Cisco Systems

CCVP QOS Quick Reference Sheets

Kevin Wallace

shortcut

Your Short Cut to Knowledge

ciscopress.com

Why You Need Quality of Service (QoS)

The networks of yesteryear physically separated voice, video, and data traffic. Literally, these traffic types flowed over separate media (for example, leased lines or fiber-optic cable plants). Today, however, network designers are leveraging the power of the data network to transmit voice and video, thus achieving significant cost savings by reducing equipment, maintenance, and even staffing costs.

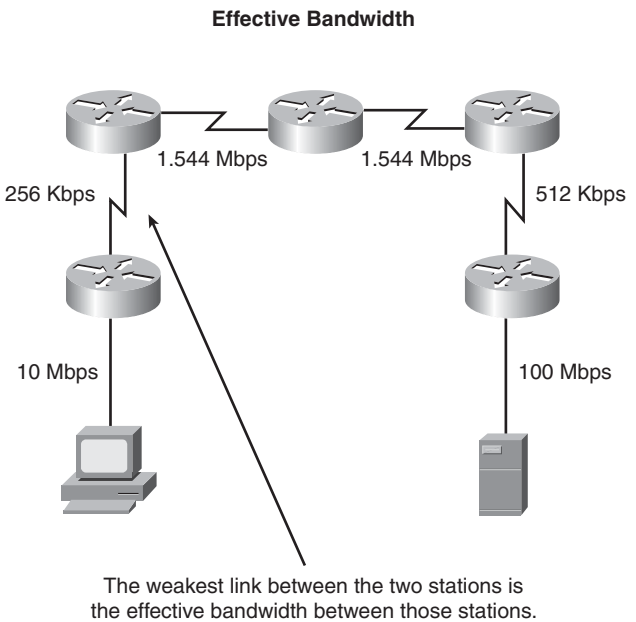
The challenge, however, with today's converged networks is that multiple applications are contending for bandwidth, and some applications such as, voice can be more intolerant of delay (that is, latency) than other applications such as, an FTP file transfer. A lack of bandwidth is the overshadowing issue for most quality problems.

When a lack of bandwidth exists, packets can suffer from one or more of the following symptoms:

- **Delay**—Delay is the time that is required for a packet to travel from its source to its destination. You might witness delay on the evening news, when the news anchor is talking through satellite to a foreign news correspondent. Because of the satellite delay, the conversation begins to feel unnatural.
- **Jitter**—Jitter is the uneven arrival of packets. For example, consider that in a Voice over IP (VoIP) conversation, packet 1 arrives. Then, 20 ms later, packet 2 arrives. After another 70 ms, packet 3 arrives, and then packet 4 arrives 20 ms behind packet 3. This variation in arrival times (that is, variable delay) is not dropping packets, but this jitter can be interpreted by the listener as dropped packets.
- **Drops**—Packet drops occur when a link is congested and a buffer overflows. Some types of traffic, such as User Datagram Protocol (UDP) traffic (for example, voice), are not retransmitted if packets are dropped.

Fortunately, quality of service (QoS) features that are available on Cisco routers and switches can recognize your "important" traffic and then treat that traffic in a special way. For example, you might want to allocate 128 kbps of bandwidth for your VoIP traffic and also give that traffic priority treatment.

Consider water that is flowing through a series of pipes with varying diameters. The water's flow rate through those pipes is limited to the water's flow rate through the pipe with the smallest diameter. Similarly, as a packet travels from its source to its destination, its effective bandwidth is the bandwidth of the slowest link along that path.



Because your primary challenge is a lack of bandwidth, the logical question is, "How do you increase available bandwidth?" A knee-jerk response to that question is often, "Add more bandwidth." Although adding more bandwidth is the best solution, it comes at a relatively high cost.

Compare your network to a highway system in a large city. During rush hour, the lanes of the highway are congested, but the lanes can be underutilized during other periods of the day. Instead of just building more lanes to accommodate peak traffic rates, the highway engineers add carpool lanes. Cars with two or more riders can use the reserved carpool lane. These cars have a higher priority on the highway. Similarly, you can use QoS features to give your mission-critical applications higher-priority treatment in times of network congestion.

Some of the QoS features that can address issues of delay, jitter, and packet loss include the following:

- **Queuing**—Queuing can send higher-priority traffic ahead of lower-priority traffic and make specific amounts of bandwidth available for those traffic types. Examples of queuing strategies that you consider later in these Quick Reference Sheets include the following:
  - Priority Queuing (PQ)
  - Custom Queuing (CQ)
  - Modified Deficit Round Robin (MDRR) queuing
  - Weighted Fair Queuing (WFQ)
  - Class-Based WFQ (CB-WFQ)
  - Low Latency Queuing (LLQ)
- **Compression**—By compressing a packet's header or payload, fewer bits are sent across the link. This effectively gives you more bandwidth.

QoS Basics

The mission statement of QoS could read something like "to categorize traffic and apply a policy to those traffic categories, in accordance with a QoS policy." Specifically, QoS configuration involves the following three basic steps:

1. Determine network performance requirements for various traffic types. For example, consider the following design rules of thumb for voice, video, and data traffic:

   Voice:

   - No more than 150 ms of one-way delay
   - No more than 30 ms of jitter
   - No more than 1 percent packet loss

   Video:

   - No more than 150 ms of one-way delay for interactive voice applications (for example, videoconferencing)
   - No more than 30 ms of jitter
   - No more than 1 percent packet loss

   Data:

   Applications have varying delay and loss characteristics. Therefore, data applications should be categorized into predefined "classes" of traffic, where each class is configured with specific delay and loss characteristics.

2. Categorize traffic into specific categories. For example, you can have a category named "Low Delay," and you decide to place voice and video packets in that category. You can also have a "Low Priority" class, where you place traffic such as music downloads from the Internet. As a rule of thumb, Cisco recommends that you create no more that ten classes of traffic.

3. Document your QoS policy and make it available to your users. Then, for example, if a user complains that his network gaming applications are running slowly, you can point him to your corporate QoS policy, which describes how applications such as network gaming have "best-effort" treatment.

QoS Deployment

Cisco offers the following four basic approaches for QoS deployment in your network:

- **Command-Line Interface (CLI)**—The CLI is the standard IOS (or Cat OS) interface that configures routers or switches. CLI QoS features such as Priority Queuing (PQ) or Custom Queuing (CQ), which are configured through the CLI, have been available for many years.
- **Modular QoS CLI (MQC)**—Instead of using the CLI to configure QoS parameters for one interface at a time, the three-step MQC process allows you to (1) place packets into different classes, (2) assign a policy for those classes, and (3) apply the policy to an interface. Because the approach is modular, you can apply a single policy to multiple interfaces.
- **AutoQoS**—AutoQoS is a script that is executed on routers or switches that automates the process of QoS configuration. Specifically, this automatic configuration helps optimize QoS performance for VoIP traffic.
- **QoS Policy Manager (QPM)**—QPM, in conjunction with CiscoWorks, centralizes QoS configuration. Policies that are created with QPM can be pushed out to routers throughout an enterprise, thus reducing the potential for misconfiguration.

QoS Components

Cisco offers a wealth of QoS resources on its switch and router platforms. These resources are classified into one of three categories, which are discussed in this section. The category of QoS resources used most often in production, however, is the Differentiated Services category, which offers greater scalability and flexibility than the resources found in the Best-Effort or Integrated Services categories.

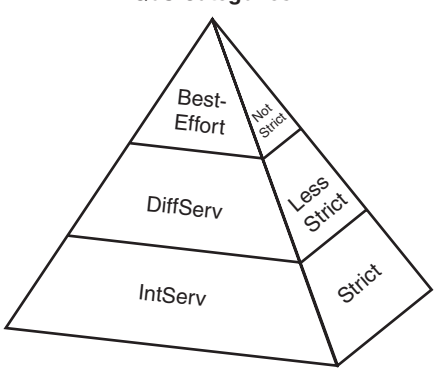
QoS Categories

All of the Cisco QoS features are categorized into one of the following three categories:

- **Best-Effort**—Best-Effort does not truly provide QoS, because there is no reordering of packets. Best-Effort uses the first-in first-out (FIFO) queuing strategy, where packets are emptied from a queue in the same order in which they entered it.
- **Integrated Services (IntServ)**—IntServ is often referred to as "Hard QoS" because it can make strict bandwidth reservations. IntServ uses signaling among network devices to provide bandwidth reservations. Resource Reservation Protocol (RSVP) is an example of an IntServ approach to QoS. Because IntServ must be configured on every router along a packet's path, the main drawback of IntServ is its lack of scalability.

- **Differentiated Services (DiffServ)**—DiffServ, as the name suggests, differentiates between multiple traffic flows. Specifically, packets are "marked," and routers and switches can then make decisions (for example, dropping or forwarding decisions) based on those markings. Because DiffServ does not make an explicit reservation, it is often called ***Soft QoS***. The focus of these Quick Reference Sheets is DiffServ, as opposed to IntServ or Best-Effort.

QoS Categories

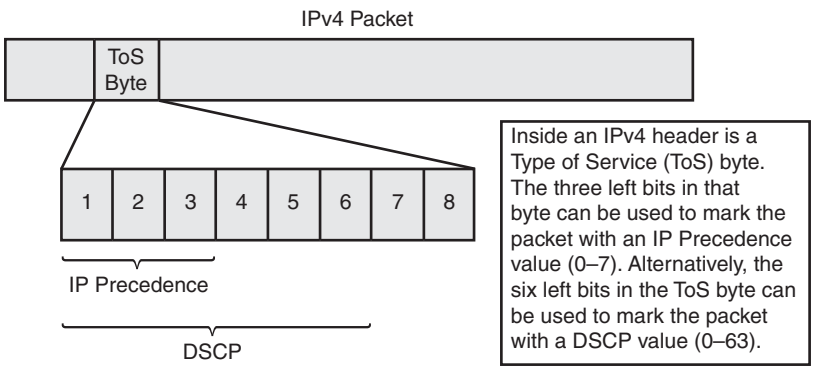


- Best-Effort does not perform reordering of packets.
- DiffServ differentiates between flows and assigns policies to those flows.
- IntServ makes a strict bandwidth reservation for an application.

DiffServ

Now that you understand the importance that marking plays in a DiffServ QoS solution, you can learn how packets can be marked. Inside an IPv4 header is a byte called the ***type of service (ToS) byte***. You can mark packets, using bits within the ToS byte, with either IP Precedence or Differentiated Service Code Point (DSCP) markings.

Type of Service (ToS) Byte



IP Precedence uses the 3 leftmost bits in the ToS byte. With 3 bits at its disposal, IP Precedence markings can range from 0 to 7. However, values 6 and 7 should not be used, because those values are reserved for network use.

For more granularity, you can choose DSCP, which uses the 6 leftmost bits in the ToS byte. Six bits yield 64 possible values (0 to 63). The challenge with so many values at your disposal is that the value you choose to represent a certain level of priority can be treated differently by a router or switch under someone else's administration.

To maintain relative levels of priority among devices, the Internet Engineering Task Force (IETF) selected a subset of those 64 values for use. These values are called ***per-hop behaviors (PHBs)*** because they indicate how packets should be treated by each router hop along the path from the source to the destination.