ISC Quality of Service Concepts

When network congestion occurs, all traffic has an equal chance of being dropped. Quality of service (QoS) provisioning categorizes network traffic, prioritizes it according to its relative importance, and provides priority treatment through congestion avoidance techniques. Implementing QoS in your network makes network performance more predictable and bandwidth utilization more effective.

QoS classifies traffic by assigning class of service (CoS) values to frames at supported ingress interfaces. QoS implements scheduling on egress interfaces with transmit queue drop thresholds and multiple transmit queues that use CoS values to give preference to higher-priority traffic.

QoS manages bandwidth to assure the desired performance for network applications. For example, email generally does not require high performance from a network, but real-time applications such as IP telephony or video streaming do. If the network is not consistently providing data flow control for these applications, the performance suffers.

Service provider network architecture contains access routers, distribution routers, core routers and ATM switches. The access routers terminate customer connections. The Cisco IP Solution Center (ISC) configures QoS at the access circuit, which involves the access router (called provider edge devices, or PEs) in the service provider network and the customer premise equipment (CPE) in the customer network. A QoS policy is applied to the selected set of access circuits using a QoS service request.

There are three ways to provision QoS using ISC:

- IP QoS—Select the device interfaces, create a QoS policy and apply it to the specified device interfaces. IP QoS can be implemented independent of VPN services and is the most common method for QoS provisioning using ISC.

  IP QoS provisioning is described in Chapter 5, “Provisioning Process for IP QoS.”.

- IP QoS for MPLS VPN—Apply an MPLS VPN-aware Qos policy to an MPLS service request.

  IP QoS MPLS VPN is described in Chapter 7, “Applying QoS Policies to VPN Services.”

- Ethernet QoS—Select an L2VPN, MPLS VPN, or VPLS service request that has already been deployed and apply QoS provisioning to that service request.

  QoS provisioning for MPLS VPN, L2VPN, and VPLS is described in Chapter 7, “Applying QoS Policies to VPN Services.”

This chapter describes the basic concepts for Quality of Service (QoS) as it is used in the ISC application.

This chapter contains the following sections:

Introduction to ISC QoS

QoS provisioning is a method for optimizing the flow of traffic in a network. If you have an enterprise network with services facilitated across a service provider MPLS infrastructure, QoS provisioning can guarantee that all applications receive the service levels required to meet expected performance in the network.

For complete QoS implementation you should identify:

- Low-latency applications (video and voice-over-IP, or VoIP) and mark them for high-priority treatment throughout the network
- Applications that require bandwidth guarantees should be marked and protected
- Applications that use more than their fair share of bandwidth can be identified and controlled

QoS is a collection of technologies that allows applications to request and receive predictable service levels in terms of bandwidth, latency variations, and delay.

Table 1-1 describes the typical QoS requirements for a multimedia network.

Table 1-1 Typical Multimedia QoS Requirements

Traffic Type	Max. Packet Loss	Max. One-way Latency	Max. Jitter	Guaranteed Priority Bandwidth Per Session
VoIP	1 percent	200 ms	30 ms	12 to 106 kbps*
Video-conferencing	1 percent	200 ms	30 ms	Size of the session plus 20 percent
Streaming Video	2 percent	5 seconds	N/A	Depends on encoding format and video stream rate.
Data	Variable	Variable	Variable	Variable
*Depending on sampling rate, codec, and Layer 2 overhead.				

Voice and video applications are less tolerant of loss, delay, and delay variation (jitter) than data, but their QoS requirements are more obvious. Data applications vary widely in their QoS requirements, and should be profiled before you determine the appropriate classification and scheduling treatment.

ISC QoS Components

There are three primary configuration components to QoS:

- Classification—Identifying and marking packets so that varying service levels can be enforced throughout the network.
- Scheduling—Assigning packets to one of multiple queues and associated service types based on classification for specific service level treatment by the network.
- Resource management—Accurately calculating the required bandwidth for all applications plus overhead.

In ISC, the QoS components used to achieve classification, scheduling, and resource management are:

- Traffic Classification, page 1-3
- Marking, page 1-4

Each of these components is described in the following sections.

Traffic Classification

Traffic classification (also called packet classification) partitions traffic into multiple priority levels, or classes of service. Traffic classification is the primary component of class-based QoS provisioning.

For example, using the three precedence bits in the type of service (ToS) field of the IP packet header, you can categorize packets into a limited set of up to six traffic classes. After you classify packets, you can use other QoS components to assign the appropriate traffic handling policies for each traffic class.

Packets can also be classified by external sources such as; by a customer, or by a downstream network provider. You can either allow the network to accept the external classification, or override it and re-classify the packet according to the QoS policy you specify in ISC.

ISC allows you to classify traffic based on source address, source port, destination port, port ranges, protocol ID, DSCP, and IP Precedence values.

IP QoS Traffic Classification

For IP QoS, ISC uses traffic classification to associate packets with a specific service level IP QoS policy. ISC provides five template service classes to use for traffic classification.

- VoIP
- Routing Protocol
- Management
- Business-Data-1
- Best Effort

A typical network uses three service classes in a QoS policy; a VoIP service class, a management service class (which is often combined with a routing protocol service class), and a data service class.

For more information on traffic classification in service classes, see Creating the Service Level IP QoS Policy, page 5-10.

Ethernet QoS Traffic Classification

For Ethernet QoS, ISC uses traffic classification to associate packets with a specific service level Ethernet QoS policy. ISC provides four template service classes for Ethernet QoS to use for traffic classification.

- Architecture for Voice, Video and Integrated Data (AVVID)
- Call Control
- Business Critical
- Best Effort

A typical network uses three service classes in a QoS policy; an AVVID service class, a call control service class, and a data service class.

For more information on traffic classification in service classes, see Service Level Ethernet QoS Policy, page 7-1.

Marking

Marking is a way to identify packet flows to differentiate them. Packet marking allows you to partition your network into multiple priority levels or classes of service.

ISC supports marking based on the following bits in the IP QoS type of service (ToS) byte for the packet:

- IP Precedence value
- IP differentiated services code point (DSCP) value
- MPLS Experimental (MPLS Exp) value

Note See Service Level Ethernet QoS Policy Entry Fields, page 7-8 for information on Ethernet QoS packet marking.

These markings can be used to identify traffic within the network, and other interfaces can match traffic based on the IP Precedence or DSCP markings. You can set up to 8 different IP precedence markings (0 through 7) and 64 different IP DSCP markings (0 through 63).

IP Precedence and DSCP markings are used in the following QoS concepts:

- Congestion Management—Used to determine how packets should be scheduled.
- Congestion Avoidance—Used to determine how packets should be treated in Weighted Random Early Detection (WRED), a packet dropping mechanism used in congestion avoidance.
- Rate Limiting—Used to set the IP precedence or DSCP values for packets entering the network. Networking devices within the network can then use the adjusted IP Precedence values to determine how the traffic should be treated based on the transmission rate.

MPLS Experimental Values

Marking with the MPLS Exp. value in addition to standard IP QoS ensures the following:

- Standard IP QoS policies are followed before the packets enter the MPLS network.
- At the ingress router to the MPLS network (PE device), the packet's DSCP or IP Precedence value is mapped to the MPLS Exp. field. These mappings are part of the QoS policy.
- The DSCP or IP Precedence value in the IP header continues to be the basis for IP QoS when the packet leaves the MPLS network.

Packet behavior for the QoS provisioning components, congestion management and congestion avoidance, are derived from the MPLS Exp. bit.

Note Marking packets with the MPLS Exp. value does not modify the DSCP/IP precedence markings in the IP header.

For more information on marking with the MPLS Exp value, see Traffic Classification Based on Variables, page 4-6 for IP QoS and Service Level Ethernet QoS Policy, page 7-1 for Ethernet QoS.

Rate Limiting

Rate limiting allows you to control the maximum rate of traffic sent or received on an interface. Rate limiting is configured on the CPE and PE device interfaces at the edge of the network and limits traffic into or out of the network. Traffic that falls within the rate parameters is sent, while traffic that exceeds the parameters is dropped or sent with a different priority.

ISC supports class-based rate limiting and interface-based aggregated rate limiting.

- Class-based rate limiting applies rate limiting parameters to an ISC service class.
- Interface-based aggregated rate limiting matches all packets, or a subset of packets, on an interface or subinterface and allows you to control the maximum rate of traffic sent or received. You can also specify traffic handling policies for traffic that either conforms to or exceeds the specified rate limits.

Rate limiting parameters in ISC include:

- Mean or peak rate
- Burst sizes
- Conform, exceed, and violate actions

For more information on configuring rate limiting QoS parameters in ISC, see Interface-Based Aggregated Rate Limiters, page 6-31 for IP QoS and Service Level Ethernet QoS Policy Entry Fields, page 7-8 for Ethernet QoS.

Traffic Shaping

Traffic shaping allows you to control the traffic exiting an interface to match its flow to the speed of the remote target interface and to ensure that traffic conforms to the policies assigned to it.

ISC supports class-based traffic shaping and aggregated traffic shaping.

- Class-based traffic shaping applies traffic shaping to an ISC service class.
- Aggregated traffic shaping applies these parameters to an interface instead of to a class of traffic.

Specifying traffic shaping allows you to make better use of available bandwidth. Traffic shaping parameters in ISC include:

- Average rate or peak rate for class-based traffic shaping
- Cell rates for ATM traffic shaping
- Rate factors for ATM traffic shaping
- Aggregated traffic shapers:
  - Frame Relay (FR) Traffic Shaper
  - FR Traffic Shaper (Non-MQC)
  - Parent-level Class-based Shaper
  - ATM Traffic Shaper (VBR-rt)
  - ATM Traffic Shaper (VBR-nrt)
  - ATM Traffic Shaper (CBR)
  - ATM Traffic Shaper (ABR)

Tip The difference between a rate limiter parameter and a traffic shaping parameter is that the a rate limiter drops traffic in the presence of congestion, while a traffic shaper delays excess traffic using a buffer, or queueing mechanism.

For more information on configuring traffic shaping parameters in ISC, see Aggregated Traffic Shapers, page 6-22.

Congestion Management

Congestion management controls congestion by determining the order in which packets are sent out an interface based on priorities assigned to those packets.

Congestion management involves:

- Creating queues
- Assigning packets to those queues based on packet classification
- Scheduling packets in a queue for transmission

With congestion management, packets are scheduled for transmission according to their assigned priority and the queueing mechanism configured for the interface. The router determines the order of packet transmission by controlling which packets are placed in which queue and how queues are serviced with respect to each other.

The congestion management component of QoS offers different types of queueing techniques, each of which allows you to specify creation of a different number of queues, with greater or lesser degrees of differentiation of traffic, and to specify the order in which that traffic is sent.

Congestion management parameters in ISC include:

- Bandwidth
- Queue limits

Congestion management parameters are configured at the service class level in ISC. For more information, see Service Level IP QoS Parameters, page 6-1 or Service Level Ethernet QoS Policy Entry Fields, page 7-8.

Congestion Avoidance

Congestion avoidance monitors network traffic loads in an effort to anticipate and avoid congestion at common network bottlenecks. Congestion management parameters provide preferential treatment for priority class traffic under congestion situations, while concurrently maximizing network throughput and capacity utilization and minimizing packet loss and delay.

ISC implements congestion avoidance parameters through packet dropping methods, such as WRED. WRED is used in combination with DSCP and IP Precedence and provides buffer management. WRED is frequently used to slow down TCP flows.

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.

Congestion avoidance parameters are configured at the service class level in ISC. For more information, see Service Level IP QoS Parameters, page 6-1.