服务质量

您是否曾用过使用卫星连接的长途电话？对话可能会不定期中断，出现短暂但可察觉的间隙。这些短暂间隙是在网络上传输的数据包到达之间的时间，即延迟。某些网络流量（例如语音和视频）不允许出现长时间延迟。通过服务质量 (QoS) 功能，您可以优先考虑重要流量、防止带宽占用及管理网络瓶颈，以防丢包。

以下主题介绍如何应用 QoS 策略。

关于 QoS

应考虑到到，在不断变化的网络环境中，QoS 不是一次性部署，而是网络设计的持续必要的部分。

本节介绍 ASA 中可用的 QoS 功能。

支持的 QoS 功能

ASA 支持以下 QoS 功能：

- 策略管制 - 要防止分类流量占用网络带宽，可以限制每个类使用的最大带宽。有关详细信息，请参阅策略管制，第 2 页。
- 优先级排队 - 对于不允许出现延迟的关键流量（例如 IP 语音 [VoIP]），可以将此流量标记为低延迟排队 (LLQ) 的流量，以便其始终在其他流量之前传输。请参阅优先级队列，第 2 页。

什么是令牌桶？

令牌桶用于管理对流量中的数据进行管制的设备，例如流量监管器。令牌桶本身不具有丢弃或优先级策略。相反，如果流量超过管制器，令牌桶会丢弃令牌，并将管理传输队列的问题留给流量。

令牌桶是传输速率的正式定义。它包含三个组成部分：突发大小、平均速率和时间间隔。虽然平均速率通常表示为位/秒，但任意两个值可以通过以下关系从第三个值中推出：

平均速率 = 突发大小/时间间隔

以下是这些术语的部分定义：

- 平均速率 - 亦称承诺信息速率 (CIR)，指定单位时间平均发送或转发的数据量。
- 突发大小 - 亦称承诺突发 (Bc) 大小，以每次突发的字节为单位指定在给定的单位时间内可以发送而不引起调度问题的流量大小。
- 时间间隔 - 亦称测量间隔，以每次突发的秒为单位指定时间量。

在令牌桶比喻中，以一定速率将令牌添加到桶中。令牌桶本身有指定的容量。如果令牌桶容量已满，新到达的令牌会被丢弃。每个令牌允许源将一定数量的位发送到网络中。要发送数据包，管制器必须从令牌桶中删除与所代表的数据包大小相等的若干令牌。

如果令牌桶内没有足够的令牌来发送数据包，数据包会一直等，直到数据包被丢弃或被降级。如果令牌桶的令牌已满，传入的令牌会溢出而不能用于后续数据包。因此，在任何时刻，源能够发送到网络中的最大突发流量都大致与令牌桶的大小成正比。

策略管制

策略管制是一种通过确保流量不超过配置的最大速率（以位/秒为单位）以保证任何一个流量类都不会沿用全部资源的方式。如果流量超过最大速率，ASA 将丢弃超额的流量。策略管制还设定了允许的单次最大突发流量。

优先级队列

LLQ 优先级排队使您可以在处理其他流量之前优先处理某些流量（例如，像语音和视频之类的延迟敏感型流量）。优先级排队使用接口上的一个 LLQ 优先级队列（请参阅为接口配置优先级队列，第 5 页），而所有其他流量进入“尽力而为”队列。由于队列大小有限制，队列可以填满和溢出。如果队列已满，任何额外的数据包无法进入队列，并将丢弃。这称为尾部丢弃。要避免队列被填满，可以增加队列缓冲区的大小。还可以优化允许进入传输队列的数据包的最大数。这些选项使您能够控制优先级排队的延迟和稳定性。LLQ 队列中的数据包始终在“尽力而为”队列中的数据包之前传输。

如何交互使用 QoS 功能

如果需要，可以为 ASA 单独配置各种 QoS 功能。不过，ASA 上通常会配置多种 QoS 功能以便可以优先排列某些流量，并防止其他流量导致带宽问题。可以配置：

优先级排队（用于特定流量）+ 策略管制（用于其余流量）。

无法对同一组的流量配置优先级排队和策略管制。

DSCP (DiffServ) 保留

通过 ASA 的所有流量上都会保留 DSCP (DiffServ) 标记。ASA 不会在本地对任何分类流量进行标记/备注。例如，您可以切断每个数据包的快速转发 (EF) DSCP 位来确定该数据包是否需要“优先”处理，并让 ASA 将这些数据包定向到 LLQ。

QoS 准则

情景模式准则

仅支持单情景模式。不支持多情景模式。

防火墙模式准则

仅支持路由防火墙模式。不支持透明防火墙模式。

IPv6 准则

不支持 IPv6。

其他准则和限制

- QoS 只能单向应用；只有流入（或流出，视 QoS 功能而定）应用了策略映射的接口的流量才会受到影响。
- 对于优先级流量，无法使用 **class-default** 类映射。
- 对于优先级排队，必须为某个物理接口配置优先级队列。
- 对于策略管制，不支持流向设备的流量。
- 对于策略管制，往返 VPN 隧道的流量会绕过接口策略管制。
- 对于策略管制，匹配隧道组类映射时，仅支持出站策略管制。

配置 QoS

按照以下顺序在 ASA 上实施 QoS。

过程

步骤 1 确定优先级队列的队列和传输环路限制，第 4 页。

步骤 2 为接口配置优先级队列，第 5 页。

步骤 3 为优先级排队和策略管制配置服务规则，第 6 页。

确定优先级队列的队列和传输环路限制

使用下列工作表确定优先级队列和传输环路限制。

队列限制工作表

下列工作表显示如何计算优先级队列大小。由于队列大小有限制，队列可以填满和溢出。当队列已满时，任何额外的数据包都无法进入队列并将被丢弃（称为尾部丢弃）。要避免队列被填满，可以根据为接口配置优先级队列，第 5 页调整队列缓冲区大小。

关于工作表的小提示：

- 出站带宽 - 例如，DSL 的上行链路速度可能为 768 Kbps。请与运营商核对。
- 平均数据包大小 - 通过编码解码器或采样量确定此值。例如，对于 VPN 上 的 VoIP，可以使用 160 字节。如果不知道使用哪种大小，我们建议使用 256 字节。
- 延迟 - 延迟取决于应用。例如，VoIP 建议的最大延迟是 200 毫秒。如果不知道使用哪种延迟，我们建议使用 500 毫秒。

表 *1:* 队列限制工作表

1	_____ 出站带宽（单位为 *Mbps* 或 *Kbps*）	Mbps	x	**125**	=	_____ 字节数/毫秒		
		kbps	x	**0.125**	=	_____ 字节数/毫秒		
2	_____ 步骤 *1* 中的字节数/毫秒		÷	_____ 平均数据包大小（字节）	x	_____ 延迟（毫秒）	=	_____ 队列限制 （数据包数）

传输环路限制工作表

下列工作表显示如何计算传输环路限制。此限制确定在以太网传输驱动器推回到接口的队列之前，允许进入驱动器的数据包的最大数量，以便缓冲数据包，直到堵塞消除为止。该设置确保基于硬件的传输环路对高优先级数据包施加有限数量的额外延迟。

关于工作表的小提示：

- 出站带宽 - 例如，DSL 的上行链路速度可能为 768 Kbps。请与运营商核对。
- 最大数据包大小 - 通常，最大数据包大小为 1538 字节（标记的以太网为 1542 字节）。如果允许超巨型帧（如果平台支持），则该数据包大小可能更大。
- 延迟 - 延迟取决于应用。例如，要控制 VoIP 的抖动，应使用 20 毫秒。

表 *2:* 传输环路限制工作表

1	_____ 出站带宽（单位为 *Mbps* 或 *Kbps*）	Mbps	x	**125**	=	_____ 字节数/毫秒		
		kbps	x	**0.125**	=	_____ 字节数/毫秒		
2	_____ 步骤 *1* 中的字节数/毫秒		÷	_____ 最大数据包大小（字节）	x	_____ 延迟（毫秒）	=	_____ 传输环路限制（数据包数）

为接口配置优先级队列

如果启用物理接口上流量的优先级排队，则需要在每个接口上创建优先级队列。每个物理接口使用两个队列：一个用于优先级流量，另一个用于所有其他的流量。对于其他流量，可以选择配置策略管制。

过程

步骤 1 创建接口的优先级队列。

priority-queue *interface_name*

示例：

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

interface_name 参数指定要启用优先级队列的物理接口名称。

步骤 2 更改优先级队列的大小。

queue-limit *number_of_packets*

默认队列限制为1024个数据包。由于队列大小有限制，队列可以填满和溢出。当队列已满时，任何额外的数据包都无法进入队列并将被丢弃（称为尾部丢弃）。以避免队列填充起来，您可以使用 **queue-limit** 命令以提高队列缓冲区大小。

用于 **queue-limit** 命令的值范围上限在运行时动态决定。要查看此限制，请在命令行中输入 **queue-limit?**。关键决定因素是支持队列所需的内存和设备上可用的内存。

指定的 **queue-limit** 对更高优先级的低延迟队列和“尽力而为”队列都有影响。

示例：

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

步骤 3 指定优先级队列的深度。

tx-ring-limit *number_of_packets*

默认 tx-ring-limit 是 511 个数据包。此命令设定在以太网传输驱动器推回到接口上的队列之前，允许进入驱动器的低延迟或正常优先级数据包的最大数量，以便缓冲数据包，直到堵塞消除为止。该设置确保基于硬件的传输环路对高优先级数据包施加有限数量的额外延迟。

用于 **tx-ring-limit** 命令的值范围上限在运行时动态决定。要查看此限制，请在命令行中输入 **tx-ring-limit ?**。关键决定因素是支持队列所需的内存和设备上可用的内存。

您指定的 **tx-ring-limit** 会影响较高优先级低延迟队列和尽力而为队列。

示例：

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

示例

以下示例在接口“outside”（GigabitEthernet0/1 接口）建立优先级队列，默认值为 queue-limit 和 tx-ring-limit：

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

以下示例在接口“outside”（GigabitEthernet0/1 接口）建立优先级队列，将 queue-limit 设置为 260 个数据包，并将 tx-ring-limit 设置为 3：

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

为优先级排队和策略管制配置服务规则

可以为同一策略映射中不同类映射配置优先级排队和策略管制。关于有效 QoS 配置的信息，请参阅如何交互使用 QoS 功能，第 2 页。

开始之前

- 对于优先级流量，无法使用 **class - default** 类映射。
- 对于策略管制，不支持流向设备的流量。
- 对于策略管制，往返 VPN 隧道的流量会绕过接口策略管制。
- 对于策略管制，匹配隧道组类映射时，仅支持出站策略管制。
- 对于优先级流量，仅识别延迟敏感型流量。
- 对于策略管制流量，可以选择对其他流量进行策略管制，也可以将流量限制到某些类型。

过程

步骤 1 创建一个 L3/L4 类映射，以标识要对其执行优先级排队的流量。

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

示例：

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

有关详细信息， 请参阅 为通过流量创建第 3/4 层类映射。

步骤 2 创建一个 L3/L4 类映射，以标识要对其执行优先级策略管制的流量。

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

示例：

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

提示
如果使用 ACL 进行流量匹配，仅在 ACL 指定的方向上应用策略管制。即从源到目标的流量受到策略管制，但是从目标到源的流量则不受策略管制。

步骤 3 添加或编辑策略映射：**policy-map** *name*

示例：

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

步骤 4 标识您为应用优先级排列的流量创建的类映射，并为该类配置优先级排队。

```
class priority_map_name
priority
```

示例：

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

步骤 5 标识您为应用策略管制的流量创建的类映射：**class** *name*

示例：

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

步骤 6 配置类的策略管制。

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

选项有：

- **output**- 为流向输出方向的流量启用策略管制。
- **input**- 为流向输入方向的流量启用策略管制。
- *conform-rate* - 为此流量类设置速率限制，范围从每秒 8000 位到 2000000000 位。例如，要将流量限制为 5Mbps，请输入 5000000。
- *conform-burst* -（可选。）指定控制到符合速率值之前连续突发中允许的最大即时字节数，该值介于 1000 到 512000000 字节之间。如果省略该变量，突发大小将按符合速率的 1/32 计算（以字节为单位）。例如，5Mbps 速率的突发大小为 156250。
- **conform-action** -（可选。）设置流量低于管制速率和突发大小时要执行的操作。可以丢弃或传输流量。默认为传输流量。
- **exceed-action** -（可选。）设置流量超过策略管制速率和突发大小时要执行的操作。可以丢弃或传输超过策略管制速率和突发大小的数据包。默认情况下会丢弃多余的数据包。

示例：

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

步骤 7 激活一个或多个接口上的策略映射。

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

示例：

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

global 选项将策略映射应用于所有接口，而 **interface** 选项将策略应用于一个接口。仅允许存在一个全局策略。可以通过向接口应用服务策略来覆盖该接口上的全局策略。每个接口只能应用一个策略映射。

监控 QoS

以下主题介绍如何监控 QoS。

QoS 策略统计信息

要查看流量策略管制的 QoS 统计信息，请使用 **show service-policy police** 命令。

```
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 优先级统计信息

要查看实施 **priority** 命令的服务策略的统计信息，请使用 **show service-policy priority** 命令。

```
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”表示此接口中的汇聚丢弃；“Aggregate transmit”表示此接口中已传输数据包的汇聚数量。

QoS 优先级队列统计信息

要显示某个接口的优先级队列统计信息，请使用 **show priority-queue statistics** 命令。结果将显示尽力而为 (BE) 队列和低延迟队列 (LLQ) 的统计信息。以下示例显示对名为 test 的接口应用 **show priority-queue statistics** 命令。

```
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#
```

在此统计报告中：

- “Packets Dropped”表示此队列中已丢弃数据包的总数量。
- “Packets Transmit”表示此队列中已传输数据包的总数量。
- “Packets Enqueued”表示此队列中已排队数据包的总数量。
- “Current Q Length”表示此队列当前的深度。
- “Max Q Length”表示此队列曾发生过的最大深度。

优先级排队和策略管制的配置示例

以下各节提供配置优先级排队和策略管制的示例。

VPN 流量的类映射示例

在以下示例中，**class-map** 命令采用一个名为 tcp_traffic 的 ACL 对所有不通过隧道传输的 TCP 流量进行分类：

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

在以下示例中，使用其他更具体的匹配条件来分类与安全相关的特定隧道组的流量。这些具体匹配标准规定隧道组（在本示例中指先前定义的隧道组 1）上的匹配需要作为第一个匹配特征以对特定隧道的流量进行分类，并且此匹配允许一个附加匹配线对流量进行分类（IP 差分服务代码点，加速转发）。

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

在以下示例中，**class-map** 命令根据流量类型对通过隧道传输和不通过隧道传输的流量分类：

```
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
```

以下示例显示一种在隧道内对流量进行策略管制的方式，前提是分类的流量未被指定为隧道，而是经过隧道。在本示例中，192.168.10.10 是远程隧道的专用端上的主机地址，且 ACL 命名为“host-over-l2l”。通过创建类映射（名为“host-specific”），可在 LAN 到 LAN 连接实施隧道策略管制前对“host-specific”类进行策略管制。在本例中，在通过隧道传输前先对“host-specific”流量应用速率限制，再对隧道应用速率限制：

```
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
```

优先级和策略管制示例

以下示例构建在上节中开发的配置之上。在上述示例中，有两个命名的类映射：tcp_traffic 和 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
```

添加第三个类映射可为定义通过隧道传输和不通过隧道传输的 QoS 策略建立基础，以下示例中为通过隧道传输和不通过隧道传输的流量创建了一个简单 QoS 策略，将 TG1-voice 类的数据包分配到低延迟队列，并设置了 tcp_traffic 和 TG1-best-effort 流量数据流上的速率限制。

在本示例中，tcp_traffic 类流量的最大速率为 56,000 位/秒，最大突发大小为 10,500 字节/秒。TC1-BestEffort 类的最大速率为 200,000 位/秒，最大突发大小为 37,500 字节/秒。TC1-voice 类流量的最大速度或突发速率不受策略管制，因为它们属于优先级类。

```
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
```

QoS 的历史记录

功能名称	平台版本	说明
优先级排队和策略管制	7.0(1)	引入了 QoS 优先级排队和策略管制。 引入了以下命令：**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** 和 **clear configure priority-queue**。
整形和分级式优先级排队	7.2(4)/8.0(4)	引入了 QoS 整形和分级式优先级排队。 引入了以下命令：**shape**、**show service-policy shape**。
ASA 5585-X 标准优先级队列支持万兆以太网	8.2(3)/8.4(1)	我们为 ASA 5585-X 支持万兆以太网接口上的标准优先级队列。

当地语言翻译版本说明

思科可能会在某些地方提供本内容的当地语言翻译版本。请注意，翻译版本仅供参考，如有任何不一致之处，以本内容的英文版本为准。