goals

The easiest way to envision Air Time Fairness is that it provides the ability to both monitor and manage the amount of time on the air on a per-client, per-SSID, or per-other basis. (This paper will discuss the per-other category after an overview of the fundamental concepts.) The effect is to control the amount of traffic a client or SSID is sending into the WLAN and receiving from the WLAN at any given time.

In other words, the primary goal of Air Time Fairness is to help avoid any one client or SSID from occupying an unfair amount of Wi-Fi airtime on a particular channel (for instance, on a given access point or radio). In accomplishing this, Air Time Fairness also provides a customer with the ability to define what fairness means within their environment with regard to the amount of on-airtime and the amount of the traffic that can be sent in any given time period.

Because airtime is itself managed and is a shared resource, airtime rate-limiting and policing applies to the sum of downlink and uplink airtime.

In summary, the goals of Air Time Fairness are to:

1. Provide the ability to allocate medium access as a percentage of available airtime instead of a specific bit rate
2. Make this ability available on a per-client, per-SSID, and per-other basis
3. Apply to **all** packets that go over the air, and not just data frame payload types (for example, Transmission Control Protocol [TCP] or User Datagram Protocol [UDP])
4. Apply to **all** clients, independent of type or current status (for instance, data rate, signal strength, etc.)

Why Rate Limiting Alone Cannot Be Applied

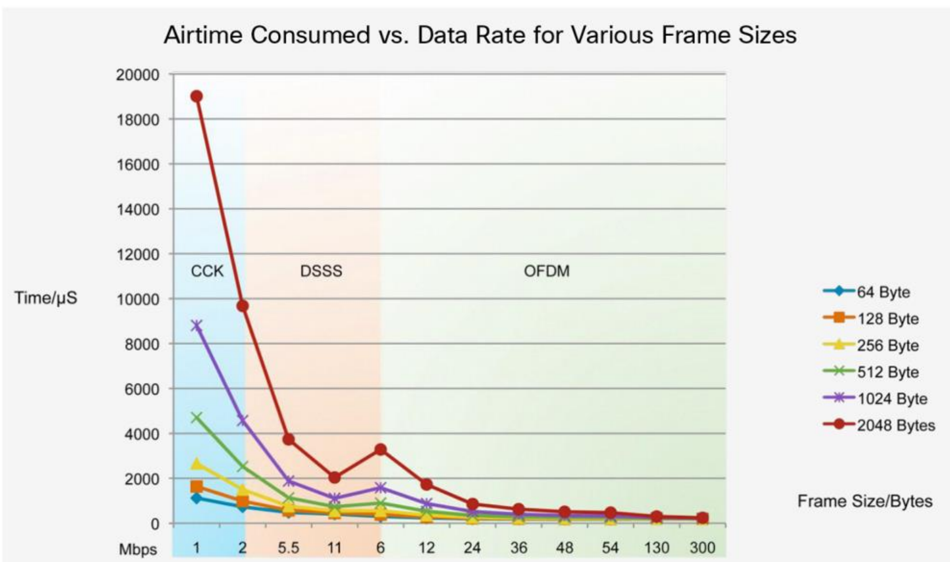
Although rate limiting controls throughput, it cannot control airtime. Policing bandwidth (capacity in bits per second [bps]) is **not** an efficient way to allocate Wi-Fi airtime. This is because clients connect at different speeds, depending on the RF link conditions between the access point and the client. Therefore, the overall capacity, being managed by a fixed-bandwidth rate limit, is dynamic. This means either we underuse the WLAN (rate limits are too restrictive) or overuse the WLAN (rate limits are too liberal). Furthermore, packet errors lead to retransmissions (to avoid packet loss), which consume additional airtime and are not accounted for in rate limiting.

Figure 1 shows that for any given packet size, the airtime consumed is dependent on the over-the-air PHY rate, which is itself dependent on RF link characteristics. As an example, the RF link characteristics are different based upon the distance of the client from the access point. A client that is nearer to the access point typically operates at a higher data rate than a client farther from the access point. This is because the access point and client are deliberately designed to adapt their transmission rates in order to maintain an optimal quality of the RF link. (This is normal behavior between the access point and clients since the client devices are neither expected to remain at a constant distance from the access point, nor are expected to be at equal distances from the access point.)

In brief, the RF environment is dynamic. Therefore, the airtime consumed by any single client will dynamically fluctuate over time.

As a consequence, two devices, each assigned a 10 Mbps rate limit and sending equally sized frames, would consume unequal airtime when their RF link characteristics are different This could mean that one device is consuming more airtime than it is entitled to, even though that device is not consuming more than its limit of bps.

Figure 1. Airtime Consumption Dependencies



Thus, SLAs made between a WLAN operator and its customers are difficult to fulfill if only rate limiting is used. Policing usage on an airtime basis is the optimal way to assure that all parties have access to the resource that they are entitled to.

Air Time Fairness - Behavior

In order to control or constrain airtime on a percentage basis, the airtime, which includes both uplink and downlink transmissions of a client or SSID is continuously measured.

However, note that only airtime in the downlink direction (access point to client) can be controlled accurately by the access point. Although airtime in the uplink direction (client to access point) can be measured, it cannot be strictly controlled on a cumulative basis. The access point can constrain the airtime for any data frame it sends to or receives from a client through configuration of transmission opportunity (TXOP). However, the access point can only measure airtime for data frames that it 'hears' from each client, as it cannot strictly limit the number of data frames any individual client sends.

As previously mentioned, a goal of Air Time Fairness is to establish airtime limits, defined as a percentage of total airtime, and to apply those limits on a per-SSID basis, where the SSID is used as parameter to define a group of clients. Other parameters can be used as well to define groups of clients, and will be discussed later in this paper.

It is important to note that if a client or SSID exceeds the airtime limit, packets will be dropped. However, only packets in the downlink direction are dropped. This is because dropping downlink packets (access point to client) frees up airtime. Dropping uplink packets (client to access point) does not do anything to free up airtime since the packet has already been transmitted over the air by the client. Once the packet has been on the air, that airtime cannot be reclaimed.

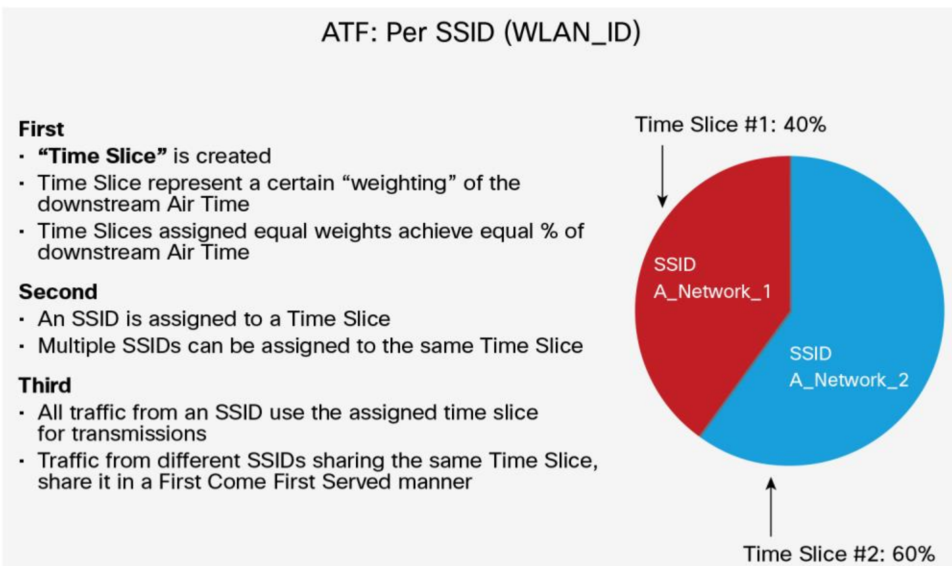
This can mean that a client that is transmitting a lot of uplink traffic can exhaust airtime. This is essentially a denial-of-service attack that is outside the scope of what Cisco Air Time Fairness addresses. Other mitigation is available for this scenario.

Cisco Air Time Fairness Per SSID

A key concept in allocating airtime is the idea of a time slice. A time slice typically represents a specified maximum, not to exceed a certain percentage of the available downstream airtime.

Figure 2 describes the basic principles underlying the time slice.

Figure 2. Time Slice



The configuration of Air Time Fairness from the network administrator’s perspective can be easily explained.

Note: The configuration procedures which follow are included as examples only. Refer to the configuration guide for supported configuration procedures.

The first step in creating a time slice is to select a policy ID and policy name, and assign a numerical weight to the policy. Assigning a numerical weight to a policy corresponds to allocating a percentage of airtime to the policy. The greater the numerical weight, the greater the percentage of allocated airtime. Equal numerical weights achieve equal allocations of airtime. But the numerical weight is not literally equal to the actual percentage allocated.

The next step is to apply the newly created Air Time Fairness policy to: all access points ; a select group of access points; or to an single access point in a given WLAN (SSID). This step effectively assigns an SSID to a time slice. It is worthwhile to note that multiple SSIDs can be assigned to the same time slice.

However, this generally assumes a single customer requires multiple SSIDs for some business-related reason. Furthermore, an access point can support multiple WLANs with multiple policies. Policies set in individual access point and access point group configurations always override per-global network policies.

All traffic from a given SSID utilizes the assigned time slice for transmissions. Downlink traffic destined for different SSIDs will share the same time slice, and will share it in a first-come-first-served manner.

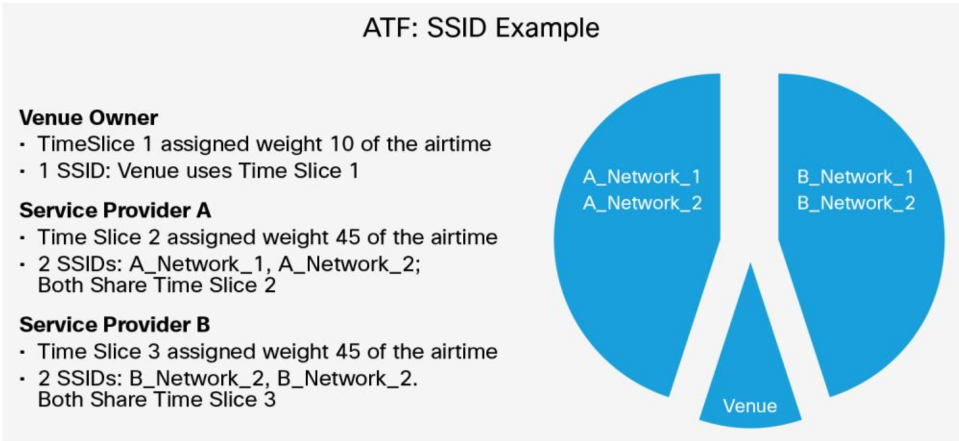
Next, consider a potential scenario that uses this mechanism (Figure 3). Assume a public Wi-Fi venue exists, where the venue WLAN provides services for itself and also for two service providers who lease the WLAN.

The network administrator assigns Time Slice 1 to the venue owner and assigns 10 percent of the available airtime to Time Slice 1. SSID 1 uses Time Slice 1 for the venue owner.

Next, the network administrator assigns Time Slice 2 to Service Provider A and assigns 45 percent of the available airtime to Time Slice 2. Note that Time Slice 2 could support two different networks (SSIDs) for Service Provider A. In this example, the SSIDs A_Network_1 and A_Network_2 both share Time Slice 2 since Service Provider A is leasing 45 percent of the available airtime.

Finally, the network administrator assigns Time Slice 3 to Service Provider B and assigns 45 percent of the available airtime to Time Slice 3. Time Slice 3 could also support two different networks (SSIDs) for Service Provider B. In this example, the SSIDs B_Network_1 and B_Network_2 both share Time Slice 3 since Service Provider B is leasing 45 percent of the available airtime.

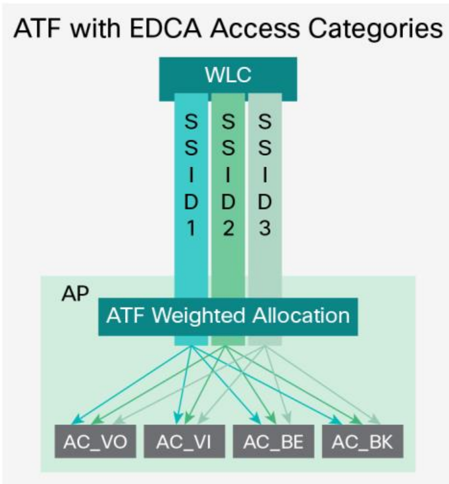
Figure 3. SSID Example



What the Air Time Fairness Solution Is Not

Before explaining Air Time Fairness in further detail, note that the solution does **not** replace Enhanced Distributed Channel Access (EDCA). Instead, it is applied in conjunction with EDCA. In the downlink direction, the queues that feed the EDCA output buffers are serviced in such a way to permit airtime to be distributed or allocated unequally per group. The queues are based on the types of users, devices, etc., instead of allowing only traffic patterns to determine medium access time.

Figure 4. Air Time Fairness with EDCA Access Categories



In this way a set of simple airtime allocation rules can be applied which still rely on EDCA rules for medium access, but which no longer use **only** EDCA methods for airtime allocation. In other words, the Air Time Fairness solution does not affect onto-the-air prioritization as EDCA is still used within each group. MAC timing is not modified.

Air Time Fairness is a "listen before talk" and Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) compatible scheduling algorithm. Contention is still used. The solution is neither a centralized nor controlled channel access coordinator, and not in any sense a rigid Time Division Multiple Access (TDMA) type scheduler.

Essentially, Air Time Fairness allows a network administrator to group devices by a predetermined type and to enable some groups to be serviced more often than other groups. As a result, some groups are entitled to more airtime than other groups.

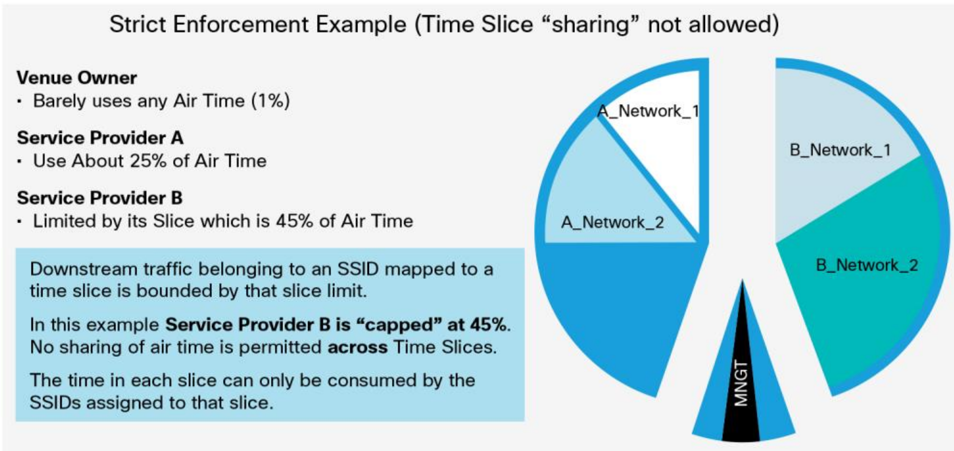
Air Time Fairness - Sharing Unused Air Time across Time Slices

In its simplest configuration, a time slice **typically** represents an assigned maximum percentage of the downstream airtime.

However, Air Time Fairness is not only configurable to share airtime within a time slice; it can also be configured to share airtime across time slices. Since it is possible that the maximum airtime entitled within a given time slice is not utilized, any unused time left over in a time slice can, at the discretion of the administrator, either be redistributed to other time slices as needed or simply remain unused.

First, consider an example **without** sharing airtime across time slices. In the example in Figure 5, the venue owner is using only one percent of an allocated 10 percent of airtime. Service Provider A is using about 25 percent of airtime out of an allocated 45 percent. Service Provider B is using its entire 45 percent of allocated airtime and is, therefore, limited by its time slice.

Figure 5. Sharing Airtime across Time Slices Not Allowed

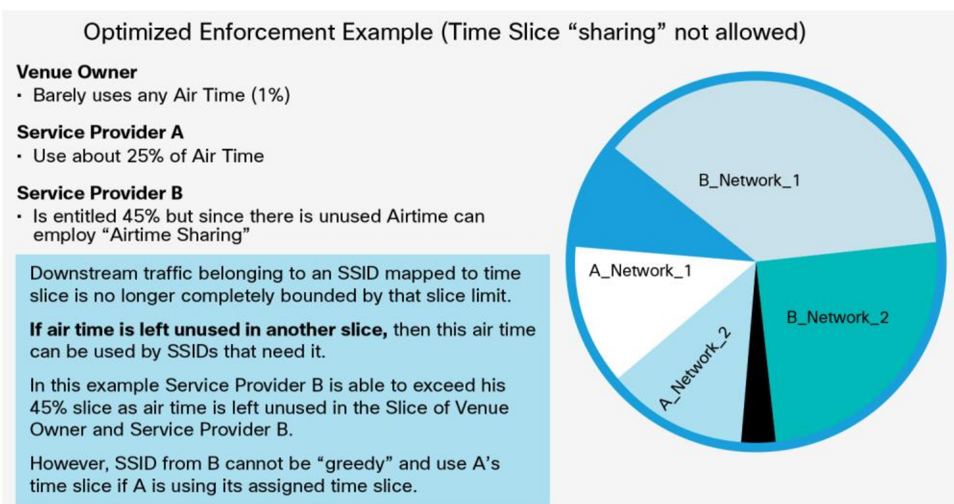


Typically, downstream traffic belonging to an SSID mapped to a time slice is bounded by that slice limit. Since sharing airtime across time slices in prohibited in the example in Figure 5, Service Provider B is capped at 45 percent of the available airtime.

However, the Air Time Fairness solution does optionally allow for downstream traffic belonging to an SSID and assigned to a time slice to no longer be completely bounded by that slice limit. If airtime is left unused in another time slice, this airtime can be temporarily and opportunistically used by time slices that need it.

Revisiting the earlier example but **with** sharing air time across time slices (Figure 6), Service Provider B is able to temporarily exceed its allocated 45 percent slice since airtime is left unused in the slices of the venue owner and Service Provider B.

Figure 6. How to Share Airtime across Time Slices



Automated protection is built into the allocation algorithms. This prevents clients connected to Service Provider B from accidentally or persistently using Service Provider A's unused allocation of the time slice if clients connected to Service Provider A begin using more of the time allocated to time slice A.

Alternate Solutions

In other airtime fairness implementations, the clients most recently consuming downlink airtime have subsequent service deferred in order to preserve fairness relative to other clients needing downlink airtime. This is not necessarily a poor approach, but it does mean that past traffic patterns are used to influence the downlink queue servicing. The network administrator still lacks direct, and accurate, control.

Summary

In conclusion, Cisco Air Time Fairness allows a network administrator to group devices by a defined category and enable some groups to receive traffic more often than other groups. As a result, some groups are entitled to more airtime than other groups.

By enabling network administrators to define what fairness means within their environments with regard to the amount of on-airtime per groups of clients, the amount of traffic that can be consumed is also controlled.

Appendix - Review of Wi-Fi QoS (Wi-Fi Multimedia [WMM] and EDCA)

A deeper understanding and increased appreciation of Cisco Air Time Fairness can be achieved through a brief review of the 802.11 medium access control (MAC).

The most important aspect of the 802.11 MAC is that it is "listen before talk" (LBT). Although this specific term is not used in the 802.11 specification, keeping this concept in mind is useful in describing how Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) behaves. As a result, the 802.11 MAC behavior is often described as "obsessively polite."

Furthermore, the original 802.11 specification did not include prioritized medium access. All devices had equal opportunity to access (transmit over) the air. Access was designed to be **distributed** across devices. That is, there was no centralized coordination required for **channel access**. All devices on the WLAN compete equally for transmitted airtime.

Consequently, LBT, or transmitting, consists of two consecutive intervals:

- A mandatory, minimum fixed wait time for all devices followed by
- A mandatory, randomly selected wait time per device

These intervals immediately follow the conclusion of the preceding transmission.

Any device intending to transmit first waits for the fixed time and then counts down from its randomly selected value. The first device to reach zero is allowed to transmit. All other devices detect that the air (the shared medium) has become busy and pause their countdown, resuming only after the current transmission finishes, and after the fixed minimum wait time elapses.

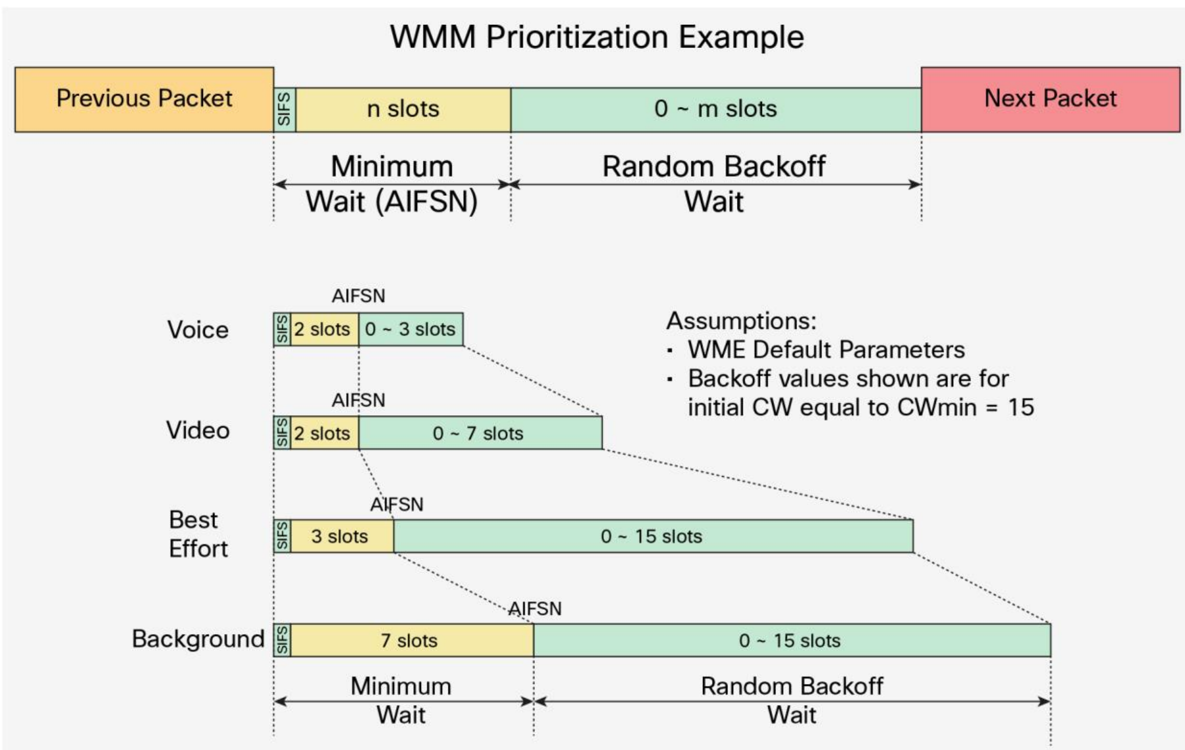
In theory, fairness occurs naturally since devices choose different random numbers. Neglecting hidden node effects, collisions only occur when two devices reach zero at the same time.

The 802.11e amendment introduced prioritization through Enhanced Distributed Channel Access (EDCA). EDCA is still obsessively polite, but introduces a controlled unfairness in order to enable an advantage in onto-the-air access for some traffic types. Essentially, the mandatory minimum wait time for some traffic types is less than that for other traffic types. Also, the subsequent mandatory randomly selected wait time is less for some traffic types compared to other traffic types. That is, the set of random numbers is fewer for some traffic types.

At a high level, EDCA creates a set of access categories based on traffic type. Traffic classified as voice waits less time than traffic classified as video, which waits less time than traffic classified as best effort, which waits less time than traffic classified as background. Best-effort traffic in EDCA has similar, but not identical, access opportunity to legacy Distributed Channel Access (DCA).

In Figure 7, WMM is the Wi-Fi Alliance certification for EDCA behavior.

Figure 7. WMM Prioritization



As can be seen, the net effect is that some traffic types get on the air sooner than other traffic types (waiting less time equals a higher priority). Each traffic type contends equally for airtime within its corresponding access category, but contention between traffic types, and access categories, is unequal.

However, this also means that the nature of the traffic itself determines access to the air. There is no way the 802.11 standard can be used alone to segment or partition airtime, and therefore, allow the network administrator to determine or define fairness. This is the issue the Cisco Air Time Fairness solution addresses.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA

C11-735947-00 11/15