# Quality of Service

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.

The following topics describe how to apply QoS policies.



# About QoS

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.

This section describes the QoS features available on the ASA.

## Supported QoS Features

The ASA supports the following QoS features:

- Policing—To prevent classified traffic from hogging the network bandwidth, you can limit the maximum bandwidth used per class. See Policing, on page 2 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 Priority Queuing, on page 2.

# What is a Token Bucket?

A token bucket is used to manage a device that regulates the data in a flow, for example, a traffic policer. 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

Here are some definitions of these terms:

- 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 bytes per burst how much traffic can be sent within a given unit of time to not create scheduling concerns.
- 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 waits until the packet is discarded or marked down. 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.

# Policing

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 class can take over the entire resource. When traffic exceeds the maximum rate, the ASA drops the excess traffic. Policing also sets the largest single burst of traffic allowed.

# Priority Queuing

LLQ priority queuing lets you prioritize certain traffic flows (such as latency-sensitive traffic like voice and video) ahead of other traffic. Priority queuing uses an LLQ priority queue on an interface (see Configure the Priority Queue for an Interface, on page 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 queuing. Packets in the LLQ queue are always transmitted before packets in the best effort queue.

## How QoS Features Interact

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

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.

## DSCP (DiffServ) Preservation

DSCP (DiffServ) markings are preserved on all traffic passing through the ASA. The ASA does not locally mark/remark any classified traffic. For example, you could key off the Expedited Forwarding (EF) DSCP bits of every packet to determine if it requires “priority” handling and have the ASA direct those packets to the LLQ.

# Guidelines for QoS

### Context Mode Guidelines

Supported in single context mode only. Does not support multiple context mode.

### Firewall Mode Guidelines

Supported in routed firewall mode only. Does not support transparent firewall mode.

### IPv6 Guidelines

Does not support IPv6.

### Additional Guidelines and Limitations

- QoS is applied unidirectionally; only traffic that enters (or exits, depending on the QoS feature) the interface to which you apply the policy map is affected.
- For priority traffic, you cannot use the **class-default** class map.
- For priority queuing, the priority queue must be configured for a physical interface.
- For policing, to-the-box traffic is not supported.
- For policing, traffic to and from a VPN tunnel bypasses interface policing.
- For policing, when you match a tunnel group class map, only outbound policing is supported.

# Configure QoS

Use the following sequence to implement QoS on the ASA.

**Procedure**

**Step 1** Determine the Queue and TX Ring Limits for a Priority Queue, on page 4.
**Step 2** Configure the Priority Queue for an Interface, on page 5.
**Step 3** Configure a Service Rule for Priority Queuing and Policing, on page 6.

# Determine the Queue and TX Ring Limits for a Priority Queue

Use the following worksheets to determine the priority queue and TX ring limits.

## Queue Limit Worksheet

The following worksheet shows how to calculate the priority queue size. 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 adjust the queue buffer size according to Configure the Priority Queue for an Interface, on page 5.

Tips on the worksheet:

- Outbound bandwidth—For example, DSL might have an uplink speed of 768 Kbps. Check with your provider.
- Average packet size—Determine this value from a codec or sampling size. For example, for VoIP over VPN, you might use 160 bytes. We recommend 256 bytes if you do not know what size to use.
- Delay—The delay depends on your application. For example, the recommended maximum delay for VoIP is 200 ms. We recommend 500 ms if you do not know what delay to use.

***Table 1: Queue Limit Worksheet***

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | _____ *Outbound bandwidth (Mbps or Kbps)* | Mbps | x | **125** | = | _____ *# of bytes/ms* | | |
| | | Kbps | x | **.125** | = | _____ *# of bytes/ms* | | |
| **2** | _____ *# of bytes/ms from Step 1* | | ÷ | _____ *Average packet size (bytes)* | x | _____ *Delay (ms)* | = | _____ *Queue limit (# of packets)* |

## TX Ring Limit Worksheet

The following worksheet shows how to calculate the TX ring limit. This limit determines the maximum number of 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. This setting guarantees that the hardware-based transmit ring imposes a limited amount of extra latency for a high-priority packet.

Tips on the worksheet:

- Outbound bandwidth—For example, DSL might have an uplink speed of 768 Kbps. Check with your provider.
- Maximum packet size—Typically, the maximum size is 1538 bytes, or 1542 bytes for tagged Ethernet. If you allow jumbo frames (if supported for your platform), then the packet size might be larger.
- Delay—The delay depends on your application. For example, to control jitter for VoIP, you should use 20 ms.

***Table 2: TX Ring Limit Worksheet***

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1** | _____ *Outbound bandwidth (Mbps or Kbps)* | Mbps | x | **125** | = | _____ *# of bytes/ms* | | |
| | | Kbps | x | **0.125** | = | _____ *# of bytes/ms* | | |
| **2** | _____ *# of bytes/ms from Step 1* | | ÷ | _____ *Maximum packet size (bytes)* | x | _____ *Delay (ms)* | = | _____ *TX ring limit (# of packets)* |

# Configure the Priority Queue for an Interface

If you enable priority queuing 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.

### Procedure

**Step 1** Choose **Configuration** > **Device Management** > **Advanced** > **Priority Queue**, and click **Add**.

**Step 2** Configure the following options:

- **Interface**—The physical interface name on which you want to enable the priority queue, or for the ASASM, the VLAN interface name.
- **Queue Limit**—The number of average, 256-byte packets that the specified interface can transmit in a 500-ms interval. The range is 0-2048, and 2048 is the default.

  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.

- **Transmission Ring Limit**—The depth of the priority queues, which is the number of maximum 1550-byte packets that the specified interface can transmit in a 10-ms interval. The range is 3-511, and 511 is the default.

  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.

**Step 3** Click **OK**, then **Apply**.

# Configure a Service Rule for Priority Queuing and Policing

You can configure priority queuing and policing for different class maps within the same policy map. See How QoS Features Interact, on page 3 for information about valid QoS configurations.

### Before you begin

- You cannot use the **class-default** class map for priority traffic.
- For policing, to-the-box traffic is not supported.
- For policing, traffic to and from a VPN tunnel bypasses interface policing.
- For policing, when you match a tunnel group class map, only outbound policing is supported.
- For priority traffic, identify only latency-sensitive traffic.
- For policing traffic, you can choose to police all other traffic, or you can limit the traffic to certain types.

### Procedure

**Step 1** Choose **Configuration** > **Firewall** > **Service Policy**, and open a rule.

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

**Step 2** Proceed through the wizard to the Rules page, selecting the interface (or global) and traffic matching criteria along the way.

For policing traffic, you can choose to police all traffic that you are not prioritizing, or you can limit the traffic to certain types.

**Tip** If you use an ACL for traffic matching, policing is applied in the direction specified in the ACL only. That is, traffic going from the source to the destination is policed, but not the reverse.

**Step 3** In the Rule Actions dialog box, click the **QoS** tab.

**Step 4** Select **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 > Device Management > Advanced > Priority Queue; for more information, see Configure the Priority Queue for an Interface, on page 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 5** Select **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 options:

- **Committed Rate**—The rate limit for this traffic class, from 8000-2000000000 bits per second. For example, to limit traffic to 5Mbps, enter 5000000.
- **Conform Action**—The action to take when the traffic is below the policing rate and burst size. You can drop or transmit the traffic. The default is to transmit the traffic.
- **Exceed Action**—The action to take when traffic exceeds the policing rate and burst size. You can drop or transmit packets that exceed the policing rate and burst size. The default is to drop excess packets.
- **Burst Rate**—The maximum number of instantaneous bytes allowed in a sustained burst before throttling to the conforming rate value, between 1000 and 512000000 bytes. Calculate the burst size calculated as 1/32 of the conform-rate in bytes. For example, the burst size for a 5Mbps rate would be 156250. The default is 1500, but the system can recalculate any value you enter as needed.

**Step 6** Click **Finish**, then **Apply**.

# Monitor QoS

The following topics explain how to monitor QoS.

To monitor QoS in ASDM, you can enter commands at the Command Line Interface tool.

## QoS Police Statistics

To view the QoS statistics for traffic policing, use the **show service-policy police** command.

```
hostname# show service-policy police

Global policy:
 Service-policy: global_fw_policy

Interface outside:
 Service-policy: qos
```

```
  Class-map: browse
   police Interface outside:
    cir 56000 bps, bc 10500 bytes
    conformed 10065 packets, 12621510 bytes; actions: transmit
    exceeded 499 packets, 625146 bytes; actions: drop
    conformed 5600 bps, exceed 5016 bps
  Class-map: cmap2
   police Interface outside:
    cir 200000 bps, bc 37500 bytes
    conformed 17179 packets, 20614800 bytes; actions: transmit
    exceeded 617 packets, 770718 bytes; actions: drop
    conformed 198785 bps, exceed 2303 bps
```

# QoS Priority Statistics

To view statistics for service policies implementing the **priority** command, use the **show service-policy priority** command.

```
hostname# show service-policy priority
Global policy:
 Service-policy: global_fw_policy
Interface outside:
 Service-policy: qos
  Class-map: TG1-voice
   Priority:
    Interface outside: aggregate drop 0, aggregate transmit 9383
```

“Aggregate drop” denotes the aggregated drop in this interface; “aggregate transmit” denotes the aggregated number of transmitted packets in this interface.

# QoS Priority Queue Statistics

To display the priority-queue statistics for an interface, use the **show priority-queue statistics** command. The results show the statistics for both the best-effort (BE) queue and the low-latency queue (LLQ). The following example shows the use of the **show priority-queue statistics** command for the interface named test.

```
hostname# show priority-queue statistics test

Priority-Queue Statistics interface test

Queue Type         = BE
Packets Dropped    = 0
Packets Transmit   = 0
Packets Enqueued   = 0
Current Q Length   = 0
Max Q Length       = 0

Queue Type         = LLQ
Packets Dropped    = 0
Packets Transmit   = 0
Packets Enqueued   = 0
Current Q Length   = 0
Max Q Length       = 0
hostname#
```

In this statistical report:

- “Packets Dropped” denotes the overall number of packets that have been dropped in this queue.
- “Packets Transmit” denotes the overall number of packets that have been transmitted in this queue.
- “Packets Enqueued” denotes the overall number of packets that have been queued in this queue.
- “Current Q Length” denotes the current depth of this queue.
- “Max Q Length” denotes the maximum depth that ever occurred in this queue.

# History for QoS

| Feature Name | Platform Releases | Description |
|---|---|---|
| Priority queuing and policing | 7.0(1) | We introduced QoS priority queuing and policing.<br>We introduced the following screens:<br>Configuration > Device Management > Advanced > Priority Queue Configuration > Firewall > Service Policy Rules |
| Shaping and hierarchical priority queuing | 7.2(4)/8.0(4) | We introduced QoS shaping and hierarchical priority queuing.<br>We modified the following screen: Configuration > Firewall > Service Policy Rules. |
| Ten Gigabit Ethernet support for a standard priority queue on the ASA 5585-X | 8.2(3)/8.4(1) | We added support for a standard priority queue on Ten Gigabit Ethernet interfaces for the ASA 5585-X. |