

# Chapter 27

## CoS Overview

For interfaces that carry IPv4 or MPLS traffic, you can configure JUNOS class-of-service (CoS) features to provide multiple classes of service for different applications. On the router, you can configure multiple output queues for transmitting packets, define which packets are placed into each output queue, schedule the transmission service level for each queue using a weighted round-robin algorithm, and manage congestion using a Random Early Detection (RED) algorithm.

The JUNOS CoS features provide a set of mechanisms that you can use to provide differentiated services when best-effort traffic delivery is insufficient. In designing CoS applications, you must give careful consideration to your service needs, and you must thoroughly plan and design your CoS configuration to ensure consistency across all routers in a CoS domain. You must consider all the routers and other networking equipment in the CoS domain to ensure interoperability among all equipment.

The Internet community has little experience with CoS and quality of service (QoS). However, because Juniper Networks routers implement CoS in hardware rather than in software, you have the ability to experiment with and deploy CoS features without any adverse affects on packet forwarding and routing.

## CoS Applications

CoS mechanisms are useful for two broad classes of applications. These applications can be referred to as *in the box* and *across the network*.

In-the-box applications use CoS mechanisms to provide special treatment for packets passing through a single node on the network. You can monitor the incoming traffic on each interface, using CoS to provide preferred service to some interfaces (that is, to some customers) while limiting the service provided to other interfaces. You can also filter outgoing traffic by the packet's destination, thus providing preferred service to some destinations.

Across-the-network applications use CoS mechanisms to provide differentiated treatment to different classes of packets across a set of nodes in a network. In these types of applications, you typically control the ingress and egress routers to a routing domain and all the routers within the domain. You can use JUNOS CoS features to modify packets traveling through the domain to indicate the packet's priority across the domain. Specifically, you modify the precedence bits in the IPv4 type-of-service (ToS) field, remapping these bits to values that correspond to levels of service. When all routers in the domain are configured to associate the precedence bits with specific service levels, packets traveling across the domain receive the same level of service from the ingress point to the egress point. For CoS to work in this case, the mapping between the precedence bits and service levels must be identical across all routers in the domain.

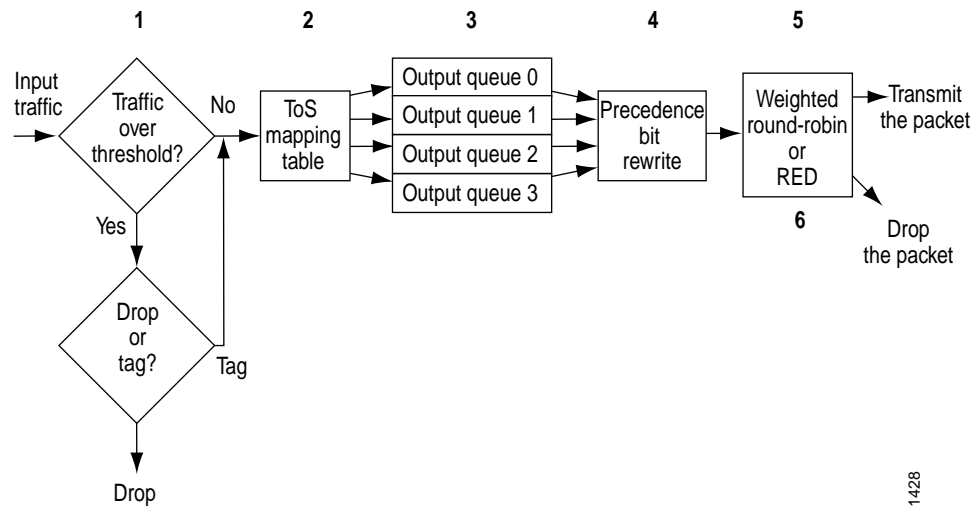
## JUNOS CoS Components

On Juniper Networks routers, a CoS mechanism is part of the basic packet-forwarding operation. By default, the router places all output traffic into one FIFO queue, and it relies on a weighted round-robin process to transmit packets and on RED to drop packets from the head of the transmission queue when congestion occurs. To modify the default CoS mechanism by adding multiple output transmission queues and fine-tuning the RED behavior, configure the JUNOS CoS features.

Figure 15 shows the components of the JUNOS CoS features, illustrating the sequence in which they interact. The components are discussed in this section:

- Hardware Monitoring of Input Traffic for Congestion (item 1 in the figure)
- Input Traffic Classification (item 2)
- Output Queues (item 3)
- Precedence Bit Rewriting (item 4)
- Servicing of Output Queues and Transmitting Output Traffic (item 5)
- Monitoring of Output Queue Congestion and Dropping Packets (item 6)

Figure 15: JUNOS CoS Components



1428

## **Hardware Monitoring of Input Traffic for Congestion**

The media-specific ASIC on each router PIC checks all input traffic levels against the link bandwidth that is configured using the leaky bucket algorithm (item 1 in Figure 15). If the flow exceeds the bucket's threshold, the packets are either dropped or tagged, depending on how the receive leaky bucket is configured. If it is configured to tag packets, the PIC sets the packet-loss priority (PLP) bit in the notification record associated with the packet to indicate that the packet encountered congestion in the router. It also indicates that the packet should have a greater probability of being dropped from the output transmission queue.

For more information on configuring the PLP bit and how this affects downstream processing, see "Rewrite the IP Precedence Bits" on page 334.

## **Input Traffic Classification**

Before being placed on a transmission queue, packets are classified based on the value of the precedence bits in the IP ToS field (item 2 in Figure 15). The priority level represents the service level to apply to the packet and corresponds to an output transmission queue. Each link can have up to four output transmission queues, and each FPC can have up to 64 queues.

For IPv4 traffic, you can map the ToS values to a priority level or you can map an input interface to a priority level.

For MPLS traffic, the mapping is static.

The least-significant bit of the precedence bits in the IPv4 ToS field is also checked to determine the status of the packet's PLP bit. (In big endian terminology, this is bit 2 of the ToS. In little endian terminology, this is bit 5.) For IPv4 traffic, if this bit is set, the PIC sets the packet's PLP bit. For MPLS traffic, if this bit is set, the low-order bit in the MPLS header's CoS field is set to 1.

## **Output Queues**

Each stream has four output transmission queues (item 3 in Figure 15), and each queue receives a configurable percentage of the stream's total available queue buffer size.

## **Precedence Bit Rewriting**

After packets are received and classified, you can rewrite the IP precedence bits in a packet's IP header (item 4 in Figure 15). For each output transmission queue, you rewrite the IP precedence bits in the IP headers of all packets headed for that queue based on whether the PLP bit is set for the packet. You can set one precedence bit value for all packets whose PLP bit is set and a second value for those whose PLP bit is not set.

## ***Servicing of Output Queues and Transmitting Output Traffic***

A weighted round-robin scheme determines the queue from which the next packet is transmitted (item 5 in Figure 15). The weighting is based on the amount of bandwidth allocated for each queue. The percentage of weight allocated to each output transmission queue determines how often weighted round-robin services the queue, with higher percentages resulting in more frequent service. For example, an output queue that is allocated 50 percent of the weight is serviced twice as often as one that is allocated 25 percent.

## ***Monitoring of Output Queue Congestion and Dropping Packets***

While weighted round-robin transmits packets from the output queues, RED attempts to manage transient and sustained congestion (item 6 in Figure 15). RED tries to anticipate incipient congestion and reacts by dropping a small percentage of packets from the head of the queue to ensure that a queue never actually becomes congested. RED examines the fullness of each output transmission queue to determine whether it is congested. It uses two values to determine whether a output transmission queue is congested:

Output queue fullness—RED calculates the fullness of the output transmission queue by dividing the buffer being used by the total buffer allocated for that queue.

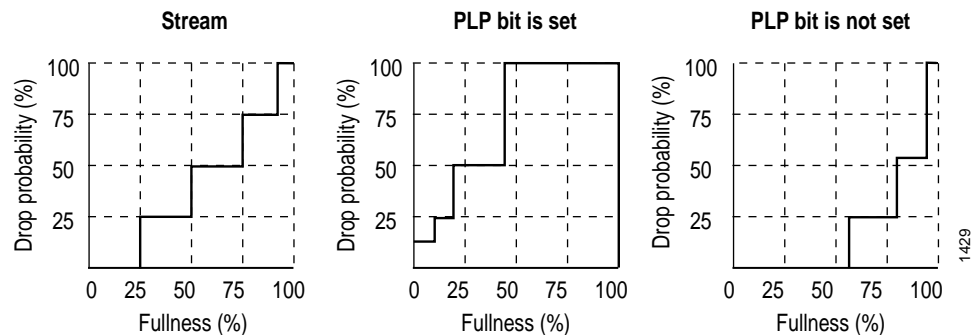
Drop profile—You specify the drop probabilities for different levels of buffer fullness.

RED uses two drop profiles to determine whether to drop a packet. One drop profile is applied to the entire traffic stream passing out through a link. The second profile is applied to individual output transmission queues depending on whether the queue is congested. A packet must pass both the stream and queue profile tests before being dropped by RED. For each output transmission queue, there are two drop profiles for queue congestion: one queue for packets whose PLP is set and one for packets whose PLP bit is not set.

Generally, you more aggressively drop packets in which the PLP bit is set because they have experienced congestion, a likely indication that there is congestion between the packet's source and the local router. When the sender discovers that the packet has been dropped (because it does not receive an acknowledgment from the destination), it throttles the rate at which it sends packets, providing some relief to the congestion on the local router.

Figure 16 shows three drop profiles—mappings between queue fullness and drop probabilities for a link's stream (on the left), packets in which the PLP bit is set (in the middle), and packets in which the PLP bit is not set (on the right). With these drop profiles, if the stream is 100 percent full and if the queue is 50 percent full, a non-PLP packet is never dropped (it matches the stream profile, but not the non-PLP profile), and a PLP packet is always dropped (it matches both the stream and PLP set profiles and in the PLP profile has a 100-percent drop probability).

Figure 16: RED Drop Profiles



To randomize the drop event, RED generates a random number for the packet in the queue and plots this number to the Y axis of the drop profile graphs shown in Figure 16. If, at the stream's or queue's congestion level, the drop probability is greater than the random number, the drop decision for the profile is taken.

The stream, PLP, and non-PLP drop profiles apply to all the output transmission queues on an FPC.

The queue fullness is the percentage of the total buffer assigned to an output transmission queue. A queue that has only 20 percent of the total buffer becomes full faster than one with 60 percent of the total buffer if an equal amount of traffic is being placed in both queues.

## CoS Packet Processing

Figure 17 illustrates the sequence in which the JUNOS CoS mechanisms process packets, showing five major steps in the process:

**Incoming classification**—When a packet is received by the router, it is classified and mapped to an output transmission queue. IPv4 packets can be classified in one of two ways:

Based on the precedence bits in the IP packet header (IPv4 packets only)—The precedence bits comprise the first three bits in the eight-bit IP ToS field, so there can be eight different precedence values (see Table 15). You map the eight precedence bit values to the four output transmission queues available on the link.

Based on the input interface (IPv4 packets only)—You can assign all packets received on an interface to one of the four output transmission queues on the link.

By default, MPLS packets are placed in output transmission queue 0. You can map MPLS packets statically based on the three experimental CoS bits in the MPLS label.

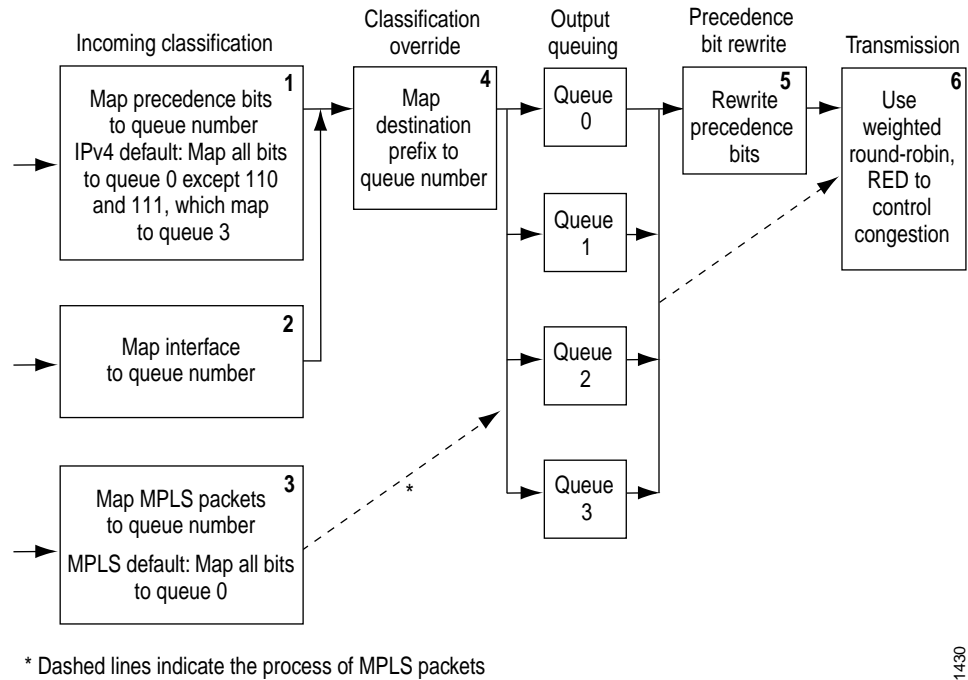
**Classification override**—For IPv4 packets, you can override the incoming classification, mapping packets to output transmission queues based on the packet's IP destination address.

**Output queuing**—Packets are placed in one of the four output transmission queues awaiting transmission.

**Precedence bit rewriting**—For IPv4 packets, you can rewrite the precedence bits in the packet's IP packet header.

Transmission—A weighted round-robin algorithm is used to schedule packets for transmission. It alternates among the four queues depending on the queues' contents and weighted round-robin credit value and transmits packets from the head of the queue. At the same time, RED attempts to manage transient and sustained congestion, dropping packets as needed from the head of the queue to control congestion.

Figure 17: Sequential Packet Processing Performed by CoS Mechanisms



1430

Table 15: IP ToS Field Precedence Values

Precedence Value (Lowest to Highest)	Description
000	Routine
001	Priority
010	Immediate
011	Flash
100	Flash override
101	Critical/ECP
110	Internetwork control
111	Network control