

Cisco CCVP - Quality of Service

Mega Guide

Prepare With Confidence

This PrepLogic Mega Guide was written by certified subject matter experts and published authors to provide you accurate, in-depth exam coverage. All exam objectives are covered in detail, giving you the knowledge and confidence you need to pass your exam.

**PrepLogic***Be Prepared. Be Confident. Get Certified.*

Sean Wilkins - Author
Gene Bagwell - Technical Editor

Domain 1 - IP QoS Fundamentals

QoS Basics

The important thing about Quality of Service (QoS) is its purpose. Networks these days are getting bigger and bigger and a larger amount of companies are using the Internet in some way to transport their traffic. Many companies have grown to the point where the differentiation of traffic over their networks is vital in order for certain traffic to get where it is going in a timely manner. Within QoS there are a number of mechanism types which are used to remedy these problems. Within networks there are two main ways to mark traffic. Marking the priority of the traffic is important because it provides some information which is part of the traffic itself, which enables the networking equipment to look for the traffic, and allows the networking equipment to perform prioritizing actions on the important traffic. The second way is a way of notating congestion in a network. This type of marking is mainly used for congestion avoidance because the routers are given some indication of when congestion starts and, from this information, have the option to use preempt congestion by dropping packets based on a congestion threshold.

The routers themselves can be configured to make QoS decisions based on the information in the traffic but there is typically no end-to-end QoS mechanism. Cisco routers use a number of different mechanisms to provide QoS. These mechanisms include congestion management solutions, queuing mechanisms, congestion avoidance mechanisms, and traffic control mechanisms.

Bottom line is that QoS is the grouping of technologies which provides a way to specify different quality parameters to different types of traffic. QoS is needed in today's networks because more and more networks are becoming data and voice converged networks, meaning that data, voice and video flow over the same network. This of course makes data, voice and video traffic influential over each other. Some traffic, like file transfers or web pages, is not as reliant on specific timing as VoIP or Video over IP. If a packet of a file arrives 30 seconds late, the transfer is just slowed down a little. If a voice or video packet arrives late it becomes useless in the audio or video stream.

The Need for QoS

In order to deal with QoS issues, there have been a number of different types of mechanisms created to help implement QoS. These different mechanisms work in different ways in order to be used in conjunction with each other so end-to-end QoS is under the requirements needed for the application.

Video and Voice

With a voice call, one-way delay budgets have been established by standards. These budgets set a total amount of delay which is considered acceptable for different levels of voice service. These are shown in Table 1:

1-way Delay (in ms)	Description
0-150	ITU G.114 Acceptable Range
0-200	Cisco Acceptable Range
150-400	ITU G.114 Degraded Range
400+	ITU G.114 Unacceptable Range

Table 1 - Delay Budget

On top of the requirements for delay, a minimal amount of sustained bandwidth is needed in order for a voice or video call to go through and be of good quality. The amount of bandwidth needed and the quality needed depend on the codec selection. The following are the main voice codecs which are used today:

Voice Codec	Acronym	Name	Bit Rate
G.711	PCM	Pulse Code Modulation	64-kbps
G.722	SB-ADPCM	Sub-Band ADPCM	48, 56, 64-kbps
G.722.1	MLT	Modulated Lapped Transform	24 and 32-kbps
G.722.2	ACELP	Algebraic Code Excited Linear Prediction Coder	6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 or 23.85-kbps
G.723.1 (5.3-kbps)	ACELP	Algebraic Code Excited Linear Prediction Coder	5.3-kbps
G.723.1 (6.3-kbps)	MP-MLQ	Multi Pulse-Maximum Likelihood Quantization	6.3-kbps
G.726	ADPCM	Adaptive Differential Pulse Code Modulation	16, 24 and 32-kbps
G.728	LDCELP	Low Delay Code Excited Linear Prediction	16-kbps
G.729	CS-ACELP	Conjugate Structure Algebraic CELP	8-kbps
G.729A	CS-ACELP Annex A	Conjugate Structure Algebraic CELP Annex A	8-kbps

Table 2 - ITU Voice Codecs

Video is also being used more and more, and the quality of this depends on the same factors as with voice, the amount of delay and the amount of bandwidth. The following shows a list of the most commonly used video codecs today:

Video Codec	Name	Bit Rate
MPEG-1	Moving Picture Experts Group, Version 1	500 to 1500 kbps
MPEG-2	Moving Picture Experts Group, Version 2	1.5 to 10 Mbps
MPEG-4	Moving Picture Experts Group, Version 4	28.8 to 400 kbps
H.261	Video Coding Experts Group	100 to 400 kbps

Table 3 - Common Video Codecs

In order for the demands of these different types of traffic to be sustained, some amount of QoS must be implemented over the network to ensure that the requirements are met.

Both voice and video can also be split into two main categories, interactive and non-interactive. Interactive traffic is communication which goes in two directions in response to each other. Interactive traffic requirements are stricter because delay, jitter and delay requirements must be low to keep the conversation ongoing and flowing. This is not true of non-interactive traffic like audio or video presentations since these are typically in one direction; the delay, loss and delay are not noticed.

TCP Windowing

There is a congestion mechanism within TCP which is used to try to limit congestion on the network. The main controlling option within this mechanism is the use of *windowing*. Windowing works by allowing a TCP connection to send a specific amount of traffic without having to receive an acknowledgement. By default, the window size is set to 536 bytes. What this means is that 536 bytes of traffic is sent and an acknowledgement is expected from the destination. If an acknowledgement is received then the window size is increased by 2 times. If this traffic is sent and an acknowledgement is received, then the window size is increased by 3 times, and so on until the maximum window size is reached. If, at any point, an acknowledgement is not received, the window size is cut in half.

If a QoS mechanism is not put in place to cut down the chance of packet loss then the window size will stay low and make TCP connections very inefficient. The mechanism which would be used for this on Cisco equipment is Weighted Random Early Detection (WRED).

Traffic Characteristics

Bandwidth

Bandwidth is simply the amount of data that can be sent over a network at one time. Bandwidth is the easiest part of QoS to understand; to use more bandwidth than is available on the network, some or all of the traffic will be affected. When thinking of bandwidth in QoS terms, it is typically available bandwidth that needs to be considered. Available bandwidth is a measurement of the minimum bandwidth available on a path from point A to point B, divided by the number of potential traffic flows. This is shown in the following figure:



Figure 1 - Bandwidth Example

Using this figure, the amount of minimum bandwidth is 10-Mbps across the whole path. If ten flows are needed, the total available bandwidth per flow is 1-Mbps.

On Cisco equipment there are two different concepts which must be clearly defined. There are two commands, **bandwidth** and **clock-rate**, on Cisco equipment. The **bandwidth** command is used to set the amount of bandwidth that the equipment will use for calculations. These include dynamic routing protocols, load and statistics. By default, the bandwidth is set to 1.544 Mbps (or a T1 Rate). The **clock-rate** command is used to set the actual line rate of the interface. A good example of this is a frame relay link. It is possible to have a physical access rate of 1.544 Mbps with a bandwidth of 512 kbps. The 512 kbps from the provider equates to a Committed Information Rate (CIR) of 512 kbps.

Delay

Delay is a crucial part of QoS management. The amount of overall delay from end-to-end is very important when dealing with voice and video over networks. There are many different things that can affect the amount of delay that is introduced from one side of a path to the other. The nine main delay factors are: processing delay, queuing delay, serialization delay, propagation delay, shaping delay, network delay, codec delay, compression delay and jitter buffer delay.

Processing or forwarding delay is the amount of time it takes the layer 3 devices (router or switch) to transfer a packet in one interface and out another. Many different things affect this, including CPU speed, CPU utilization, total memory, available memory, and bus speed, among others.

Queuing delay is the amount of time a packet spends in the queue of a layer 3 device. Queues are used in equipment to store data when the bandwidth is currently completely utilized. This information is stored for a short time in a queue until the bandwidth opens up. The amount of time that the packet spends in these queues is the queuing delay.

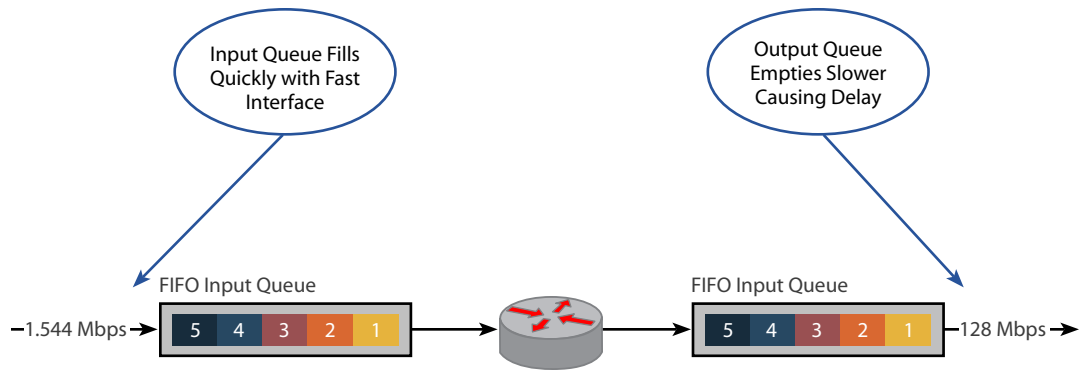


Figure 2 - Queuing Delay Example

Serialization delay is the amount of time it takes for a packet to be broken down into layer 2 frames, then into layer 1 electric or optical signals. The amount of time taken for a packet to be broken down for a specific link can be calculated with the formula from Figure 3.

$$\frac{\text{\# of bits sent}}{\text{Link Speed}}$$

Figure 3 - Serialization Delay Formula

Propagation delay is the amount of time it takes for a packet to cross the physical medium. The amount of time that is taken for a packet to be transmitted over media can be roughly calculated with the formula in Figure 4.

$$\frac{\text{Length of Link (meters)}}{2.1 \times 10^8 \text{ meters per second}}$$

Figure 4 - Speed of Light over Copper/Fiber

Shaping delay is the delay that is introduced when a piece of equipment shapes the flow of traffic. As an example, think of what happens when traffic comes into a fast interface and has to go out a slow interface. Without shaping, there is a high potential that some of this traffic will be queuing delayed then tail dropped, meaning that the traffic would fill the queue and then overflow and drop future packets. In order to try to stop this, traffic shaping can be configured to average out the traffic flow. In this case, traffic shaping would try to queue the faster traffic and slow down the flow until incoming traffic slows. This slowing in the traffic introduces shaping delay. Assuming adequate queue depth, this is quite useful because it tries to allow the largest amount of packets to reach their destination.

Network delay is the delay that is introduced by the cloud. The cloud is typically a part of the network which is not under direct control, so trying to find the exact delay can be challenging. Typically this number is given as a maximum delay that is given as part of the network contract.