



CHAPTER 25

Configuring QoS

Have you ever participated in a long-distance phone call that involved a satellite connection? The conversation might be interrupted with brief, but perceptible, gaps at odd intervals. Those gaps are the time, called the latency, between the arrival of packets being transmitted over the network. Some network traffic, such as voice and video, cannot tolerate long latency times. Quality of Service (QoS) is a feature that lets you give priority to critical traffic, prevent bandwidth hogging, and manage network bottlenecks to prevent packet drops.

This chapter describes how to apply QoS policies, and includes the following sections:

- [QoS Overview, page 25-1](#)
- [Creating the Standard Priority Queue for an Interface, page 25-5](#)
- [Creating a Policy for Standard Priority Queueing and/or Policing, page 25-6](#)
- [Creating a Policy for Traffic Shaping and Hierarchical Priority Queueing, page 25-7](#)

QoS Overview

You should consider that in an ever-changing network environment, QoS is not a one-time deployment, but an ongoing, essential part of network design.



Note

QoS is only available in single context mode.

This section describes the QoS features supported by the security appliance, and includes the following topics:

- [Supported QoS Features, page 25-2](#)
- [What is a Token Bucket?, page 25-2](#)
- [Policing Overview, page 25-3](#)
- [Priority Queueing Overview, page 25-3](#)
- [Traffic Shaping Overview, page 25-4](#)
- [DSCP and DiffServ Preservation, page 25-5](#)

Supported QoS Features

The security appliance supports the following QoS features:

- **Policing**—To prevent individual flows from hogging the network bandwidth, you can limit the maximum bandwidth used per flow. See the [“Policing Overview” section on page 25-3](#) for more information.
- **Priority queuing**—For critical traffic that cannot tolerate latency, such as Voice over IP (VoIP), you can identify traffic for Low Latency Queuing (LLQ) so that it is always transmitted ahead of other traffic. See the [“Priority Queueing Overview” section on page 25-3](#) for more information.
- **Traffic shaping**—If you have a device that transmits packets at a high speed, such as a security appliance with Fast Ethernet, and it is connected to a low speed device such as a cable modem, then the cable modem is a bottleneck at which packets are frequently dropped. To manage networks with differing line speeds, you can configure the security appliance to transmit packets at a fixed slower rate. See the [“Traffic Shaping Overview” section on page 25-4](#) for more information.

What is a Token Bucket?

A token bucket is used to manage a device that regulates the data in a flow. For example, the regulator might be a traffic policer or a traffic shaper. A token bucket itself has no discard or priority policy. Rather, a token bucket discards tokens and leaves to the flow the problem of managing its transmission queue if the flow overdrives the regulator.

A token bucket is a formal definition of a rate of transfer. It has three components: a burst size, an average rate, and a time interval. Although the average rate is generally represented as bits per second, any two values may be derived from the third by the relation shown as follows:

average rate = burst size / time interval

These terms are defined as follows:

- **Average rate**—Also called the committed information rate (CIR), it specifies how much data can be sent or forwarded per unit time on average.
- **Burst size**—Also called the Committed Burst (Bc) size, it specifies in bits or bytes per burst how much traffic can be sent within a given unit of time to not create scheduling concerns. (For traffic shaping, it specifies bits per burst; for policing, it specifies bytes per burst.)
- **Time interval**—Also called the measurement interval, it specifies the time quantum in seconds per burst.

In the token bucket metaphor, tokens are put into the bucket at a certain rate. The bucket itself has a specified capacity. If the bucket fills to capacity, newly arriving tokens are discarded. Each token is permission for the source to send a certain number of bits into the network. To send a packet, the regulator must 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 traffic shaping) or the packet is discarded or marked down (in the case of policing). If the bucket is already full of tokens, incoming tokens overflow 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.

Note that the token bucket mechanism used for traffic shaping has both a token bucket and a data buffer, or queue; if it did not have a data buffer, it would be a policer. For traffic shaping, packets that arrive that cannot be sent immediately are delayed in the data buffer.

For traffic shaping, a token bucket permits burstiness but bounds it. It guarantees that the burstiness is bounded so that the flow will never send faster than the token bucket capacity, divided by the time interval, plus the established rate at which tokens are placed in the token bucket. See the following formula:

$$(\text{token bucket capacity in bits} / \text{time interval in seconds}) + \text{established rate in bps} = \text{maximum flow speed in bps}$$

This method of bounding burstiness also guarantees that the long-term transmission rate will not exceed the established rate at which tokens are placed in the bucket.

Policing Overview

Policing is a way of ensuring that no traffic exceeds the maximum rate (in bits/second) that you configure, thus ensuring that no one traffic flow or class can take over the entire resource. When traffic exceeds the maximum rate, the security appliance drops the excess traffic. Policing also sets the largest single burst of traffic allowed.

Priority Queueing Overview

LLQ priority queueing lets you prioritize certain traffic flows (such as latency-sensitive traffic like voice and video) ahead of other traffic.

The security appliance supports two types of priority queueing:

- Standard priority queueing—Standard priority queueing uses an LLQ priority queue on an interface (see the [“Creating the Standard Priority Queue for an Interface”](#) section on page 25-5), while all other traffic goes into the “best effort” queue. Because queues are not of infinite size, they can fill and overflow. When a queue is full, any additional packets cannot get into the queue and are dropped. This is called *tail drop*. To avoid having the queue fill up, you can increase the queue buffer size. You can also fine-tune the maximum number of packets allowed into the transmit queue. These options let you control the latency and robustness of the priority queueing. Packets in the LLQ queue are always transmitted before packets in the best effort queue.
- Hierarchical priority queueing—Hierarchical priority queueing is used on interfaces on which you enable a traffic shaping queue. A subset of the shaped traffic can be prioritized. The standard priority queue is not used. See the following guidelines about hierarchical priority queueing:
 - Priority packets are always queued at the head of the shape queue so they are always transmitted ahead of other non-priority queued packets.
 - Priority packets are never dropped from the shape queue unless the sustained rate of priority traffic exceeds the shape rate.
 - For IPSec-encrypted packets, you can only match traffic based on the DSCP or precedence setting.
 - IPSec-over-TCP is not supported for priority traffic classification.

Traffic Shaping Overview

Traffic shaping is used to match device and link speeds, thereby controlling packet loss, variable delay, and link saturation, which can cause jitter and delay.

**Note**

Traffic shaping is not supported on the ASA 5580.

- Traffic shaping must be applied to all outgoing traffic on a physical interface or in the case of the ASA 5505, on a VLAN. You cannot configure traffic shaping for specific types of traffic.
- Traffic shaping is implemented when packets are ready to be transmitted on an interface, so the rate calculation is performed based on the actual size of a packet to be transmitted, including all the possible overhead such as the IPSec header and L2 header.
- The shaped traffic includes both through-the-box and from-the-box traffic.
- The shape rate calculation is based on the standard token bucket algorithm. The token bucket size is twice the Burst Size value. See the [“What is a Token Bucket?”](#) section on page 25-2.
- When bursty traffic exceeds the specified shape rate, packets are queued and transmitted later. Following are some characteristics regarding the shape queue (for information about hierarchical priority queuing, see the [“Priority Queuing Overview”](#) section on page 25-3):
 - The queue size is calculated based on the shape rate. The queue can hold the equivalent of 200-milliseconds worth of shape rate traffic, assuming a 1500-byte packet. The minimum queue size is 64.
 - When the queue limit is reached, packets are tail-dropped.
 - Certain critical keep-alive packets such as OSPF Hello packets are never dropped.
 - The time interval is derived by $time_interval = burst_size / average_rate$. The larger the time interval is, the burstier the shaped traffic might be, and the longer the link might be idle. The effect can be best understood using the following exaggerated example:

Average Rate = 1000000

Burst Size = 1000000

In the above example, the time interval is 1 second, which means, 1 Mbps of traffic can be bursted out within the first 10 milliseconds of the 1-second interval on a 100 Mbps FE link and leave the remaining 990 milliseconds idle without being able to send any packets until the next time interval. So if there is delay-sensitive traffic such as voice traffic, the Burst Size should be reduced compared to the average rate so the time interval is reduced.

How QoS Features Interact

You can configure each of the QoS features alone if desired for the security appliance. Often, though, you configure multiple QoS features on the security appliance so you can prioritize some traffic, for example, and prevent other traffic from causing bandwidth problems.

See the following supported feature combinations per interface:

- Standard priority queuing (for specific traffic) + Policing (for the rest of the traffic).

You cannot configure priority queuing and policing for the same set of traffic.

- Traffic shaping (for all traffic on an interface) + Hierarchical priority queuing (for a subset of traffic).

You cannot configure traffic shaping and standard priority queueing for the same interface; only hierarchical priority queueing is allowed. For example, if you configure standard priority queueing for the global policy, and then configure traffic shaping for a specific interface, the feature you configured last is rejected because the global policy overlaps the interface policy.

Typically, if you enable traffic shaping, you do not also enable policing for the same traffic, although the security appliance does not restrict you from configuring this.

DSCP and DiffServ Preservation

- DSCP markings are preserved on all traffic passing through the security appliance.
- The security appliance does not locally mark/re-mark any classified traffic, but it honors the Expedited Forwarding (EF) DSCP bits of every packet to determine if it requires “priority” handling and will direct those packets to the LLQ.
- DiffServ marking is preserved on packets when they traverse the service provider backbone so that QoS can be applied in transit (QoS tunnel pre-classification).

Creating the Standard Priority Queue for an Interface

If you enable standard priority queueing for traffic on a physical interface, then you need to also create the priority queue on each interface. Each physical interface uses two queues: one for priority traffic, and the other for all other traffic. For the other traffic, you can optionally configure policing.



Note

The standard priority queue is not required for hierarchical priority queueing with traffic shaping; see the [“Priority Queueing Overview”](#) section on page 25-3 for more information.

To create the priority queue, perform the following steps:

- Step 1** Go to Configuration > Device Management > Advanced > Priority Queue, and click **Add**.
The Add Priority Queue dialog box displays.
- Step 2** From the Interface drop-down list, choose the physical interface name on which you want to enable the priority queue, or for the ASA 5505, the VLAN interface name.
- Step 3** To change the size of the priority queues, in the Queue Limit field, enter the number of average, 256-byte packets that the specified interface can transmit in a 500-ms interval.
A packet that stays more than 500 ms in a network node might trigger a timeout in the end-to-end application. Such a packet can be discarded in each network node.
Because queues are not of infinite size, they can fill and overflow. When a queue is full, any additional packets cannot get into the queue and are dropped (called *tail drop*). To avoid having the queue fill up, you can use this option to increase the queue buffer size.
The upper limit of the range of values for this option is determined dynamically at run time. The key determinants are the memory needed to support the queues and the memory available on the device.
The Queue Limit that you specify affects both the higher priority low-latency queue and the best effort queue.
- Step 4** To specify the depth of the priority queues, in the Transmission Ring Limit field, enter the number of maximum 1550-byte packets that the specified interface can transmit in a 10-ms interval.

This setting guarantees that the hardware-based transmit ring imposes no more than 10-ms of extra latency for a high-priority packet.

This option sets the maximum number of low-latency or normal priority packets allowed into the Ethernet transmit driver before the driver pushes back to the queues on the interface to let them buffer packets until the congestion clears.

The upper limit of the range of values is determined dynamically at run time. The key determinants are the memory needed to support the queues and the memory available on the device.

The Transmission Ring Limit that you specify affects both the higher priority low-latency queue and the best-effort queue.

Creating a Policy for Standard Priority Queueing and/or Policing

You can configure standard priority queueing and policing rules for the same interface. See the [“How QoS Features Interact”](#) section on page 25-4 for information about valid QoS configurations.

To configure a QoS service policy, perform the following steps:

-
- Step 1** To configure priority queueing, configure a service policy rule in the Configuration > Firewall > Service Policy Rules pane according to [Chapter 22, “Configuring Service Policy Rules.”](#)
- You can configure QoS as part of a new service policy rule, or you can edit an existing service policy.
- For priority traffic, identify only latency-sensitive traffic. You can match traffic based on many characteristics, including access lists, tunnel groups, DSCP, precedence, and more. You cannot use the **class-default** class map for priority traffic. You cannot configure priority queueing for the global policy if you also enable traffic shaping on any interfaces.
- Step 2** In the Rule Actions dialog box, click the **QoS** tab.
- Step 3** Click **Enable priority for this flow**.
- If this service policy rule is for an individual interface, ASDM automatically creates the priority queue for the interface (Configuration > Properties > Priority Queue; for more information, see the [“Creating the Standard Priority Queue for an Interface”](#) section on page 25-5). If this rule is for the global policy, then you need to manually add the priority queue to one or more interfaces *before* you configure the service policy rule.
- Step 4** Click **Finish**. The service policy rule is added to the rule table.
- Step 5** To configure policing, configure a service policy rule for the same interface in the Configuration > Firewall > Service Policy Rules pane according to [Chapter 22, “Configuring Service Policy Rules.”](#)
- For policing traffic, you can choose to police all traffic that you are not prioritizing, or you can limit the traffic to certain types.
- Step 6** In the Rule Actions dialog box, click the **QoS** tab.
- Step 7** Click **Enable policing**, then check the **Input policing** or **Output policing** (or both) check boxes to enable the specified type of traffic policing. For each type of traffic policing, configure the following fields:
- Committed Rate—The rate limit for this traffic flow; this is a value in the range 8000-2000000000, specifying the maximum speed (bits per second) allowed.

- **Conform Action**—The action to take when the rate is less than the conform-burst value. Values are transmit or drop.
- **Exceed Action**—Take this action when the rate is between the conform-rate value and the conform-burst value. Values are transmit or drop.
- **Burst Rate**—A value in the range 1000-512000000, specifying the maximum number of instantaneous bytes allowed in a sustained burst before throttling to the conforming rate value.

Step 8 Click **Finish**. The service policy rule is added to the rule table.

Step 9 Click **Apply** to send the configuration to the device.

Creating a Policy for Traffic Shaping and Hierarchical Priority Queueing

You can configure traffic shaping for all traffic on an interface, and optionally hierarchical priority queueing for a subset of latency-sensitive traffic. See the [“How QoS Features Interact” section on page 25-4](#) for information about valid QoS configurations.



Note

One side-effect of priority queueing is packet re-ordering. For IPSec packets, out-of-order packets that are not within the anti-replay window generate warning syslog messages. These warnings are false alarms in the case of priority queueing. You can configure the IPSec anti-replay window size to avoid possible false alarms. See the Configuration > VPN > IPSec > IPSec Rules > Enable Anti-replay window size option in the [“Crypto Maps” section on page 34-9](#).

Traffic shaping is not supported on the ASA 5580.

To configure a QoS service policy, perform the following steps:

Step 1 Configure a service policy on the Configuration > Firewall > Service Policy Rules pane according to [Chapter 22, “Configuring Service Policy Rules.”](#)

You can configure QoS as part of a new service policy rule, or you can edit an existing service policy.

For traffic shaping, all traffic on an interface must be shaped. You can only use the **class-default** class map, which is automatically created by the security appliance, and which matches all traffic.

You cannot configure a separate traffic shaping rule on the same interface for which you configure a priority queueing rule (see the [“Creating a Policy for Standard Priority Queueing and/or Policing” section on page 25-6](#)); you can, however, configure priority queueing for a subset of shaped traffic under the traffic shaping rule. You also cannot configure traffic shaping for the global policy if you also enable priority queueing on any interfaces.

Step 2 In the Rule Actions dialog box, click the **QoS** tab.

Step 3 Click **Enable traffic shaping**, and configure the following fields:

- **Average Rate**—Sets the average rate of traffic in bits per second over a given fixed time period, between 64000 and 154400000. Specify a value that is a multiple of 8000.

- **Burst Size**—Sets the average burst size in bits that can be transmitted over a given fixed time period, between 2048 and 154400000. Specify a value that is a multiple of 128. If you do not specify the Burst Size, the default value is equivalent to 4-milliseconds of traffic at the specified Average Rate. For example, if the average rate is 1000000 bits per second, 4 ms worth = $1000000 * 4/1000 = 4000$.

Step 4 (Optional) To configure priority queueing for a subset of shaped traffic:

- a. Click **Enforce priority to selected shape traffic**.
- b. Click **Configure** to identify the traffic that you want to prioritize.
You are prompted to identify the traffic for which you want to apply priority queueing.
- c. After you identify the traffic (see the “[Adding a Service Policy Rule for Through Traffic](#)” section on [page 22-6](#)), click **Next**.
- d. Click **Enable priority for this flow**.
- e. Click **Finish**.

You return to the QoS tab.



Note For this type of priority queueing, you do *not* need to create a priority queue on an interface (**Configuration > Properties > Priority Queue**).

Step 5 Click **Finish**. The service policy rule is added to the rule table.

Step 6 Click **Apply** to send the configuration to the device.
