

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, on page 1
- Guidelines for QoS, on page 3
- Configure QoS, on page 3
- Monitor QoS, on page 9
- Configuration Examples for Priority Queuing and Policing, on page 10
- History for QoS, on page 12

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 Create the priority queue for the interface.

priority-queue *interface_name*

Example:

```
hostname(config)# priority-queue inside
```

The *interface_name* argument specifies the physical interface name on which you want to enable the priority queue.

Step 2 Change the size of the priority queues.

queue-limit *number_of_packets*

The default queue limit is 1024 packets. 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 the **queue-limit** command to increase the queue buffer size.

The upper limit of the range of values for the **queue-limit** command is determined dynamically at run time. To view this limit, enter **queue-limit ?** on the command line. 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.

Example:

```
hostname(config-priority-queue)# queue-limit 260
```

Step 3 Specify the depth of the priority queues.

tx-ring-limit *number_of_packets*

The default tx-ring-limit is 511 packets. This command 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. This setting guarantees that the hardware-based transmit ring imposes a limited amount of extra latency for a high-priority packet.

The upper limit of the range of values for the **tx-ring-limit** command is determined dynamically at run time. To view this limit, enter **tx-ring-limit ?** on the command line. The key determinants are the memory needed to support the queues and the memory available on the device.

The **tx-ring-limit** that you specify affects both the higher priority low-latency queue and the best-effort queue.

Example:

```
hostname(config-priority-queue)# tx-ring-limit 3
```

Examples

The following example establishes a priority queue on interface “outside” (the GigabitEthernet0/1 interface), with the default queue-limit and tx-ring-limit:

```
hostname(config)# priority-queue outside
```

The following example establishes a priority queue on the interface “outside” (the GigabitEthernet0/1 interface), sets the queue-limit to 260 packets, and sets the tx-ring-limit to 3:

```
hostname(config)# priority-queue outside
hostname(config-priority-queue)# queue-limit 260
hostname(config-priority-queue)# tx-ring-limit 3
```

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 Create an L3/L4 class map to identify the traffic for which you want to perform priority queuing.

```
class-map name
match parameter
```

Example:

```
hostname(config)# class-map priority_traffic
hostname(config-cmap)# match access-list priority
```

See Create a Layer 3/4 Class Map for Through Traffic for more information.

Step 2 Create an L3/L4 class map to identify the traffic for which you want to perform priority policing.

```
class-map name
match parameter
```

Example:

```
hostname(config)# class-map policing_traffic
hostname(config-cmap)# match access-list policing
```

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 Add or edit a policy map: **policy-map** *name*

Example:

```
hostname(config)# policy-map QoS_policy
```

Step 4 Identify the class map you created for prioritized traffic and configure priority queuing for the class.

```
class priority_map_name
```

priority

Example:

```
hostname(config-pmap)# class priority_class
hostname(config-pmap-c)# priority
```

Step 5 Identify the class map you created for policed traffic: **class** *name*

Example:

```
hostname(config-pmap)# class policing_class
```

Step 6 Configure policing for the class.

police {**output** | **input**} *conform-rate* [*conform-burst*] [**conform-action** [**drop** | **transmit**]] [**exceed-action** [**drop** | **transmit**]]

The options are:

- **output**—Enables policing of traffic flowing in the output direction.
- **input**—Enables policing of traffic flowing in the input direction.
- *conform-rate*—Sets the rate limit for this traffic class, from 8000 and 2000000000 bits per second. For example, to limit traffic to 5Mbps, enter 5000000.
- *conform-burst*—(Optional.) Specifies the maximum number of instantaneous bytes allowed in a sustained burst before throttling to the conforming rate value, between 1000 and 512000000 bytes. If you omit the variable, the burst size is calculated as 1/32 of the conform-rate in bytes. For example, the burst size for a 5Mbps rate would be 156250.
- **conform-action**—(Optional.) Sets 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**—(Optional.) Sets 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.

Example:

```
hostname(config-pmap-c)# police output 56000 10500
```

Step 7 Activate the policy map on one or more interfaces.

service-policy *policymap_name* {**global** | **interface** *interface_name*}

Example:

```
hostname(config)# service-policy QoS_policy interface inside
```

The **global** option applies the policy map to all interfaces, and **interface** applies the policy to one interface. Only one global policy is allowed. You can override the global policy on an interface by applying a service policy to that interface. You can only apply one policy map to each interface.

Monitor QoS

The following topics explain how to monitor QoS.

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.

Configuration Examples for Priority Queuing and Policing

The following sections provide examples of configuring priority queuing and policing.

Class Map Examples for VPN Traffic

In the following example, the **class-map** command classifies all non-tunneled TCP traffic, using an ACL named tcp_traffic:

```
hostname(config)# access-list tcp_traffic permit tcp any any
hostname(config)# class-map tcp_traffic
hostname(config-cmap)# match access-list tcp_traffic
```

In the following example, other, more specific match criteria are used for classifying traffic for specific, security-related tunnel groups. These specific match criteria stipulate that a match on tunnel-group (in this

case, the previously-defined Tunnel-Group-1) is required as the first match characteristic to classify traffic for a specific tunnel, and it allows for an additional match line to classify the traffic (IP differential services code point, expedited forwarding).

```
hostname(config)# class-map TG1-voice
hostname(config-cmap)# match tunnel-group tunnel-grp1
hostname(config-cmap)# match dscp ef
```

In the following example, the **class-map** command classifies both tunneled and non-tunneled traffic according to the traffic type:

```
hostname(config)# access-list tunneled extended permit ip 10.10.34.0 255.255.255.0
192.168.10.0 255.255.255.0
hostname(config)# access-list non-tunneled extended permit tcp any any
hostname(config)# tunnel-group tunnel-grp1 type IPsec_L2L

hostname(config)# class-map browse
hostname(config-cmap)# description "This class-map matches all non-tunneled tcp traffic."
hostname(config-cmap)# match access-list non-tunneled

hostname(config-cmap)# class-map TG1-voice
hostname(config-cmap)# description "This class-map matches all dscp ef traffic for
tunnel-grp 1."
hostname(config-cmap)# match dscp ef
hostname(config-cmap)# match tunnel-group tunnel-grp1

hostname(config-cmap)# class-map TG1-BestEffort
hostname(config-cmap)# description "This class-map matches all best-effort traffic for
tunnel-grp1."
hostname(config-cmap)# match tunnel-group tunnel-grp1
hostname(config-cmap)# match flow ip destination-address
```

The following example shows a way of policing traffic within a tunnel, provided the classed traffic is not specified as a tunnel, but does go *through* the tunnel. In this example, 192.168.10.10 is the address of the host machine on the private side of the remote tunnel, and the ACL is named “host-over-l2l”. By creating a class-map (named “host-specific”), you can then police the “host-specific” class before the LAN-to-LAN connection polices the tunnel. In this example, the “host-specific” traffic is rate-limited before the tunnel, then the tunnel is rate-limited:

```
hostname(config)# access-list host-over-l2l extended permit ip any host 192.168.10.10
hostname(config)# class-map host-specific
hostname(config-cmap)# match access-list host-over-l2l
```

Priority and Policing Example

The following example builds on the configuration developed in the previous section. As in the previous example, there are two named class-maps: tcp_traffic and TG1-voice.

```
hostname(config)# class-map TG1-best-effort
hostname(config-cmap)# match tunnel-group Tunnel-Group-1
hostname(config-cmap)# match flow ip destination-address
```

Adding a third class map provides a basis for defining a tunneled and non-tunneled QoS policy, as follows, which creates a simple QoS policy for tunneled and non-tunneled traffic, assigning packets of the class TG1-voice to the low latency queue and setting rate limits on the tcp_traffic and TG1-best-effort traffic flows.

In this example, the maximum rate for traffic of the tcp_traffic class is 56,000 bits/second and a maximum burst size of 10,500 bytes per second. For the TC1-BestEffort class, the maximum rate is 200,000 bits/second, with a maximum burst of 37,500 bytes/second. Traffic in the TC1-voice class has no policed maximum speed or burst rate because it belongs to a priority class.

```
hostname(config)# access-list tcp_traffic permit tcp any any
hostname(config)# class-map tcp_traffic
hostname(config-cmap)# match access-list tcp_traffic

hostname(config)# class-map TG1-voice
hostname(config-cmap)# match tunnel-group tunnel-grp1
hostname(config-cmap)# match dscp ef

hostname(config-cmap)# class-map TG1-BestEffort
hostname(config-cmap)# match tunnel-group tunnel-grp1
hostname(config-cmap)# match flow ip destination-address

hostname(config)# policy-map qos
hostname(config-pmap)# class tcp_traffic
hostname(config-pmap-c)# police output 56000 10500

hostname(config-pmap-c)# class TG1-voice
hostname(config-pmap-c)# priority

hostname(config-pmap-c)# class TG1-best-effort
hostname(config-pmap-c)# police output 200000 37500

hostname(config-pmap-c)# class class-default
hostname(config-pmap-c)# police output 1000000 37500

hostname(config-pmap-c)# service-policy qos global
```

History for QoS

Feature Name	Platform Releases	Description
Priority queuing and policing	7.0(1)	We introduced QoS priority queuing and policing. We introduced the following commands: **priority-queue**, **queue-limit**, **tx-ring-limit**, **priority, police, show priority-queue statistics, show service-policy police, show service-policy priority, show running-config priority-queue, clear configure priority-queue** .
Shaping and hierarchical priority queuing	7.2(4)/8.0(4)	We introduced QoS shaping and hierarchical priority queuing. We introduced the following commands: **shape**, **show service-policy shape**.
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.