Products & Services

# QoS Frequently Asked Questions

**Document ID: 22833**

Questions


Introduction
General
Classification and Marking
Queueing and Congestion Management
Congestion Avoidance Weighted Random Early Detection (WRED)
Policing and Shaping
Quality of Service (QoS) Frame Relay
Quality of Service (QoS) Over Asynchronous Transfer Mode (ATM)
Voice and Quality of Service (QoS)
Related Information


## Introduction

This document addresses the most Frequently Asked Questions (FAQs) related to Quality of Service (QoS).

## General

### Q. What is Quality of Service (QoS)?

**A.** QoS refers to the ability of a network to provide better service to selected network traffic over various underlying technologies including Frame Relay, Asynchronous Transfer Mode (ATM), Ethernet and 802.1 networks, SONET, and IP-routed networks.

QoS is a collection of technologies which allows applications to request and receive predictable service levels in terms of data throughput capacity (bandwidth), latency variations (jitter), and delay. In particular, QoS features provide better and more predictable network service by the following methods:

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

The Internet Engineering Task Force (IETF) defines the following two architectures for QoS:

- Integrated Services (IntServ)
- Differentiated Services (DiffServ)

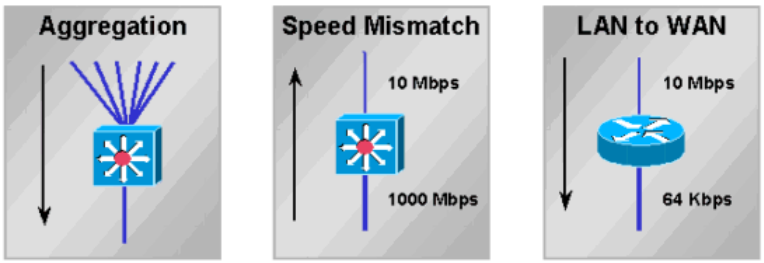
IntServ uses the Resource Reservation Protocol (RSVP) to signal explicitly the QoS needs of an application's traffic along the devices in the end-to-end path through the network. If every network device along the path can reserve the necessary bandwidth, the originating application can begin transmitting. Request for Comments (RFC) 2205 defines RSVP, and RFC 1633 defines IntServ.

DiffServ focuses on aggregated and provisioned QoS. Instead of signaling an application's QoS requirements, DiffServ uses a DiffServ Code Point (DSCP) in the IP header to indicate the required QoS levels. Cisco IOS® Software Release 12.1(5)T introduced DiffServ compliance on Cisco routers. For more information, refer to the following documents:

- Integrated Service in Cisco IOS 12.1
- Implementing DiffServ for End-to-End Quality of Service
- Implementing Quality of Service Policies with DSCP

### Q. What are congestion, delay, and jitter?

**A.** An interface experiences congestion when it is presented with more traffic than it can handle. Network congestion points are strong candidates for Quality of Service (QoS) mechanisms. The following is an example of typical congestion points:




Network congestion results in delay. A network and its devices introduce several kinds of delays, as explained in Understanding Delay in Packet Voice Networks. Variation in delay is known as jitter, as explained in Understanding Jitter in Packet Voice Networks (Cisco IOS Platforms). Both delay and jitter need to be controlled and minimized to support real-time and interactive traffic.

### Q. What is the MQC?

**A.** MQC stands for modular Quality of Service (QoS) Command Line Interface (CLI). It is designed to simplify the configuration of QoS on Cisco routers and switches by defining a common command syntax and resulting set of QoS behaviors across platforms. This model replaces the previous model of defining unique syntaxes for each QoS feature and for each platform.

The MQC contains the following three steps:

1. Define a traffic class by issuing the **class-map** command.
2. Create a traffic policy by associating the traffic class with one or more QoS features by issuing the **policy-map** command.
3. Attach the traffic policy to the interface, subinterface, or Virtual Circuit (VC) by issuing the **service-policy** command.

**Note:** You implement the traffic conditioning functions of DiffServ, such as marking and shaping, using the MQC syntax.

For more information, refer to Modular Quality of Service Command-Line Interface.

### Q. What does the `service-policy is supported only on VIP interfaces with DCEF enabled` message mean?

**A.** On Versatile Interface Processors (VIPs) in a Cisco 7500 Series, only distributed Quality of Service (QoS) features are supported as of Cisco IOS 12.1(5)T, 12.1(5)E, and 12.0(14)S. Enabling distributed Cisco Express Forwarding (dCEF) automatically enables distributed QoS.

Non-VIP interfaces, known as legacy Interface Processors (IPs), support central QoS features as enabled on the Route Switch Processor (RSP). For more information, refer to the following documents:

- Distributed Class-Based Weighted Fair Queueing and Distributed Weighted Random Early Detection
- Distributed Low Latency Queueing
- Distributed Traffic Shaping
- Versatile Interface Processor-Based Distributed FRF.11 and FRF.12 for Cisco IOS Release 12.1 T

### Q. How many classes does a Quality of Service (QoS) policy support?

**A.** In Cisco IOS versions earlier than 12.2 you could define a maximum of only 256 classes, and you could define up to 256 classes within each policy if the same classes are reused for different policies. If you have two policies, the total number of classes from both policies should not exceed 256. If a policy includes Class-Based Weighted Fair Queueing (CBWFQ) (meaning it contains a bandwidth [or priority] statement within any of the classes), the total number of classes supported is 64.

In Cisco IOS versions 12.2(12),12.2(12)T, and 12.2(12)S, this limitation of 256 global class-maps was changed, and it is now possible to configure up to 1024 global class-maps and to use 256 class-maps inside the same policy-map.

### Q. How are routing updates and Point-to-Point Protocol (PPP) / High-Level Data Link Control (HDLC) keepalives processed when a service policy is applied?

**A.** Cisco IOS routers use the following two mechanisms to prioritize control packets:

- IP precedence
- pak_priority

Both mechanisms are designed to ensure that key control packets are not dropped or are dropped last by the router and the queuing system when an outbound interface is congested. For more information, refer to Understanding How Routing Updates and Control Packets Are Queued on an Interface with a QoS Service Policy.

**Q. Is Quality of Service (QoS) supported on interfaces configured with Integrated Routing and Bridging (IRB)?**

**A.** No. You cannot configure QoS featues when the interface is configured for IRB.

**Classification and Marking**

**Q. What is Quality of Service (QoS) pre-classification?**

**A.** QoS pre-classification enables you to match on and classify the original IP header contents of packets undergoing tunnel encapsulation and/or encryption. This feature does not describe the process of copying the original value of the Type of Service (ToS) byte from the original packet header to the tunnel header. For more information, refer to the following documents:

- Configuring QoS for Virtual Private Networks
- Quality of Service for Virtual Private Networks, 12.2(2)T Feature Module

**Q. Which packet header fields can be remarked? What values are available?**

**A.** The class-based marking feature allows you to set or mark the layer 2, layer 3 or Multiprotocol Label Switching (MPLS) header of your packets. For more information, refer to the following documents:

- Configuring Class-Based Packet Marking
- When Does a Router Set the CLP Bit in an ATM Cell?
- Configuring Packet Marking on Frame Relay PVCs

**Q. Can I prioritize traffic based on the URL?**

**A.** Yes. Network Based Application Recognition (NBAR) allows you to classify packets by matching on fields at the application layer. Prior to the introduction of NBAR, the most granular classification was layer 4 Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) port numbers. For more information, refer to the following documents:

- Network-Based Application Recognition Q&A
- NBAR Application Networking
- Using Network-Based Application Recognition and Access Control Lists for Blocking the Code Red Worm
- How to Protect Your Network Against the Nimda Virus

**Q. What platforms and Cisco IOS software versions support Network Based Application Recognition (NBAR)?**

**A.** Support for NBAR is introduced in the following versions of Cisco IOS software:

| Platform | Minimum Cisco IOS Software Version |
|---|---|
| 7200 | 12.1(5)T |
| 7100 | 12.1(5)T |
| 3660 | 12.1(5)T |
| 3640 | 12.1(5)T |
| 3620 | 12.1(5)T |
| 2600 | 12.1(5)T |
| 1700 | 12.2(2)T |

**Note:** You need to enable Cisco Express Forwarding (CEF) in order to use NBAR.

Distributed NBAR (DNBAR) is available on the following platforms:

| Platform | Minimum Cisco IOS Software Version |
|---|---|
| 7500 | 12.2(4)T, 12.1(6)E |
| FlexWAN | 12.1(6)E |

**Note:** NBAR is not supported on Catalyst 6000 Multilayer Switch Feature Card (MSFC) VLAN interfaces, the Cisco 12000 Series, or the Route Switch Module (RSM) for the Catalyst 5000 Series. If you do not see a particular platform listed above, please contact your Cisco technical representative.

### Queueing and Congestion Management

#### Q. What is the purpose of queueing?

**A.** Queueing is designed to accommodate temporary congestion on a network device's interface by storing excess packets in buffers until bandwidth becomes available. Cisco IOS routers support several queueing methods to meet the varying bandwidth, jitter, and delay requirements of different applications.

The default mechanism on most interfaces is First In First Out (FIFO). Some traffic types have more demanding delay/jitter requirements. Thus, one of the following alternative queueing mechanisms should be configured or is enabled by default:

- Weighted Fair Queueing (WFQ)
- Class-Based Weighted Fair Queueing (CBWFQ)
- Low Latency Queueing (LLQ), which is in fact CBWFQ with a Priority Queue (PQ) (known as PQCBWFQ)
- Priority Queueing (PQ)
- Custom Queueing (CQ)

Queueing generally happens on outbound interfaces only. A router queues packets that are going out an interface. You can police inbound traffic, but usually you cannot queue inbound (an exception is receive-side buffering on a Cisco 7500 Series router using distributed Cisco Express Forwarding (dCEF) to forward packets from the ingress to the egress interface; for more information, refer to Understanding VIP CPU Running at 99% and Rx-Side Buffering. On high-end distributed platforms such as the Cisco 7500 and 12000 Series, the inbound interface may use its own packet buffers to store excess traffic switched to a congested outbound interface following the inbound interface's switching decision. In rare conditions, typically when the inbound interface is feeding a slower outbound interface, the inbound interface can experience incrementing ignored errors when it runs out of packet memory. Excessive congestion can lead to output queue drops. Input queue drops have a different root cause most of the time . For more info on troubleshooting drops, refer to the following document:

- Troubleshooting Input Queue Drops and Output Queue Drops

For more information, refer to the following documents:

- Troubleshooting "Ignored" Errors on an ATM Port Adapter
- Troubleshooting Ignored Errors and No Memory Drops on the Cisco 12000 Series Internet Router

#### Q. How do Weighted Fair Queueing (WFQ) and Class Based Weighted Fair Queueing (CBWFQ) operate?

**A.** Fair queueing seeks to allocate a fair share of an interface's bandwidth among active conversations or IP flows. It classifies packets into subqueues, identified by a conversation identification number, using a hashing algorithm based on several fields of the IP header and the length of the packet. The following is how the weight is calculated:

- **W=K/(precedence +1)**

K= 4096 with Cisco IOS 12.0(4)T and earlier releases, and 32384 with 12.0(5)T and later releases.

The lower the weight, the higher the priority and the share of the bandwidth. In addition to the weight, the length of the packet is taken into account.

CBWFQ allows you to define a class of traffic and assign it a minimum bandwidth guarantee. The algorithm behind this mechanism is WFQ, which explains the name. To configure CBWFQ, you define specific classes in map-class statements. Then you assign a policy to each class in a policy-map. This policy-map will then be attached outbound to an interface. For more information, refer to the following documents:

- Understanding Class Based Weighted Fair Queuing on ATM
- Understanding Weighted Fair Queuing on ATM

#### Q. If a class in Class Based Weighted Fair Queueing (CBWFQ) is not using its bandwidth, can other classes use the bandwidth?

**A.** Yes. Although the bandwidth guarantees provided by issuing the **bandwidth** and **priority** commands have been described with words like "reserved" and "bandwidth to be set aside", neither command implements a true reservation. Meaning, if a traffic class is not using its configured bandwidth, any unused bandwidth is shared among the other classes.

The queueing system imposes an important exception to this rule with a priority class. As noted above, the offered load of a priority class is metered by a traffic policer. During congestion conditions, a priority class cannot use any excess bandwidth. For more information, refer to Comparing the bandwidth and priority Commands of a QoS Service Policy.

#### Q. Is Class Based Weighted Fair Queueing (CBWFQ) supported on subinterfaces?

**A.** Cisco IOS logical interfaces do not inherently support a state of congestion and do not support the direct application of a service policy that applies a queueing method. Instead, you first need to apply shaping to the subinterface using either Generic Traffic Shaping (GTS) or class-based shaping. For more information, refer to Applying QoS Features to Ethernet Subinterfaces.

### Q. What is the difference between the `priority` and `bandwidth` statements in a policy-map?

**A.** The **priority** and **bandwidth** commands differ in both functionality and in which applications they typically support. The following table summarizes these differences:

| Function | bandwidth Command | priority Command |
|---|---|---|
| Minimum bandwidth guarantee | Yes | Yes |
| Maximum bandwidth guarantee | No | Yes |
| Built-in policer | No | Yes |
| Provides low latency | No | Yes |

For more information, refer to Comparing the **bandwidth** and **priority** Commands of a QoS Service Policy.

### Q. How is the queue limit calculated on the FlexWAN and Versatile Interface Processors (VIP)?

**A.** Assuming sufficient SRAM on the VIP or FlexWAN, the queue limit is calculated based on a maximum delay of 500ms with average packet size of 250 bytes. The following is an example of a class with one Mbps of bandwidth:

Queue limit = 1000000 / (250 x 8 x 2) = 250

Smaller queue limits are assigned as the amount of available packet memory decreases and with a larger number of Virtual Circuits (VCS).

In the following example, a PA-A3 is installed in a FlexWAN card for the Cisco 7600 Series and is supporting multiple subinterfaces with 2 MB Permanent Virtual Circuits (PVCs). The service policy is applied to each VC.

```
class-map match-any XETRA-CLASS
  match access-group 104
class-map match-any SNA-CLASS
  match access-group 101
  match access-group 102
  match access-group 103
policy-map POLICY-2048Kbps
  class XETRA-CLASS
    bandwidth 320
  class SNA-CLASS
    bandwidth 512

interface ATM6/0/0
 no ip address
 no atm sonet ilmi-keepalive
 no ATM ilmi-keepalive
!
interface ATM6/0/0.11 point-to-point
 mtu 1578
 bandwidth 2048
 ip address 22.161.104.101 255.255.255.252
 pvc ABCD
  class-vc 2048Kbps-PVC
  service-policy out POLICY-2048Kbps
```

The Asynchronous Transfer Mode (ATM) interface gets a queue limit for the entire interface. The limit is a function of total available buffers, the number of physical interfaces on the FlexWAN, and the maximum queueing delay allowed on the interface. Each PVC gets a portion of the interface limit based on the PVC's Sustained Cell Rate (SCR) or Minimum Cell Rate (MCR), and each class gets a portion of the PVC limit based on its bandwidth allocation.

The following sample output of the **show policy-map interface** command is derived from a FlexWAN with 3687 global buffers. Issue the **show buffer** command to see this value. Each two Mbps PVC is allocated 50 packets based on the PVC bandwidth of two Mbps (2047/149760 x 3687 = 50). Each class is then allocated a portion of the 50, as shown in the following output:

```
service-policy output: POLICY-2048Kbps
    class-map: XETRA-CLASS (match-any)
      687569 packets, 835743045 bytes
```

```
    5 minute offered rate 48000 bps, drop rate 6000 BPS
    match: access-group 104
      687569 packets, 835743045 bytes
      5 minute rate 48000 BPS
    queue size 0, queue limit 7
    packets output 687668, packet drops 22
    tail/random drops 22, no buffer drops 0, other drops 0
    bandwidth: kbps 320, weight 15

  class-map: SNA-CLASS (match-any)
    2719163 packets, 469699994 bytes
    5 minute offered rate 14000 BPS, drop rate 0 BPS
    match: access-group 101
      1572388 packets, 229528571 bytes
      5 minute rate 14000 BPS
    match: access-group 102
      1146056 packets, 239926212 bytes
      5 minute rate 0 BPS
    match: access-group 103
      718 packets, 245211 bytes
      5 minute rate 0 BPS
    queue size 0, queue limit 12
    packets output 2719227, packet drops 0
    tail/random drops 0, no buffer drops 0, other drops 0
    bandwidth: kbps 512, weight 25
    queue-limit 100

  class-map: class-default (match-any)
    6526152 packets, 1302263701 bytes
    5 minute offered rate 44000 BPS, drop rate 0 BPS
    match: any
      6526152 packets, 1302263701 bytes
      5 minute rate 44000 BPS
    queue size 0, queue limit 29
    packets output 6526840, packet drops 259
    tail/random drops 259, no buffer drops 0, other drops 0
```

If your traffic streams use large packet sizes, the **show policy-map** interface command output may report an incrementing value for the `no buffer drops` field since you may run out of buffers before reaching the queue limit. In this case, try manually tuning down the queue-limit in non-priority classes. For more information, refer to Understanding the Transmit Queue Limit With IP to ATM CoS.

**Q. How do you verify the queue-limit value?**

**A.** On non-distributed platforms, the queue limit is 64 packets by default. The following example output was captured on a Cisco 3600 Series router:

```
november# show policy-map interface s0
   Serial0

   Service-policy output: policy1

     Class-map: class1 (match-all)
       0 packets, 0 bytes
       5 minute offered rate 0 BPS, drop rate 0 BPS
       Match: ip precedence 5
       Weighted Fair Queueing
         Output Queue: Conversation 265
         Bandwidth 30 (kbps) Max Threshold 64 (packets)

!--- Max Threshold is the queue-limit.

         (pkts matched/bytes matched) 0/0
         (depth/total drops/no-buffer drops) 0/0/0

       Class-map: class2 (match-all)
         0 packets, 0 bytes
         5 minute offered rate 0 BPS, drop rate 0 BPS
         Match: ip precedence 2
         Match: ip precedence 3
         Weighted Fair Queueing
           Output Queue: Conversation 266
           Bandwidth 24 (kbps) Max Threshold 64 (packets)
```

```
    (pkts matched/bytes matched) 0/0
    (depth/total drops/no-buffer drops) 0/0/0

  Class-map: class-default (match-any)
    0 packets, 0 bytes
    5 minute offered rate 0 BPS, drop rate 0 BPS
    Match: any
```

**Q. Can I enable fair queueing inside a class?**

**A.** The Cisco 7500 Series with distributed Quality of Service (QoS) supports fair queueing per class. Other platforms, including the Cisco 7200 Series and Cisco 2600/3600 Series, support Weighted Fair Queueing (WFQ) in the class-default class; all bandwidth classes use First In First Out (FIFO).

**Q. What commands can I use to monitor queueing?**

**A.** Use the following commands to monitor queueing:

- **show queue** *{interface}{interface number}* - On Cisco IOS platforms other than the Cisco 7500 Series, this command displays the active queues or conversations. If the interface or Virtual Circuit (VC) is not congested, no queues will be listed. On the Cisco 7500 Series, the **show queue** command is not supported.
- **show queueing interface interface-number [vc [[vpi/] vci]** - This displays the queueing statistics of an interface or a VC. Even when there is no congestion, you will still be able to see some hits here. The reason for this is that process switched packets are always counted regardless of congestion being present. Cisco Express Forwarding (CEF) and fast-switched packets are not being counted unless there is congestion. The legacy queueing mechanisms like Priority Queueing (PQ), Custom Queueing (CQ), and Weighted Fair Queueing (WFQ), will not provide classification statistics. Only modular Quality of Service Command Line Interface (MQC)-based features in images later than 12.0(5)T provide these statistics.
- **show policy interface** *{interface}{interface number}* - The `packets` counter counts the number of packets matching the criteria of the class. This counter increments whether or not the interface is congested. The `packets matched` counter indicates the number of packets matching the criteria of the class when the interface was congested. For more information on packet counters, refer to the following document:

  Understanding Packet Counters in show policy-map interface Output

- Cisco Class-Based QoS Configuration and Statistics MIB - Provides Simple Network Management Protocol (SNMP) monitoring capabilities.

**Q. RSVP can be used in conjunction with Class Based Weighted Fair Queueing (CBWFQ). When both Resource Reservation Protocol (RSVP) and CBWFQ are configured for an interface, do RSVP and CBWFQ act independently, exhibiting the same behavior that they would if each were running alone? RSVP seems to behave as if CBWFQ is not configured regarding bandwidth availability, assessment, and allocation.**

**A.** When using RSVP and CB-WFQ in Cisco IOS Software Release 12.1(5)T and later, the router can operate such that RSVP flows and CBWFQ classes share the available bandwidth on an interface or PVC, without oversubscription.

IOS Software Release 12.2(1)T and later, allows RSVP to do admission control using its own "ip rsvp bandwidth" pool, while CBWFQ does classification, policing, and scheduling of RSVP packets. This assumes premarked packets by the sender, and that non-RSVP packets are marked differently.

**Congestion Avoidance Weighted Random Early Detection (WRED)**

**Q. Can I enabled Weighted Random Early Detection (WRED) and Low Latency Queueing (LLQ), or Class Based Weighted Fair Queueing (CBWFQ) at the same time?**

**A.** Yes. Queueing defines the order of packets leaving a queue. This means, it defines a packet-scheduling mechanism. It also can be used to provide fair bandwidth allocation and minimum bandwidth guarantees. In contrast, Request for Comments (RFC) 2475 defines dropping as the "process of discarding packets based on specified rules." The default drop mechanism is tail drop, in which the interface drops packets when the queue is full. An alternative drop mechanism is Random Early Detection (RED) and Cisco's WRED, which begins dropping packets randomly before the queue is full and seeks to maintain a consistent average queue depth. WRED uses the IP precedence value of packets to make a differentiated drop decision. For more information, refer to Weighted Random Early Detection (WRED).

**Q. How can I monitor Weighted Random Early Detection (WRED) and see it actually taking effect?**

**A.** WRED monitors the average queue depth and begins to drop packets when the calculated value goes above the minimum threshold value. Issue the **show policy-map interface** command and monitor the mean queue depth value, as shown in the following example:

```
Router# show policy interface s2/1

  Serial2/1
```

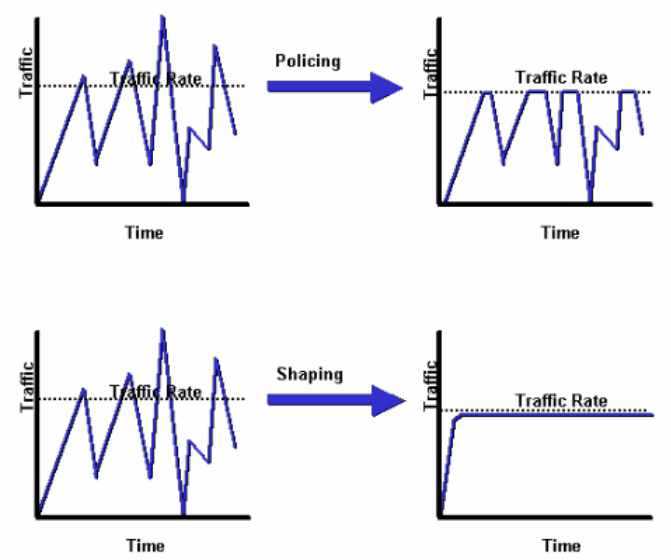
```
output : p1
 Class c1
  Weighted Fair Queueing
      Output Queue: Conversation 265
        Bandwidth 20 (%)
        (pkts matched/bytes matched) 168174/41370804
        (pkts discards/bytes discards/tail drops) 20438/5027748/0
        mean queue depth: 39

   Dscp     Random drop       Tail drop        Minimum   Maximum     Mark
   (Prec)    pkts/bytes        pkts/bytes      threshold threshold probability
     0(0)    2362/581052        1996/491016        20        40       1/10
     1          0/0                0/0             22        40       1/10
     2          0/0                0/0             24        40       1/10
     [output omitted]
```

### Policing and Shaping

**Q. What is the difference between policing and shaping?**

**A.** The following diagram illustrates the key difference. Traffic shaping retains excess packets in a queue and then schedules the excess for later transmission over increments of time. The result of traffic shaping is a smoothed packet output rate. In contrast, traffic policing propagates bursts. When the traffic rate reaches the configured maximum rate, excess traffic is dropped (or remarked). The result is an output rate that appears as a saw-tooth with crests and troughs.




For more information, refer to Policing and Shaping Overview.

**Q. What is a token bucket and how does the algorithm work?**

**A.** A token bucket itself has no discard or priority policy. The following is an example of how the token bucket metaphor works:

- Tokens are put into the bucket at a certain rate.
- Each token is permission for the source to send a certain number of bits.
- To send a packet, the traffic regulator must be able to remove from the bucket a number of tokens equal in representation to the packet size.
- If not enough tokens are in the bucket to send a packet, the packet either waits until the bucket has enough tokens (in the case of a shaper) or the packet is discarded or marked down (in the case of a policer).
- The bucket itself has a specified capacity. If the bucket fills to capacity, newly arriving tokens are discarded and are not available to future packets. Thus, at any time, the largest burst a source can send into the network is roughly proportional to the size of the bucket. A token bucket permits burstiness, but bounds it.

**Q. With a traffic policer such as class-based policing, what do Committed Burst (BC) and Excess Burst (Be) mean and how should I select these values?**

**A.** A traffic policer does not buffer excess packets and transmit them later, as is the case for a shaper. Instead, the policer executes a simple send or do not send policy without buffering. During periods of congestion, since you cannot buffer, the best you can do is drop packets less aggressively by properly configuring extended burst. Therefore, it is important to understand the policer uses the normal burst and extended burst values to ensure the configured Committed Information Rate (CIR) is reached.

The burst parameters are loosely modeled on the generic buffering rule for routers. The rule recommends configuring buffering equal to the round-trip time bitrate to accommodate the outstanding Transmission Control Protocol (TCP) windows of all connections in times of congestion.

The following table describes the purpose and the recommended formula for the normal and extended burst values:

| Burst Parameter | Purpose | Recommended Formula |
|---|---|---|
| normal burst | • Implements a standard token bucket.<br>• Sets the maximum size of the token bucket (although tokens can be borrowed if Be is greater than BC).<br>• Determines how large the token bucket can be since newly arriving tokens are discarded and are not available to future packets if the bucket fills to capacity. | **`CIR [BPS] * (1 byte)/(8 bits) * 1.5 seconds`**<br>**Note:** 1.5 seconds is the typical round trip time. |
| extended burst | • Implements a token bucket with extended burst capability.<br>• Disabled by setting BC = Be.<br>• When BC is equal to Be, the traffic regulator cannot borrow tokens and simply drops the packet when insufficient tokens are available. | **`2 * normal burst`** |

Not all platforms use or support the same range of values for a policer. Refer to the following document to learn the supported values for your specific platform:

- Policing and Shaping Overview

**Q. How does Committed Access Rate (CAR) or class-based policing decide if a packet conforms or exceeds the Committed Information Rate (CIR)? The router is dropping packets and reporting an exceeded rate even though the conformed rate is less than the configured CIR.**

**A.** A traffic policer uses the normal burst and extended burst values to ensure the configured CIR is reached. Setting sufficiently high burst values is important to ensuring good throughput. If the burst values are configured too low, the achieved rate may be much lower than the configured rate. Punishing temporary bursts can have a strong adverse impact on throughput of Transmission Control Protocol (TCP) traffic. With CAR, issue the **show interface rate-limit** command to monitor the current burst and determine whether the displayed value is consistently close to the limit (BC) and extended limit (Be) values.

```
rate-limit 256000 7500 7500 conform-action continue exceed-action drop
rate-limit 512000 7500 7500 conform-action continue exceed-action drop

router# show interfaces virtual-access 26 rate-limit
    Virtual-Access26 Cable Customers
      Input
        matches: all traffic
          params:  256000 BPS, 7500 limit, 7500 extended limit
          conformed 2248 packets, 257557 bytes; action: continue
          exceeded 35 packets, 22392 bytes; action: drop
          last packet: 156ms ago, current burst: 0 bytes
          last cleared 00:02:49 ago, conformed 12000 BPS, exceeded 1000 BPS
      Output
        matches: all traffic
          params:  512000 BPS, 7500 limit, 7500 extended limit
          conformed 3338 packets, 4115194 bytes; action: continue
          exceeded 565 packets, 797648 bytes; action: drop
          last packet: 188ms ago, current burst: 7392 bytes
          last cleared 00:02:49 ago, conformed 194000 BPS, exceeded 37000 BPS
```

For more information, refer to the following documents:

- Policing and Shaping Overview
- QoS Policing on the Catalyst 6000
- Quality of Service on Catalyst 4000 Frequently Asked Questions
- Catalyst G-L3 Series Switches and WS-X4232-L3 Layer 3 Modules QoS FAQ

### Q. Are the burst and queue limit independent of each other?

**A.** Yes, policer burst and queue limit are separate and independent of each other. You can view the policer as a gate that allows a certain number of packets (or bytes) and the queue as a bucket of size *queue limit* that holds the admitted packets prior to transmission on the network. Ideally, you want your bucket to be large enough to hold a *burst* of bytes/packets admitted by the gate (policer).

## Quality of Service (QoS) Frame Relay

### Q. What values should I select for Committed Information Rate (CIR), Committed Burst (BC), Excess Burst (Be), and Minimum CIR (MinCIR)?

**A.** Frame Relay Traffic Shaping, which you enable by issuing the **frame-relay traffic-shaping** command, supports several configurable parameters. These parameters include `frame-relay cir`, `frame-relay mincir`, and `frame-relay BC`. Refer to the following documents for more information on selecting these values and understanding related show commands:

- Configuring Frame Relay Traffic Shaping on 7200 Routers and Lower Platforms
- show Commands for Frame Relay Traffic Shaping
- VoIP over Frame Relay with Quality of Service (Fragmentation, Traffic Shaping, IP RTP Priority)

### Q. Does Priority Queueing on the Frame Relay main interface work in Cisco IOS 12.1?

**A.** Frame Relay interfaces support both interface queueing mechanisms and per-Virtual Circuit (VC) queueing mechanisms. As of Cisco IOS 12.0(4)T, the interface queue supports First In First Out (FIFO) or Per Interface Priority Queueing (PIPQ) only when you configure Frame Relay Traffic Shaping (FRTS). Therefore, the following configuration will no longer work if you upgrade to Cisco IOS 12.1.

```
interface Serial0/0
  frame-relay traffic-shaping
  bandwidth 256
  no ip address
  encapsulation frame-relay IETF
  priority-group 1

 !
 interface Serial0/0.1 point-to-point
```

```
 bandwidth 128
 ip address 136.238.91.214 255.255.255.252
 no ip mroute-cache
 traffic-shape rate 128000 7936 7936 1000
 traffic-shape adaptive 32000
 frame-relay interface-dlci 200 IETF
```

If FRTS is not enabled, you can apply an alternative queueing method, such as Class Based Weighted Fair Queueing (CBWFQ), to the main interface, which is acting like a single bandwidth pipe. In addition, as of Cisco IOS 12.1.1(T), you can enable Frame Relay Permanent Virtual Circuits (PVC) Priority Interface Queueing (PIPQ) on a Frame Relay main interface. You can define high, medium, normal, or low priority PVCs and issue the **frame-relay interface-queue priority** command on the main interface, as shown in the following example:

```
interface Serial3/0
 description framerelay main interface
 no ip address
 encapsulation frame-relay
 no ip mroute-cache
 frame-relay traffic-shaping
 frame-relay interface-queue priority

interface Serial3/0.103 point-to-point
 description frame-relay subinterface
 ip address 1.1.1.1 255.255.255.252
 frame-relay interface-dlci 103
  class frameclass

map-class frame-relay frameclass
 frame-relay adaptive-shaping becn
 frame-relay cir 60800
 frame-relay BC 7600
 frame-relay be 22800
 frame-relay mincir 8000
 service-policy output queueingpolicy
 frame-relay interface-queue priority low
```

**Q. Does Frame Relay Traffic Shaping (FRTS) work with Distributed Cisco Express Forwarding (dCEF) and Distributed Class Based Weighted Fair Queueing (dCBWFQ)?**

**A.** As of Cisco IOS 12.1(5)T, only the distributed version of QoS features are supported on VIPs in the Cisco 7500 Series. To enable traffic shaping on Frame Relay interfaces, use Distributed Traffic Shaping (DTS). For more information, refer to the following documents:

- Versatile Interface Processor-Based Distributed FRF.11 and FRF.12 for Cisco IOS Release 12.1 T
- Frame Relay Traffic Shaping With Distributed QoS on the Cisco 7500 Series

**Quality of Service (QoS) Over Asynchronous Transfer Mode (ATM)**

**Q. Where do I apply a service policy with Class Based Weighted Fair Queueing (CBWFQ) and Low Latency Queueing (LLQ) on an Asynchronous Transfer Mode (ATM) interface?**

**A.** As of Cisco IOS 12.2, ATM interfaces support service policies at three levels or logical interfaces: main interface, subinterface, and Permanent Virtual Circuit (PVC). Where you apply the policy is a function of the Quality of Service (QoS) feature that you are enabling. Queueing policies should be applied per Virtual Circuit (VC) since the ATM interface monitors the congestion level per VC, and maintains queues for excess packets per VC. For more information, refer to the following documents:

- Where Do I Apply a QoS Service Policy on an ATM Interface?
- Understanding Per-VC Transmit Queuing on the PA-A3 and NM-1A ATM Interfaces

**Q. What bytes are counted by IP to Asynchronous Transfer Mode (ATM) Class of Service (COs) queueing?**

**A.** The bandwidth and priority commands configured in a service policy to enable class-based weighted fair queueing (CBWFQ) and low latency queueing (LLQ), respectively, use a Kbps value that counts the same overhead bytes as are counted by the show interface command output. Specifically, the layer 3 queueing system counts Logical Link Control / Subnetwork Access Protocol (LLC/SNAP). It does not count the following:

- ATM Adaptation Layer 5 (AAL5) Trailer
- Padding to make last cell an even multiple of 48 bytes
- Five-byte cell header

- What Bytes Are Counted by IP to ATM COs Queueing?

**Q. How many Virtual Circuits (VCS) can support a service policy simultaneously?**

**A.** The following document provides useful guidelines on the number of Asynchronous Transfer Mode (ATM) VCS that can support. About 200 to 300 VBR-nrt Permanent Virtual Circuits (PVCs) have been safely deployed:

- IP to ATM Class of Service Design Guide

Also consider the following:

- Use a powerful processor capable. For example, a VIP4-80 provides significantly higher performance than a VIP2-50.
- Amount of available packe memory. On the NPE-400, up to 32 MB (in a system with 256 MB) is set aside for packet buffer. For an NPE-200, up to 16 MB is set aside for packet buffers on a system with 128 MB.
- Configurations with per-VC Weighted Random Early Detection (WRED) operating simultaneously on up to 200 ATM PVCs have been extensively tested. The amount of packet memory on the VIP2-50 that can be used for the per-VC queues is finite. For instance, a VIP2-50 with 8-MB SRAM provides 1085 packet buffers available for IP to ATM COs per-VC queueing on which WRED operates. If 100 ATM PVCs were configured and if all VCS were experiencing excessive congestion simultaneously (as could be simulated in test environments where non-TCP flow controlled source would be used), then on average each PVC would have about 10 packets worth of buffering, which may be too short for WRED to operate successfully. VIP2-50 devices with large SRAM are thus strongly recommended in designs with a high number of ATM PVCs running per-VC WRED and that can experience congestion simultaneously.
- The higher the number of configured active PVCs, the lower their Sustained Cell Rate (SCR) would need to be, and consequently the shorter the queue required by WRED to operate on the PVC. Thus, as is the case when using the default WRED profiles of the IP to ATM Class of Service (COs) Phase 1 feature, configuring lower WRED drop thresholds when per-VC WRED is activated on a very large number of low-speed congested ATM PVCs would minimize the risk of buffer shortage on the VIP. Buffer shortage on the VIP does not result in any malfunction. In the case of buffer shortage on the VIP, the IP to ATM COs Phase 1 feature simply degrades to First In First Out (FIFO) tail drop during the period of buffer shortage (that is, the same drop policy that would occur if the IP to ATM COs feature were not activated on this PVC).
- Maximum number of simultaneous VCS that can be reasonably supported.

**Q. Which Asynchronous Transfer Mode (ATM) hardware supports IP to ATM Class of Service (COs) features including Class Based Weighted Fair Queueing (CBWFQ) and Low Latency Queueing (LLQ)?**

**A.** IP to ATM COs refers to a set of features that are enabled on a per-Virtual Circuit (VC) basis. Given this definition, IP to ATM COs is not supported on the ATM Interface Processor (AIP), PA-A1 or 4500 ATM network processors. This ATM hardware does not support per-VC queueing as the PA-A3 and most network modules (other than the ATM-25) define it. For more information, refer to the following document:

- Understanding ATM Hardware Support for IP to ATM COs
- Per-VC Class-Based, Weighted Fair Queuing on RSP-Based Platforms
- Per-VC Class-Based, Weighted Fair Queuing (Per-VC CBWFQ) on the Cisco 7200, 3600, and 2600 Routers
- Per-VC Queuing on the PA-A3-8T1/E1 IMA ATM Port Adapter
- Configuring ATM Per-VC Queueing on the MC3810

**Voice and Quality of Service (QoS)**

**Q. How does Link Fragmentation and Interleaving (LFI) work?**

**A.** Interactive traffic such as Telnet and Voice over IP is susceptible to increased latency when the network processes large packets such as File Transfer Protocol (FTP) transfers over a WAN. Packet delay for interactive traffic is significant when the FTP packets are queued on slower WAN links. A method was devised for fragmenting larger packets, and queueing the smaller (voice) packets between the fragments of the larger packets (FTP) packets. Cisco IOS routers support several layer 2 fragmentation mechanisms. For more information, refer to the following documents:

- Link Efficiency Mechanisms Overview
- VoIP over Frame Relay with Quality of Service (Fragmentation, Traffic Shaping, IP RTP Priority)
- VoIP over PPP Links with Quality of Service (LLQ / IP RTP Priority, LFI, cRTP)

**Q. What tools can I use to monitor Voice over IP performance?**

**A.** Cisco currently offers several options for monitoring Quality of Service (QoS) in networks using Cisco's Voice over IP solutions. These solutions do not measure voice quality using Perceptual Speech

Quality Measurement (PSQM) or some of the new proposed algorithms for voice quality measurement. Tools from Agilent (HP) and NetIQ are available for this purpose. However, Cisco does offer tools that provide some idea of the voice quality you are experiencing by measuring delay, jitter and packet loss. For more information, refer to Using Cisco Service Assurance Agent and Internetwork Performance Monitor to Manage Quality of Service in Voice over IP Networks.

**Q. %SW_MGR-3-CM_ERROR_FEATURE_CLASS: Connection Manager Feature Error: Class SSS: (QoS) - install error, ignore.**

**A.** The feature install error observed is an expected behavior when an invalid configuration is applied to a template. It indicates that the service policy was not applied due to a conflict. In general, you should not configure shaping on class-default of the child policy in hierarchical policy maps, instead configure it on the parent policy of the interface. This message along with the traceback is printed out as a consequence.

With session based policies, shaping on class-default has to be done only at the sub-interface or PVC level. Shaping at the physical interface is not supported. If the configuration is done on the physical interface, the occurance of this error message is an expected behavior.

In caseof LNS, another reason might be that the service policy could be provisioned via the radius server when the sessions are brought up. Issue the **show tech** command in order to view the radius server configuration and in order to view any illegal service policies that are installed via the radius server when the session comes up or flaps.

## Related Information

- **Performance Tuning Basics**
- **Quality of Service (QoS) Support**
- **Technical Support & Documentation - Cisco Systems**

---

Updated: Jun 04, 2009 Document ID: 22833

---

Contacts | Feedback | Help | Site Map
© 1992-2010 Cisco Systems, Inc. All rights reserved. Terms & Conditions | Privacy Statement | Cookie Policy | Trademarks of Cisco Systems, Inc.