

Chapter 35

CoS Overview

For interfaces that carry IPv4, IPv6, or MPLS traffic, you can configure JUNOS class-of-service (CoS) features to provide multiple classes of service for different applications. On the routing platform, you can configure multiple forwarding classes for transmitting packets, define which packets are placed into each output queue, schedule the transmission service level for each queue, and manage congestion using a Random Early Detection (RED) algorithm.



NOTE: JUNOS CoS features are not supported on ATM interfaces. ATM has traffic-shaping capabilities that would override CoS, because ATM traffic shaping is performed at the ATM layer and CoS is performed at the IP layer. For more information about ATM traffic shaping, see “Defining the ATM Traffic-Shaping Profile” on page 214 and “Configuring ATM2 IQ VC Tunnel CoS Components” on page 236.

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 routing platforms in a CoS domain. You must also consider all the routing platforms 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 routing platforms implement CoS in hardware rather than in software, you can experiment with and deploy CoS features without adversely affecting packet forwarding and routing performance.

The standards are defined in the following RFCs:

RFC 2474, *Definition of the Differentiated Services Field in the IPv4 and IPv6 Headers*

RFC 2597, *Assured Forwarding PHB Group*

RFC 2598, *An Expedited Forwarding PHB*

This chapter discusses the following topics:

Non-CoS-Configurable Interfaces on page 800

CoS Applications on page 801

JUNOS CoS Components on page 802

Hardware Capabilities and Limitations on page 811

For information about CoS components that you apply to the ATM2 intelligent queuing (IQ) interface specifically, see “Configuring ATM2 IQ VC Tunnel CoS Components” on page 236.

Non-CoS-Configurable Interfaces

You can configure CoS on all interfaces, except the following. For channelized interfaces, you can configure CoS on channels, but not at the controller level.

ae—Aggregated Ethernet interface.

as—Aggregated SONET/SDH interface.

cau4—Channelized STM1 IQ interface (configured on the Channelized STM1 IQ PIC).

coc1—Channelized OC1 IQ interface (configured on the Channelized OC12 IQ PIC).

coc12—Channelized OC12 IQ interface (configured on the Channelized OC12 IQ PIC).

cstm-1—Channelized STM1 IQ interface (configured on the Channelized STM1 IQ PIC).

ct1—Channelized T1 IQ interface (configured on the Channelized DS3 IQ PIC or Channelized OC12 IQ PIC).

ct3—Channelized T3 IQ interface (configured on the Channelized DS3 IQ PIC or Channelized OC12 IQ PIC).

ce1—Channelized E1 IQ interface (configured on the Channelized E1 IQ PIC or Channelized STM1 IQ PIC).

dsc—Discard interface.

fxp—Management and internal Ethernet interfaces.

gr—Generic routing encapsulation tunnel interface.

ip—IP-over-IP encapsulation tunnel interface.

lo—Loopback interface. This interface is internally generated.

mt—Multicast tunnel interface (internal routing platform interface for VPNs).

pe—Encapsulates packets destined for the rendezvous point routing platform. This interface is present on the first-hop routing platform.

pd—De-encapsulates packets at the rendezvous point. This interface is present on the rendezvous point.

vt—Virtual loopback tunnel interface.



NOTE: For original Channelized OC12 PICs, limited CoS functionality is supported. For more information, contact Juniper Networks customer support.

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 routing platforms to a routing domain and all the routing platforms 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 routing platforms 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 routing platforms in the domain.

JUNOS CoS applications support the following range of mechanisms:

Differentiated Services—The CoS application supports DiffServ as well as six-bit IPv4 and IPv6 header ToS byte settings. The configuration uses DiffServ code points (DSCPs) in the IP and IPv6 ToS fields to determine the forwarding class associated with each packet.

Layer 2 to Layer 3 CoS Mapping—The CoS application supports mapping of Layer 2 (IEEE 802.1p) packet headers to routing platform forwarding class and loss-priority values.

Layer 2 to Layer 3 CoS mapping involves setting the forwarding class and loss priority based on information in the Layer 2 header. Output involves mapping the forwarding class and loss priority to a Layer 2-specific marking. You can mark the Layer 2 and Layer 3 headers simultaneously.

MPLS EXP—Supports configuration of mapping of MPLS experimental (EXP) bit settings to routing platform forwarding classes and vice versa.

VPN Outer-Label Marking—Supports setting of outer-label EXP bits, also known as CoS bits, based on MPLS EXP mapping.

JUNOS CoS Components

You can configure CoS features to meet your application needs. Because the components are generic, you can use a single CoS configuration syntax across multiple platforms. The JUNOS CoS features include:

Classifiers—Allow you to associate incoming packets with a forwarding class and loss priority and, based on the associated forwarding class, assign packets to output queues. Two general types of classifiers are supported:

Behavior aggregate (BA) or code point traffic classifiers—Code points determine each packet's forwarding class and loss priority. BA classifiers allow you to set the forwarding class and loss priority of a packet based on DiffServ code point (DSCP) bits, DSCP IPv6, IP precedence bits, MPLS EXP bits, and IEEE 802.1p bits. The default classifier is based on IP precedence bits.

Multifield (MF) traffic classifiers—Allow you to set the forwarding class and loss priority of a packet based on firewall filter rules. For more information about configuring MF classifiers, see the *JUNOS Policy Framework Configuration Guide*.

Forwarding classes—Also known as ordered aggregates in the IETF's DiffServ architecture. Affect the forwarding, scheduling, and marking policies applied to packets as they transit a routing platform. The forwarding class plus the loss priority define the per-hop behavior. Four categories of forwarding class are supported: best effort, assured forwarding, expedited forwarding, and network control. For M-series routing platforms, four forwarding classes are supported; you can configure up to one each of the four types of forwarding class. For M320 and T-series platforms, eight forwarding classes are supported, thus allowing you to classify packets more granularly. For example, you can configure multiple classes of EF traffic: EF, EF1, and EF2.

Loss priorities—Allow you to set the priority of dropping a packet. Typically you mark packets exceeding some service level with a high loss priority. Loss priority affects the scheduling of a packet without affecting the packet's relative ordering. You set loss priority by configuring a classifier or a policer.

Forwarding policy options—Allow you to associate forwarding classes with next hops. Forwarding policy also allows you to create classification overrides, which assign forwarding classes to sets of prefixes.

Transmission scheduling and rate control—Provide you with a variety of tools to manage traffic flows:

Schedulers—Allow you to define the priority, bandwidth, delay buffer size, rate control status, and RED drop profiles to be applied to a particular forwarding class for packet transmission.

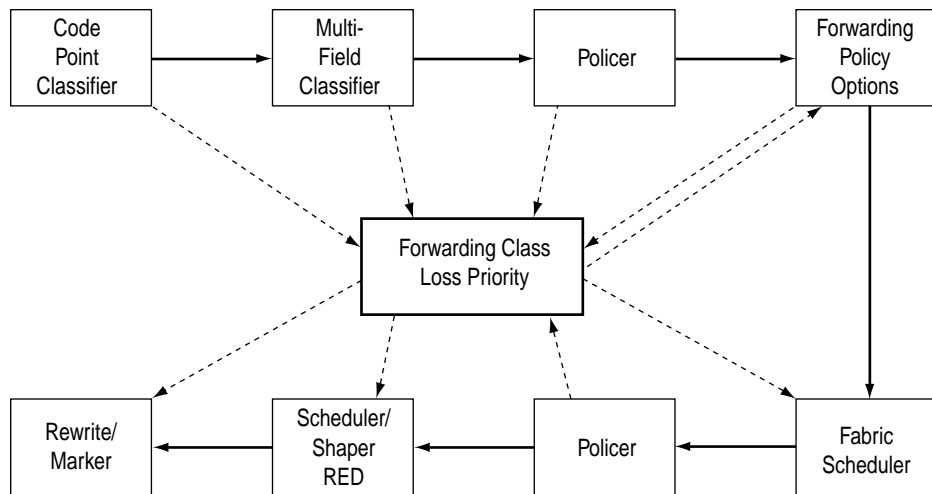
Fabric schedulers—For M320 and T-series platforms only, fabric schedulers allow you to identify a packet as high or low priority based on its forwarding class, and to associate schedulers with the fabric priorities.

Policers for traffic classes—Allow you to limit traffic of a certain class to a specified bandwidth and burst size. Packets exceeding the policer limits can be discarded, or can be assigned to a different forwarding class or to a different loss priority, or to both. You define policers with filters that can be associated with input or output interfaces. For information about configuring policers, see the *JUNOS Policy Framework Configuration Guide*.

Rewrite markers—Allow you to redefine the code-point value of outgoing packets. Rewriting or *marking* outbound packets is useful when the routing platform is at the border of a network and must alter the code points to meet the policies of the targeted peer.

Figure 36 shows the components of the JUNOS CoS features, illustrating the sequence in which they interact.

Figure 36: Packet Flow Through CoS Configurable Components



1750

The JUNOS CoS components are discussed in the following sections:

CoS Inputs and Outputs on page 804

Traffic Classifiers on page 805

Forwarding Classes on page 807

Transmission Scheduling and Rate Control on page 807

Rewrite Markers on page 810

CoS Inputs and Outputs

Some CoS components map one set of values to another set of values. Each mapping contains one or more inputs and one or more outputs. When you configure a mapping, you set the outputs for a given set of inputs, as shown in Table 45.

Table 45: CoS Mappings—Inputs and Outputs

CoS Mappings	Inputs	Outputs	Comments
drop-profile-map	loss-priority protocol	drop-profile	The map sets the drop profile for a specific PLP and protocol type. See “Configuring the Scheduler Drop Profile” on page 838.
rewrite-rules	forwarding-class loss-priority	code-points	The map sets the code-point aliases and bit patterns for a specific forwarding class and PLP. See “Rewriting Packet Header Information” on page 854.
classifiers	code-points	forwarding-class loss-priority	The map sets the forwarding class and PLP for a specific set of code-point aliases and bit patterns. See “Classifying Packets by Behavior Aggregate” on page 830.

Traffic Classifiers

By default, all logical interfaces are assigned an IP precedence *classifier* for incoming IP packets. For all PICs except PICs mounted on non-enhanced M-series FPCs, if you enable the MPLS protocol family on a logical interface, the default MPLS EXP classifier is applied to that logical interface.

At the core router, the JUNOS software matches the classifier to a *code point* to determine each packet's forwarding class and loss priority. This classifier is called the behavior aggregate (BA) classifier. Supported code points include the DiffServ code point (DSCP) for IP DiffServ, DSCP for IPv6 DiffServ, IP precedence bits, MPLS EXP bits, and IEEE 802.1p CoS bits.

In an edge router, a multifield (MF) *classifier* provides the filtering functionality that scans through a variety of packet fields to determine the forwarding class for a packet. Typically, a classifier performs matching operations on the selected fields against a configured value.

For M-series routing platforms, only four classes can forward traffic independently. For M320 and T-series platforms, only eight classes can forward traffic independently. Therefore, you must configure additional classes to be aggregated into one of these classes. You use the BA classifier to configure class aggregation. For more information, see "Forwarding Classes" on page 807.

We do not recommend classifying packets into a forwarding class that has no associated scheduler on the egress interface. Such a configuration can cause unnecessary packet drops because an unconfigured scheduling class might lack adequate buffer space. For example, if you configure a custom scheduler map that does not define queue 0, and the default classifier assigns incoming packets to the best-effort class (queue 0), the unconfigured egress queue for the best-effort forwarding class might not have enough space to accommodate even short packet bursts.

The following sections discuss classifiers in more detail:

- Default Classifier on page 805

- Behavior Aggregate Classifier on page 806

- Multifield Classifier on page 806

Default Classifier

When you install a classifier, it becomes effective on any interface for which you configure it.

By default, all logical interfaces are assigned an IP precedence classifier. For all PICs except PICs mounted on non-enhanced M-series FPCs, if you enable the MPLS protocol family on a logical interface, the default MPLS EXP classifier is applied to that logical interface.

The default IP precedence classifier maps IP precedence bits to forwarding classes and loss priorities as shown in Table 46.

Table 46: Default IP Precedence Classifier

IP Precedence Code Point	Forwarding Class	Loss Priority
000	best-effort	low
001	best-effort	high
010	best-effort	low
011	best-effort	high
100	best-effort	low
101	best-effort	high
110	network-control	low
111	network-control	high

Behavior Aggregate Classifier

A behavior aggregate classifier uses IP DSCPs, IPv6 DSCPs, IP precedence bits, the MPLS EXP field, or Layer 2 CoS indication (IEEE 802.1p) to determine the forwarding treatment for each packet, called a per-hop behavior (PHB). A PHB defines how a particular routing platform in a DiffServ domain treats a packet. A BA classifier can aggregate multiple DiffServ PHBs into a single one if the routing platform cannot support multiple simultaneous PHBs.

The BA classifier maps a code point to a loss priority. The loss priority is used later in the work flow to select one of the two drop profiles used by RED.

Decoding the EXP header field can also determine the packet loss priority (PLP) status.



NOTE: For a specified interface, you can configure both an MF classifier and a BA classifier without conflicts. Because the classifiers are always applied in sequential order, the BA classifier followed by the MF classifier, any BA classification result is overridden by an MF classifier if they conflict.

For information about configuring BA classifiers, see “Classifying Packets by Behavior Aggregate” on page 830 and “Examples: Configuring Class of Service” on page 873.

Multifield Classifier

A multifield classifier examines one or more packet fields to determine the forwarding treatment that a packet receives. An MF classifier typically matches one or more of the six packet header fields: destination address, source address, IP protocol, source port, destination port, and DSCP. MF classifiers are used when a simple BA classifier is insufficient to classify a packet.

From a CoS perspective, MF classifiers (or firewall filter rules) provide the following services:

- Classify packets to a forwarding class and loss priority.

Police traffic to a specific bandwidth and burst size. Packets exceeding the policer limits can be discarded, or can be assigned to a different forwarding class or to a different loss priority, or to both.

To activate an MF classifier, you must configure it on a logical interface. There is no restriction on the number of MF classifiers you can configure.

For information about configuring MF classifiers, see the *JUNOS Policy Framework Configuration Guide*.

Forwarding Classes

For a classifier to assign an output queue to each packet, it must associate the packet with one of the following forwarding classes:

Expedited Forwarding (EF)—Provides a low loss, low latency, low jitter, assured bandwidth, end-to-end service.

Assured Forwarding (AF)—Provides a group of values you can define and includes four subclasses, AF1, AF2, AF3, and AF4, each with three drop probabilities, low, medium, and high.

Best Effort (BE)—Provides no service profile. For the BE forwarding class, loss priority is typically not carried in a code point and RED drop profiles are more aggressive.

Network Control (NC)—The NC forwarding class is typically high priority because it supports protocol control.

For M320 and T-series platforms, eight forwarding classes are supported, thus allowing you to classify packets more granularly. For example, you can configure multiple classes of EF traffic: EF, EF1, and EF2.

By default, the loss priority is low. For each forwarding class, you can configure high or low loss priority. For J-series platforms only, you can configure high, low, medium-high, or medium-low loss priority. For information about configuring forwarding classes, see “Configuring Forwarding Classes” on page 821, “Configuring up to Eight Forwarding Classes” on page 824, and “Examples: Configuring Class of Service” on page 873.

Transmission Scheduling and Rate Control

You use *schedulers* to configure transmission scheduling and rate control parameters. Schedulers define the priority, bandwidth, delay buffer size, rate control status, and RED drop profiles to be applied to a particular class of traffic.

You associate the schedulers with forwarding classes by means of *scheduler maps*. You can then associate each scheduler map with an interface, thereby configuring the hardware queues, packet schedulers, and RED processes that operate according to this mapping.

The following sections describe these processes in more detail:

Priority Scheduling on page 808

Fabric Priority Queuing on page 809

Transmission Rate Control on page 809

Allocation of Leftover Bandwidth on page 809

Default Congestion and Transmission Control on page 810

RED Congestion Control on page 810

Priority Scheduling

Priority scheduling determines the order in which an output interface transmits traffic from the queues. The JUNOS software supports multiple levels of transmission priority, which in order of increasing priority are low, medium-low, medium-high, and high.

Higher-priority queues transmit packets ahead of lower priority queues as long as the higher-priority forwarding classes retain enough bandwidth credit. When you configure a higher-priority queue with a significant fraction of the transmission bandwidth, the queue might lock out lower priority traffic.

You can also configure one queue per interface to have strict-high priority, which works the same as high priority, but provides unlimited transmission bandwidth. As long as the queue with strict-high priority has traffic to send, it receives precedence over all other queues, except queues with high priority. Queues with strict-high and high priority take turns transmitting packets until the strict-high queue is empty, the high priority queues are empty, or the high priority queues run out of bandwidth credit. Only then can lower priority queues send traffic.

The high priority allows you to protect traffic classes from being underserved. For example, a network-control queue might require a small bandwidth allocation (say, 5 percent). You can assign high priority to this queue to prevent it from being underserved.

A queue with strict-high priority supersedes bandwidth guarantees for queues with lower priority; therefore, we recommend that you use the strict-high priority to ensure proper ordering of special traffic, such as voice traffic. You can preserve bandwidth guarantees for queues with lower priority by allocating to the queue with strict-high priority only the amount of bandwidth that it generally requires. For example, consider the following allocation of transmission bandwidth:

Q0 BE—20 percent, low priority

Q1 EF—30 percent, strict-high priority

Q2 AF—40 percent, low priority

Q3 NC—10 percent, low priority

This bandwidth allocation assumes that, in general, the EF forwarding class requires only 30 percent of an interface's transmission bandwidth. However, if short bursts of traffic are received on the EF forwarding class, 100 percent of the bandwidth is given to the EF forwarding class because of the strict-high setting.



NOTE: For 8-port, 12-port, and 48-port Fast Ethernet PICs, transmission scheduling is not supported.

For more information about configuring scheduling priority, see “Configuring Scheduling Maps” on page 835 and “Examples: Configuring Class of Service” on page 873.

Fabric Priority Queuing

On M320 and T-series platforms, the default behavior is for fabric priority queuing on egress interfaces to match the scheduling priority you assign. High-priority egress traffic is automatically assigned to high-priority fabric queues. Likewise, low-priority egress traffic is automatically assigned to low-priority fabric queues.

For information about overriding automatic fabric priority queuing, see “Overriding Fabric Priority Queuing” on page 823 and “Associating a Scheduler with a Fabric Priority” on page 851.

Transmission Rate Control

The transmission rate control determines the actual traffic bandwidth from each of the forwarding classes you configure. The rate is specified in bits per second. You can limit the transmission bandwidth to the exact value you configure, or allow it to exceed the configured rate if additional bandwidth is available from other queues.

For information about configuring transmission rate control, see “Configuring Scheduling Maps” on page 835.

Allocation of Leftover Bandwidth

When a forwarding class fails to fully use the allocated transmission bandwidth, the remaining bandwidth can be taken by other forwarding classes if they receive a larger amount of offered load than the bandwidth allocated. This use of leftover bandwidth is the default behavior. If you want a forwarding class to not take any extra bandwidth, you must configure it for strict allocation. With rate control in place, the specified bandwidth is strictly observed.

When more than one forwarding class can use leftover bandwidth, the higher-priority forwarding class takes the bandwidth first. When several forwarding classes of equal priority are contending for the leftover bandwidth, more of the leftover bandwidth is given to the queues configured for lower transmission rates.

For information about configuring leftover bandwidth allocation, see “Configuring Scheduling Maps” on page 835.

Default Congestion and Transmission Control

A default congestion and transmission control mechanism is needed when an output interface is not configured for a certain forwarding class, but receives packets destined for that unconfigured forwarding class. This default mechanism uses the delay buffer and WRR credit allocated to the designated forwarding class, with a default drop profile. Because the buffer and WRR credit allocation is minimal, packets might be lost if a larger number of packets are forwarded without configuring the forwarding class for the interface.

RED Congestion Control

You can configure two parameters to control congestion at the output stage. The first parameter defines the delay-buffer bandwidth, which provides packet buffer space to absorb burst traffic up to the specified duration of delay. Once the specified delay buffer becomes full, packets with 100 percent drop probability are dropped from the head of the buffer.

The second parameter defines the drop probabilities across the range of delay-buffer occupancy, supporting the RED process. Depending on the drop probabilities, RED might drop packets aggressively long before the buffer becomes full, or it might drop only a few packets even if the buffer is almost full.

You specify the delay-buffer size for each scheduler associated with an output interface configuration in temporal units of 1 through 200,000 microseconds, or as a percentage of the entire interface buffer space. For the temporal setting, the queueing algorithm starts dropping packets when it queues more than a computed number of bytes. This maximum is computed by multiplying the logical interface speed by the configured temporal value.

You specify drop probabilities in the drop profile section of the CoS configuration hierarchy and reference them in each scheduler configuration. For each scheduler, you can configure four separate drop profiles, one for each combination of loss priority (low or high) and IP transport protocol (TCP or non-TCP).

You can configure a maximum of 32 different drop profiles.

For information about configuring delay buffers and drop profiles, see “Configuring Scheduling Maps” on page 835 and “Configuring RED Drop Profiles” on page 853.

Rewrite Markers

A marker reads the current forwarding class and loss priority information associated with a packet and finds the chosen code point from a table. It then writes the code point information into the packet header. Entries in a marker configuration represent the mapping of the current forwarding class into a new forwarding class, to be written into the header.

You define markers in the rewrite rules section of the CoS configuration hierarchy and reference them in the logical interface configuration. This model supports marking on the DSCP, DSCP IPv6, IP Precedence, and MPLS EXP bits CoS indications.

When an interface is not associated with any marker, the ingress classifier decodes the ingress CoS bits into a forwarding class and PLP combination, which in turn determines the egress CoS bits. This means the egress CoS information is entirely dependent on forwarding class and PLP and has nothing to do with ingress CoS values. For example, unless you apply custom classifiers or rewrite markers, EXP values ranging 0 through 7 on ingress do not result in EXP values 0 through 7 on egress.

For information about configuring rewrite markers, see “Rewriting Packet Header Information” on page 854 and “Examples: Configuring Class of Service” on page 873.

Hardware Capabilities and Limitations

Juniper Networks T-series platforms and M-series platforms with an enhanced FPC have more CoS capabilities than M-series platforms that use the earlier FPC model. Table 47 on page 816 lists the differences between the FPC and the enhanced FPC.

