Chapter 3

QUALITY OF SERVICE

3.1 Introduction

QoS can be defined as the ability of the network to support good services in order to accept good customers. In other words, QoS measures the degree of user satisfactions and network performance. Applications like FTP, HTTP, video conferencing and e-mail are not sensitive to delay of transmitted information, while other applications like voice and video are more sensitive to loss, delay and jitter of the information.

It is a very vital factor to allocate the resources for improving the QoS (Quality of service) for any network carrying various types of traffic. Real-time applications, such as video conferences, are in the most important to get the benefit of QoS adaptation. Several scheduling disciplines are employed at the router to guarantee the QoS of the network. Each scheduling discipline has its advantages and disadvantages[2].Applications differ in the way they use bandwidth and in their QoS requirements. The unpredictable mix of applications running on a network and the conflicts that occur when several user try to run their applications at once causes QoS problems. Dealing with these conflicts is the key challenge of QoS management[18].

3.2 Definition of QoS

QoS is the ability of a networking equipment to differentiate among different classes of traffic and to give each class different priority over the network when there is congestion in the network based on the traffic significance. QoS is not something that will be configured on a router or switch; rather it is a term that refers to a wide variety of mechanism used to influence traffic patterns on a network. It gives network administrators the ability to give some traffic more priority over others[17].

The Internet was traditionally built to carry traffic on a best-effort service model. In a best-effort service model the networks and the underlying network elements and protocols transported the users' traffic from the source to the destination but did not provide the user with any guarantee of packet delivery. In the event of congestion introduced due to the lack of bandwidth the Internet would drop the packets. Additionally the Internet would not provide any guarantees with respect to the time in which the datais transported. This default behavior of the Internet is not suitable for real time traffic such as voice and multimedia traffic. A certain assurance needs to be provided to the user in case of a network congestion or delay. QoS mechanisms allow the application or the user to request for a certain guarantee thereby classifying the packets of this transmission into a separate class or flow. QoS mechanisms also allow the network administrator to administer and control the resources necessary for the successful transmission of these packets. Finally, QoS mechanisms allow the intervening routers to process this request by reserving network resources such as bandwidth and controlling delay and packet loss [3].

3.3 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:

3.3.1 Determine network performance

There are four service characteristics that are commonly used as QoS metrics – bandwidth requirements, latency, jitter, and packet loss. In addition, resource availability must be considered with regards to tiered QoS.

3.3.1.1 Bandwidth

Application bandwidth requirements can be categorized as sustained, bursty, and interactive. Streaming media applications, such as Microsoft's NetShow, require a sustained amount of bandwidth in order to provide users with high quality audio and video information. Some are designed to run over low-speed dial-up connections and require no more than 56 kbps while others, such as high qualityMoving Picture Experts Group-2(MPEG-2) video, may require up to 10 Mbps. These applications can become virtually unusable if the required bandwidth is not available – even for very short periods of time.

Other applications, such as file transfers, are bursty. These applications attempt to grab as much bandwidth as they can to speed data delivery. This bursty traffic must be controlled because it is the most common cause of network congestion that adversely affects the performance of other applications. These bursty traffic characteristics are often the result of the end-to-end TCP protocols used by many applications[18].

3.3.1.2 Latency (delay types)

Packet delay is the time it takes a packet to reach the receiving end of an endpoint after it has been transmitted from the source. This is called end-to-end delay. It consists of two components: fixed network delay and variable delay. Packet delay can cause degradation of voice quality due to the end-to-end voice latency or packet loss if the delay is variable. Queuing, waiting for packets being transmitted and serialization are the main causes of delay and are the ones we can do something about while the other causes of delay, like jitter buffers, codecs, are causes we can do nothing about [17].

Some applications are sensitive to the latency, or delay, in transmitting data across a network. End-to-end latency is due to the latency of physical transmission media and delays introduced by intermediate routers and switches. Significant delays can be introduced when packets are queued for long periods of time. Some queuing mechanisms are designed to control these delays while others can magnify the problem. Real-time audio and video applications, including voice-over IP, fall into this category. Latency also degrades the response times of interactive applications[18].

Fixed-network delay: Includes encoding and decoding time and the latency required for the electrical and optical signals to travel the media en route to the receiver. Generally, applying

QoS does not affect fixed-network delay because fixed-network delay is a property of the medium. Upgrading to higher-speed media such as 10 Gigabit Ethernet and newer network hardware with lower encoding and decoding delays, depending on application, may result in lower fixed-network delay.

In brief, the following list details the types of delay that induce end-to-end latency:

- **Packetization delay:** Amount of time that it takes to segment, sample, and encode signals, process data, and turn the data into packets.
- **Serialization delay**: It is the time required to place the packet on the transmission line, this delay depends on the line speed, and the packet size.
- **Propagation delay**: the time taken by the packet to reach the receiver, this delay depends also on the transmission line speed.
- **Processing delay**: Amount of time it takes for a network device to take the frame from an input interface, place it into a receive queue, and place it into the output queue of the output interface.
- **Queuing delay:** It is the delay caused at different switching and transmission points in the network, such as router, when packets have to wait behind other packets waiting to be transmitted over the same link. Because the number of packets waiting to be served in a queue depends on the statistical arrival process, queuing delay is of statistical nature and it can vary greatly from one packet to another.

3.3.1.3 Jitter

Jitter is related to latency because it refers to the time variability of delay the data experiences in networks. Variable queuing delays in routers and switches can cause jitter, and some queuing techniques differ in the amount of jitter that they introduce. Excessive jitter can disrupt real-time video, audio, and voice over IP traffic flows [18].

3.3.1.4 Packet loss

Packet loss is typically caused by network congestion and it affects all applications because packet retransmission reduces the overall efficiency of networks and, therefore, the amount of bandwidth available to applications. The impact of packet loss differs from application to application. Some multimedia applications can become unusable when packet loss occurs while most business applications simply experience degraded performance[18].

Application Types	QoS Requirements			
	Bandwidth	Latency	Jitter	Packet Loss
E-Mail	Low to Moderate	-	-	-
File Transfer	Bursty High	-	-	-
Thin Clients (Citrix, etc.)	Low to Moderate	Low	-	Low
Videoconferencing	Sustained High	Low	Low	Low
Voice over IP	Sustained Moderate	Low	Low	Low
Streaming Media	Sustained Moderate to High	Low	Low	Low
Server Load Balancing	QoS requirements are application and server dependent			

Figure 3.1 QoS Requirements[18].

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, video conferencing).
- No more than 30 ms of jitter.
- No more than 1 percent packet loss.

Data:

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

3.3.2 Categorize traffic

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 than ten classes of traffic.

3.3.3Document your QoS policy

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. A "Low Priority" class,is where to place traffic such as music downloads from the Internet.

3.4 QoS Components

The purpose of QoS usage is to make sure that minimum bandwidth is guaranteed for identified traffic, that jitter and latency are being controlled and that packet loss is improved. This can be carried out using several tools either QoS congestion management, congestion avoidance or policing and traffic shaping. Tools chosen depend largely on the goal of the network administrator[17].

QoS can be divided into three different models. These models describe a set of end-to-end QoS capabilities. In order to facilitate end-to-end QoS on an IP network, the Internet Engineering Task Force (IETF) defined two models: Integrated Services (IntServ) and Differentiated Services (DiffServ). A default model that comes with all networking devices is the Best Effort model. It does not require any QoS configuration. The IntServ model follows the end-to-end signaling process whereby the end-hosts tells the network their QoS needs in advance while DiffServ follows the provisioned-QoS model whereby the network elements set up multiple classes of traffic with different degrees of QoS needs[17].

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[13].Figure 3.2.

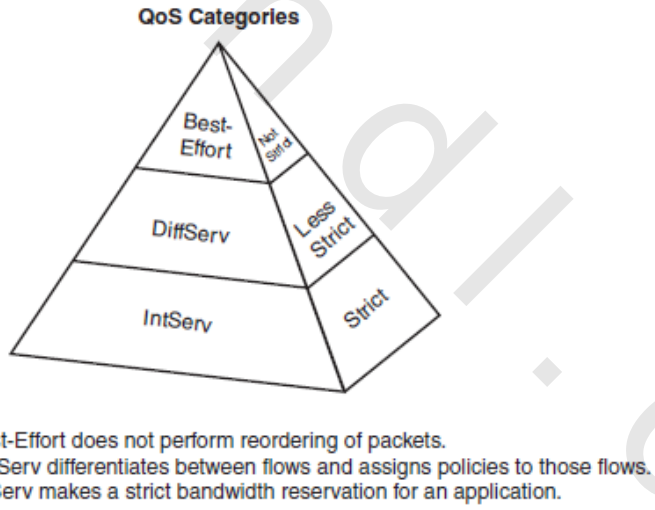


Figure 3.2 QoS Categories[13].

3.4.1 Best-Effort service

Best effort is a single service model in which an application sends data whenever it must, in any quantity, and without requesting permission or first informing the network. For best-effort service, the network delivers data if it can, without any assurance of reliability, delay bounds, or throughput.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. Best-effort service is suitable for a wide range of networked applications such as general file transfers or e-mail[13].

3.4.2 Integrated services (IntServ)

IntServ provides high QoS for IP packets but it requires special QoS to be made available by the network for a period and that bandwidth should be reserved. This method uses a protocol to reserve bandwidth on a per flow basis. This protocol is called the Resource Reservation Protocol (RSVP). RSVP is a signaling mechanism that is used by IntServ architecture to carry out its function[17].

IntServ is often referred to as “Hard QoS” because it can make strict bandwidth reservations.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[13].

IntServ represents the application-signaled approach of the QoS architecture. The application frames its service request within the RSVP protocol and then passes this request into the network. The network can either respond positively in terms of its agreement to commit to this service profile, or it can reject the request. If the network commits to the request with a resource reservation, the application can send traffic into the network. The reservation remains in force until the application explicitly requests termination of the reservation, or the network signals to the application that it is unable to continue with a service commitment. The essential feature of the IntServ model is "all or nothing" nature. In other words, the IntServ framework either provides all the QoS guarantees or replies negatively.
The IntServ QoS model requires that the network must maintain the remembered state to describe the resources that have been reserved and the network path over which the reserved service will operate. In addition, each active network element within the network must maintain a local state that allows incoming IP packets to be correctly classified into a reservation class[15].The application is expected to send data only after it gets a confirmation from the network. It is also expected to send data that lies within its described traffic profile.

The network performs admission control on the basis of information from the application and available network resources. It also commits to meeting the QoS requirements of the application as long as the traffic remains within the profile specifications. The network fulfills its commitment by maintaining per-flow state and then performing packet classification, policing, and intelligent queuing based on that state.

Cisco IOS QoS includes the following features that provide controlled load service, which is a kind of integrated service:

- The Resource Reservation Protocol (RSVP), which can be used by applications to signal their QoS requirements to the router.
- Intelligent queuing mechanisms, which can be used with RSVP to provide the following kinds of services:

- Guaranteed rate service, which allows applications to reserve bandwidth to meet their requirements. For example, VoIP application can reserve the required amount of bandwidth end-to-end using this kind of service. Cisco IOS QoS uses weighted fair queuing (WFQ) with RSVP to provide this kind of service.
- Controlled load service, which allows applications to have low delay and high throughput even during times of congestion. For example, adaptive real-time applications, such as playback of a recorded conference, can use this kind of service. Cisco IOS QoS uses RSVP with Weighted Random Early Detection (WRED) to provide this kind of service[16].

3.4.3 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.

Though classification is carried in the IP header, it can also be carried in the Layer 2 frame. These special bits in the Layer 2 frame or Layer 3 packet are described below.

- Prioritization bits in Layer 2 frames: Layer 2 Inter-Switch Link (ISL) frame headers carries the IEEE 802.1p class of service(CoS) in the 1-byte User field while the Layer 2 802.1Q frame headers carries the CoS value in a 2-byte Tag Control Information field. The Tag Control Information (TAG) field carries the CoS value in the 3 most-important bits, which is normally called User Priority bits in a Layer 2 802.1Q header frame. Layer 2 CoS values range from 0 for low priority to 7 for high priority. [17]. Figure 3.3.

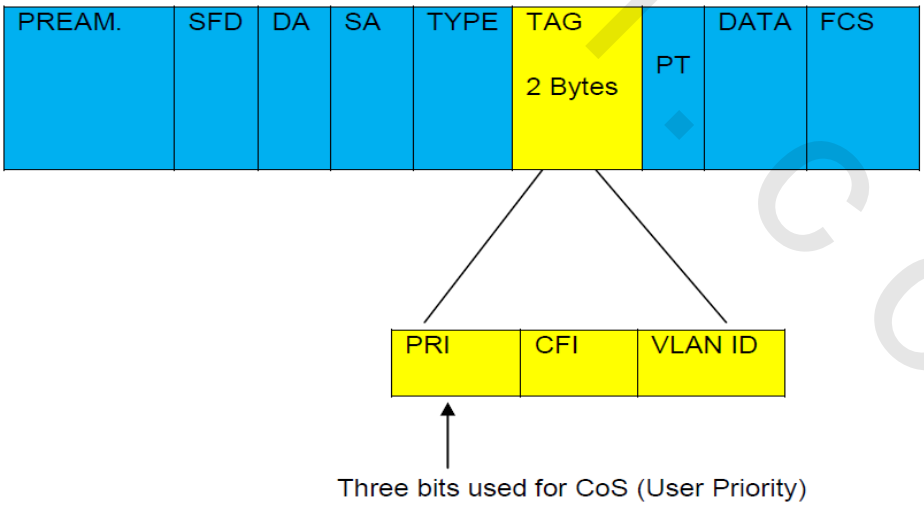


Figure 3.3 Overview of the TAG and User Priority in a Layer 2 802.1Q frame[17].

- Prioritization bits in Layer 3 packets: Layer 3 IP packets can carry either (DSCP) value or IP precedence value. This is possible because DSCP is backward compatible with IP precedence value. IP precedence values range from 0 to 7 while DSCP values range from 0 to 63. However, values 6 and 7 should not be used, because those values are reserved for network use [17]. Figure 3.4.

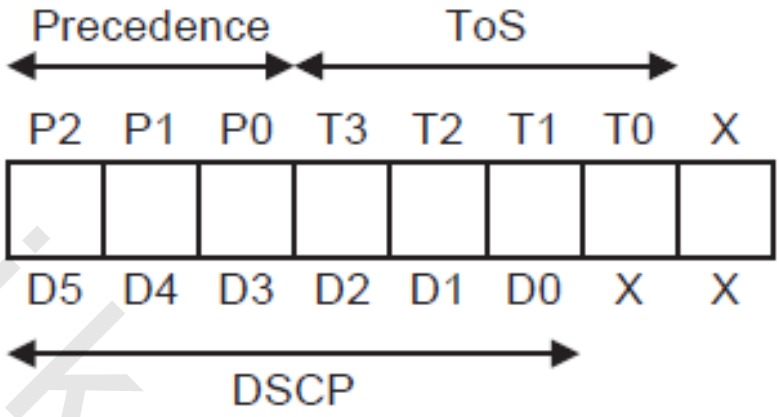


Figure 3.4 Structure of DSCP[13].

DiffServ provides the need for simple and coarse methods of putting traffic into classes, called Class of Service (CoS). It does not specify that a specific protocol should be used for providing QoS but specifies an architectural framework for carrying out its function. DiffServ carries out its major function through a small, well-defined set of building blocks from which different aggregates of behaviors can be built. The packet Type of Service (ToS) byte in the IP header is marked in order for the packets to be divided into different classes which forms the aggregate behaviors.

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 under someone else's administration.

To maintain relative levels of priority among devices, 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.

The DiffServframework has a different approach comparing to the IntServ. Instead of maintaining the per-flow information in each router, flows with similar characteristics are grouped into a class, which is referred to as an aggregate. As a packet enters the DiffServ domain, the appropriate DiffServ code point (DSCP) is written into the IP header, which determines an aggregate a packet belongs to. Now, routers classify and schedule packets based on their DSCP values, not on their source/destination IP addresses, protocol, and the port numbers.Figure 3.5.

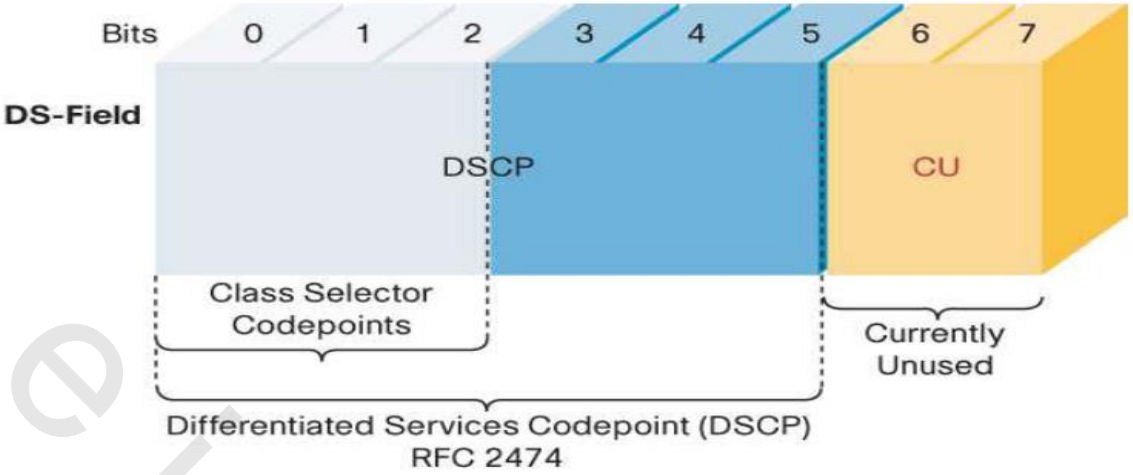
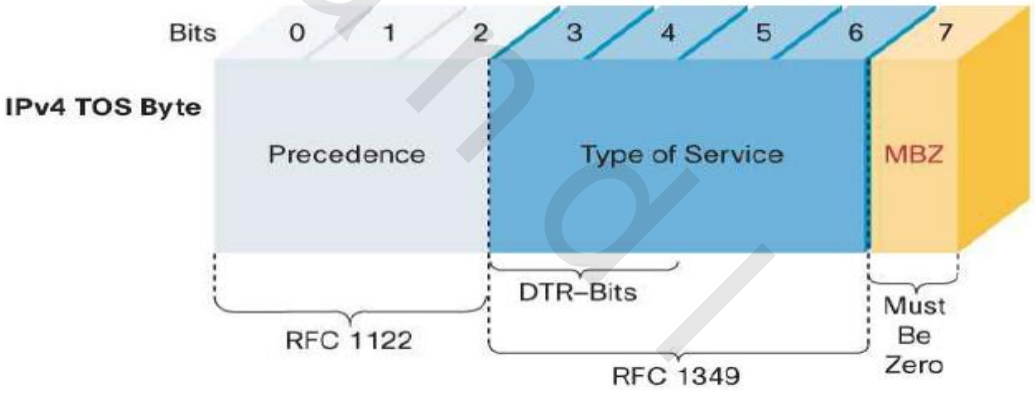


Figure 3.5 DS field as is presently used today in the IPv4 header[17].

The ToS field was renamed to the Differentiated Services field. 6-bits of the DS field is used as the code points for selecting the per-hop behavior(PHB) while the last 2-bits are unused just like the original ToS field[17].Figure 3.6.



Bits (0–2): IP-Precedence Defined
111 – Network Control
110 – Internetwork Control
101 – CRITIC/ECP
100 – Flash Override
011 – Flash
101 – Immediate
001 – Priority
000 – Routine

Bits (3–6): The Type of Service Defined
0000 – [all normal]
1000 – [minimize delay]
0100 – [maximize throughput]
0010 – [maximize reliability]
0001 – [minimize monetary cost]

Figure 3.6 ToS field before it was renamed to be the DS field[17].

The DSCP field occupies six bits and overlap with the ToS octet of the IPv4 packets. Figure 3.5 illustrates the structure of the IPv4 ToS octet and its interpretation for the DiffServ[15]. The remaining two bits are reserved for the Explicit Congestion Notification (ECN) technology. It bears mention that the DSCP value is backward compatible with the

IPv4 precedence field. As follows from Figure 3.6, three higher bits of DSCP overlap with the IP precedence bits. It must be noted that the higher bits of DSCP determine a class, while the three lower bits determine a subclass, if any.Thus, packets marked according to the DSCP value will be given the sufficient level of QoS if a provider has the old switching equipment that relies upon the IP precedence value. Consequently, packets marked according to the IP precedence field will be forwarded appropriately within the DiffServ domain. The DiffServ model allows network traffic to be broken down into small flows for appropriate marking. This small flow is called a class. Thus, the network recognizes traffic as a class instead of the network receiving specific QoS request from an application. The devices along the path of the flow are able to recognize the flow because these flows are marked. So the marked flow is given appropriate treatment by the various devices on the network.

When the packet has been properly marked and identified by the router or switch, it is given special treatment called Per Hop Behavior (PHB) by these devices. PHB identifies how the packets are treated at each hop. There are three standardized PHB currently in use: [17]

- **Default**
  - Traffic that only needs best-effort treatment can be marked with the Default PHB, which simply means that the 6 leftmost bits in the packet's ToS byte (that is, the DSCP bits) are all 0 (that is, a DSCP value of 0).
- **Expedited Forwarding (EF)**
  - The EF PHB has a DSCP value of 46. Latency-sensitive traffic, such as voice, typically has a PHB of EF.
- **Assured Forwarding (AF)**
  - The broadest category of PHBs is the AF PHB. Specifically, 12 AF PHBs exist, as shown in Table 3.1[13].

Table 3.1 Assured Forwarding (AF).

PHB	Low Drop Preference	Medium Drop Preference	High Drop Preference
Class 1	AF 11 (10) 001010	AF12 (12) 001100	AF13 (14) 001110
Class 2	AF21 (18) 010010	AF22 (20) 010100	AF23 (22) 010110
Class 3	AF31 (26) 011010	AF32 (28) 011100	AF33 (30) 011110
Class 4	AF41 (34) 100010	AF42 (36) 100100	AF43 (38) 100110

Notice that the Assured Forwarding PHBs are grouped into four classes. Examining these DSCP values in binary reveals that the 3 leftmost bits of all the Class 1 AF PHBs are 001 (that is, a decimal value of 1); the 3 leftmost bits of all the Class 2 AF PHBs are 010 (that is, a decimal value of 2); the 3 leftmost bits of all the Class 3 AF PHBs are 011 (that is, a decimal value of 3); and the 3 leftmost bits of all the Class 4 AF PHBs are 100 (that is, a decimal value of 4). Because IP Precedence examines these 3 leftmost bits, all Class 1 DSCP

values would be interpreted by an IP Precedence–aware router as an IP Precedence value of 1. The same applies to the Class 2, 3, and 4 PHB values.Table 3.1.

Within each AF PHB class are three distinct values, which indicate a packet's "drop preference." Higher values in an AF PHB class are more likely to be discarded during periods of congestion. For example, an AF13 packet is more likely to be discarded than an AF11 packet.

3.5 QoS Tools

Now that you understand how markings can be performed with the DiffServ QoS model, realize that marking alone does not alter the behavior of packets. You must have a QoS tool that references those marking and alters the packets' treatment based on those markings.

3.5.1 Classification

Classification is the process of placing traffic into different categories. Multiple characteristics can be used for classification. For example, Post Office Protocol (POP3), Internet Message Access Protocol (IMAP), Simple Mail Transfer Protocol(SMTP), and Exchange traffic could all be placed in an "E-MAIL" class. Classification does not, however, alter bits in the frame or packet.

3.5.2 Marking

Marking alters bits (for example, bits in the ToS byte) within a frame, cell, or packet to indicate how the network should treat that traffic. Marking alone does not change how the network treats a packet. Other tools (for example, queuing tools) can, however, reference those markings and make decisions based on them.

3.5.3 Congestion management

When you hear the term congestion management, think queuing. These concepts are the same. When an interface's output software queue contains packets, the interface's queuing strategy determines how the packets are emptied from the queue. For example, some traffic types can be given priority treatment, and bandwidth amounts can be made available for specific classes of traffic.

3.5.4 Congestion avoidance

If an interface's output queue fills to capacity, newly arriving packets are discarded (that is, "tail-dropped"), regardless of the priority that is assigned to the discarded packet. To prevent this behavior, Cisco uses a congestion avoidance technique called Weighted Random Early Detection (WRED). After the queue depth reaches a configurable level (that is, the minimum threshold) for a particular priority marking (for example, IP Precedence or DSCP), WRED introduces the possibility of discard for packets with those markings. As the queue depth continues to increase, the possibility of discard increases until a configurable maximum threshold is reached. After the queue depth has exceeded the maximum threshold for traffic with a specific priority, there is a 100 percent chance of discard for those traffic types[13].

3.6 Basic QoS Configuration

Cisco continues to improve the ease and efficiency with which QoS mechanisms can be configured. This section addresses two of the Cisco more recent developments: Modular QoS Command Line Interface (MQC) and Automatic Quality of ServiceAutoQoS.

3.6.1 Using MQC

One of the most powerful approaches to QoS configuration is the Modular Quality of Service Command-Line Interface (MQC). After you master the three basic steps of MQC, you can use them to configure a wide range of QoS tools, including queuing, policing, shaping, header compression, WRED, and marking.

Following is an MQC example in which you are classifying various types of e-mail traffic (for example, SMTP, IMAP, and POP3) into one class-map. The KaZaa protocol, which is used frequently for music downloads, is placed in another class-map. VoIP traffic is placed in yet another class-map. Then, the policy-map assigns bandwidth allocations or limitations to these traffic types. The MQC example is as follows:

```
Router(config)#class-map match-any EMAIL
Router(config-cmap)#match protocol pop3
Router(config-cmap)#match protocol imap
Router(config-cmap)#match protocol smtp
Router(config-cmap)#exit

Router(config)#class-map MUSIC
Router(config-cmap)#match protocol kazaa2
Router(config-cmap)#exit
Router(config)#class-map VOICE
Router(config-cmap)#match protocol rtp
Router(config-cmap)#exit

Router(config)#policy-map QOS-STUDY
Router(config-pmap)#class EMAIL
Router(config-pmap-c)#bandwidth 128
Router(config-pmap-c)#exit

Router(config-pmap)#class MUSIC
Router(config-pmap-c)#police 32000
Router(config-pmap-c)#exit

Router(config-pmap)#class-map VOICE
Router(config-pmap-c)#priority 256
Router(config-pmap-c)#exit
```

```
Router(config-pmap)#exit

Router(config)#interface serial 0/1
Router(config-if)#service-policy output QOS-STUDY
```

Notice that the **QOS-STUDY** policy-map makes 128 kbps of bandwidth available to e-mail traffic. However, KaZaa version 2 traffic bandwidth is limited to 32 kbps. Voice packets not only have access to 256 kbps of bandwidth, but they also receive "priority" treatment, meaning that they are sent first (that is, ahead of other traffic) up to the 256-kbps limit[13].

3.6.2 Using AutoQoS

Fortunately, Cisco added a feature called AutoQoSto many of its router and switch platforms to automatically generate router-based or switch based VoIP QoS configurations.
The following router platforms support AutoQoS:

- 1700 Series
- 2600 Series
- 3600 Series
- 3700 Series
- 7200 Series

Cisco also supports the AutoQoS feature on the following Catalyst switch series:

- 2950
- 3550
- 4500
- 6500

3.7 Traffic Classification and Marking

Classification and marking allow QoS-enabled networks to identify traffic types near the source and assign specific markings to those traffic types. This section addresses the need for and various approaches used to perform classification and marking.

3.7.1 Classification and marking basics

One of the first QoS mechanisms that you apply to your traffic is classification, which recognizes the types of traffic that are flowing across the network. For example, you might recognize Telnet, FTP, and HTTP traffic and categorize those applications together in a specific class of traffic.

On an Ethernet trunk, you can mark frames with a CoS value. A CoS value can be ranged from 0 through 7, although Cisco recommends that you never use 6 or 7. The bits that create the CoS marking depend on the type of trunk that is being used, as follows:

- IEEE 802.1Q trunk—Uses 3 bits in a Tag Control byte to mark a CoS value. (Note: This method is referred to as IEEE 802.1p.)
- ISL trunk: Uses 3 bits in the ISL header to mark a CoS value[13]. Figure 3.7.

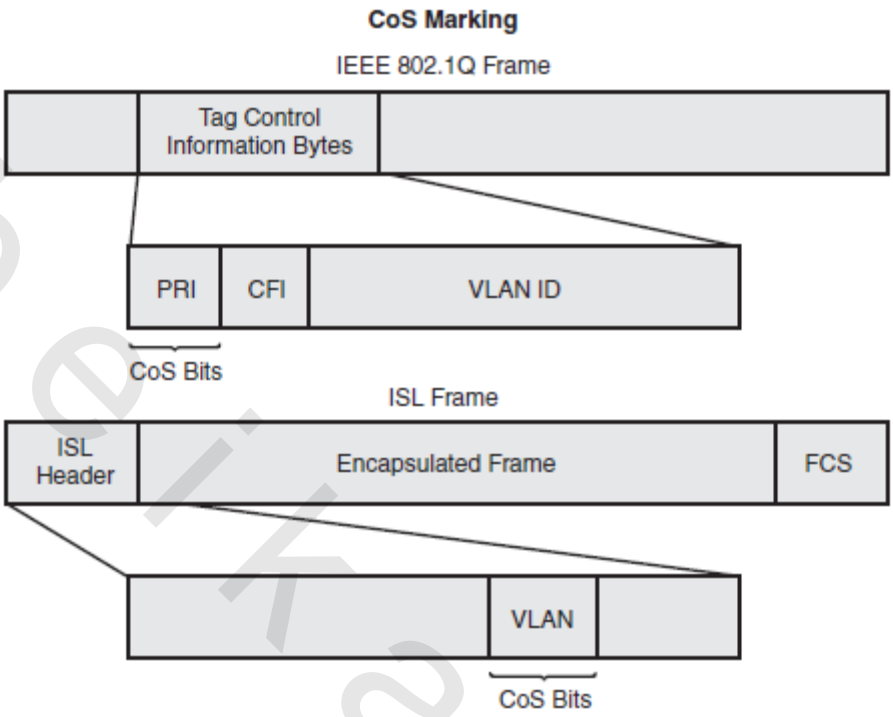


Figure 3.7CoS Marking[13].

3.7.2 Modular classification with MQC

In class-map configuration mode, you can use the match command to categorize traffic based on any of the following criteria:

- Access control lists (ACLs).
- Existing markings (for example, CoS, IP Precedence, or DSCP).
- QoS group (a locally significant grouping of packets).
- Traffic matching another class-map.
- Incoming interface.
- Media access control address (MAC) (source or destination).
- Range of UDP port numbers.

In the following example, you are matching traffic based on a variety of the preceding criteria:

```
Router(config)#class-map match-any INT
Router(config-cmap)#match input-interface ethernet 0/0
Router(config-cmap)#match input-interface ethernet 0/1
Router(config-cmap)#exit

Router(config)#class-map ACL
Router(config-cmap)#match access-group 101
Router(config-cmap)#exit
```

```
Router(config)#class-map COS
Router(config-cmap)#match cos 0 1 2 3
Router(config-cmap)#exit

Router(config)#access-list 101 permit tcp any anyeq 23
```

In this example, the INT class-map matches traffic that came into the router on any of the specified interfaces. The ACL class-map matches traffic that is matched by access-list 101. Finally, the COS class-map categorizes traffic with a CoS marking of 0, 1, 2, or 3.

3.7.3 Modular marking with MQC

After you have classified your traffic using class-maps, you can use a policy-map to mark the traffic. Following is a listing of supported markings and the corresponding syntax:

- IP Precedence (set ip precedence value).
- DSCP (set ipdscp value).
- QoS group(set ip precedence value).
- CoS value (set cos value).

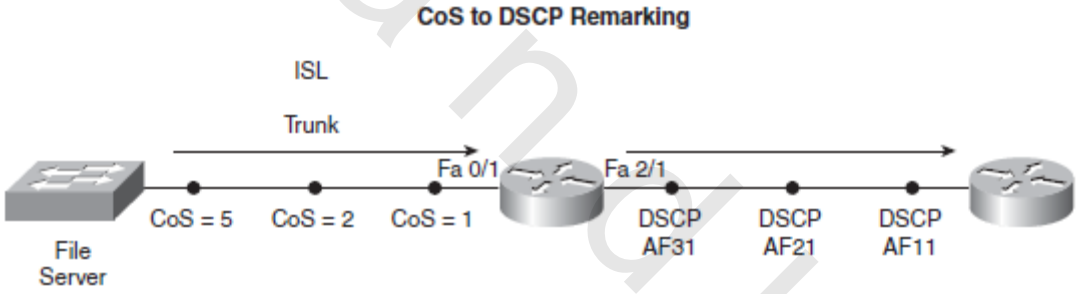


Figure 3.8CoS to DSCP Remarking[13].

```
Router(config)#class-map HI-PRI
Router(config-cmap)#match cos 5 6 7
Router(config-cmap)#exit
Router(config)#class-map MED-PRI
Router(config-cmap)#match cos 2 3 4
Router(config-cmap)#exit
Router(config)#class-map LOW-PRI
Router(config-cmap)#match cos 0 1
Router(config-cmap)#exit
Router(config)#policy-map REMARK
Router(config-pmap)#class HI-PRI
Router(config-pmap-c)#set ipdscp af31
Router(config-pmap-c)#exit
Router(config-pmap)#class MED-PRI
```

```
Router(config-pmap-c)#set ipdscp af21
Router(config-pmap-c)#exit
Router(config-pmap)#class-map LOW-PRI
Router(config-pmap-c)#set ipdscp af11
Router(config-pmap-c)#exit
Router(config-pmap)#exit
Router(config)#interface fastethernet 0/1
Router(config-if)#service-policy input REMARK
```

In this example, traffic marked with CoS values of 5, 6, or 7 is classified in the HI-PRI class-map, whereas traffic with CoS values of 2, 3, or 4 goes into the MED-PRI class-map. Finally, CoS values of 0 and 1 are placed in the LOW-PRI class-map. The REMARK policy-map assigns a DSCP value of AF31 to the HI-PRI traffic, a DSCP value of AF21 to the MED-PRI traffic, and a DSCP value of AF11 to the LOWPRI traffic. The third step of MQC applies a policy-map to an interface. In this case, you are applying the REMARK policy-map to the FastEthernet 0/1 interface in the inbound direction. It is critical that you apply the policy-map in the inbound direction. By doing so, you are remarking the CoS values before the route processor strips them.

3.7.4 Catalyst-Based classification and marking

You can perform classification and marking functions, not just on router platforms, but also on many Catalyst series switches. Even though QoS is considered primarily a WAN technology, proper QoS configuration in an enterprise network's infrastructure is also critical. For example, a switch might have interfaces that run at different speeds (for example, 10 Mbps and 1 Gbps). Such a scenario could lead to a switch queue overflowing. Also, traffic can enter a switch marked with a Layer 2 CoS value. These Layer 2 values do not pass through a route processor. Therefore, a Catalyst switch is an excellent place to perform CoS-to-DSCP remarking. So, when the traffic reaches the route processor, it has a Layer 3 marking, which can pass successfully through the route processor.

Applying QoS features at the edge of the network (for example, in a wiring closet) offers the following benefits:

- Provides immediate traffic classification.
- Reduces congestion within the remainder of the network.
- Eases the processor burden on the distribution and core routers.

Table 3.2 Default Queue Assignment

CoS Value	Queue
0 and 1	1
2 and 3	2
4 and 5	3
6 and 7	4

For internal QoS processing, all traffic (even non-IP traffic) is assigned an internal DSCP number. The default CoS-to-DSCP mappings are shown Table 3.3.

Table 3.3CoS-to-DSCP Mappings.

Cos Value	DSCP Value
0	0
1	8
2	16
3	24
4	32
5	40
6	48
7	56