CHAPTER 5

Quality of Service and Call Admission Control

Revised: March 30, 2012, OL-27011-01

Data packets that make up the video streams are transmitted over the network. The packets reach their destination based on the order in which they are placed into the queue. In a simple network behavior, the first packet in is the first packet out. For a single LAN switch this process is fairly simple because the switch will have some ports that are sending the media and some that are receiving it. As the network grows, the scenario changes from being an ideal world of ordered packets to a mass of unordered packets, where there are more packets being generated at the same time than can be sent through network links that do not have sufficient capacity to handle the load. In these real-world scenarios, some methods are needed to control the flow of packets across the network links.

Quality of Service (QoS)

Quality of Service (QoS) is used to identify certain types of packets that can be processed ahead of others. The QoS information is inserted into the packets that need a different priority.

QoS can be compared to a freeway system and the vehicles that use it. The network is similar to the freeway system, which provides the vehicles (data packets in the case of QoS) with a way to travel from their starting point to their destination. As long as the freeway has sufficient lanes and there is no incident, traffic flows smoothly in most cases and the travel time is acceptable to most users. However, during peak traffic times, things might not be as good. Carpool lanes can help. Cars that meet certain criteria have the privilege to use the carpool lanes and bypass the traffic congestion. In addition, emergency services such as ambulances have an even higher priority to bypass other traffic. On the other hand, large or heavily loaded vehicles might use more lanes and can slow down traffic. QoS is similar in that it allows certain packets to have preferred access to the network and to be transmitted ahead of other packets in the queue.

Traditionally IP Precedence or Type of Service (ToS) (RFC 791) was specified using three bits in the IP packet. The Differentiated Services (DiffServ) (RFC 2474 and RFC 2475) model uses six bits and also maintains the IP Precedence values. The DiffServ model uses assured forwarding (RFC 2597) that defines various classes of traffic with a drop probability and expedited forwarding (RFC 2598) to provide for low loss, low latency, and low jitter service. The class in assured forwarding is used to group different types of traffic, and the drop probability is used to group the traffic that will experience dropped packets last. The expedited forwarding is used for traffic such as voice that is sensitive to packet drop and delays.

Each type of traffic can have a different QoS value, and the network then provides preference when it identifies packets that have a higher QoS value. Table 5-1 lists some of the standard values used for various types of voice and video traffic.

Table 5-1 Differentiated Services Code Point (DSCP) Values for Various Types of Traffic

Traffic Type	Layer 2 Class of Service	Layer 3 IP Precedence	Layer 3 DSCP
Call signaling	3	3	CS3 (24)
Voice	5	5	EF (46)
Video	4	4	AF41 (34)
TelePresence	4	4	CS4 (32)

Voice calls have only one stream of packets. Video calls have two streams, one for video and another for voice, and it is important for both streams of the call to have the same QoS marking.

Cisco Unified Communications Manager (Unified CM) supports endpoints that mark QoS for their media packets. Voice packets are marked as EF (DSCP value 46), while video devices mark media packets as AF41 (DSCP value 34) and TelePresence endpoints mark their traffic as CS4 (DSCP value 32). All call signaling is marked as CS3 (DSCP value 24).

QoS should be configured on the Cisco TelePresence Video Communication Server (VCS) because it processes the media and the call signaling. The endpoints that register to the VCS (such as the Cisco TelePresence System EX Series, C Series, Cisco IP Video Phone E20, or others) should be configured so that the call signaling uses DSCP CS3 and the media from those endpoints is marked as DSCP AF41.

Trust Boundary

Traffic on the network needs to be marked so that the network can trust it. The network elements such as the switches can be the trust boundary based on the switch that trusts the packets. It is important to establish a trust boundary so that the rest of the network does not have to remark packets for QoS. Access switches can trust IP phones based on association with them. Cisco switches use Cisco Discovery Protocol (CDP), a Layer 2 protocol, to associate phones with the switches. The switches use CDP to put IP phones in their respective voice VLANs. Access switches can then trust such phones for marking their packets with appropriate QoS and thus establish a trust boundary. When IP phones cannot be associated with the switch or when the trust boundary needs to be the access or distribution switches, the switch can build the trust boundary by enforcing the marking of packets based on criteria such as the IP addresses of the devices or the common ports used for signaling or media for calls.

Packet Queuing

While QoS helps the network distinguish different types of traffic and then prioritize it, the network uses a queuing mechanism to orderly move packets and control their flow. Queuing is widely used for networks that have low-capacity links between them, such as MAN or WAN networks.

Networks use the following common queuing mechanisms:

- First In, First Out (FIFO)

  This type of queuing is simple and gives the same preference to all packets based on the time they arrive in the queue. The packets are sent out through the switch or the network in the same order they arrive. This type of queuing is useful in networks that do not see a large change in the volume of packets.

- Priority Queuing (PQ)

  This type of queuing gives preference to packets with higher priority over packets with lower priority. The Priority Queue is commonly used for low-bandwidth traffic that is very sensitive to delay, such as voice calls.

- Weighted Fair Queuing (WFQ)

  This type of queuing uses multiple queues for priority and non-priority traffic. It provides for priority traffic without starving lower-priority traffic. This mechanism is used where networks have traffic that needs priority (such as voice traffic) as well as other important traffic such as business applications that should not be dropped.

- Class-Based Weighted Fair Queuing (CBWFQ)

  This type of queuing uses different classes to group traffic and then uses the WFQ mechanism while also providing dedicated bandwidth for some custom queues. This type of mechanism is used widely for interactive voice and video traffic when combined with business applications. This mechanism provides the advantage of flexibility for various types of enterprise deployments.

Queuing mechanisms provide Class of Service (CoS) for packets in the organization. The class of service is used to guarantee latency, jitter, and delay requirements. It also uses the link bandwidth more efficiently so that organizations can estimate the traffic they can send through their WAN links or plan their links to support desired traffic.

Policing traffic is important to prevent certain queues from using all the bandwidth. Policing prevents a certain type of traffic from exceeding its set use limit through a link. Traffic shaping and policing is needed to avoid packet drops and to allow servicing of non-critical traffic.

With video calls it is important for both the video and voice streams of the call to be sent through the same queue in order to avoid lip-sync issues. When the video and voice streams use different queues, one arrives later than the other and causes the video to be shown while the voice associated with that video may lag, or vice versa.

Call Admission Control

To ensure that the voice and video traffic does not use all the bandwidth in the link and cause other important data such as business applications to experience dropped packets, organizations can use calls admission control. Call admission control limits the number of calls allowed through a particular link between sites.

There are two main methods for limiting the number of calls on a link:

- Call counting — With this method, the call control agent counts the number of calls allowed between locations. Only calls of the same type are counted together, so a set voice codec and a set video codec make counting such calls easy.
- Bandwidth — This method is similar to call counting, but here the count consists of the amount of bandwidth used by the calls. The type of voice codec and the type of bandwidth for video calls are used to count the bandwidth.

Preserving the call quality is important. When calls traverse WAN links, oversubscribing the link can cause call quality to degrade. Call admission control is important because it can prevent calls from filling up the link. When routing calls, call control agents know if a call should be allowed or if the link cannot handle that call. This provides a consistent call experience to users.

Cisco Unified CM supports call admission control using the bandwidth method, so calls with different codec types and video bandwidth types can be supported on the enterprise network. Unified CM uses the mechanism of regions to specify per-call parameters for codec and video bandwidth. Unified CM also

uses the mechanism of locations to specify the bandwidth value used to limit voice and video calls for a particular site. Cisco TelePresence VCS uses similar mechanisms, where Links set the per-call bandwidth and Pipes set the bandwidth for calls to a site.

Call counting and bandwidth methods for call accounting use static configured values, but the actual call may use less or more bandwidth. To find the actual bandwidth used for calls, Resource Reservation Protocol (RSVP) is used. This protocol looks at the actual path used by the media to determine if sufficient bandwidth is available for a call. When a device in the path does not have the bandwidth to service the call, that condition gets reported through RSVP and the call might not be allowed.

Cisco Unified CM uses a separate media device called Cisco RSVP Agent to negotiate the network bandwidth, using RSVP on behalf of the video endpoints. This allows for a more accurate accounting of calls through it.

TelePresence calls do not use call admission control because the network is designed to allow the needed traffic so as to guarantee the user experience through such technologies. TelePresence traffic is marked differently than voice and video; thus, it can be queued separately and can provide low latency and delay, thereby preserving the user experience.

When there is more than one call control agent in the organization, each agent does its own call admission control. They work in parallel and do not know about calls through each other. If a call occurs between two different call control agents at the same site, both agents may count the bandwidth for that call even though it does not use any WAN bandwidth. Therefore, it is important to have only one call control agent doing call admission control for the devices in the organization.