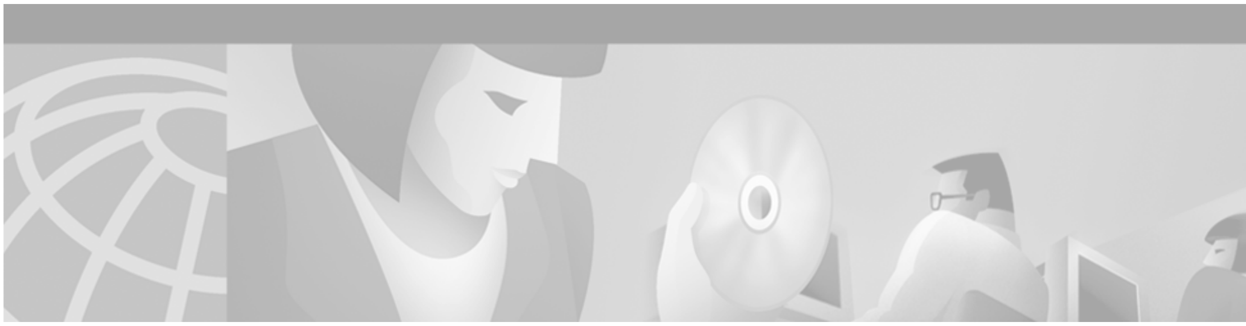

Quality of Service Overview

This chapter explains quality of service (QoS) and the service models that embody it. It also suggests benefits you can gain from implementing Cisco IOS QoS in your network. Then it focuses on the Cisco IOS QoS features and the technologies that implement them.

To identify the hardware platform or software image information associated with a feature, use the Feature Navigator on Cisco.com to search for information about the feature or refer to the software release notes for a specific release. For more information, see the "Identifying Supported Platforms" section in the "Using Cisco IOS Software" chapter in this book.

What Is Quality of Service?

QoS refers to the ability of a network to provide improved service to selected network traffic over various underlying technologies including Frame Relay, ATM, Ethernet and 802.1 networks, SONET, and IP-routed networks. In particular, QoS features provide improved and more predictable network service by providing the following services:

- Supporting dedicated bandwidth
- Improving loss characteristics
- Avoiding and managing network congestion
- Shaping network traffic
- Setting traffic priorities across the network

About QoS Architecture

You configure QoS features throughout a network to provide for end-to-end QoS delivery. The following three components are necessary to deliver QoS across a heterogeneous network:

- QoS within a single network element, which includes queueing, scheduling, and traffic shaping features.
- QoS signalling techniques for coordinating QoS for end-to-end delivery between network elements.
- QoS policing and management functions to control and administer end-to-end traffic across a network.

Not all QoS techniques are appropriate for all network routers. Because edge routers and backbone routers in a network do not necessarily perform the same operations, the QoS tasks they perform might differ as well. To configure an IP network for real-time voice traffic, for example, you would need to consider the functions of both edge and backbone routers in the network, then select the appropriate QoS feature or features.

In general, edge routers perform the following QoS functions:

- Packet classification
- Admission control
- Configuration management

In general, backbone routers perform the following QoS functions:

- Congestion management
- Congestion avoidance

Who Could Benefit from Using Cisco IOS QoS?

All networks can take advantage of aspects of QoS for optimum efficiency, whether the network is for a small corporation, an enterprise, or an Internet service provider (ISP). Different categories of networking users—such as major enterprises, network service providers, and small and medium-sized business networking users—have their own QoS requirements; in many areas, however, these requirements overlap. The Cisco IOS QoS features described in the section "Cisco QoS Features" later in this chapter address these diverse and common needs.

Enterprise networks, for example, must provide end-to-end QoS solutions across the various platforms comprising the network; providing solutions for heterogeneous platforms often requires that you take a different QoS configuration approach for each technology. As enterprise networks carry more complex, mission-critical applications and experience increased traffic from Web multimedia applications, QoS serves to prioritize this traffic to ensure that each application gets the service it requires.

ISPs require assured scalability and performance. For example, ISPs that long have offered best-effort IP connectivity now also transfer voice, video, and other real-time critical application data. QoS answers the scalability and performance needs of these ISPs to distinguish different kinds of traffic, thereby enabling them to offer service differentiation to their customers.

In the small and medium-sized business segment, managers are experiencing firsthand the rapid growth of business on the Internet. These business networks must also handle increasingly complex business applications. QoS lets the network handle the difficult task of utilizing an expensive WAN connection in the most efficient way for business applications.

Why Deploy Cisco IOS QoS?

The Cisco IOS QoS features enable networks to control and predictably service a variety of networked applications and traffic types. Implementing Cisco IOS QoS in your network promotes the following features:

- Control over resources. You have control over which resources (bandwidth, equipment, wide-area facilities, and so on) are being used. For example, you can limit bandwidth consumed over a backbone link by File Transfer Protocol (FTP) transfers or give priority to an important database access.
- Tailored services. If you are an ISP, the control and visibility provided by QoS enables you to offer carefully tailored grades of service differentiation to your customers.
- Coexistence of mission-critical applications. Cisco QoS features make certain of the following conditions:
  - That your WAN is used efficiently by mission-critical applications that are most important to your business.
  - That bandwidth and minimum delays required by time-sensitive multimedia and voice applications are available.
  - That other applications using the link get their fair service without interfering with mission-critical traffic.

Moreover, in implementing QoS features in your network, you put in place the foundation for a future fully integrated network.

End-to-End QoS Models

A service model, also called a level of service, describes a set of end-to-end QoS capabilities. End-to-end QoS is the ability of the network to deliver service required by specific network traffic from one end of the network to another. Cisco IOS QoS software supports three types of service models: best effort, integrated, and differentiated services.

Note QoS service models differ from one another in how they enable applications to send data and in the ways in which the network attempts to deliver that data. For instance, a different service model applies to real-time applications, such as audio and video conferencing and IP telephony, than a model that applies to file transfer and e-mail applications.

Consider the following factors when deciding which type of service to deploy in the network:

- The application or problem you are trying to solve. Each of the three types of service—best effort, integrated, and differentiated—is appropriate for certain applications.
- The kind of ability you want to allocate to your resources.
- Cost-benefit analysis. For example, the cost of implementing and deploying differentiated service is certain to be more expensive than the cost for a best-effort service.

The following sections describe the service models supported by features in Cisco IOS software:

- Best-Effort Service
- Integrated Service
- Differentiated Service

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.

The Cisco IOS QoS feature that implements best-effort service is FIFO queueing. Best-effort service is suitable for a wide range of networked applications such as general file transfers or e-mail.

Integrated Service

Integrated service is a multiple service model that can accommodate multiple QoS requirements. In this model the application requests a specific kind of service from the network before it sends data. The request is made by explicit signalling; the application informs the network of its traffic profile and requests a particular kind of service that can encompass its bandwidth and delay requirements. 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, based on 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 queueing 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 queueing 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, a Voice over IP (VoIP) application can reserve the required amount of bandwidth end-to-end using this kind of service. Cisco IOS QoS uses weighted fair queueing (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.

Differentiated Service

Differentiated service is a multiple service model that can satisfy differing QoS requirements. However, unlike in the integrated service model, an application using differentiated service does not explicitly signal the router before sending data.

For differentiated service, the network tries to deliver a particular kind of service based on the QoS specified by each packet. This specification can occur in different ways, for example, using the IP Precedence bit settings in IP packets or source and destination addresses. The network uses the QoS specification to classify, mark, shape, and police traffic, and to perform intelligent queueing.

The differentiated service model is used for several mission-critical applications and for providing end-to-end QoS. Typically, this service model is appropriate for aggregate flows because it performs a relatively coarse level of traffic classification.

Cisco IOS QoS includes the following features that support the differentiated service model:

- Committed access rate (CAR), which performs packet classification through IP Precedence and QoS group settings. CAR performs metering and policing of traffic, providing bandwidth management.
- Intelligent queueing schemes such as WRED and WFQ and their equivalent features on the Versatile Interface Processor (VIP), which are distributed WRED (DWRED) and distributed WFQ. These features can be used with CAR to deliver differentiated services.

For more information on how to implement Differentiated Services using the components of Cisco IOS software, see the chapter "Implementing DiffServ for End-to-End Quality of Service Overview" in this book.

Cisco QoS Features

The Cisco IOS QoS software provides the major features described in the following sections. Some of which have been previously mentioned, and all of them are briefly introduced in this chapter.

- Classification
- Congestion Management
- Congestion Avoidance
- Policing and Shaping
- Signalling
- Link Efficiency Mechanisms
- QoS Solutions
- Modular QoS Command-Line Interface
- Security Device Manager

The features listed are described more fully in the overview chapters of this book, which is organized into parts, one for each of the major features listed. Each book part contains an overview chapter and one or more configuration chapters.

Classification

Packet classification features provide the capability to partition network traffic into multiple priority levels or classes of service. For example, by using the three precedence bits in the Type of service (ToS) field of the IP packet header—two of the values are reserved for other purposes—you can categorize packets into a limited set of up to six traffic classes. After you classify packets, you can utilize other QoS features to assign the appropriate traffic handling policies including congestion management, bandwidth allocation, and delay bounds for each traffic class.

Packets can also be classified by external sources, that is, by a customer or by a downstream network provider. You can either allow the network to accept the classification or override it and reclassify the packet according to a policy that you specify.

Packets can be classified based on policies specified by the network operator. Policies can be set that include classification based on physical port, source or destination IP or MAC address, application port, IP protocol type, and other criteria that you can specify by using access lists or extended access lists.

You can use Cisco IOS QoS policy-based routing (PBR) and the classification features of Cisco IOS QoS CAR to classify packets. You can use Border Gateway Protocol (BGP) policy propagation to propagate destination-based packet classification policy throughout a large network via BGP routing updates. This section gives a brief description of these features.

In addition, you can use the QoS for Virtual Private Networks (VPNs) feature to classify packets before tunneling and encryption occur. The process of classifying features before tunneling and encryption is called preclassification.

The Class-Based Packet Marking feature provides users with a user-friendly command-line interface (CLI) for efficient packet marking by which users can differentiate packets based on the designated markings.

For more complete conceptual information on packet classification, see the chapter "Classification Overview" in this book.

For information on how to configure the various protocols that implement classification, see the following chapters:

- "Configuring Policy-Based Routing"
- "Configuring QoS Policy Propagation via Border Gateway Protocol"
- "Configuring Committed Access Rate"
- "Configuring Class-Based Packet Marking"
- "Configuring QoS for Virtual Private Networks"
- "Configuring Network-Based Application Recognition"

For complete command syntax information, refer to the *Cisco IOS Quality of Service Solutions Command Reference*.

IP Precedence

The IP Precedence feature allows you to specify the class of service of a packet using the three precedence bits in the ToS field of the IP version 4 (IPv4) header. Other features configured throughout the network can then use these bits to determine how to treat the packet in regard to the type of service to grant it. For example, although IP Precedence is not a queueing method, other queueing methods such as WFQ can use the IP Precedence setting of the packet to prioritize traffic.

Policy-Based Routing

Cisco IOS QoS PBR allows you to perform the following tasks:

- Classify traffic based on extended access list criteria.
- Set IP Precedence bits.
- Route specific traffic to engineered paths, which may be required to allow a specific QoS service through the network.

Classification of traffic through PBR allows you to identify traffic for different classes of service at the perimeter of the network and then implement QoS defined for each class of service in the core of the network using priority queueing, custom queueing, or WFQ techniques. This process obviates the need to classify traffic explicitly at each WAN interface in the core-backbone network.

Some possible applications for policy routing are to provide equal access, protocol-sensitive routing, source-sensitive routing, routing based on interactive versus batch traffic, or routing based on dedicated links.

BGP Policy Propagation

BGP provides a powerful, scalable means of utilizing attributes, such as community values, to propagate destination-based packet classification policy throughout a large network via BGP routing updates. Packet classification policy can be scalably propagated via BGP without writing and deploying complex access lists at each of a large number of routers. BGP ensures that return traffic to customers is handled as premium traffic by the network.

Committed Access Rate (Packet Classification)

CAR is the main feature supporting packet classification. CAR uses the ToS bits in the IP header to classify packets. You can use the CAR classification commands to classify and reclassify a packet.

Here are some example packet classification policies:

- All packets received on a particular T1 line are classified as high priority (port-based classification).
- All HTTP traffic is classified as medium priority (application classification).
- Video traffic from a specified IP address is classified as medium priority.
- Packets bound for particular destinations are classified as high priority traffic (for example, international traffic or traffic bound for a premium customer).
- Some packets are classified for subrate IP services. The network operator delivers a physical T1/E1 or T3/E3 line to the customer, but offers a less expensive subrate service, for example, 1 Mbps on an E1 line or 10 Mbps on a T3 line. The customer pays for the subrate bandwidth and may be upgraded to additional access bandwidth over time based on demand. CAR limits the traffic rate available to the customer and delivered to the network to the agreed-upon rate limit (with the ability to temporarily burst over the limit). The network operator may upgrade the service without any physical network arrangement.
- Traffic is classified for exchange point traffic control. An ISP offers transit services to downstream ISPs via exchange point connectivity provided by a Layer 2 switch. The upstream provider utilizes MAC-address rate limits provided by CAR to enforce bandwidth usage limitations on the downstream ISPs.

Note CAR also implements rate-limiting services, which are described later in this chapter.

Class-Based Packet Marking

The Class-Based Packet Marking feature provides users with a means for efficient packet marking by which users can differentiate packets based on the designated markings. The Class-Based Packet Marking feature allows users to perform the following tasks:

- Mark packets by setting the IP Precedence bits or the IP differentiated services code point (DSCP) in the IP ToS byte.
- Mark packets by setting the Layer 2 class of service (CoS) value.
- Associate a local QoS group value with a packet.
- Set the cell loss priority (CLP) bit setting in the ATM header of a packet from 0 to 1.

QoS for Virtual Private Networks

When packets are encapsulated by tunnel or encryption headers, QoS features are unable to examine the original packet headers and correctly classify the packets. Packets traveling across the same tunnel have the same tunnel headers, so the packets are treated identically if the physical interface is congested.

With the growing popularity of VPNs, the need to classify traffic within a traffic tunnel is gaining importance. QoS features have historically been unable to classify traffic within a tunnel. With the introduction of the QoS for VPNs feature, packets can now be classified before tunneling and encryption occur. The process of classifying features before tunneling and encryption is called preclassification.

The QoS for VPNs feature is designed for tunnel interfaces. When the feature is enabled, the QoS features on the output interface classify packets before encryption, allowing traffic flows to be adjusted in congested environments. The result is more effective packet tunneling.

Network-Based Application Recognition

The Network-Based Application Recognition (NBAR) feature provides intelligent network classification to network infrastructures. NBAR is a classification engine that recognizes a wide variety of applications, including web-based and other difficult-to-classify protocols that utilize dynamic TCP/User Datagram Ports (UDP) port assignments. When an application is recognized and classified by NBAR, a network can invoke services for that specific application.

Congestion Management

Congestion management features operate to control congestion once it occurs. One way that network elements handle an overflow of arriving traffic is to use a queueing algorithm to sort the traffic, then determine some method of prioritizing it onto an output link. Each queueing algorithm was designed to solve a specific network traffic problem and has a particular effect on network performance.

The Cisco IOS software congestion management, or queueing, features include the following:

- FIFO
- Priority queueing (PQ)
- Frame Relay permanent virtual circuit (PVC) interface priority queueing (FR PIPQ)
- Custom queueing (CQ)
- Flow-based, class-based, and distributed WFQ
- Distributed class-based WFQ

- IP RTP Priority and Frame Relay IP RTP Priority
- Low latency queueing (LLQ), Distributed LLQ, and LLQ for Frame Relay

For more complete conceptual information on packet classification, see the chapter "Congestion Management Overview" in this book.

For information on how to configure the various protocols that implement congestion management, see the following chapters:

- "Configuring Weighted Fair Queueing"
- "Configuring Custom Queueing"
- "Configuring Priority Queueing"

For complete command syntax information, refer to the *Cisco IOS Quality of Service Solutions Command Reference*.

What Is Congestion in Networks?

To give you a more definite sense of congestion in networks, this section briefly describes some of its characteristics, drawing on the explanation presented by V. Paxson and S. Floyd in a paper titled *Wide Area Traffic: The Failure of Poisson Modeling*.

What does congestion look like? Consideration of the behavior of congested systems is not simple and cannot be dealt with in a simplistic manner, because traffic rates do not simply rise to a level, stay there a while, then subside. Periods of traffic congestion can be quite long, with losses that are heavily concentrated. In contrast to Poisson traffic models, linear increases in buffer size do not result in large decreases in packet drop rates; a slight increase in the number of active connections can result in a large increase in the packet loss rate. This understanding of the behavior of congested networks suggests that because the level of busy period traffic is not predictable, it would be difficult to efficiently size networks to reduce congestion adequately. Observers of network congestion report that in reality, traffic "spikes," which causes actual losses that ride on longer-term ripples, which in turn ride on still longer-term swells.

FIFO Queueing

FIFO provides basic store and forward capability. FIFO is the default queueing algorithm in some instances, thus requiring no configuration. See "WFQ and Distributed WFQ" later in this section for a complete explanation of default configuration.

PQ

Designed to give strict priority to important traffic, PQ ensures that important traffic gets the fastest handling at each point where PQ is used. PQ can flexibly prioritize according to network protocol (such as IP, IPX, or AppleTalk), incoming interface, packet size, source/destination address, and so on.

Frame Relay PVC PQ

The FR PIPQ provides an interface-level PQ scheme in which prioritization is based on destination PVC rather than packet contents. For example, FR PIPQ allows you to configure PVC transporting voices traffic to have absolute priority over a PVC transporting signalling traffic, and a PVC transporting signalling traffic to have absolute priority over a PVC transporting data.

FR PIPQ provides four levels of priority: high, medium, normal, and low. The Frame Relay packet is examined at the interface for the data-link connection identifier (DLCI) value. The packet is then sent to the correct priority queue based on the priority level configured for that DLCI.

CQ

CQ reserves a percentage of the available bandwidth of an interface for each selected traffic type. If a particular type of traffic is not using the bandwidth reserved for it, then other traffic types may use the remaining reserved bandwidth.

WFQ and Distributed WFQ

WFQ applies priority (or weights) to identified traffic to classify traffic into conversations and determine how much bandwidth each conversation is allowed relative to other conversations. WFQ classifies traffic into different flows based on such characteristics as source and destination address, protocol, and port and socket of the session.

To provide large-scale support for applications and traffic classes requiring bandwidth allocations and delay bounds over the network infrastructure, Cisco IOS QoS includes a version of WFQ that runs only in distributed mode on VIPs. This version is called VIP-distributed WFQ (DWFQ). It provides increased flexibility in terms of traffic classification, weight assessment, and discard policy, and delivers Internet-scale performance on the Cisco 7500 series platforms.

For serial interfaces at E1 (2.048 Mbps) and below, WFQ is used by default. When no other queueing strategies are configured, all other interfaces use FIFO by default.

CBWFQ and Distributed CBWFQ

The class-based WFQ (CBWFQ) and distributed class-based WFQ (DCBWFQ) features extend the standard WFQ functionality to provide support for user-defined traffic classes. They allow you to specify the exact amount of bandwidth to be allocated for a specific class of traffic. Taking into account available bandwidth on the interface, you can configure up to 64 classes and control distribution among them.

DCWFQ is intended for use on the VIP-based Cisco 7000 series routers with the Route Switch Processors (RSPs), and the Cisco 7500 series routers except those with PA-A3-8T1IMA modules.

IP RTP Priority

The IP RTP Priority feature provides a strict priority queueing scheme that allows delay-sensitive data such as voice to be dequeued and sent before packets in other queues are dequeued. This feature can be used on serial interfaces and Frame Relay PVCs in conjunction with either WFQ or CBWFQ on the same outgoing interface. In either case, traffic matching the range of UDP ports specified for the priority queue is guaranteed strict priority over other CBWFQ classes or WFQ flows; packets in the priority queue are always serviced first.

Frame Relay IP RTP Priority

The Frame Relay IP RTP Priority feature provides a strict priority queueing scheme on a Frame Relay PVC for delay-sensitive data such as voice. Voice traffic can be identified by its Real-Time Transport Protocol (RTP) port numbers and classified into a priority queue configured by the **frame-relay ip rtp priority** command. The result of using this feature is that voice is serviced as strict priority in preference to other nonvoice traffic.

LLQ

LLQ provides strict priority queueing on ATM VCs and serial interfaces. This feature allows you to configure the priority status for a class within CBWFQ, and is not limited to UDP port numbers, as is IP RTP Priority. LLQ and IP RTP Priority can be configured at the same time, but IP RTP Priority takes precedence.

Additionally, the functionality of LLQ has been extended to allow you to specify the Committed Burst (Bc) size in LLQ and to change (or vary) the number of packets contained in the hold queue per-VC (on ATM adapters that support per-VC queueing). For more information, see the chapter "Congestion Management Overview" in this book.

Distributed LLQ

The Distributed LLQ feature provides the ability to specify low latency behavior for a traffic class on a VIP-based Cisco 7500 series router except those with PA-A3-8T1IMA modules. LLQ allows delay-sensitive data such as voice to be dequeued and sent before packets in other queues are dequeued

The Distributed LLQ feature also introduces the ability to limit the depth of a device transmission ring.

LLQ for Frame Relay

LLQ for Frame Relay is provides strict PQ for voice traffic and WFQs for other classes of traffic. Before the release of this feature, LLQ was available at the interface and ATM VC levels. It is now available at the Frame Relay VC level when Frame Relay Traffic Shaping is configured.

Strict PQ improves QoS by allowing delay-sensitive traffic such as voice to be pulled from the queue and sent before other classes of traffic.

LLQ for Frame Relay allows you to define classes of traffic according to protocol, interface, or access lists. You can then assign characteristics to those classes, including priority, bandwidth, queue limit, and WRED.

Congestion Avoidance

Congestion avoidance techniques monitor network traffic loads in an effort to anticipate and avoid congestion at common network and internetwork bottlenecks before it becomes a problem. These techniques are designed to provide preferential treatment for premium (priority) class traffic under congestion situations while concurrently maximizing network throughput and capacity utilization and minimizing packet loss and delay. WRED and DWRED are the Cisco IOS QoS congestion avoidance features.

Router behavior allows output buffers to fill during periods of congestion, using the tail drop feature to resolve the problem when WRED is not configured. During tail drop, a potentially large number of packets from numerous connections are discarded because of lack of buffer capacity. This behavior can result in waves of congestion followed by periods during which the transmission link is not fully used. WRED obviates this situation proactively by providing congestion avoidance. That is, instead of waiting for buffers to fill before dropping packets, the router monitors the buffer depth and performs early discards on selected packets sent over selected connections.

WRED is the Cisco implementation of the RED class of congestion avoidance algorithms. When RED is used and the source detects the dropped packet, the source slows its transmission. RED is primarily designed to work with TCP in IP internetwork environments.

WRED can also be configured to use the DSCP value when it calculates the drop probability of a packet, enabling WRED to be compliant with the DiffServ standard being developed by the Internet Engineering Task Force (IETF).

For more complete conceptual information, see the chapter "Congestion Avoidance Overview" in this book.

For information on how to configure WRED, DWRED, flow-based WRED, and DiffServ Compliant WRED, see the chapter "Configuring Weighted Random Early Detection" in this book.

For complete command syntax information, refer to the *Cisco IOS Quality of Service Solutions Command Reference*.

WRED

WRED, the Cisco implementation of RED, combines the capabilities of the RED algorithm with IP Precedence to provide preferential traffic handling for higher priority packets. It can selectively discard lower priority traffic when the interface begins to get congested and provide differentiated performance characteristics for different classes of service. WRED is also RSVP-aware. WRED is available on the Cisco 7200 series RSP.

DWRED

DWRED is the Cisco high-speed version of WRED. The DWRED algorithm was designed with ISP providers in mind; it allows an ISP to define minimum and maximum queue depth thresholds and drop capabilities for each class of service.

DWRED, which is available on the Cisco 7500 series routers or the Cisco 7000 series router with RSPs is analogous in function to WRED, which is available on the Cisco 7200 series RSP.

Flow-Based WRED

The Flow-based WRED feature forces WRED to afford greater fairness to all flows on an interface in regard to how packets are dropped.

To provide fairness to all flows, flow-based WRED has the following features:

- It ensures that flows that respond to WRED packet drops by backing off packet transmission are protected from flows that do not respond to WRED packet drops.
- It prohibits a single flow from monopolizing the buffer resources at an interface.

DiffServ Compliant WRED

The DiffServ Compliant WRED feature extends the functionality of WRED to enable support for Differentiated Services (DiffServ) and Assured Forwarding (AF) Per Hop Behavior (PHB). This feature enables customers to implement AF PHB by coloring packets according to DSCP values and then assigning preferential drop probabilities to those packets.

The DiffServ and the AF PHB standards are supported by this feature.

Policing and Shaping

Cisco IOS QoS includes traffic policing capabilities implemented through the rate-limiting aspects of CAR, and the Traffic Policing feature.

For traffic shaping, Cisco IOS QoS includes Generic Traffic Shaping (GTS), Class-Based Shaping, Distributed Traffic Shaping (DTS), and FRTS protocols.

For more complete conceptual information, see the chapter "Policing and Shaping Overview" in this book.

For information on how to configure the Traffic Policing feature, see the chapter "Configuring Traffic Policing" in this book.

For information on how to configure GTS, Class-Based Shaping, and DTS, see the following chapters in this book:

- "Configuring Generic Traffic Shaping"
- "Configuring Class-Based Shaping"
- "Configuring Distributed Traffic Shaping"

Note For information on how to configure Frame Relay and Frame Relay Traffic Shaping, refer to the *Cisco IOS Wide-Area Networking Configuration Guide*.

Note For complete command syntax information on the commands related to Traffic Policing, GTS, Class-Based Shaping, and DTS, refer to the *Cisco IOS Quality of Service Solutions Command Reference*.

CAR Rate Limiting

The rate-limiting feature of CAR provides the network operator with the means to define Layer 3 aggregate or granular access, or egress bandwidth rate limits, and to specify traffic handling policies when the traffic either conforms to or exceeds the specified rate limits. Aggregate access or egress matches all packets on an interface or subinterface. Granular access or egress matches a particular type of traffic based on precedence. You can designate CAR rate-limiting policies based on physical port, packet classification, IP address, MAC address, application flow, and other criteria specifiable by access lists or extended access lists. CAR rate limits may be implemented either on input or output interfaces or subinterfaces including Frame Relay and ATM subinterfaces.

An example of the use of the rate-limiting capability of CAR is application-based rates limiting HTTP World Wide Web traffic to 50 percent of link bandwidth, which ensures capacity for non-Web traffic including mission-critical applications.

Traffic Policing

The Traffic Policing feature performs the following functions:

- Limits the input or output transmission rate of a class of traffic based on user-defined criteria
- Marks packets by setting the IP Precedence value, the QoS group, or the DSCP value

Shaping

Cisco IOS QoS software the following traffic shaping features that manage traffic and congestion on the network:

- GTS, which provides a mechanism to control the flow of outbound traffic on a particular interface. It reduces outbound traffic flow to avoid congestion by constraining specified traffic to a particular bit rate. Traffic adhering to a particular profile can be shaped to meet downstream requirements, eliminating bottlenecks in topologies with data rate mismatches.
- Class-Based Shaping, which provides the means for configuring GTS on a class, rather than only on an access control list (ACL).

  Using the Class-Based Shaping feature, you can perform the following tasks:

  - Configure GTS on a traffic class
  - Specify average rate or peak rate traffic shaping
  - Configure CBWFQ inside GTS

  Class-Based Shaping can be enabled on any interface that supports GTS.
- DTS, which provides the means for managing the bandwidth of an interface to avoid congestion, to meet remote site requirements, and to conform to a service rate that is provided on that interface. DTS uses queues to buffer traffic surges that can congest a network.
- FRTS, which provides parameters such as the following that are useful for managing network traffic congestion:

  - Committed information rate (CIR)
  - Forward and backward explicit congestion notification (FECN/BECN)
  - The discard eligible (DE) bit

  For some time Cisco has provided support for FECN for DECnet and OSI, BECN for SNA traffic using direct Logical Link Control, type 2 (LLC2) encapsulation via RFC 1490, and DE bit support. The FRTS feature builds upon this Frame Relay support by providing additional capabilities that improve the scalability and performance of a Frame Relay network by increasing the density of VCs and improving response time.

  FRTS applies only to Frame Relay permanent PVCs and switched virtual circuits (SVCs).

Signalling

Cisco IOS QoS signalling provides a way for an end station or network node to signal its neighbors to request special handling of certain traffic. QoS signalling is useful for coordinating the traffic handling techniques provided by other QoS features. It plays a key role in configuring successful overall end-to-end QoS service across your network.

Cisco IOS QoS signalling takes advantage of IP. Either in-band (IP Precedence, 802.1p) or out-of-band (RSVP) signalling is used to indicate that a particular QoS service is desired for a particular traffic classification. Together, IP Precedence and RSVP provide a robust combination for end-to-end QoS signalling: IP Precedence signals for differentiated QoS and RSVP for guaranteed QoS.

To achieve the end-to-end benefits of IP Precedence and RSVP signalling, Cisco IOS QoS software offers ATM User Network Interface (UNI) signalling and the Frame Relay Local Management Interface (LMI) to provide signalling into their respective backbone technologies.

To achieve centralized monitoring and control of RSVP signalling, Cisco IOS software offers Common Open Policy Service (COPS) with RSVP.

To enable admission control over IEEE 802-styled networks, Cisco IOS QoS software offers Subnetwork Bandwidth Manager (SBM).

To provide support for Controlled Load Service using RSVP over an ATM core network, Cisco IOS QoS software offers the RSVP-ATM QoS Interworking feature.

Cisco also provides RSVP support for Low Latency Queueing (LLQ) and Frame Relay.

For more complete conceptual information, see the chapter "Signalling Overview" in this book.

For information on how to configure the various protocols that implement signalling, see the following chapters:

- "Configuring RSVP"
- "Configuring RSVP Support for LLQ"
- "Configuring RSVP Support for Frame Relay"
- "Configuring COPS for RSVP"
- "Configuring Subnetwork Bandwidth Manager"
- "Configuring RSVP-ATM QoS Interworking"

For complete command syntax information, refer to the *Cisco IOS Quality of Service Solutions Command Reference*.

Link Efficiency Mechanisms

Cisco IOS QoS software offers three link efficiency mechanisms that work in conjunction with queueing and traffic shaping to improve efficiency and predictability of the application services levels: Link Fragmentation and Interleaving (LFI), Compressed Real-Time Protocol (CRTP), and Distributed Compressed Real-Time Protocol (dCRTP).

For more complete conceptual information, see the chapter "Link Efficiency Mechanisms Overview" in this book.

For complete command syntax information, refer to the *Cisco IOS Quality of Service Solutions Command Reference*.

Link Fragmentation and Interleaving

Interactive traffic, such as Telnet and VoIP, is susceptible to increased latency and jitter when the network processes large packets, such as LAN-to-LAN FTP Telnet transfers traversing a WAN link. This susceptibility increases as the traffic is queued on slower links. Cisco IOS QoS LFI reduces delay and jitter on slower speed links by breaking up large datagrams and interleaving low-delay traffic packets with the resulting smaller packets.

For information on how to configure LFI, see the chapter "Configuring Link Fragmentation and Interleaving for Multilink PPP," or the chapter "Configuring Link Fragmentation and Interleaving for Frame Relay and ATM Virtual Circuits" in this book.

Compressed Real-Time Protocol

RTP is a host-to-host protocol used for carrying newer multimedia application traffic, including packetized audio and video, over an IP network. RTP provides end-to-end network transport functions intended for applications sending real-time requirements, such as audio, video, or simulation data multicast or unicast network services.

To avoid the unnecessary consumption of available bandwidth, the RTP header compression feature, referred to as CRTP, is used on a link-by-link basis.

For information on how to configure CRTP, see the chapter "Configuring Compressed Real-Time Protocol" in this book.

Distributed Compressed Real-Time Protocol

The dCRTP feature compresses the combined 40-byte IP/UDP/RTP packet headers to 2 to 4 bytes on packets traveling on a Cisco 7500 series router with a VIP in distributed fast-switching and distributed Cisco Express Forwarding (dCEF) environments. This compression reduces the packet size, improves the speed of packet transmission, and reduces packet latency.

For information on how to configure the dCRTP feature, see the chapter "Configuring Distributed Compressed Real-Time Protocol" in this book.

QoS Solutions

IP to ATM CoS

IP to ATM CoS is a feature suite that maps QoS characteristics between IP and ATM, making it possible to support differential services in network service provider environments.

Network managers can use existing features such as CAR or PBR to classify and mark different IP traffic by modifying the IP Precedence field in the IPv4 packet header. Subsequently, WRED or DWRED can be configured on a per-VC basis so that the IP traffic is subject to different drop probabilities (and therefore priorities) as IP traffic coming into a router competes for bandwidth on a particular VC.

IP to ATM CoS provides support for ATM VC bundle management, allowing you to configure multiple VCs that have different QoS characteristics between any pair of ATM-connected routers.

IP to ATM CoS also provides for per-VC WFQ and CBWFQ, which allows you to apply CBWFQ functionality—normally applicable at the interface or subinterface levels only—to an individual VC configured for IP to ATM CoS. You can use this feature to apply either CBWFQ or flow-based WFQ on a per-VC basis.

For more complete conceptual information, see the chapter "IP to ATM Class of Service Overview" in this book.

For information on how to configure IP to ATM CoS, see the chapter "Configuring IP to ATM Class of Service" in this book.

QoS Features for Voice

Many of the QoS features already mentioned in this chapter are useful for voice applications. For a high-level overview of Cisco IOS QoS features for voice, see the chapter "Introduction to QoS Features for Voice" in this book.

Differentiated Services Implementations

Many of the QoS features mentioned in this book can be used to implement Differentiated Services on your network. For a high-level overview of how to use the Cisco IOS components to implement Differentiated Services, see the chapter "Implementing DiffServ for End-to-End Quality of Service Overview" in this book.

Modular QoS Command-Line Interface

The Modular CLI is a CLI structure that allows users to create traffic policies and attach these policies to interfaces. For conceptual information about the Modular QoS CLI, see the chapter "Modular Quality of Service Command-Line Interface Overview" in this book.

The Modular QoS CLI contains the following three steps:

- Define a traffic class with the **class-map** command.
- Create a traffic policy by associating the traffic class with one or more QoS features (using the **policy-map** command).
- Attach the traffic policy to the interface with the **service-policy** command.

For information on how to configure the Modular QoS CLI, see "Configuring the Modular Quality of Service Command-Line Interface" in this book.

Security Device Manager

The Cisco Router and Security Device Manager (SDM) provides an intuitive, graphical user interface for configuring and monitoring advanced IP-based QoS functionality within Cisco routers.

For a high-level overview of SDM, see the chapter "Security Device Manager Overview" in this book.