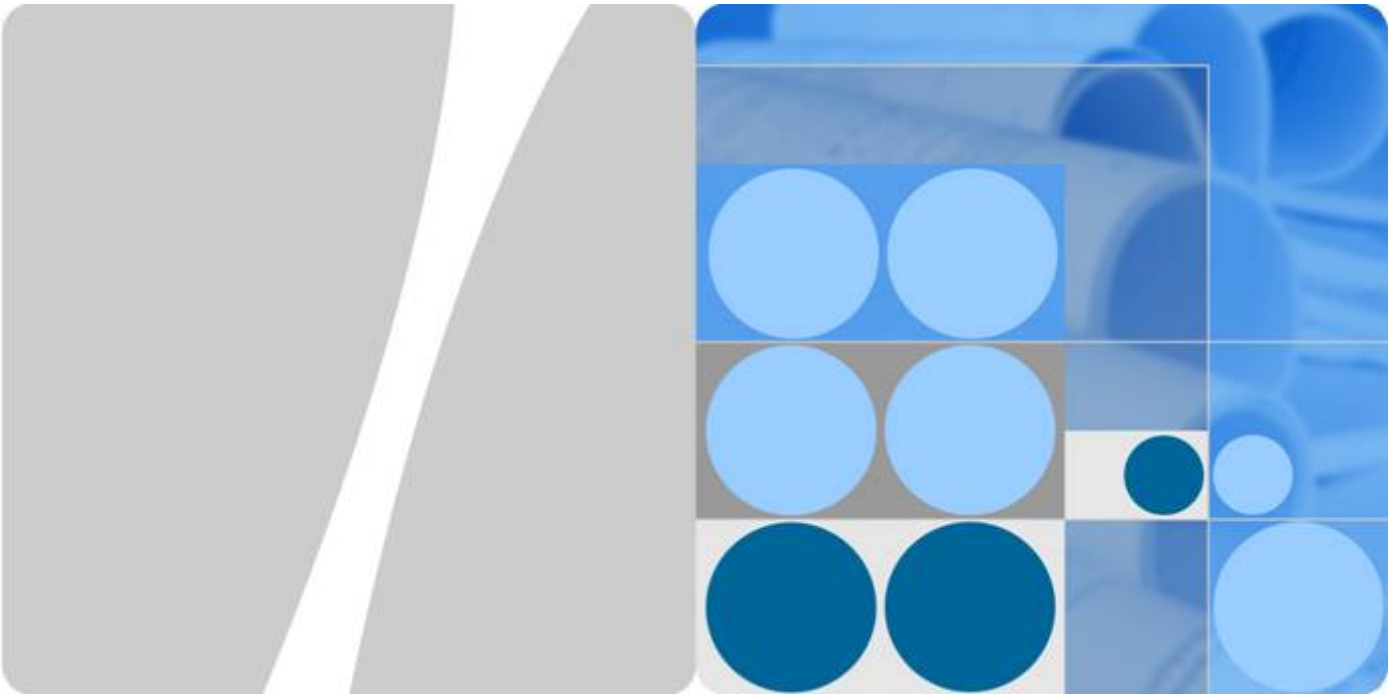


HUAWEI NetEngine20E-X6 高端业务路由器

V600R003C00

特性描述-QoS

文档版本　01

发布日期　2011-05-15

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 0755-28560000 4008302118

客户服务传真： 0755-28560111

前 言

读者对象

本文档针对分流和引流特性，从简介、原理描述和应用三个方面介绍了 QoS 特性。

本文档与其它类型手册相结合，便于读者深入掌握 QoS 特性的实现原理。

本文档主要适用于以下工程师：

- 网络规划工程师
- 调测工程师
- 数据配置工程师
- 系统维护工程师

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
危险	以本标志开始的文本表示有高度潜在危险，如果不能避免，会导致人员死亡或严重伤害。
警告	以本标志开始的文本表示有中度或低度潜在危险，如果不能避免，可能导致人员轻微或中等伤害。
注意	以本标志开始的文本表示有潜在风险，如果忽视这些文本，可能导致设备损坏、数据丢失、设备性能降低或不可预知的结果。
窍门	以本标志开始的文本能帮助您解决某个问题或节省您的时间。
说明	以本标志开始的文本是正文的附加信息，是对正文的强调和补充。

修订记录

修改记录累积了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 01 (2011-05-15)

第一次正式归档。

目 录

插图目录

表格目录

1 流量监管和流量整形

关于本章

1.1 介绍

定义

- 流量监管

  流量监管 TP（Traffic Policing）是一种在入接口或出接口应用的对进入路由器的某流量进行限制的流量管理技术。

- 流量整形

  流量整形采用的技术叫做 Generic Traffic Shaping（通用流量整形，简称 GTS），可以对不规则或不符合预定流量特性的流量进行整形，以利于网络上下游之间的带宽匹配。

目的

当网络发生拥塞的时候，利用流量监管（采用 CAR 技术）可以控制报文的流量特性，对流量加以限制，将不符合流量特性的报文进行丢弃。有时，为了减少报文的丢弃，可以先将超出规格的报文进行缓冲，然后在令牌桶的控制下再均匀地发送，这就是流量整形。流量整形既可以减少报文的丢弃，同时又能满足报文的流量特性。

流量整形的典型作用是限制流出某一网络的某一连接的流量与突发，使这类报文以比较均匀的速度向外发送。

受益

- 运营商受益

  监督并对超出的部分流量进行“惩罚”，以保护网络资源和运营商的利益。

1.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC 2697	A Single Rate Three Color Marker.txt	无
RFC 2698	A Two Rate Three Color Marker	无

1.3 原理描述

1.3.1 流量监管的基本原理

流量监管的实现机制

- 单令牌桶监管

  在单令牌桶监管中，采用一个令牌桶，容量是 CBS (Committed Burst Size)，称为 C 桶。用 Tc 表示桶中的令牌数量，Tc 初始化值等于 CBS。桶中的填充令牌速率为 CIR (Committed Information Rate)。当有报文传过来的时候，根据令牌桶的当前容量来对这个报文进行处理（着色、抛弃、传送等）。

  首先以 CIR 的速率向令牌桶 C 中注入令牌，直到令牌桶满（达到 CBS），多余令牌被丢弃。

  Tc 在每秒钟内更新 CIR 次，每次更新时遵循以下规则：

  - 如果 Tc<CBS，则 Tc 增加 1。
  - 否则，Tc 不增加。

  具体处理流程又分为两种情况：

  - 一种是没有预先着色的（Color-Blind Mode），即色盲方式。
  - 一种是已经预先着过色的（Color-Aware Mode），即非色盲方式。

  色盲模式下，在对到达报文（假设报文大小为 B）进行测量时，遵循以下规则：

  - 如果 Tc-B >= 0，则报文被标记为绿色，且 Tc 降低 B。
  - 如果 Tc-B < 0，则报文被标记为红色，且 Tc 不降低。

  非色盲模式下，在对到达报文（假设报文大小为 B）进行测量时，遵循以下规则：

  - 如果报文已被标记为绿色，且 Tc-B >= 0，则报文被标记为绿色，且 Tc 降低 B。
  - 否则，报文被标记为红色，且 Tc 不降低。

- 双令牌桶监管
  - 单速率监管

    在单速率监管中，采用两个令牌桶，一个桶的容量是 CBS (Committed Burst Size)，一个桶的容量是 PBS (Peak Burst Size)，分别称为 C 桶和 P 桶。用 Tc 和 Tp 表示桶中的令牌数量，Tc 和 Tp 初始化时分别等于 CBS 和 PBS。两个桶采用同一个填充令牌的速率 CIR(Committed Information Rate)。当有报文传过来的时候，根据两个桶的当前容量来对这个报文进行处理（着色、抛弃、传送等）。

    首先以 CIR 的速率向令牌桶 C 和 P 中注入令牌，直到令牌桶满（分别达到 CBS 和 PBS），多余令牌被丢弃。Tc 和 Tp 初始化时分别等于 CBS 和 PBS。

    Tc 和 Tp 在每秒钟内更新 CIR 次，每次更新时遵循以下规则：

    - 如果 Tc<CBS，则 Tc 增加 1。
    - 如果 Tp<PBS，则 Tp 增加 1。
    - 否则，Tc 和 Tp 都不增加。

    具体处理流程又分为两种情况：

    - 一种是没有预先着色的（Color-Blind Mode），即色盲方式。
    - 一种是已经预先着过色的（Color-Aware Mode），即非色盲方式。

    色盲模式下，在对到达报文（假设报文大小为 B）进行测量时，遵循以下规则：

    - 如果 Tc-B >= 0，则报文被标记为绿色，且 Tc 降低 B。
    - 如果 Tc-B < 0 并且 Tp-B >= 0，则报文被标记为黄色，且 Tp 降低 B。
    - 如果 Tp-B < 0，则报文被标记为红色且 Tc 和 Tp 都不降低。

非色盲模式下，在对到达报文（假设报文大小为 B）进行测量时，遵循以下规则：

- 如果报文已被标记为绿色，且 Tc-B >= 0，则报文被标记为绿色，且 Tc 降低 B。
- 如果报文已被标记为绿色或黄色且 Tc-B < 0 且 Tp-B >= 0，则报文被标记为黄色，且 Tp 降低 B。
- 否则，报文被标记为红色且 Tc 和 Tp 都不降低。

- 双速率监管

当网络流量情况比较复杂的情况下，可以采用双速率监管。

在双速率监管中，采用两个令牌桶，一个桶的容量是 CBS (Committed Burst Size)，一个桶的容量是 PBS (Peak Burst Size)，分别称为 C 桶和 P 桶。用 Tc 和 Tp 表示桶中的令牌数量，Tc 和 Tp 初始化时分别等于 CBS 和 PBS。CBS 比 PBS 要小。

两个令牌桶的填充令牌的速率不同，分别为承诺的平均速率 CIR（Committed Information Rate）和峰值速率 PIR（Peak Information Rate）。当有报文传过来的时候，根据两个桶的当前容量来对这个报文进行处理（着色、抛弃、传送等）

首先分别以 CIR 和 PIR 的速率向令牌桶 C 和 P 中注入令牌，直到令牌桶满（分别达到 CBS 和 PBS），多余令牌被丢弃。Tc 和 Tp 初始化时分别等于 CBS 和 PBS。

Tc 和 Tp 在每秒钟内分别更新 CIR 和 PIR 次，每次更新时遵循以下规则：

- 如果 Tc<CBS，则 Tc 增加 1。
- 如果 Tp<PBS，则 Tp 增加 1。
- 否则，Tc 和 Tp 都不增加。

具体处理流程也分为两种情况：色盲方式和非色盲方式。

色盲模式下，在对到达报文（假设报文大小为 B）进行测量时，遵循以下规则：

- 如果 Tp-B < 0，则报文被标记为红色。
- 如果 Tp-B >= 0 且 Tc-B < 0，则报文被标记为黄色，且 Tp 降低 B。
- 否则，报文被标记为绿色且 Tc 和 Tp 都降低 B。

在非色盲模式下，在对到达报文（假设报文大小为 B）进行测量时，遵循以下规则：

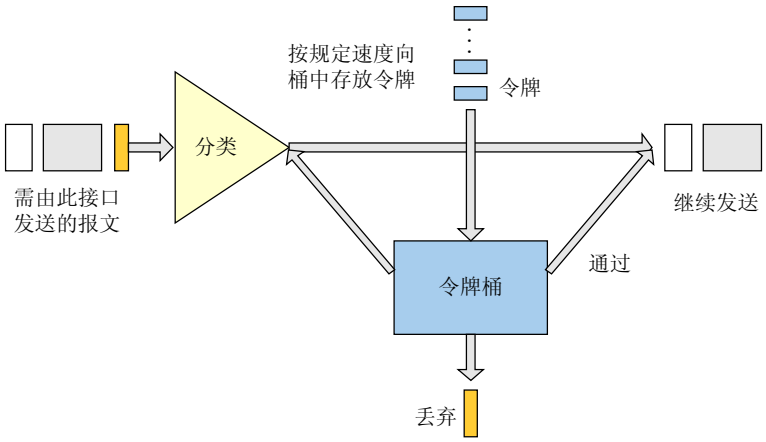
- 如果报文已被标记为红色或者 Tp-B < 0，则报文被标记为红色。
- 如果报文已被标记为黄色或者 Tp-B 〉= 0 且 Tc-B < 0，则报文被标记为黄色，且 Tp 降低 B。
- 否则，报文被标记为绿色且 Tc 和 Tp 都降低 B。

流量监管的处理流程

流量监管采用承诺访问速率 CAR（Committed Access Rate）来对流量进行控制。首先，根据预先设置的匹配规则来对报文进行分类，如果是符合流量规定的报文，则报文通过继续发送；如果是超出流量规定的报文，可以选择丢弃报文或重新设置报文的优先级再发送。

CAR 技术对流量进行控制采用令牌桶 TB（Token Bucket）实现。具体处理流程如**图 1-1** 所示。

图 1-1 CAR 处理过程示意图



- 令牌桶按用户设定的速度向桶中放置令牌，并且令牌桶有用户设定的容量，当桶中令牌的数量超出桶的容量的时候，令牌的数量不再增加。
- 当报文到来时，首先根据 IP precedence 或源或目的地址等信息进行分类。满足某类特征的报文进入令牌桶中进行处理。
- 如果令牌桶中有足够的令牌可以用来发送报文，则报文直接通过并继续发送，同时令牌桶中的令牌量按报文的长度做相应的减少；如果令牌桶中的令牌数量不足或为空，则无法得到足够转发令牌的报文将被丢弃或进入标记器进行 IP Precedence、DSCP 或 EXP 值的重标记然后再发送，此时令牌桶中的令牌数量不发生变化。

从上述可以看出，CAR 技术不仅可以做到流量控制，还可以进行报文的标记（mark）或重新标记（remark）。

速率限制功能是 CAR 的主要功能。主要是通过使用令牌桶对流经端口的数据流进行度量，使得在特定时间内只有得到令牌的流量通过，从而实现限速功能。也就是说它可以限制接口入和出两个方向的流量的最大速率。同时我们还可以根据特定的数据特征来对特定的数据流进行速率控制，如针对数据的 IP 地址，端口号，优先级等。不符合条件的数据流设备将不进行限速处理，以端口原速率转发。

CAR 技术主要应用于网络边缘，从而保证核心设备的正常数据处理。

1.3.2 流量整形的基本原理

流量整形的基本概念

- 令牌桶

  流量整形采用令牌桶和缓存队列一起实现。令牌桶可以看作是一个存放令牌的容器，预先设定一定的容量。系统按用户设定的速度向桶中放置令牌，当桶中令牌满时，多出的令牌溢出，且令牌的量不再增加。

  具体处理过程如下：

  - 当令牌桶中有足够的令牌可以用来发送报文时，则报文可以通过并被继续发送下去，同时令牌桶中的令牌量按报文的长度做相应的减少。
  - 当令牌桶内的令牌数量不足或为空时，报文缓存进入 GTS 队列。
- GTS 队列

如果报文需要进行 GTS 处理，并且令牌桶中的令牌不足时，报文进入缓存队列，称为 GTS 队列。

当 GTS 队列中有报文的时候，GTS 按一定的周期从队列中取出报文进行发送。每次发送都会与令牌桶中的令牌数作比较，直到令牌桶中的令牌数减少到队列中的报文不能再发送或是队列中的报文全部发送完毕为止。

对于参与流量整形的分组，当有 GTS 队列存在的时候，直接进入队列，等待 GTS 队列按照一定周期对队列中的分组进行调度转发。缓存报文队列采用 PQ 或 WFQ 调度算法。采用这种调度优势在于，既能对时延敏感的实时业务得到保证，对优先业务的报文的带宽占用可以绝对优先，又可以为不同优先级的流根据配置的权重分配不同的带宽。

当分组到来而发现 GTS 队列已满时，分组被丢弃。

流量整形的实现

目前，NE20E-X6 只支持接口出方向报文的流量整形。并且只支持对端口所有报文进行流量整形。

- 在端口上，对于参与流量整形的报文可以根据不同的服务等级（EF、AF1、AF2、AF3、AF4、BE、CS6 或 CS7）配置不同的整形参数。
- GTS 队列的调度模式可以采用 PQ 调度也可以采用 WFQ 调度。系统对于 GTS 队列中的不同服务级别报文的调度模式有默认值，如下：
- 对 AF1 ～ AF4 以及 BE 队列默认配置成 WFQ 调度，根据配置的权重参数按比例分配带宽。
- 对于 EF、CS6、CS7 队列默认配置 PQ 调度，根据绝对优先级调度，一般适用于对时延敏感的业务。
- 当 GTS 队列采用 WFQ 调度时，可以配置 weight 值，表示 WFQ 队列的不同优先级业务之间的权值或每类流所占的带宽比例。该参数对于 PQ 队列没有意义，不必配置。
- 可以配置端口的 shaping 值，即向令牌桶中存放令牌的速率。如果报文的速率超过该值，将会进入 GTS 队列缓存。

在 NE20E-X6 的实现中，令牌桶的深度由系统设置好，无需人工配置。

流量整形的处理流程

为减少报文的无谓丢失，应在上游路由器出口对报文进行 GTS 处理。对于超出 GTS 流量特性的报文，缓存在上游路由器的接口缓冲区中。当网络拥塞消除时，GTS 再从缓冲队列中取出报文，继续发送。这样，发向下游路由器的报文都符合路由器的流量规定，从而减少报文在下游路由器被丢弃的概率。若不在上游路由器出口做 GTS 处理，则所有超出下游路由器的 CAR 规定流量的报文将被下游路由器丢弃。

流量整形和流量监管的区别

流量整形与流量监管的主要区别在于：

- 利用流量监管进行报文流量控制时，对不符合流量特性的报文直接进行丢弃。而流量整形对流量监管中超出流量规格的报文进行缓存，当令牌桶有足够的令牌时，再均匀的向外发送这些被缓存的报文。
- 流量整形可能会增加延迟，而流量监管几乎不引入额外的延迟。

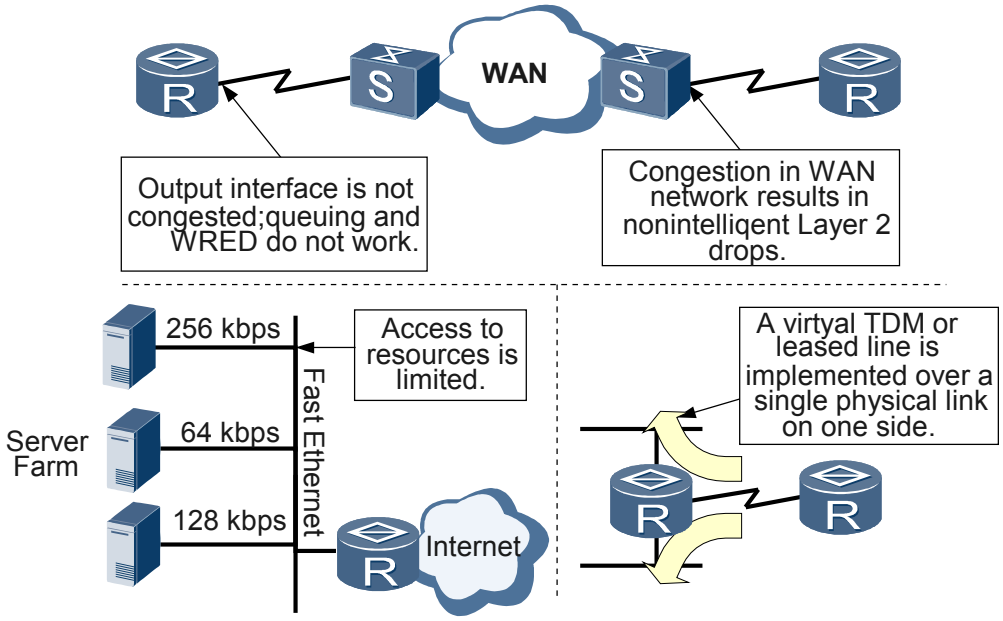
表 1-1 流量整形和流量监管的比较

类型	优点	缺点
流量整形	较少丢弃报文	引入时延和抖动，需要较多的 buffer 资源缓存报文。
流量监管	支持标记，不需使用额外的 buffer。	丢包可能引发重传。

1.4 应用

流量监管的应用

图 1-2 流量监管典型应用示意图



当从高速链路向低速链路传输数据时，带宽会在低速链路接口处出现瓶颈，导致数据丢失严重，特别是会影响到低延时要求的数据传输如语音传输等。这时我们除了要限制大流量数据的速率外，还需要保证语音数据的优先传输。这种情况下，可以通过流分类功能给语音数据流分配较高的优先级，采用 CAR 监管与队列调度配合来进行通信保证。

- 基于接口的流量监管

  基于接口的流量监管是指对进入该接口的所有流量进行控制，而不区分具体报文的类型，一般应用于网络核心路由器。可以同时应用于流量的入方向和出方向。

- 基于流的 CAR 策略

  基于流的 CAR 策略是指对进入该接口的满足特定条件的某一类或几类报文进行流量控制，而非所有报文。具体实现时，首先根据五元组（源地址、源端口号、协议号码、目的地址、目的端口号）等报文信息对报文进行分类，对于不同类的报文配

置不同的 CAR 行为，通过流策略将流分类与流量控制行为相结合，然后在接口应用流策略。

一个端口上可以有多个 CAR 策略，用于处理不同的数据流。例如，若想针对几个特定源地址的数据流进行速率限制分别为 1.1.1.1 ，2.2.2.1 ，3.3.3.1，可以针对这三个地址进行 CAR 的设置，生成三个 CAR 策略。CAR 策略生效的顺序是按照配置的顺序排列的，最先配置的策略会在数据流到达端口时最先生效。

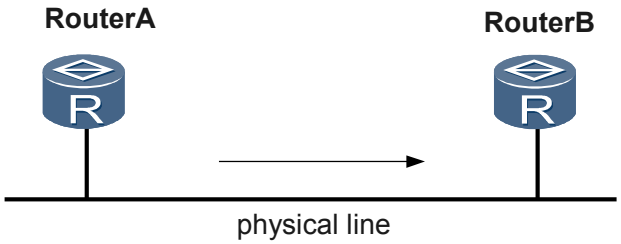
CAR 往往是和其他的 QOS 策略共同来提供全网 QOS 保障的。当 CAR 与其他 QOS 策略共同工作时，各个策略之间的发生顺序如下：

- 当策略是配置在入端口方向的时候，所有的策略都将在数据被作出转发决定之前生效，其中当设置了 CAR 与其他的多种策略如 PQ、WFQ 等，则其他策略会在 CAR 生效之后生效
- 当策略是配置在出端口方向的时候，所有的策略都会在数据被作出转发决定后的瞬间生效，其中当设置了 CAR 与其他的多种策略如 PQ、WFQ 等，则其他策略会在 CAR 生效之前生效

流量整形的应用

典型应用如图 1-3 所示。

图 1-3 流量整形典型应用示意图



如图 1-3 所示的应用中，RouterA 向 RouterB 发送报文，为了减少报文的丢失，可以在 RouterA 的出接口对报文进行 GTS 处理，对于超出 GTS 流量特性的报文，将在 RouterA 中缓存。当可以继续发送下一批报文时，GTS 再从缓冲队列中取出报文进行发送。这样，发往 RouterB 的报文将都符合 RouterB 的流量规定，从而减少报文在 RouterB 上的丢弃。

1.5 术语与缩略语

术语

术语	解释
Committed Access Rate（CAR）	流量监管的一个实例。CAR 可以定义四个流量参数：承诺信息速率 CIR（Committed Information Rate）、承诺突发尺寸 CBS（Committed Burst Size）、峰值信息速率 PIR（Peak Information Rate）、峰值突发尺寸 PBS（Peak Burst Size），依据它们对流量进行评估。CAR 还包括了对监管对象的流分类及监管动作的定义
Generic Traffic Shaping（GTS）	通用流量整形，一种流量整形措施的实例，它采用的排队策略是 WFQ。
Traffic policing	流量监管，一种监督特定流量进入路由器的规格的机制，当流量超出规格时，可以采取限制或惩罚措施，以保护运营商的商业利益和网络资源不受损害。
Traffic Shaping	流量整形，是一种主动调整流的输出速率的流控措施，通常是为了使流量适配下游路由器可供给的网络资源，避免不必要的报文丢弃和拥塞。

缩略语

缩略语	英文全称	中文全称
CAR	Committed Access Rate	承诺访问速率
CBS	Committed Burst Size	承诺突发尺寸
CIR	Committed Information Rate	承诺信息速率
TP	Traffic Policing	流量监管
TS	Traffic Shaping	流量整形
WFQ	Weighted Fair Queuing	加权公平队列调度

2 拥塞避免和拥塞管理

关于本章

2.1 介绍

定义

- 拥塞避免

  拥塞避免，是指通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载的一种流控机制。

- 拥塞管理

  拥塞管理是指网络在发生拥塞时，如何进行管理和控制。处理的方法是使用队列调度技术。将所有要从一个接口发出的报文进入多个队列，按照各个队列的优先级进行处理。通过适当的队列调度机制，可以优先保证某种类型的报文的 QoS 参数，例如带宽、时延、抖动等。我们这里所说的队列是指出队列，其作用是在接口有能力发送报文之前先将报文在队列中保留下来。所以队列调度机制都是在出端口发生拥塞情况下产生作用，另外一个主要作用就是将报文重新排序（先进先出队列除外）。

目的

拥塞避免和拥塞管理都是为了用来解除网络流量过载的一种机制。

受益

- 用户受益

  当网络拥塞时，用户的高优先级流量被优先保证。

2.2 参考标准和协议

无

2.3 原理描述

2.3.1 拥塞避免基本原理

拥塞避免是指网络在发生拥塞之前根据队列状态进行有选择性的丢包，这样可以提高拥塞流量的 QoS 性能。

拥塞避免传统的处理方法是尾丢弃，当网络发生拥塞时，所有到来的报文都被丢弃。对于 TCP 报文，由于大量的报文被丢弃，将造成 TCP 超时，从而引发 TCP 慢启动，使得 TCP 减少报文的发送。当队列同时丢弃多个 TCP 连接的报文时，将造成多个 TCP 连接同时进入慢启动和拥塞避免，称之为“TCP 全局同步”。这样多个 TCP 连接发往队列的报文将同时减少，使得队列中的报文数量达不到线路发送的速度，降低了线路的带宽利用率。

为了避免这些问题，必须在队列将要发生拥塞之前进行丢弃。WRED 就是在队列拥塞前进行报文丢弃的一种拥塞避免机制。WRED 会概率丢弃持续增长可能造成拥塞的报文，使TCP 连接所占用的输出带宽缓慢降低，不会引起大量的 TCP 慢同步，并减低平均队列的长度，降低流量的延时。

路由器同时支持尾丢弃和 WRED 两种算法作为拥塞避免机制。对于 DiffServ 模型，系统为每个端口预留 8 个业务队列，分别对应 BE，AF1 至 AF4，EF，CS6，CS7 等业务类别，对 AF1~AF4 以及 BE 队列默认配置成 WFQ 调度，根据配置的权重参数按比例分配带宽。EF，CS6，CS7 队列默认配置 PQ 调度。

WRED 算法

RED 算法可以很好地解决 TCP 全局同步问题。但其不能感知任何 QoS 信令，所有类别的报文被同等对待，灵活性较差。为了对不同的报文队列采取有区别的丢弃策略，引入了加权随机早期检测 WRED（Weighted Random Early Detection）丢弃策略。

WRED 算法与 RED 算法类似，也为每个队列都设定了一对低限值和高限值，并规定：

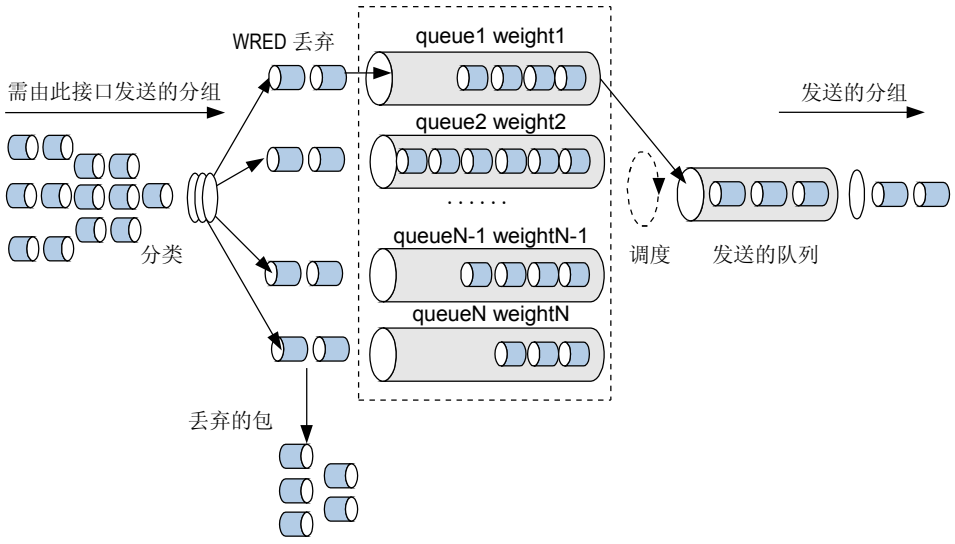
- 当队列的长度小于低限值时，不丢弃报文
- 当队列的长度超过高限值时，丢弃所有到来的报文。
- 当队列的长度在低限值和高限值之间时，WRED 开始随机丢弃到来的报文。方法是为每个到来的报文赋予一随机数，并用该随机数与当前队列的丢弃概率比较，如果大于丢弃概率则被丢弃。队列越长，丢弃概率越高，但有一个最大丢弃概率。
- 并且，为了避免对突发性的数据流造成不公正的待遇，也采用队列的平均长度与低限值和高限值做比较。

但与 RED 算法不同，WRED 生成的随机数是基于优先权的。它引入 IP 优先权 DSCP 值来区别丢弃策略，可以为不同 DSCP 值的报文设定不同的队列长度、队列阈值、丢弃概率，从而对不同优先级的报文提供不同的丢弃策略。这是 WRED 的重要特点。

- 当队列机制采用 WFQ 时，可以为不同优先级（precedence）的报文设定不同的低限值、高限值、丢弃概率。从而对不同优先级的报文提供不同的丢弃特性。
- 当队列机制采用 FIFO、PQ、CQ 时，可以为每个队列设定不同的低限值、高限值、丢弃概率，为不同类别的报文提供不同的丢弃特性。

WRED 和队列关系如图 2-1 所示。

图 2-1 WRED 与队列关系图



拥塞避免的实现机制

- PQ 队列

  对于 PQ 队列，其丢弃策略可以采用尾丢弃或者 WRED。对于实时性要求比较高的业务一般使用的是尾丢弃。因为这种报文要提供最大限度的保证。采用尾丢弃是只有当报文队列达到最大长度时才会丢弃，由于使用 PQ 调度，抢占其他业务的带宽，所以当发生拥塞时，实时性的业务带宽能够得到最大的保证。

- WFQ 队列

  对于 WFQ 队列，其丢弃策略默认采用尾丢弃，但一般采用 WRED。WFQ 调度通常针对优先级比较低，对时延不敏感的报文。WFQ 与 WRED 配合使用，可以针对不同的流配置不同的丢弃参数，达到不同的效果。

  路由器支持采用配置模板的方式实现 WRED。首先定义 WRED 模板，设置不同颜色报文的高、低门限值和丢弃概率，然后在端口上为不同服务等级的报文应用 WRED 模板。端口上队列 WRED 模板最多配置 8 个，每个模板最多支持 3 种颜色报文的处理，定义为红、黄、绿。一般绿色报文设置的丢弃概率比较小，门限值比较大；红色反之。每种颜色报文的门限值和丢弃概率都是可配置的，非常灵活。

  当拥塞发生时，队列开始缓存报文。根据报文分类，红色报文由于设置的低门限值比较小，丢弃概率比较大，红色报文最先开始丢包；随着队列的长度逐渐增长，最后才是绿色报文开始丢包。如果队列长度达到相应颜色的最大门限，这种颜色开始作尾丢弃。

  由于 WFQ 队列是按比例分享带宽，很容易发生拥塞，采用 WRED 策略有效的避免了 TCP 全局同步现象。

  目前，设备只支持在出方向（outbound）采用 WRED 策略。

- 流队列拥塞避免

  上行和下行的 FQ 队列支持 WRED 拥塞避免机制和尾丢弃机制。

- 端口队列拥塞避免

  下行 CQ 队列支持 WRED 拥塞避免机制和尾丢弃机制。

2.3.2 拥塞管理基本原理

常见的队列调度技术

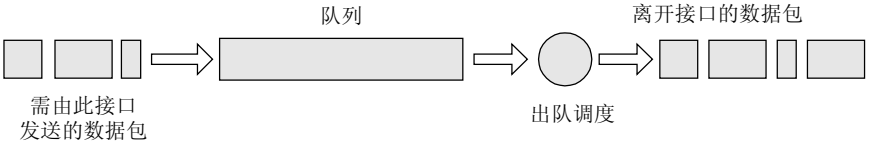
用户无法通过命令行设置 PQ 和 WFQ 队列的长度，PQ 队列长度由系统根据 PIR 参数和接口速率计算得出，WFQ 队列的长度由系统根据 CIR 参数和接口速率计算得出。

- FIFO

  FIFO 是最简单的队列机制。每个接口上只有一个 FIFO 队列。因此 FIFO 无需考虑流分类，也不存在队列如何调度问题，报文的入队列顺序和出队列顺序一致。 FIFO 只关心队列长度：队列的长度大小影响时延和丢包率。

  FIFO 使用尾丢弃（Tail Drop）机制。尾丢弃机制简单的说就是如果该队列已经满了，那么后续进入的报文将被丢弃，而没有什么机制来保证后续的报文可以挤掉已经在队列内的报文。

  图 2-2 FIFO 示意图



如图 2-2 所示，FIFO 不对报文进行分类，当报文进入接口的速度大于接口能发送的速度时，FIFO 按报文到达接口的先后顺序让报文进入队列。同时，FIFO 在队列的出口让报文按进队的顺序出队，先进的报文将先出队，后进的报文将后出队。

FIFO 如果定义了较长的队列长度，那么队列不容易填满，被丢弃的报文就少，但是队列长度太长会出现时延的问题；如果定义了较短的队列，时延的问题可以得到解决，但是发生尾丢弃的报文就会变多。在具体部署业务时，需要折衷考虑以得到较满意的效果。类似的问题在其他的队列调度机制中也存在。

- PQ

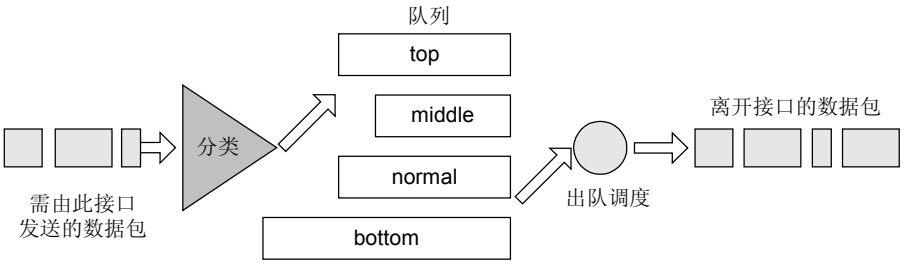
  通常情况下，PQ 将优先级队列分为四类：分别为高优先队列 Top、中优先队列 Middle、正常优先队列 Normal 和低优先队列 Bottom。它们的优先级依次降低。

  说明

  产品实现了 8 个优先级队列（0 ~ 7）。

  如图 2-3 所示。当报文到来时，PQ 首先对报文进行分类，将所有报文最多分成 4 类，分别属于 PQ 的 4 个队列中的一个。然后，按报文的类别将报文送入相应的队列。

图 2-3 PQ 队列示意图



在报文出队的时候，PQ 首先让高优先队列中的报文出队并发送，只要高优先级队列有报文，就一直从高优先级队列取报文。直到高优先队列中的报文发送完，然后才发送中优先队列中的报文，同样，直到发送完，然后依次是正常优先队列和低优先队列。

这样，分类时属于较高优先级队列的报文将会得到优先发送，而较低优先级的报文将会在发生拥塞时被较高优先级的报文抢先，使得关键业务（如 ERP）的报文能够得到优先处理，非关键业务（如 E-Mail）的报文在网络处理完关键业务后的空闲中得到处理，既保证了关键业务的优先，又充分利用了网络资源。

PQ 具有如下特征：

- 可以使用 ACL 对报文进行分类，根据需要将报文入队列。
- 报文丢弃策略采用 Tail Drop 机制，且只有这一种机制。
- 队列长度可以设置为 0，表示该队列无穷大，即进入该队列的报文不会被 Tail Drop 机制丢弃，除非内存耗尽了。
- 队列内部使用 FIFO 逻辑。
- 当从队列调度报文时，先从高优先级的队列调度报文。

PQ 的优缺点是很明显的。

- 优点是可以保证高优先级队列的报文可以得到较大带宽、较低的时延、较小的抖动。

－ 缺点是低优先级队列的报文不能得到及时的调度，会出现“饿死”现象。

为了解决 PQ“饿死”的问题，提出了 CQ 队列调度机制。CQ 有 0 ～ 16 个队列，其中 0 队列是优先级队列，只有 0 队列的报文处理完才会去处理 1 ～ 16 队列，所以 0 队列一般用做系统队列，通常把实时性要求高的交互式协议报文放到 0 号队列。1 到 16 号队列可以按用户的定义分配它们能占用接口带宽的比例，在报文出队的时候，CQ 按定义的带宽比例分别从 1 到 16 号队列中取一定量的报文在接口上发送出去。

CQ 采用 Round Robin 调度方式，能够使所有的队列都得到服务。

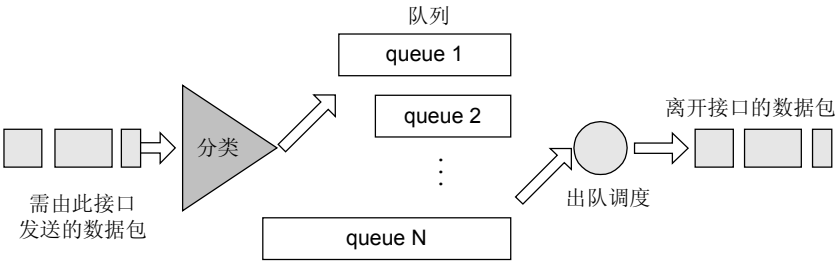
- WFQ

  加权公平队列（以下简称 WFQ）是一个复杂的排队过程，可以保证相同优先级业务间公平，不同优先级业务间加权。队列的数目可以预先配置，范围是 16 ～ 4096。

  WFQ 在保证公平（带宽、延迟）的基础上体现权值，权值大小依赖于 IP 报文头中携带的 IP 优先级（precedence）。WFQ 对报文按五元组信息（或者 TOS 域值）进行动态的流分类。相同源 IP 地址、目的 IP 地址、源端口号、目的端口号、协议号和 TOS 域值的报文属于同一个流。每一个流被分配到一个队列，该过程称为散列。WFQ 在入队过程采用 HASH 算法来自动完成，尽量将不同的流分入不同的队列。在出队的时候，WFQ 按流的优先级（precedence）来分配每个流应占有出口的带宽。优先级的数值越小，所分得的带宽越少。 优先级的数值越大，所分得的带宽越多。这样就保证了相同优先级之间业务的公平，体现了不同优先级业务之间的权值。

  WFQ 对报文的处理流程如图 2-4 所示。

图 2-4 WFQ 队列示意图



假设在当前接口中有 8 类流，它们的优先级分别为 0，1，2，3，4，5，6，7。则带宽的总配额将是所有 (流的优先级 + 1) 的和。即：1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36

每类流所占带宽比例为：（自己的优先级数 + 1）/（所有 (流的优先级 + 1) 的和）。即，每类流可得的带宽分别为：1/36，2/36，3/36，4/36，5/36，5/36，6/36，7 / 36，8/36。

又比如：假设当前接口中共有 4 个流，其中 3 个流的优先级为 4，1 个流的优先级为 5。则带宽的总配额将是：

(4 + 1) * 3 + (5 + 1) = 21

那么，其中 3 个优先级为 4 的流获得的带宽比例均为 5/21，优先级为 5 的流获得的带宽比例为 6/21。

由此可见，WFQ 在保证公平的基础上对不同优先级的业务体现权值，而权值依赖于 IP 报文头中所携带的 IP 优先级。

WFQ 的主要特征如下。

－ 基于五元组对报文进行流分类，不支持用户自定义的分类。

- 采用 WFQ 丢弃机制，是对 Tail Drop 的改进。
- WFQ 基于流的，每个流占有一个队列，一个接口最多可支持 4096 个队列。
- 对不同的队列采用 WFQ 调度机制；在队列内部采用 FIFO。

此外还有 CBWFQ 队列调度机制。CBWFQ 有些类似 CQ，可以为每个队列保留最小带宽，使用和 CQ 类似的报文分类，但是与 CQ 不同的是：用户可以配置 CBWFQ 实际占有的流量百分比，而不是字节数。

拥塞管理的实现机制

对于队列配置，用户无须关心采用什么抽象的调度算法，只需关心队列所承载业务的外在流量参数特征，比如保证多少兆的带宽、峰值最多多少兆的带宽、要占剩余带宽的比例权重等。根据配置的流量参数选用不同的调度算法来严格保证用户的配置。

设备的队列调度的构架是一个两级调度模式，即端口的 shapping 和端口队列 CQ 调度。对不同的队列参数配置，其内部将自动采用：

- PBS 峰值带宽保证算法，基于时间片的调度
- PQ 算法，基于优先级调度
- WFQ 算法，基于 weight 值的调度

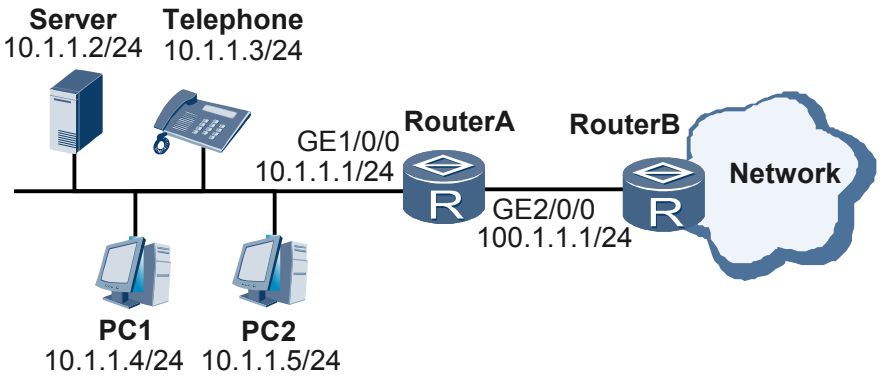
端口队列调度采用 PQ 或 WFQ 调度算法，采用这种调度优势在于，既能对时延敏感的实时业务得到保证，对优先业务的报文的带宽占用可以绝对优先，又可以为不同优先级的流根据配置的权重分配不同的带宽。

对于 DiffServ 模型，系统为每个端口预留 8 个业务队列，分别对应 BE，AF1 至 AF4，EF，CS6，CS7 等业务类别，对 AF1~AF4 以及 BE 队列默认配置成 WFQ 调度，根据配置的权重参数按比例分配带宽。EF，CS6，CS7 队列默认配置 PQ 调度，这种按照绝对优先级调度，一般是时延敏感的业务采用 PQ 调度。

2.4 应用

拥塞避免的组网及应用

图 2-5 拥塞避免典型应用示意图



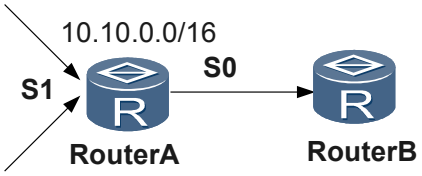
如图 2-5 所示，Server、Telephone、PC1 和 PC2 通过 RouterA 向网络发送数据，其中 Server 发送关键业务数据，Telephone 发送语音数据，PC1 和 PC2 发送非关键业务数

据。由于 RouterA 入接口 GE1/0/0 的速率大于出接口 GE2/0/0 的速率，在 GE2/0/0 接口处可能发生拥塞，并且可能出现拥塞加剧现象。

要求在网络拥塞时保证 Server 和 Telephone 发送的业务数据得到优先发送。但由于 PC1 和 PC2 是 VIP 用户，他们的数据在发送的过程中也需要一定的带宽保证，可以有少量延迟，但不希望延迟过大。在 RouterA 上配置 WFQ 和 WRED 配合调度和丢弃，在拥塞加剧的时候根据优先权来丢弃报文。

拥塞管理的组网及应用

图 2-6 拥塞管理典型应用示意图



如图 2-6，当处于公司局域网的 RouterA 通过 S0 接口向处于广域网中的 RouterB 发送数据时，由于广域网的带宽小于局域网的带宽，会使 RouterA 的接口 S0 发生拥塞。

为了对拥塞的接口进行管理和控制，需要使用队列技术，将所有要从 S0 接口发出的报文进行分类，送入多个不同的队列，按照各个队列不同的优先级分别进行处理。优先级高的报文会得到优先处理。

以 PQ 队列为例，可以设置满足访问控制列表 1 的报文进入 top 队列，从接口 S1 进入的报文进入 normal 队列，缺省队列为 middle 队列。设置好各个队列的最大长度后，把 PQ 规则组 1 应用到接口 S0 上。这样 PQ 队列将对不同的业务进行区别对待，既保证了高优先级的报文的正常转发，又对拥塞进行了管理。

2.5 术语与缩略语

术语

术语	解释
Congestion	是指在分组投递过程中由于供给资源的相对不足而造成服务速率下降的一种现象。拥塞会引发一系列的负面影响，影响 QoS。
Congestion Avoidance	拥塞避免，是通过监视网络资源的使用情况，在拥塞已经产生并且有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载的一种流控机制。
Congestion Management	拥塞管理，一种解决网络资源竞争的流控措施。它在网络发生拥塞时将报文放入队列中缓存，并采取某种调度策略决定报文的转发次序。
First In First Out Queueing（FIFO）	先进先出的排队策略，其特点是可为先到来的报文分配资源

术语	解释
Priority Queue（PQ）	优先队列，根据优先级进行排队的策略。特点是如果同时存在多种优先级的报文，高优先级的报文先被分配资源。
Weighted Fair Queue（WFQ）	加权公平排队策略，其特点是可以自动进行流分类，并且均衡各个流的延迟和延迟抖动。WFQ 与 FQ（公平队列）相比，考虑了优先级报文的利益。
Weighted Random Early Detection（WRED）	加权随机早期检测，一种用于拥塞避免的丢包算法，可以避免传统的尾部丢包（Tail-Drop）所带来的 TCP 全局同步现象，并在计算报文的丢包概率时，考虑了高优先级报文的利益。

缩略语

缩略语	英文全称	中文全称
CBQ	Class-based Queue	基于类的队列
FIFO	First In First Out	先入先出
FQ	Fair Queue	公平队列
PQ	Priority Queue	优先队列
WFQ	Weighted Fair Queuing	加权公平排队
WRED	Weighted Random Early Detection	加权随机早期检测

3 基于类的 QoS

关于本章

3.1 介绍

定义

流分类是指根据报文的某些信息定义一些匹配规则对报文进行分类，匹配不同规则的报文实施不同的 QoS 策略。

根据分类规则参考信息的不同，流分类分为简单流分类 BA（Behavior Aggregation）和复杂流分类 MF（Multiple Field）。

- 简单流分类

  简单流分类是将数据报文划分为多个优先级或多个服务类，如使用 IP 报文头的服务类型 ToS（Type of Service）字段的前三位（即 IP 优先级）来标记报文，可以将报文最多分成 8 类；若使用区分服务编码点 DSCP（Differentiated Services Code Point，ToS 域的前 6 位），则最多可分成 64 类。在报文分类后，就可以将其它的 QoS 特性应用到不同的分类，实现基于类的拥塞管理、流量整形等。

  网络管理者可以设置报文简单流分类的策略，这个策略可以包括 IP 报文的 IP 优先级或 DSCP 值、MPLS 报文的 EXP 域值、VLAN 报文的 802.1p 值等。

- 复杂流分类

  复杂流分类是指根据五元组（源 IP 地址、源端口号、协议号码、目的 IP 地址、目的端口号）、TCP SYN 等报文信息对报文进行分类（一般的分类依据都局限在封装报文的头部信息，使用报文内容作为分类的标准比较少见），缺省应用于网络的边缘位置。报文进入边缘节点时，网络管理者可以灵活配置分类规则。分类的结果是没有范围限制的，它可以是一个由五元组（源 IP 地址、源端口号、协议号码、目的 IP 地址、目的端口号）确定的狭小范围，也可以是匹配某网段的所有报文。

  简单流分类可以直接根据优先级字段去检索数据，查找表项和执行流行为，操作简单不影响转发性能；复杂流分类则需要提取报文信息，构造 key 值，先查找匹配然后再根据匹配后得到的索引去获取数据执行流行为，对转发性能有一定的影响。

目的

进行流分类是为了在 Diffserv 域为用户和业务有区别地提供服务。

由于 IP 网络流量模型和业务模型的特点，使得 Internet 骨干网同时要为成千上万的业务流提供服务，因此为单个流的路径预留资源的解决思路在 Internet 骨干网上无法扩展，这严重制约了 IntServ 在实际网络中的应用。当然还存在其他一些限制 IntServ 应用的因素，包括 RSVP 信令大规模的部署、不同厂商设备之间的互通以及基于业务的管理包括认证、计费等。IntServ 从 1994 年推出至今并没有获得任何规模的商用。

DiffServ 解决方案，是一种基于类的 QoS 技术。使用 DiffServ，在网络入口处根据服务要求对业务进行分类、流量控制，同时设置报文的 ToS 域；在网络中根据 QoS 机制来区分每一类通信即依据分组的 ToS 域的值，并为之提供包括资源分配、队列调度、分组丢弃策略等 QoS 服务，这些行为统称为 PHB。DiffServ 域中的所有节点都将根据分组的 DSCP 字段来遵守 PHB。DiffServ 通过将业务定义为有限的类、可以很好地解决扩展性的问题。

Diffserve 模型的核心思想是为不同类型的流提供有差别的服务。因此在 Diffserv 方案中，首先需要根据服务要求对业务流进行分类，这是有区别地进行服务的前提和基础，这也是流分类技术的由来。

受益

- 运营商受益
  可以区分用户和业务从而提供不同的 QoS 服务。

3.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC 2597	Assured Forwarding PHB	-
RFC2598	Expedited Forwarding PHB	-
RFC 2697	A Single Rate Three Color Marker.txt	-
RFC 2698	A Two Rate Three Color Marker	-
FRC2474	Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers	-

3.3 原理描述

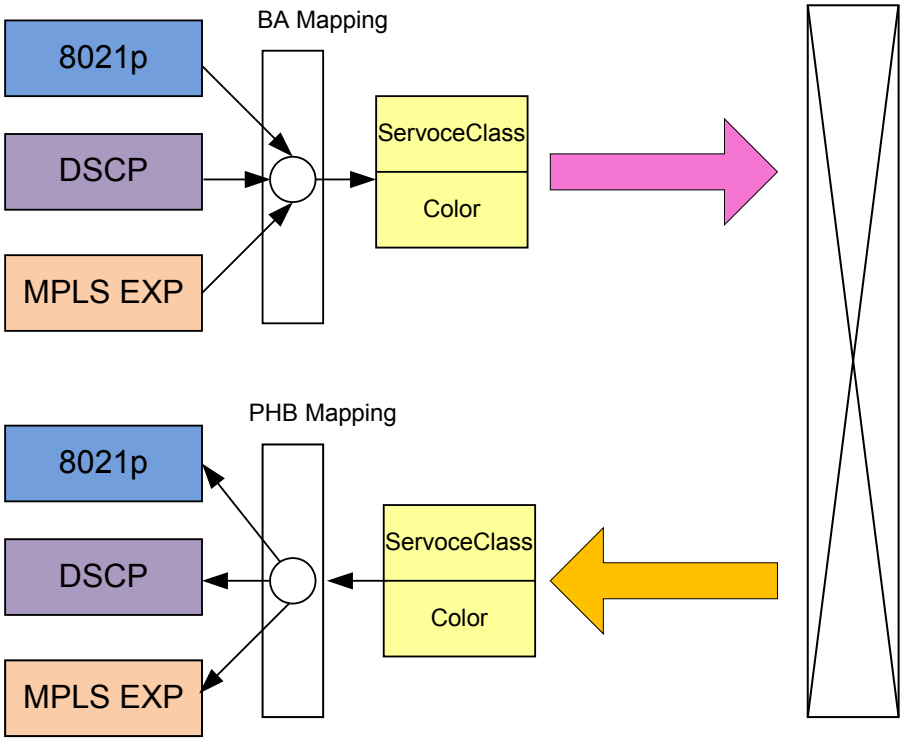
3.3.1 简单流分类的基本原理

简单流分类实现外部优先级和内部优先级之间的映射。首先根据 IP 报文的 DSCP 域值、MPLS 报文的 EXP 域值、VLAN 报文的 802.1P 值对报文进行分类，建立不同网络间报文优先级的映射关系。采用相同的服务提供策略和实现了相同 PHB（Per-Hop Behavior）集合的相连 DS（DiffServ）节点组成 DS 域。在 DS 域中定义路由器的流量策略，并在接口上通过命令与 DS 域（或 802.1p）绑定，就可以实现简单流分类功能。

简单流分类又分为上行和下行：

- 上行简单流分类，根据 IP DSCP、MPLS EXP 或 802.1P 将报文分为八种业务类型（CS7、CS6、EF、AF4 ～ AF1、BE）、三种颜色（green、yellow、red）。其中，当报文的业务类型为 EF、BE、CS6 或者 CS7 时，报文只能标记为绿色。通过上行做简单流分类从而区分不同的业务（如语音、视频、数据等）。拥塞管理、队列调度时，不同业务进入不同的队列，从而得到差异化的调度。例如语音可以进入高优先级的 PQ 队列，保证低延时。上行若不做简单流分类，报文业务类型都为 BE。
- 下行简单流分类，根据内部业务类型（CS7、CS6、EF、AF4 ～ AF1、BE）、三种颜色（green、yellow、red），重新设置 IP DSCP、MPLS EXP 或 802.1p。下行做简单流分类，实现了重标记的功能，重新标记 IP DSCP、MPLS EXP 或 802.1p。

图 3-1 简单流分类上下行映射关系图



IP 报文 DSCP 的上行映射

IP 报文的 DSCP 和内部优先级间的映射关系，根据报文的 DSCP 值指定报文在路由器内部调度的优先级和颜色，保证报文在路由器内部的合理调度。

表 3-1 DSCP 与服务等级之间缺省的映射表

DSCP	Service	Color	DSCP	Service	Color
00	BE	Green	32	AF4	Green
01	BE	Green	33	BE	Green
02	BE	Green	34	AF4	Green
03	BE	Green	35	BE	Green
04	BE	Green	36	AF4	Yellow
05	BE	Green	37	BE	Green
06	BE	Green	38	AF4	Red
07	BE	Green	39	BE	Green
08	AF1	Green	40	EF	Green
09	BE	Green	41	BE	Green

DSCP	Service	Color	DSCP	Service	Color
10	AF1	Green	42	BE	Green
11	BE	Green	43	BE	Green
12	AF1	Yellow	44	BE	Green
13	BE	Green	45	BE	Green
14	AF1	Red	46	EF	Green
15	BE	Green	47	BE	Green
16	AF2	Green	48	CS6	Green
17	BE	Green	49	BE	Green
18	AF2	Green	50	BE	Green
19	BE	Green	51	BE	Green
20	AF2	Yellow	52	BE	Green
21	BE	Green	53	BE	Green
22	AF2	Red	54	BE	Green
23	BE	Green	55	BE	Green
24	AF3	Green	56	CS7	Green
25	BE	Green	57	BE	Green
26	AF3	Green	58	BE	Green
27	BE	Green	59	BE	Green
28	AF3	Yelloe	60	BE	Green
29	BE	Green	61	BE	Green
30	AF3	Red	62	BE	Green
31	BE	Green	63	BE	Green

VLAN 报文 802.1p 的上行映射

VLAN 报文的 802.1p 和内部优先级间的映射关系，根据报文的 802.1p 值指定报文在路由器内部调度的优先级和颜色。保证报文在路由器内部的合理调度。

表 3-2 VLAN 报文中 802.1p 与服务等级之间缺省的映射表

802.1p	Service	Color	802.1p	Service	Color
0	BE	Green	4	AF4	Green

802.1p	Service	Color	802.1p	Service	Color
1	AF1	Green	5	EF	Green
2	AF2	Green	6	CS6	Green
3	AF3	Green	7	CS7	Green

MPLS 报文 exp 的上行映射

MPLS 报文的 exp 和内部优先级间的映射关系，根据报文的 exp 值指定报文在路由器内部调度的优先级和颜色。保证报文在路由器内部的合理调度。

表 3-3 MPLS EXP 与服务等级之间缺省的映射表

Exp	Service	Color	Exp	Service	Color
0	BE	Green	4	AF4	Green
1	AF1	Green	5	EF	Green
2	AF2	Green	6	CS6	Green
3	AF3	Green	7	CS7	Green

出口报文的优先级映射

报文在入口根据报文的 DSCP、802.1p 或 EXP 获取在路由器内部调度的优先级和颜色。在经过路由器内部调度后，报文可以根据报文在路由器内部调度的优先级和颜色获取出口需要封装的报文优先级 Field 字段如 DSCP、802.1p 和 EXP。

3.3.2 复杂流分类的基本原理

复杂流分类是根据一定的报文特征对报文进行分类，对于属于不同分类的报文，进行预定义的转发动作处理。

配置流分类

流分类是定义的对报文分类条件的集合，通过指定报文的一些特征字段来区分出一类报文。

在一个流分类中可以定义多个匹配规则，这些规则之前默认关系为 or，即报文只要匹配这些规则中的一个，就实施相应的动作；规则之间的关系可以通过参数 operator 设置。

配置流行为

进行流分类是为了有区别地提供服务，它必须与某种流量控制或资源分配动作关联起来才有意义。对于复杂流分类的动作包括以下几类（这些动作可以组合使用）：

- 禁止或允许

禁止或允许是最简单的流控动作。通过对报文的通过或丢弃处理，来达到控制网络流量的目的。

- 流量监管

  流量监管也就是我们通常所说的 CAR，是流分类之后的动作之一。通过 CAR，运营商可以限制从网络边缘入的各类业务的最大流量，控制网络整体资源的使用，从而保证网络整体的 QoS。运营商合用之间都签有服务水平协议（SLA），其中包含每种业务流的承诺速率、峰值速率、承诺突发流量、峰值突发流量等流量参数，对超出 SLA 约定的流量报文可指定给予通过（pass）、直接丢弃（drop）或者重新指定报文优先级处理。

- 重标记

  所谓标记就是根据 SLA 以及流分类的结果对业务流打上类别标记。目前 RFC 定义了六类标准业务即：EF、AF1-AF4、BE，并且通过定义各类业务的 PHB（Per-hop Behavior）明确了这六类业务的服务实现要求，即设备处理各类业务的具体实现要求。从业务的外在表现看，基本上可认为 EF 流要求低时延、低抖动、低丢包率，对应于实际应用中的 Video、语音、会议电视等实时业务；AF 流要求较低的延迟、低丢包率、高可靠性，对应于数据可靠性要求高的业务如电子商务、企业 VPN 等；对 BE 流则不保证最低信息速率和时延，对应于传统 Internet 业务。产品可以重标记报文的 DSCP、IP-Precedence、802.1p 和 mpls-exp。还可以直接指定报文的服务类型（EF、AF1-AF4、BE）。

- 重定向

  重定向是指将不按报文原始的目的地址进行路由转发，而是将报文重定向到指定的下一跳地址或 LSP（Label Distribution Path）下一跳标签，实现策略路由。重定向动作目前只能对三层转发报文才能生效。

  产品实现了多种类型的重定向动作：

  - IPV4/IPV6 强重定向

    用户指定下一跳 IP 地址和出接口，报文在转发过程中不查 FIB 表，直接把报文送到用户指定的出接口，在出接口取用户配置的 IP 地址和出接口封装 ARP 信息后发送报文。当出接口状态为 Down 时，报文就会被丢弃，不会重新按报文原始的目的地址进行路由转发。

  - IPV4 弱重定向

    用户指定下一跳 IP 地址不指定出接口，报文在转发过程中取用户配置的下一跳 IP 地址查 FIB 表转发。如果能转发，则按照这条路径转发报文。如果用户指定的路径转发不通，则重新按报文原始的目的地址进行路由转发。

  - IPV4 多下一跳强重定向

    用户指定多个下一跳和出接口。最多支持 4 个下一跳和 4 个出接口，流量转发时按照下一跳的配置顺序选取一个能走通的路径转发报文。在当前的下一跳路径失效后，流量自动选择一个能通的下一跳路径转发。如果所有的端口都失效，报文丢弃。

  - 重定向到 IPv4 L3VPN

    固定从某一个 VPN 主机来的流量可以通过策略访问其他的 VPN，在策略中直接限定报文的 32 位源 IP 地址、指定五元组或指定源网段，将从此源地址来的报文重定向到一个或多个 VPN。

  - 重定向到 SI

    在 MACinMAC 隧道中，支持报文重定向到一个点到点的服务实例中。

  - 重定向到公网的 LSP 隧道

    在 L3VPN 网络中，支持报文重定向到一个公网的 LSP 隧道中。

- 安全

  安全动作是指对报文实施安全检查 URPF（Unicast Reverse Path Forwarding）、镜像或者对报文进行流量采样等措施。安全动作本身不是 QoS 措施，但可以和其他 QoS 动作组合使用，以提高网络和报文的安全性。

- 队列调度

  产品支持上行基于流的 HQOS。在上行通过 ACL 来区分出用户，然后给每一个用户分配 SQ。通过这样方式能更加灵活的区分用户，使用户参与层次化的队列调度保证用户的 SLA。

配置流策略

流量策略是将流分类和 QoS 动作关联后形成的完整的 QoS 策略，流量策略可以应用到接口、全局或者用户的业务策略中，从而将流量策略中定义的流分类和动作应用到这些地方。

流策略支持共享和非共享属性的配置。共享是指在同一块接口板上不同接口使用相同的流策略共享同一套流分类和流行为表项。非共享是指同一块接口板上不同接口使用相同的流策略会根据接口和 VLAN 衍生出多套流分类和流行为表项。

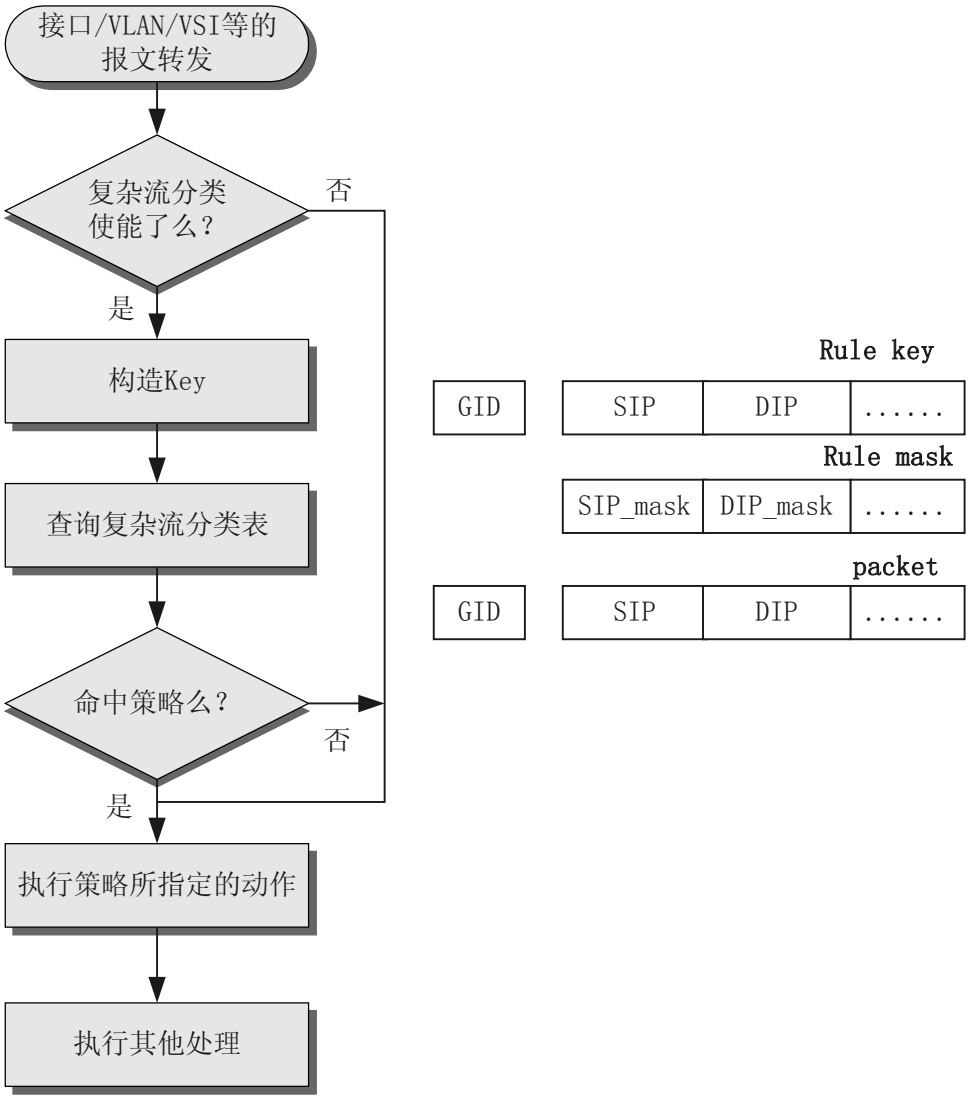
当在同单板的不同接口使用了相同的 policy 时，如果 policy 的属性是共享属性则两个接口使用同一套规则和行为。如果行为中配置了流量限速 car，两个接口的流量将共同限速。

如果 policy 的属性是非共享属性则两个接口使用两套规则和行为，规则内容完全一样但是流行为的内容不一样。如果行为中配置了流量限速 car，两个接口的流量将独立限速。

产品支持流策略规则的动态修改，但是不允许流策略共享和非共享属性的动态修改。在接口应用了某流策略后，可以动态添加、删除和修改该流策略的规则和行为，但是不能修改该流策略的共享属性。必须在接口去使能流策略应用后才能修改策略的共享属性。

复杂流分类的处理流程

图 3-2 复杂流分类的处理流程



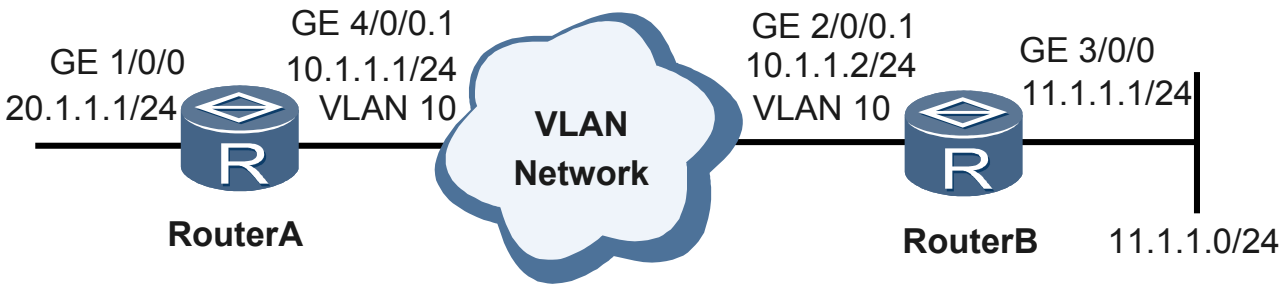
如图所示，这是报文执行复杂流分类的基本处理流程。当报文中的源 IP 地址和目的 IP 地址和规则中的掩码进行与操作，得到的值和规则中构造的值相同时则命中策略，然后执行其策略指定的动作。

3.4 应用

简单流分类映射实例

- VLAN 报文的优先级映射

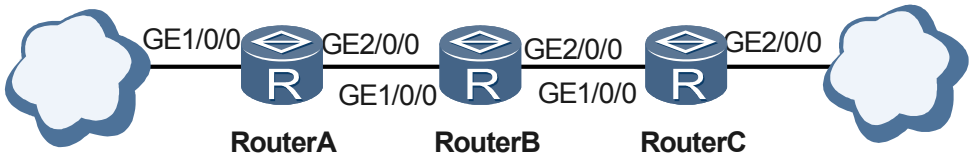
图 3-3 VLAN 报文的优先级映射



如图所示，RouterA 和 RouterB 通过 VLAN 网络互联，当有 IP 数据流从 RouterA 进入 VLAN 网络时，根据缺省的 DS 域映射关系，IP 报文中的优先级直接映射为 VLAN 帧的优先级，当从 VLAN 网络进入到 RouterB 时，通过在 RouterB 上配置 DS 域的优先级映射，设置从 RouterB 转发出去的 IP 数据包的 IP 优先级。

- MPLS 网络中简单流分类的应用

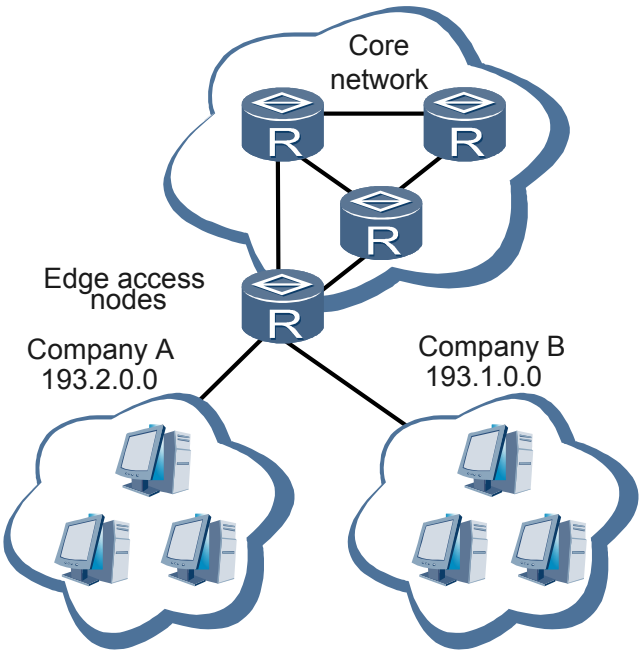
图 3-4 MPLS 网络中简单流分类的应用



如图所示，三台路由器之间建立 MPLS 邻居，从 RouterA 上进入的 IP 流量，在 RouterA 至 RouterC 之间走 MPLS 转发。从 RouterC 流出后，又恢复成 IP 流。在 RouterA 的入接口 GE1/0/0 配置 IP DSCP 到 MPLS EXP 的映射，在 RouterC 的入接口 GE1/0/0 配置 MPLS EXP 到 IP DSCP 的映射，并使能简单流分类。要求流量在 RouterA 上，能任意修改成为 MPLS 流量之后的优先级，在 RouterC 上能任意修改恢复 IP 流之后的优先级。

复杂流分类的使用实例

图 3-5 复杂流分类的应用



如图所示，假设 A 公司购买的带宽为 200M，B 公司购买的带宽为 400 M。为了实现带宽保证，可以在边缘接入节点上配置复杂流分类，根据 IP 地址区分 A，B 公司，然后执行不同的流量监管策略。

3.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
Diff-Serv	Differentiated Service	差别服务
DSCP	Differentiated Services CodePoint	有差别服务编码点
QoS	Quality of Service	服务质量

4 MPLS HQoS

关于本章

4.1 介绍

定义

MPLS HQoS 是 QoS 特性在 VPN 网络中的具体应用，包括 VPN 网络侧基于对端 PE/VPN 实例的层次化 QoS 以及用户接入侧基于接口的 QoS。

MPLS HQoS 特性用来解决 MPLS VPN 网络中各种业务的具体 QoS 需求。

目的

随着 VPN 业务的发展和成熟，运营商和用户对在 VPN 网络中实现类似物理专线的 QoS 保证的需求日益紧迫。

运营商为一个 VPN 用户提供 L2VPN/L3VPN 接入业务，需要和用户签订服务协议，主要包括下面这些内容：

- 用户接入到 MPLS VPN 网络的总带宽。
- 用户业务在 MPLS 网络中的业务优先级。

以上两点确定了用户可以将多大的流量接入到运营商的网络。接入到运营商网络后，面临的问题是，运营商对于用户接入的这些流量分别提供什么样的 QoS 保证：

- 保证到达某个指定对端 PE 用户流量的带宽要求。
- 到达某个指定 PE 对端的用户流量中有多种业务类型，比如语音业务、视频业务、关键数据业务和普通上网业务等，需要保证对每种业务的带宽和时延要求。

MPLS HQoS 提供一个完整的 MPLS L2VPN/L3VPN QoS 解决方案，借助各种 QoS 技术，来满足 VPN 用户多样化和精细化的 QoS 需求。

受益

运营商受益

- 可以给用户提供多样化的 QoS 业务，增强竞争力
- 保证关键客户的业务服务质量

用户受益

通过 MPLS VPN 网络获得同传统物理专线类似的服务质量，减少投资。

4.2 参考标准和协议

与 MPLS HQoS 特性相关的参考标准与协议如下：

- IETF RFC3270 Multi-Protocol Label Switching (MPLS) Support of Differentiated Services

4.3 原理描述

4.3.1 实现原理

基于 VPN 实例＋对端 PE 在公网侧实现层次化 QoS(L2VPN/L3VPN)

在 MPLS VPN 网络中，运营商的 PE 设备之间可能需要签订带宽约定，对两个 PE 之间的流量进行一定限制/保证，这时可以使用 VPN ＋对端 PE 的公网侧层次化 QoS 特性解决。

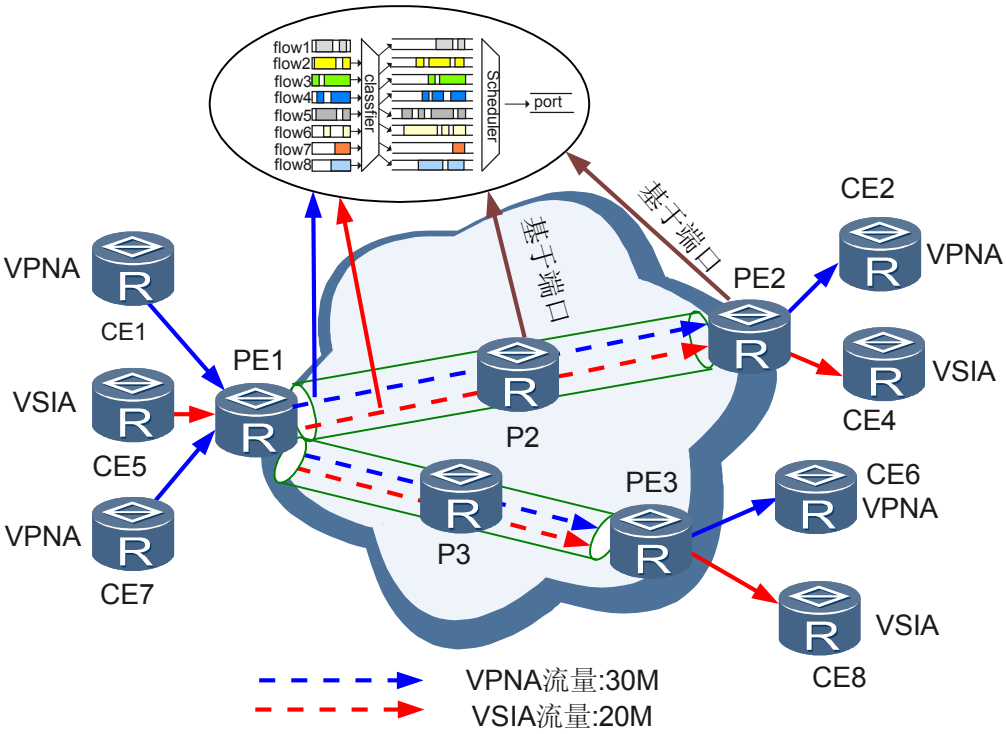
如图 4-1 所示，在 MPLS VPN 网络侧的 PE 之间指定特定的带宽和业务服务优先级约定，比如对于 VPNA，PE1 和 PE2 之间带宽约定为 30M，并支持在这 30M 带宽中优先保证高优先级的业务。

说明

由于 LDP LSP/静态 LSP 并不在其转发路径上预留资源，如果要对业务进行 PE 到 PE 整条路径的服务质量保证，需要使用有带宽保证的 TE Tunnel，因为 TE Tunnel 支持在其转发路径上预留资源。使用 LDP LSP/静态 LSP 隧道时无法保证端到端的 QoS 业务，这时 VPN 业务只支持在 Ingress 节点的 PE 设备上实现优先级调度和带宽限制。

如果只需要在网络侧实现带宽限制而不需要带宽保证，则直接指定 QoS 参数 CIR 为 0，PIR 为限定的带宽即可。

图 4-1 基于 VPN 实例＋对端 PE 在公网侧实现层次化 QoS



其流量模型描述如下：

1. 隧道为 LDP 或未配带宽的 TE

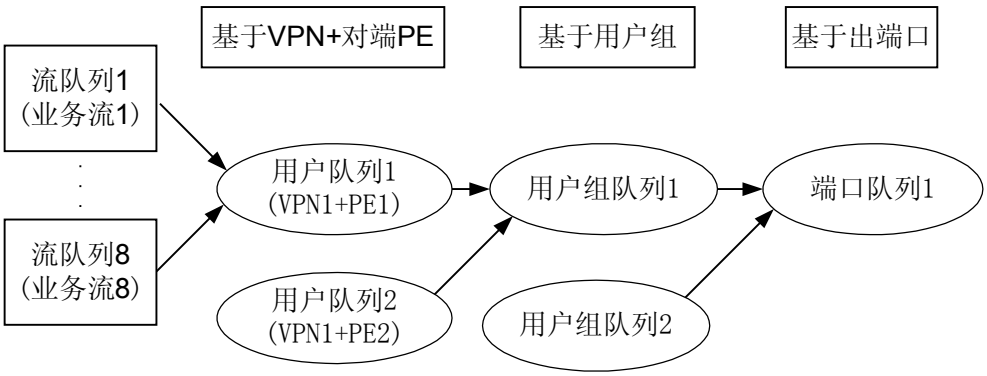
   基于 VPN 实例＋对端 PE 配置 QoS 实现网络侧带宽限制。

   如图 4-2 所示，流量在 PE 出接口根据优先级信息被映射到用户队列中的 8 个流队列，属于同一 VPN 实例＋对端 PE 的流队列被映射到同一个用户队列。

根据用户需要，可以将多个 VPN ＋对端 PE 配置为一个用户组，进行用户组队列（GQ）调度。

在出接口上，流量经过端口队列调度后被发送出去。

图 4-2 流量层次化调度模型图



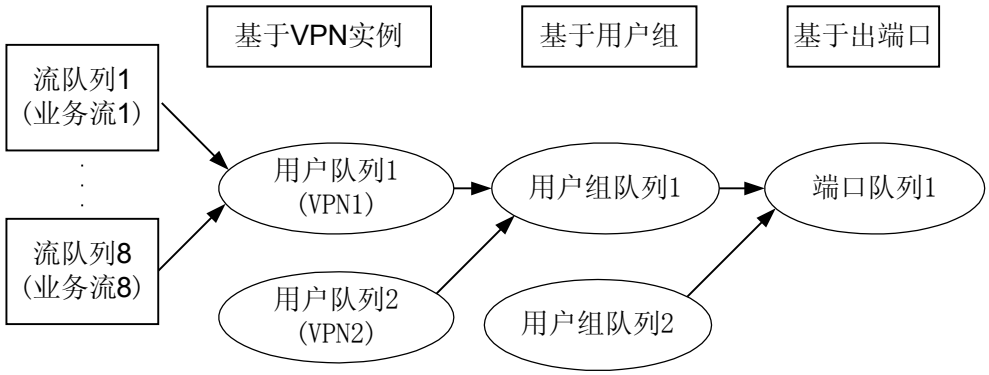
这里流队列支持优先级调度，用户队列支持带宽限制和带宽保证，用户组队列只支持 shapping 功能。

2. 隧道为配置了带宽的 TE

同隧道为 LDP 或未配带宽的 TE 比较，流量调度机制有以下不同：

- 流量经过用户队列调度后，缺省再经过基于 TE Tunnel 的用户组队列调度，缺省情况下 TE Tunnel 承载的流量为一个用户组；也可以根据用户需要，配置基于 VPN 实例＋对端 PE 进行用户组队列调度，这时不再基于 TE Tunnel 进行用户组队列调度。
- 流量支持 PE 到 PE 的带宽保证，因为 TE 隧道已预留带宽资源。

图 4-3 流量层次化调度模型图



如果对等体中的 TE 存在负载分担，则所有负载分担的总流量在端口出方向采用与图 4-3 相同的流量调度规则实现优先级调度和带宽限制。

说明

DS-TE 只能承载符合其 CT 的流量，其余流量将被丢弃。详细描述请参见 NE20E-X6 特性描述-MPLS 中的 DS-TE。

本场景推荐使用配置带宽的 TE Tunnel 作为隧道以实现 PE 到 PE 全路径的带宽保证。

基于 VPN 实例在公网侧实现层次化 QoS（L2VPN/L3VPN/MVPN）

在 MPLS VPN 网络中，当 VPN 实例作为一个用户，多数情况为企业用户，运营商希望在 PE 设备上对该用户进行流量限制或保证。

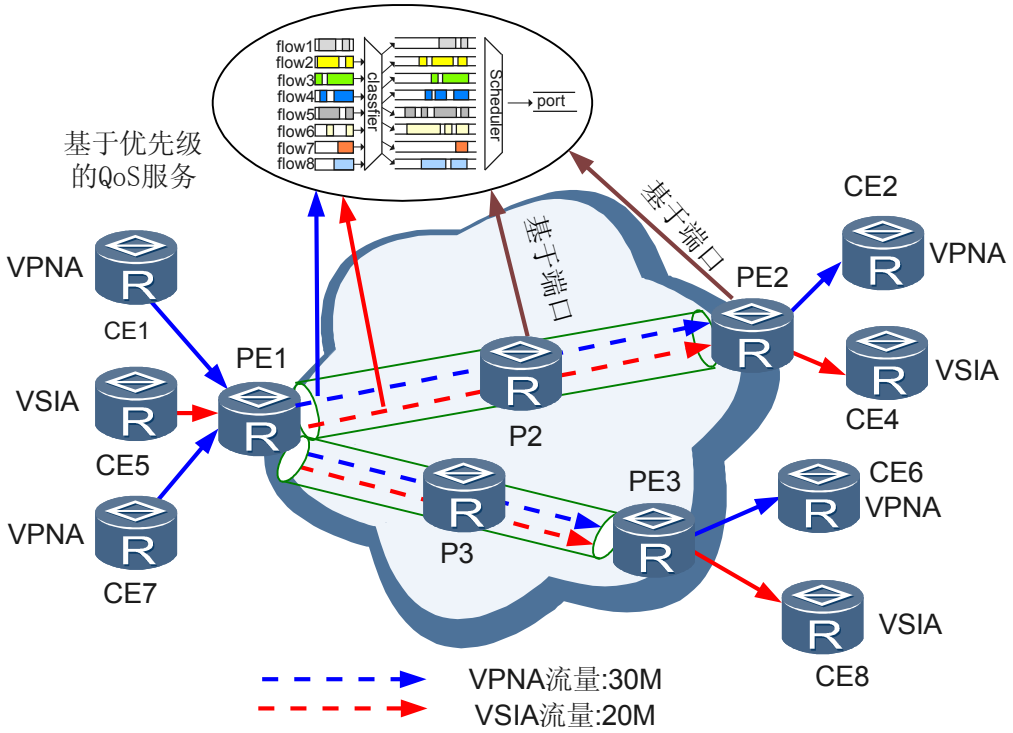
一般情况下，用户接入侧会有多个接口接入，且这些接口可能分布于不同单板上，如果在接入侧接口上进行 QoS 控制，不仅部署复杂，而且涉及不同单板上 QoS 的协同处理问题。这时可以使用 VPN ＋对端 PE 的公网侧层次化 QoS 特性解决。

如图 4-4 所示，在入口 PE 上针对 VPN 实例指定特定的带宽和业务服务优先级约定，如对于 VPNA，带宽约定为 30M，并支持在这 30M 带宽中优先保证高优先级的业务，在 PE 网络侧流量按 VPN 做队列调度，保证和限制整个 VPN 的带宽。

说明

如果只需要在网络侧实现带宽限制而不需要带宽保证，则直接指定 QoS 参数 CIR 为 0，PIR 为限定的带宽即可。

图 4-4 L2VPN/L3VPN 网络侧基于 VPN 进行 QoS 服务



其流量模型描述如下：

1. 隧道为 LDP 或未配带宽的 TE

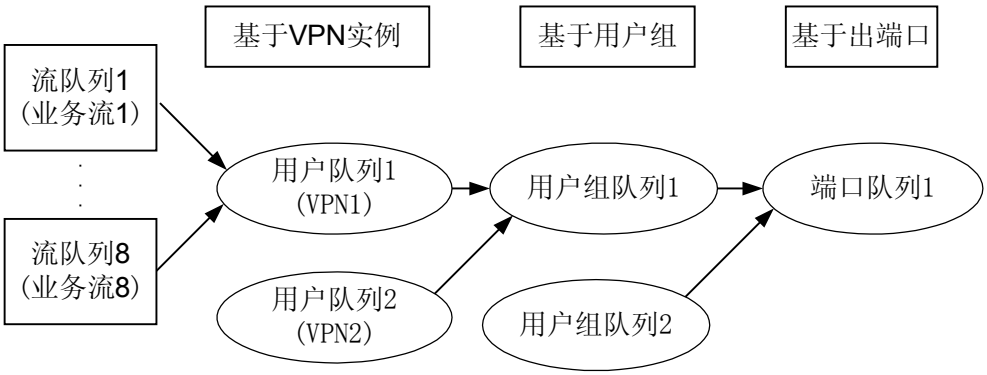
   基于 VPN 实例配置 QoS 实现网络侧带宽限制。

如图 4-5 所示，流量在入口 PE 的出接口根据优先级信息被映射到用户队列中的 8 个流队列，其中流队列支持优先级调度，属于同一 VPN 实例的流队列被映射到同一个用户队列。

根据用户需要，可以将多个 VPN 配置为一个用户组进行用户组调度。

在出接口上，流量经过端口队列调度后被发送出去。

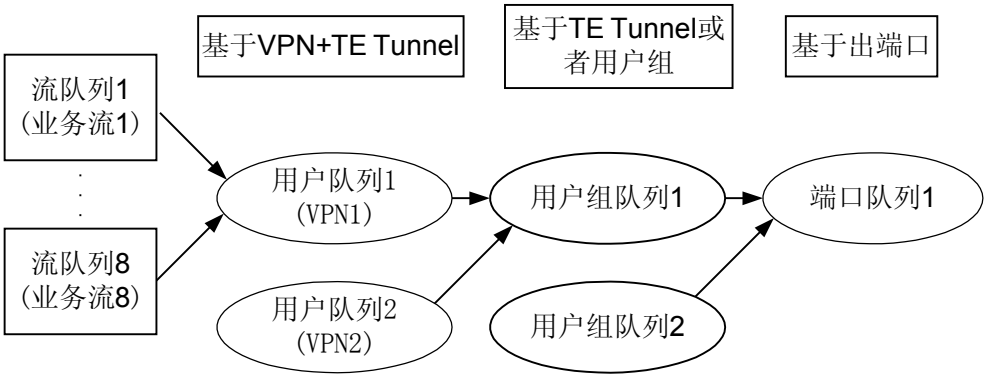
图 4-5 流量层次化调度模型图



这里流队列支持优先级调度，用户队列（SQ）支持带宽限制和带宽保证，用户组队列（GQ）只支持 shapping 功能。

2. 隧道为配置带宽的 TE

图 4-6 流量层次化调度模型图



同隧道为 LDP 或未配带宽的 TE 比较，流量调度机制有以下不同：

- 用户队列基于 VPN ＋ TE Tunnel 分配，而隧道为 LDP 或未配带宽的 TE 时为基于 VPN 分配，如果 VPN 选定了 N 条这样的 TE Tunnel 作为隧道，其总的带宽将扩大为 N 倍。
- 流量经过用户队列调度后，缺省再经过基于 TE Tunnel 的用户组队列调度，缺省情况下 TE Tunnel 承载的流量为一个用户组。也可以根据用户需要，配置多个 VPN 实例为一个组进行用户组队列调度，这时不再基于 TE Tunnel 进行用户组队列调度。
- 流量支持 PE 到 PE 的带宽保证，因为 TE 隧道已预留带宽资源。

说明

本场景不推荐使用配置带宽的 TE Tunnel 作为隧道，因为基于 VPN 的控制更多关注的是 PE 上的带宽限制和保证，不是 PE 到 PE 这种端到端的处理。

基于 VPN 实例配置 VPN 公网侧 QoS 时，如果该 VPN 实例同时应用了组播 VPN（MVPN）业务，则 MVPN 业务同单播业务共享 VPN 实例下的 QoS 配置。

基于 VPN 实例的 CE 侧接口实现层次化 QoS（L2VPN/L3VPN）

绑定 VPN 的接口也支持接口相关的 QoS 特性，可以利用基于接口的 QoS 实现 CE 发送和接收的流量的 QoS 处理。

基于 VPN 实现流量统计

在 VPLS 网络中，PE 设备支持对 AC/PW 的出入流量进行统计，AC 侧统计直接使用接口流量进行统计，PW 侧基于 PW 表进行统计。配置 MPLS HQoS 后 PE 设备还支持 PW 侧基于优先级的发送报文的流量统计。

在 L3VPN 网络中，PE 设备支持对 VPN 用户的出入流量进行统计，通过基于私网侧的接口统计来实现。配置 MPLS HQoS 后 PE 设备还支持入口 PE 公网侧基于优先级的发送报文的流量统计。

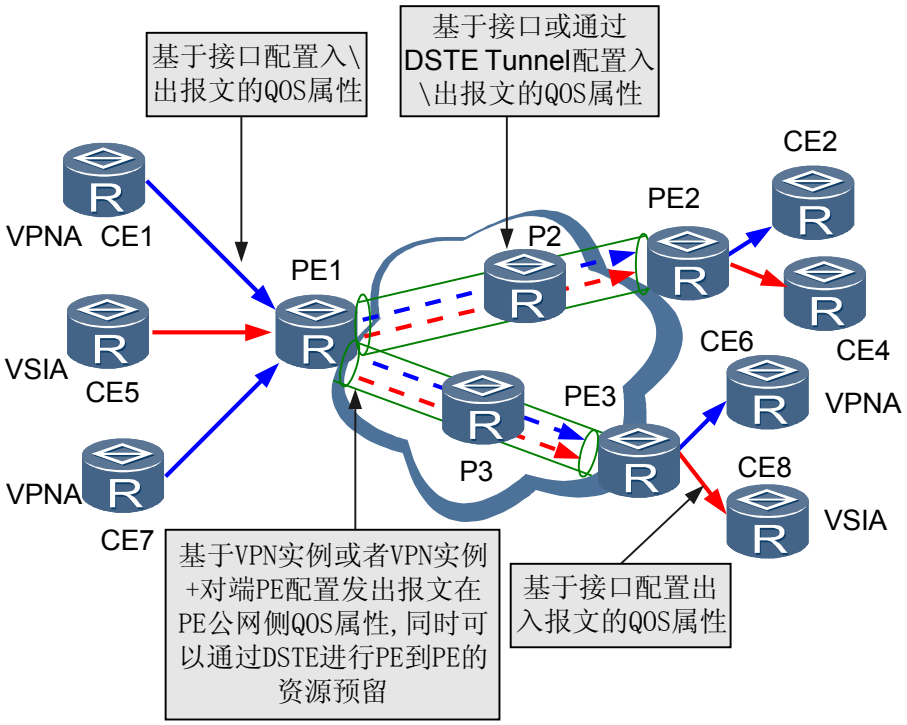
在 VLL 网络中，PE 设备支持对 VLL 用户的出入流量进行统计，通过基于私网侧的接口统计来实现。配置 MPLS HQoS 后 PE 设备还支持入口 PE 公网侧基于优先级的发送报文的流量统计。

4.4 应用

端到端的 MPLS HQoS 解决方案

如图 4-7 所示，为实现端到端的 MPLS HQoS，可以通过如下步骤实现。

图 4-7 L2VPN/L3VPN 实现端到端 QoS 服务



在 PE 的 CE 侧接口上配置基于接口的 QoS 业务，对 CE 发送和接收的报文进行 QoS 处理。

在入口 PE 即 PE1 上配置基于 VPN 实例或 VPN 实例＋对端 PE 对发送到公网侧的报文进行 QoS 处理。为实现流量端到端的保证，可以配置有带宽保证的 TE Tunnel 作为 VPN 流量的承载隧道；在 PE 上还可以通过配置 QPPB 实现 QoS 策略传递以及配置 MPLS DiffServ 模型实现 MPLS VPN 业务私网和公网转发时不同 QoS 模型的互通。

在 P 设备上配置基于接口/TE Tunnel 进行 QoS 业务处理，P 设备不感知 VPN 业务。

4.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
MPLS	Multi-Protocol Label Switching	多协议标签交换
VPN	Virtual Private Network	虚拟专用网
L2VPN	Layer 2 Virtual Private Network	二层 VPN
VLL	Virtual lease line	虚拟租赁线路

缩略语	英文全称	中文全称
VPLS	Virtual Private LAN Service	虚拟专用局域网业务
PW	Pseudo-Wire	虚连接
PE	Provider Edge	供应商边缘设备
CE	Customer Edge	客户边缘设备
QOS	Quality Of Service	服务质量
DS-TE	DiffServ-Traffic Engineering	区分业务流量工程
CIR	Committed Information Rate	承诺信息速率
PIR	Peak information rate	峰值信息速率
FRR	Fast ReRoute	快速重路由

5 HQoS

关于本章

5.1 介绍

定义

HQoS 即层次化 QoS，是 Hierarchical Quality of Service 的简称，是一种通过队列调度机制，解决 Diffserv 模型下多用户多业务带宽保证的技术。

Diffserv 即差分服务，是 Differentiated Service 的简称，是一种基于类的 QoS 技术。Diffserv 的核心思想是为不同类型的流提供有差别的服务，因此在 Diffserv 方案中，首先需要根据服务需求对业务流进行分类。流分类是 HQoS 技术对多用户多业务进行带宽保证的前提和基础。关于 Diffserv 的介绍请参见《流分类特性描述》，这里不再赘述。

家庭用户的 HQOS，就是具有某些特性的业务属于同一个家庭，共享一个 SQ 队列。目前支持的家庭用户识别方式有：NONE、C-VLAN、P-VLAN、P+C VLAN 和 Option82，也支持同一家庭的不同业务从同一主接口的不同子接口上线。

企业专线用户即运营商把设备端口统一租用给某一企业，对该企业的所有用户统一计费、统一调度和管理。

目的

随着 IP 网络上新应用的不断出现，IP 网络的服务质量也有了新的要求。例如 VoIP 等实时业务就对报文的传输延迟提出了较高要求，报文传送延时太长将为用户所不能接受（相对而言，E-Mail 和 FTP 业务对时间延迟并不敏感）。为了支持不同服务需求的语音、视频以及数据等业务，网络首先需要能区分出不同的业务，进而为之提供相应的服务。为此，QoS 技术应运而生。但是随着现在网络设备高速发展，单端口容量变大，接入用户增多，传统的 QoS 在应用中遇到了新问题：

- 传统流量管理是基于端口带宽进行调度的，这样导致流量管理对用户不敏感，只对服务等级敏感，适合网络核心侧，但不适合业务接入侧。
- 传统流量管理很难做到同时对多个用户的多个业务进行控制。

为解决以上问题，提供更好的 QoS 解决方案，迫切需要一种既能控制用户流量又能根据用户业务的优先级进行调度的 QoS 技术。HQoS 结合 Diffserv 解决方案，对多个用户的多个业务采用五级调度的方式，在现有的硬件环境下使设备具备内部资源的控制策略，既能够为高级用户提供质量保证，又能够从整体上节约网络构造成本。

通过配置家庭用户的 HQoS 实现对家庭用户的业务进行统一调度管理。

企业专线用户的 HQoS 实现了对企业用户的业务统一调度管理。

5.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
TR-059	DSL 论坛的层次化 QoS 模型	无

5.3 原理描述

5.3.1 QoS 队列基本原理

队列是报文在转发过程中的一种存储形式，当流量的速率超过接口带宽或超过为该流量设置的带宽时，报文就以队列的形式存储在缓存中。报文离开队列的时间、顺序，以及各个队列之间报文离开的相互关系则由调度策略决定。QoS 队列结构分为上行队列结构和下行队列结构，包括流队列、普通队列和端口队列，如图 5-1 和图 5-2。

图 5-1 QoS 上行队列调度体系结构

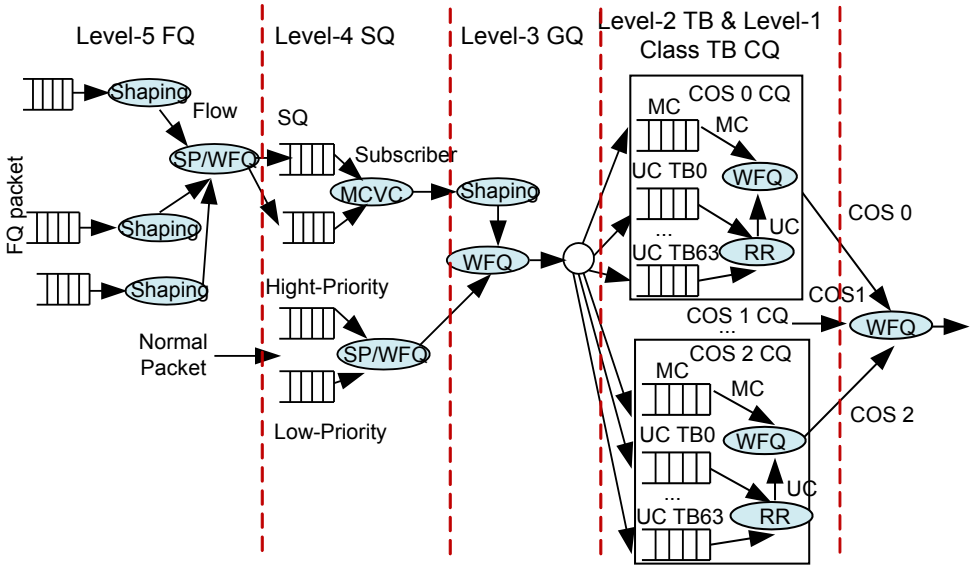
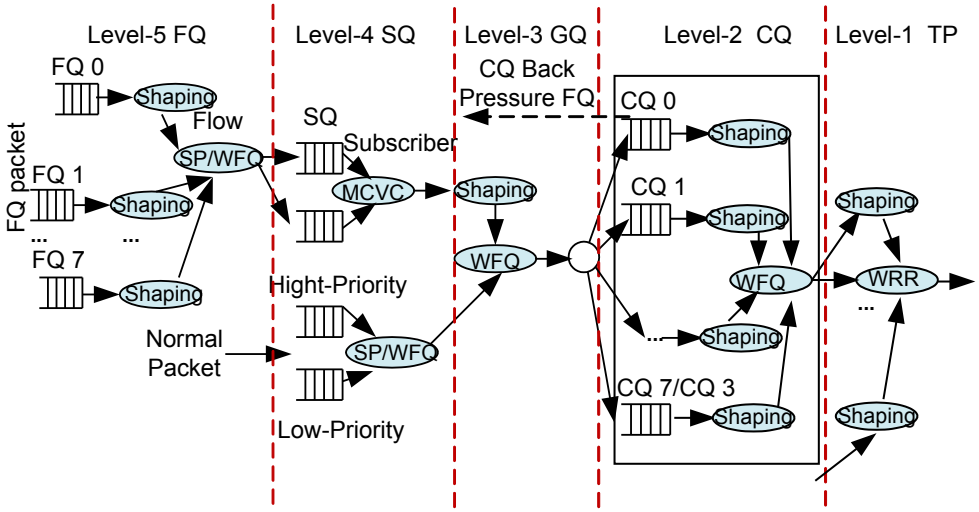


图 5-2 QoS 下行队列调度体系结构



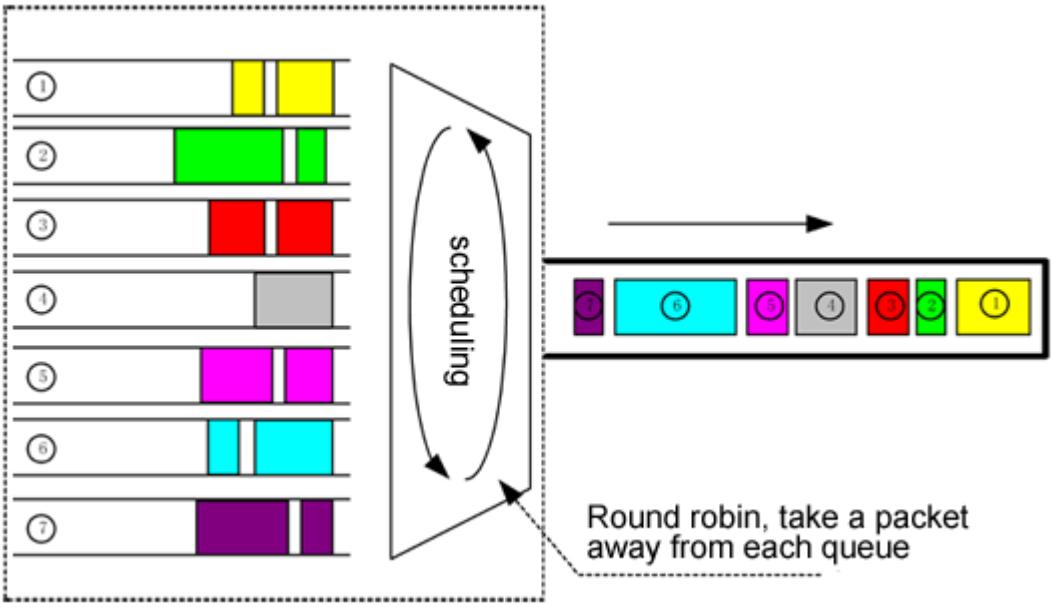
5.3.2 队列调度技术

队列调度机制是 QoS 中非常重要的一个技术，当流量发生拥塞时，通过适当的队列调度机制，可以优先保证某种类型的报文的 QoS 参数，例如带宽、时延、抖动等。队列调度机制都是在流量发生拥塞情况下产生作用。我们常用的队列调度技术包括：WFQ、WRR、PQ 等。

RR 调度技术

RR 是 Round Robin 的缩写，是一种简单的调度方法，采用轮循的方式，对多个队列进行调度。

图 5-3 RR 队列调度示意图



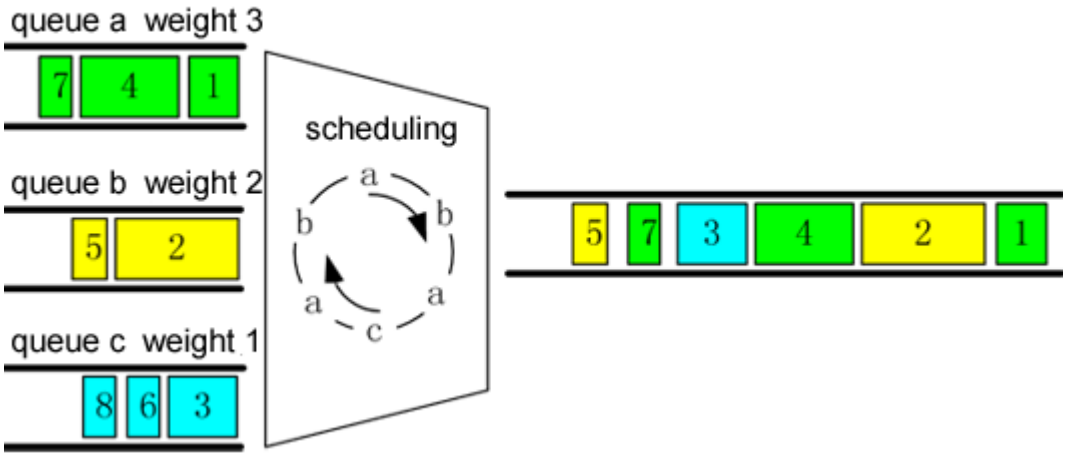
RR 以环形的方式轮循多个队列。如果轮循的队列不为空，则从该队列取走一定长度的报文；如果该队列为空，则直接跳过该队列，调度器不等待。

WRR 调度技术

WRR 主要解决 RR 不能设置权重的不足，在轮循的时候，每个队列享受的服务时间和该队列的权重成比例。

WRR 的实现方法是根据权重的不同，为每个队列分配不同的服务时间。每次轮循到一个队列时，该队列按其分配的服务时间输出报文。如图 5-4 所示。

图 5-4 WRR 队列调度示意图



WRR 对于空的队列直接跳过，循环调度的周期变短，因此当某个队列流量小的时候，剩余带宽能够被其他队列按照比例占用。

WFQ+SP 调度技术

SP 是绝对优先级调度。

WFQ 是加权公平调度算法，WFQ 算法根据参与调度的队列权重按比例分配带宽。当权值配置为 0 时表示拥有无限带宽，也就是配置为 SP，所以称为 WFQ+SP。WFQ 算法中如果有未被使用的带宽会被再分配，WFQ 可以根据配置的权重保证队列的最小带宽。

各调度技术的优缺点对比

调度算法	复杂度	时延/抖动	公平性
RR	实现简单、复杂度低	当调度速率低的时候，时延和抖动的问题比较突出	依赖报文长度
WRR	实现简单、复杂度低	当调度速率低的时候，时延和抖动的问题比较突出	依赖报文长度

调度算法	复杂度	时延/抖动	公平性
WFQ	复杂度高	时延控制的好，抖动小。	按照字节粒度进行调度，调度公平

5.3.3 QoS 队列调度

HQoS 通过 QoS 的五级队列调度来实现多用户多业务的带宽保证。如 **5.3.1 QoS 队列基本原理**所示，HQoS 区分上、下行五级队列调度。

- 上行 QoS 的五级队列调度顺序：流队列（FQ）->用户队列（SQ）->用户组队列（GQ）->目的板（TB）->类队列（CQ）。
- 下行 QoS 的五级队列调度顺序：流队列（FQ）->用户队列（SQ）->用户组队列（GQ）->类队列（CQ）->端口队列（PQ）。

QoS 队列调度包括流队列调度、普通队列调度、端口队列调度；其中端口队列调度还分为端口队列入队前调度、上行端口队列调度、下行端口队列调度。

流队列调度

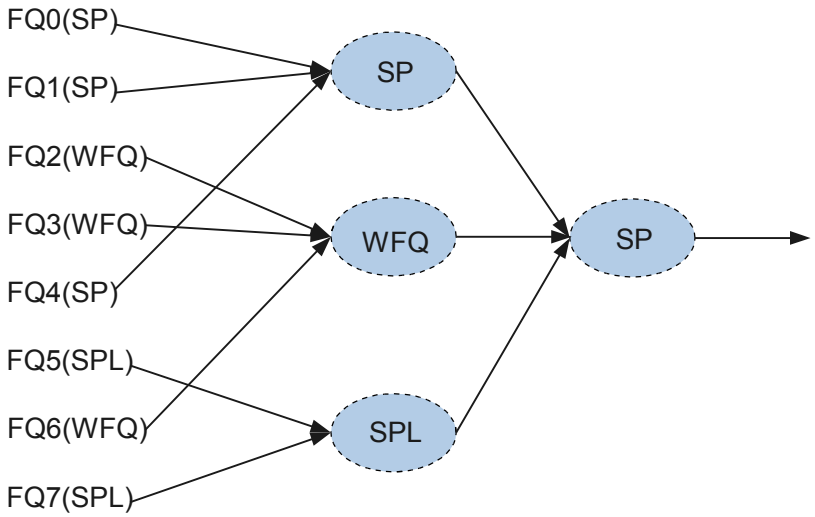
流队列调度包括 GQ 流量整形、SQ 调度、FQ 调度和 FQ 流量整形。

GQ 不支持保证带宽的调度算法，仅采用流量整形实现带宽限制，即配置流量整形器速率。流量整形采用单令牌桶形式，根据流量整形器速率定时填充令牌。当一个 SQ 队列完成调度后，如果 SQ 不属于 GQ，则 SQ 在 GQ 这级直接通过；否则进行 GQ 流量整形调度。

一个 SQ 对应一个用户，SQ 调度用于保证用户的 CIR 和 PIR 带宽。

FQ 调度采用 SP/WFQ/SPL 队列调度技术，每个 SQ 固定对应的 8 个 FQ 共享该 SQ 的带宽，且每个 FQ 可以定义自己 PIR。FQ 队列的层次化调度结构如**图 5-5** 所示：

图 5-5 FQ 队列的层次化调度结构



每个 FQ 都提供了流量整形以实现速率限制，即 FQ 的 PIR。流量整形使用单令牌桶算法实现。

普通队列调度

普通队列通过 WFQ 调度后选择进入 CQ 的报文。

端口队列调度

为确保可用带宽的分配，上行端口队列调度算法分成以下 3 层调度结构：

- 相同 COS 的端口先进行 RR 调度。
- 相同 COS 的单播和多播进行 WFQ 调度。
- 不同 COS 之间进行 SP/WFQ 调度。

下行端口队列调度算法分成以下 3 层调度结构：

- 端口内不同 COS 之间进行 WFQ 调度。
- 端口间进行 WRR 调度。
- CQ 的整形调度和端口的整形调度。

5.3.4 队列映射的基本原理

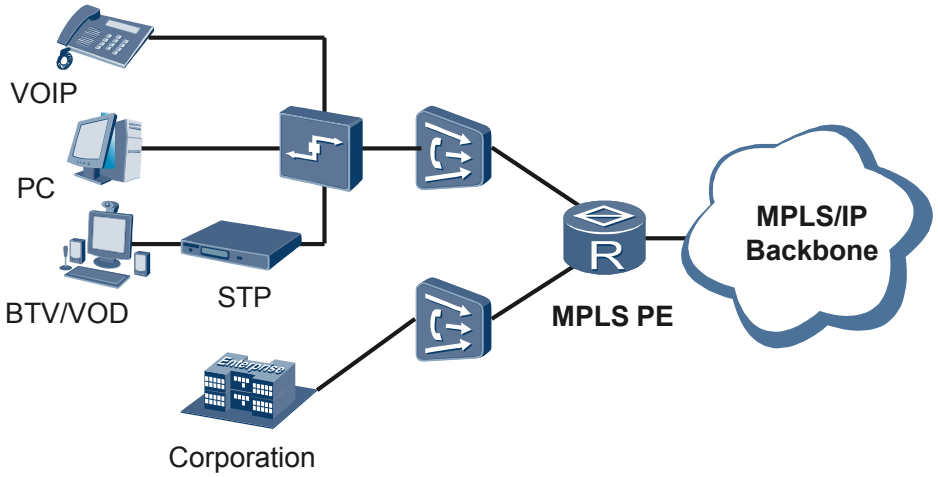
队列映射是 HQoS 所提供的五级调度队列中第一级 FQ 的 8 个优先级队列（BE、AF1、AF2、AF3、AF4、EF、CS6、CS7）在入第五级队列 CQ 下的 8 个队列（BE、AF1、AF2、AF3、AF4、EF、CS6、CS7）时的一个入队列映射功能，通过建立 FQ->CQ 队列的一个映射，可以灵活的控制 FQ 某一服务等级队列中的业务流量入 CQ 的某一服务等级队列。

5.4 应用

基于 VPN 用户的 HQoS 应用举例

多个 VPN 用户通过同一条物理链路接入 Internet，用户 VPN A 有语音业务和数据业务两种需求，语音业务在峰值时带宽为 0.5M，数据业务占用剩余带宽。用户希望在有语音时保证语音通话带宽，在没有语音业务时，数据业务能抢占所有带宽，达到 2M。其他的 VPN 用户也有类似的需求。应用组网场景如图 5-6 所示，通过在 PE 设备接入侧上部署 HQoS，保证了每个 VPN 用户的业务与带宽。

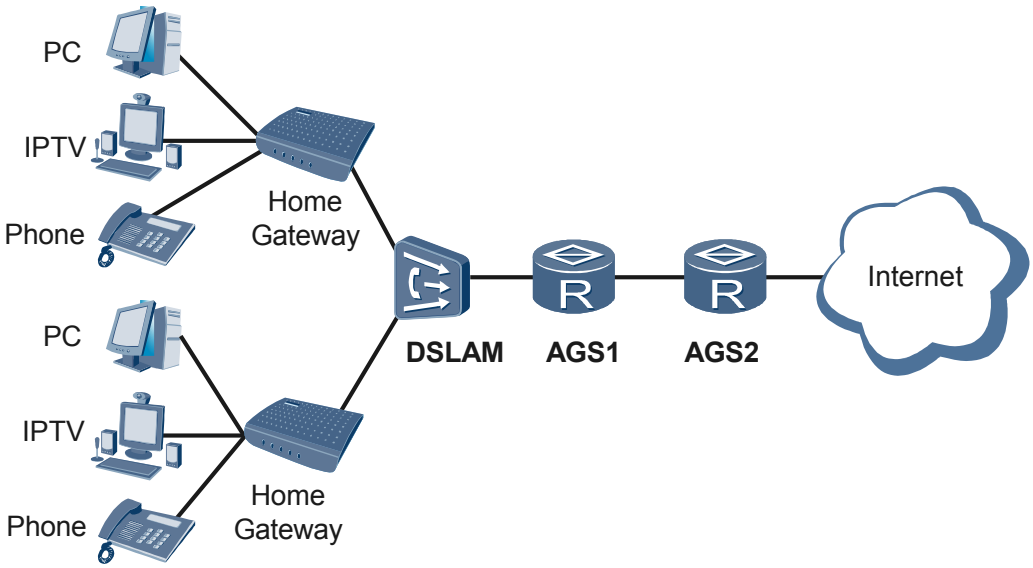
图 5-6 基于 VPN 用户的 HQoS 的组网图



家庭用户的 HQOS 典型组网

如图 5-7 所示，用户有数字电视(IPTV)、电话（VoIP）和 PC 上网业务（HSI）三种业务，通过 AGS1 接入 Internet，对用户的所有业务进行家庭识别和家庭整体调度。

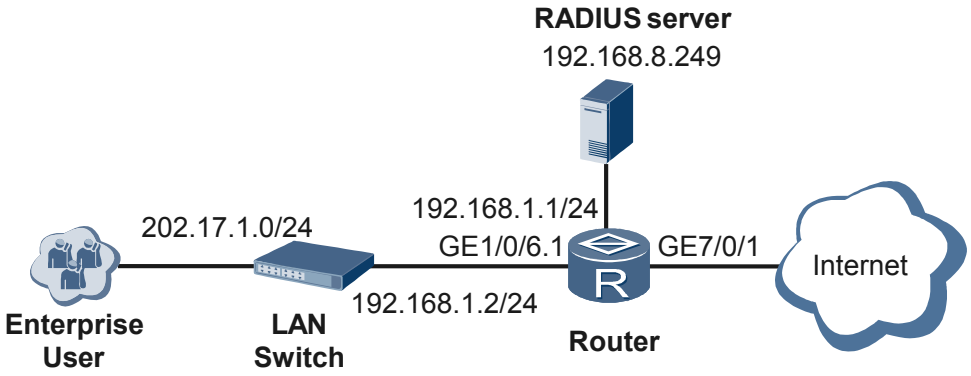
图 5-7 家庭用户的 HQOS 组网图



企业专线用户的 HQoS 应用

企业用户采用以太网三层专线方式接入组网，配置企业专线用户的 HQoS 调度保证企业总带宽及业务的带宽得到保证。

图 5-8 企业专线用户 HQoS 应用组网图



5.5 术语与缩略语

术语

术语	解释
Differentiated Service	区分服务模型，一种在网络上提供 QoS 服务的模型，通过 IP 包的优先级位（IP Precedence、DSCP），报文的源地址，目的地址等组合来对报文进行服务等级划分，对不同等级的报文进行有区别的 QoS 服务。常用来为特定的应用程序提供端到端的 QoS。
Fair Queue	公平队列，尽可能使队列公平的分享网络资源，使所有的流的延迟和延迟抖动达到最优的队列调度机制。
Priority Queue（PQ）	优先队列，根据优先级进行排队的策略。特点是如果同时存在多种优先级的报文，高优先级的报文先被分配资源。
QoS	服务质量，指对 IP 网络投递分组的服务能力的评估。通常以对延迟、延迟抖动、丢包率等服务需求提供支持的能力作为核心评估对象。为了满足这些核心需求，需要有一定的支撑技术
Weighted Fair Queue（WFQ）	加权公平排队策略，其特点是可以自动进行流分类，并且均衡各个流的延迟和延迟抖动。WFQ 与 FQ（公平队列）相比，考虑了优先级报文的利益。

缩略语

缩略语	英文全称	中文全称
QoS	Quality of Service	服务质量
HQoS	Hierarchical QoS	层次化 QoS
RR	Round Robin	轮转算法

缩略语	英文全称	中文全称
WRR	Weighted Round Robin	加权轮转算法
WFQ	Weighted Fair Queuing	加权公平队列
SP	Strict Priority	严格优先级
SPL	Strict Priority Low	严格低优先级
COS	Class of Service	服务等级

6 VLAN HQoS 特性

关于本章

6.1 介绍

定义

HQoS 即层次化 QoS，是 Hierarchical Quality of Service 的简称，是一种通过队列调度机制，解决 Diffserv 模型下多用户多业务带宽保证的技术。

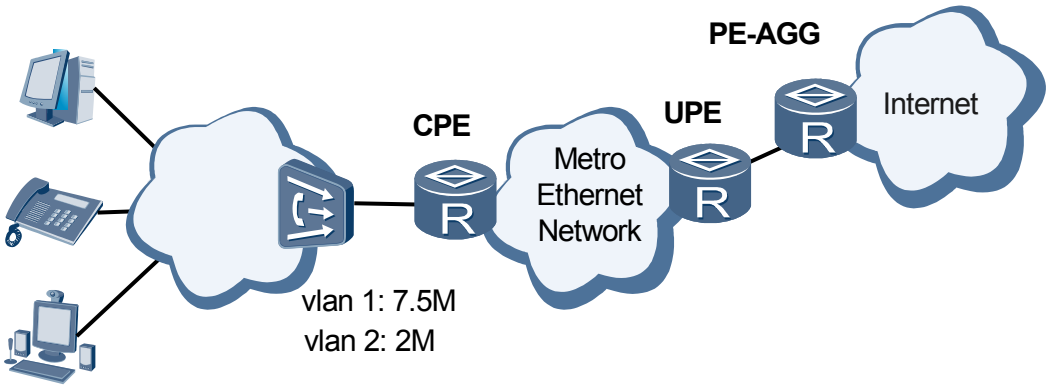
基于 IEEE 802.1Q 标准实现的 VLAN（Virtual Local Area Network）技术，对标准 Ethernet 帧格式进行了修改，在源 MAC 地址字段和协议类型字段之间加入 VLAN Tag 字段，使用 VLAN Tag 表示广播域，将一个物理的 LAN 在逻辑上划分成多个广播域（多个 VLAN），VLAN 内的主机间可以直接通信，而 VLAN 间不能直接互通，以实现广播域的隔离。除用于广播域隔离外，VLAN 技术还可以满足更复杂的网络应用，VLAN Tag 作为网络设备间的约定的一个字段，主要表示以下几种意义，如业务类型，用户 ID 和链路通道化实例 ID。当网络设备支持 QinQ(IEEE 802.1ad)技术时，使用两层 Tag 报文来表示对应业务，运营商在网络业务表示和部署上就更加灵活了，也更加复杂。

VLAN HQoS 是 NE20E-X6 HQoS 特性在 Vlan 和 QinQ 特性上的重要应用，在 VLAN/QinQ 接入的转发业务中，无论是 VLAN 透传，还是 VLAN 终结，HQoS 都可以基于 VLAN 粒度部署，以适应 VLAN/QinQ 复杂的网络的部署需求。

目的

在城域网部署中，一般使用 VLAN 区分业务或用户，用户侧流量经过汇聚设备，通过共享链路接入运营商边缘设备，多个用户属于不同的 VLAN，共享同一条链路接入 Internet，在同一个链路上，可能同时存在语音、视频和数据三种需求，三种业务对带宽和实时性要求不同，比如语音峰值带宽为 0.5M，视频带宽为 2M，数据为 7.5M。用户希望在数据流大于 7.5M 时能保证语音和视频，在没有语音和视频时数据流量可以达到 10M 带宽，同时要对多个用户进行带宽保证。如下图如示：

图 6-1 基于用户 VLAN 的 HQoS 的组网图



VLAN HQoS 有效地解决了以太链路多用户多业务的带宽保证，最大限度的保护运营商的投资，完善了现有设备硬件环境下的 QoS 特性。随着未来电话网、有线电视网和 Internet 的三网合一，城域网越来越多的用于承载语音、视频等实时和时延敏感业务，VLAN HQoS 在以太网业务中,必将作为一项重要的 QoS 特性担任越来越重要的角色。

受益

- 运营商受益
  - 增强运营商的网络业务部署的灵活性，运营商可基于 VLAN 配置不同的 HQoS，为用户提供差异化服务。
  - 在 VLAN 透传的二层转发业务中，设备可以基于透传 VLAN 作不同的带宽保障。运营商可在使用简单转发模型同时，使用灵活的用户细分的带宽管理策略。
- 用户受益

  用户各种业务的带宽保证相互独立、可靠，并享有用户间的差异化服务。

6.2 参考标准和协议

本特性的参考资料清单如下：

IEEE 802.1ad/D6.0. Virtual Bridged Local Area Networks - Amendment 4: Provider Bridges

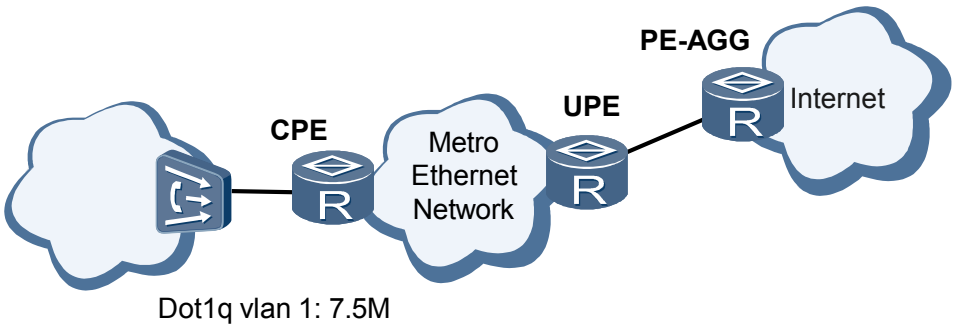
6.3 原理描述

6.3.1 VLAN HQoS 的基本原理

6.3.1 VLAN HQoS 的基本原理

NE20E-X6 在二层接口与子接口下都实现了 VLAN 和 QinQ，HQoS 的配置目前通过接口下配置 Qos-profile 指定具体 VLAN 来实现，对于二层交换端口和配置多个 VLAN 的子接口，可以同时配置多条 Qos-profile 命令，指定不同的 VLAN，每个 VLAN 都可以实现不同的 HQoS，如下图所示:

图 6-2 基于用户 VLAN 的 HQoS 的组网图



路由器对接口输入的报文的处理流程如下：识别报文的优先级->根据优先级映射 FQ->根据 VLAN 映射 SQ->进行 QoS 内部调度。VLAN HQoS 的实现分为两种，基于单层 Tag 和基于两层 Tag，目前 VLAN/QinQ 的接入的用户识别归类如下：

表 6-1 VLAN/QinQ 接入用户识别

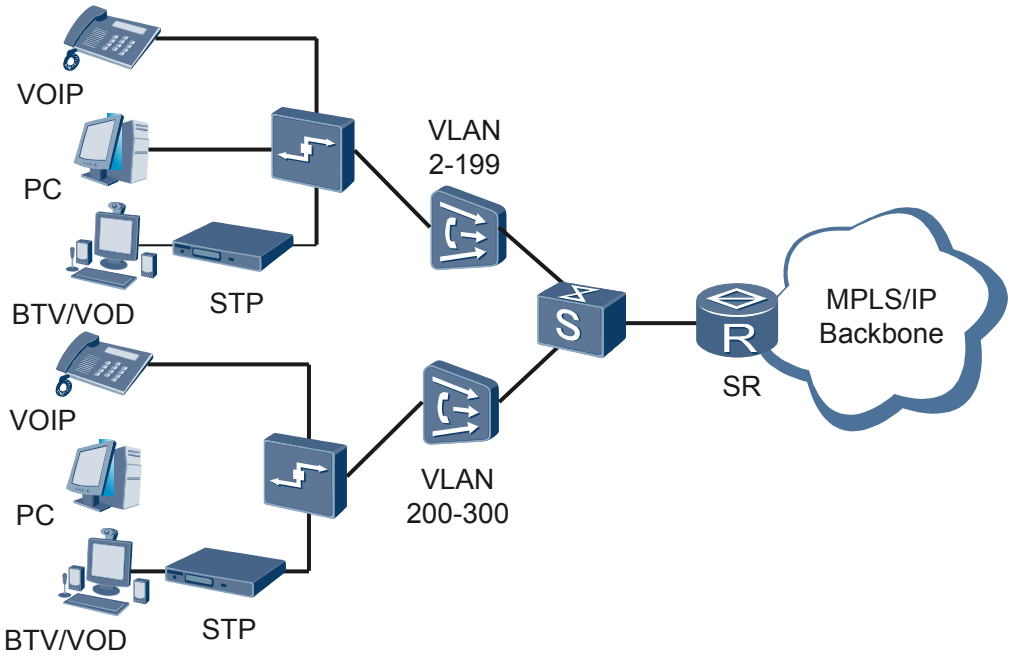
Vlan 接入方式	接入 VLL/VPLS 时的 Tag 处理
Trunk 端口 Vlan 接入	单层 Tag 识别用户
Dot1Q Tunnel 端口接入	
Stacking/Mapping 端口接入	
Stacking 子接口接入	
Dot1q 接入	
QinQ 非对称方式接入	两层 Tag 识别用户
QinQ 对称方式接入	
QinQ 内 Tag Any 方式接入	
动态 QinQ 接入	

6.4 应用

基于 VLAN 用户的 HQoS 应用举例

多个用户属于不同的 VLAN，共享同一条链路接入 Internet，每个用户有语音、视频和数据三种需求，语音峰值带宽为 0.5M，视频带宽为 2M，数据为 7.5M。用户希望在数据流大于 7.5M 时能保证语音和视频，在没有语音和视频时数据流量可以达到 10M 带宽，同时要对多个用户进行带宽保证。组网场景如图 6-3 所示，通过在 SR/BRAS 设备的用户接入侧部署 HQoS，保证每个 VLAN 用户带宽和业务。

图 6-3 基于 VLAN 用户的 HQoS 的组网图

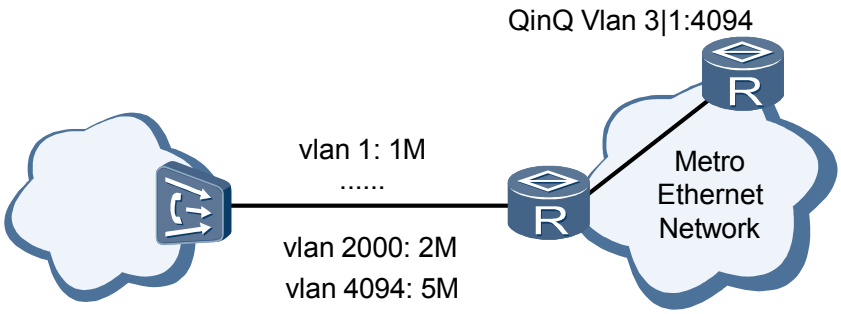


HQoS 有效地解决了多用户多业务的带宽保证，最大限度的保护运营商的投资，完善了现有设备硬件环境下的 QoS 特性。随着未来电话网、有线电视网和 Internet 的三网合一，数据设备越来越多的用于承载语音，视频等实时和时延敏感业务，HQoS 必将作为一项重要的 QoS 特性担任越来越重要的角色。

单层 Tag 用户识别的 VLAN HQoS

以 Dot1Q Tunnel 为例，此时路由器对所有收到的报文统一加外层 Tag，再进入 Vlan 的广播域转发，当二层报文从另一个配置为 Vlan Trunk 类型的二层接口出去时，输出报文时就是 QinQ 报文，组网类似如下：

图 6-4 基于二层端口 Dot1q Tunnel HQoS 典型组网图

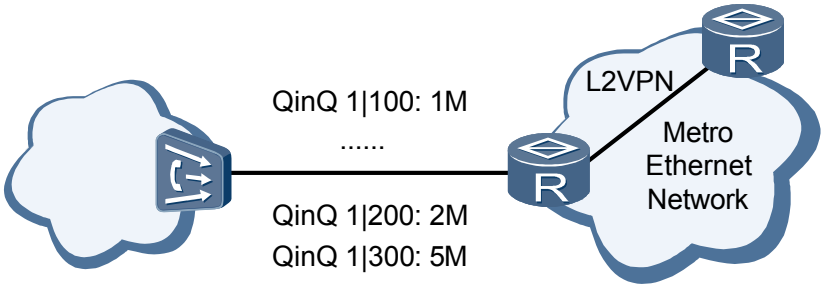


在图 6-4 中，在转发模型上，连接 CE 的 PE 端口被当成一个用户处理，但 HQoS 配置则可以基于每个 Vlan 用户进行配置。

两层 Tag 用户识别 VLAN HQoS

以 QinQ 对称方式接口接入为例，路由器在收到 QinQ 报文后，作两层 Tag 的用户识别，内层 Tag 保持不变，透传入 L2VPN 转发，下行回程从 QinQ 接口出报文时，再封装上接口的外层 Tag 信息，组网类似如下：

图 6-5 基于二层端口 QinQ HQoS 典型组网图



在图 6-5 中，在转发模型上，连接 CE 的 PE 上的 QinQ 子接口被当成一个用户处理，但 HQoS 配置则可以基于每个 Tag 对(PE-VID+CE-VID)进行配置。

6.5 术语与缩略语

缩略语

缩略语	英文全称	中文全称
VPLS	Virtual Private LAN Service	虚拟专用局域网业务
VLAN	Virtual Local Area Network	虚拟局域网
PE	Provider Edge	运营商边缘
VLL	Virtual Leased Line	虚拟租用专线
P-Tag	Provider-Tag	运营商 Tag
SVLAN	Service Vlan	业务 Vlan
CVLAN	Customer Vlan	用户 Vlan
PVLAN	Provider Vlan	运营商 Vlan
HQoS	Hierarchical Quality of Service	层次化质量服务

7 联合流量整形

关于本章

7.1 介绍

定义

联合流量整形，也称为 FGQ（Flow Group Queue） shaping，是在流队列模板中配置多个流队列的联合流量整形技术。

流队列模版中，支持对一组 FQ 进行联合流量整形，至少需要有 2 个 FQ 队列。

目的

在一个用户队列里面实现两级调度，保证几个特定业务带宽的同时，避免其它业务流量饿死，从而实现用户队列的层次化调度。

7.2 参考标准和协议

不涉及

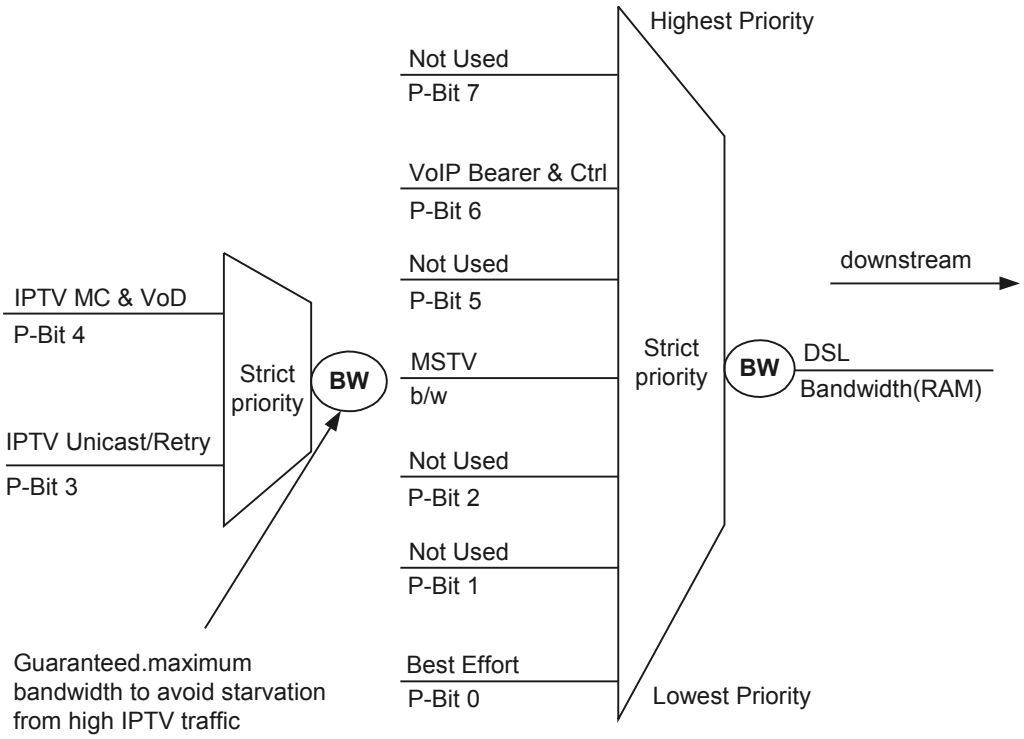
7.3 原理描述

7.3.1 联合流量整形的基本原理

联合流量整形，实现了用户级队列的层次化调度。首先对联合流量整形中的队列做整形后，再与其它的用户队列一起做调度。

例如：用户队列中有 HSI、IPTV 和 VoIP 三种业务，其中 IPTV 业务包含 IPTV UC 和 IPTV MC 两种，两种 IPTV 业务分别入不同的 FQ，用户想在保证 IPTV 业务的同时又不影响其它业务，可以采用联合流量整形功能来实现。即把 IPTV MC 和 IPTV UC 业务加入联合流量整形中，用户流量会先对 IPTV MC 和 IPTV UC 做整形，然后再参与用户的 HQOS 调度。

图 7-1 联合流量整形基本原理图



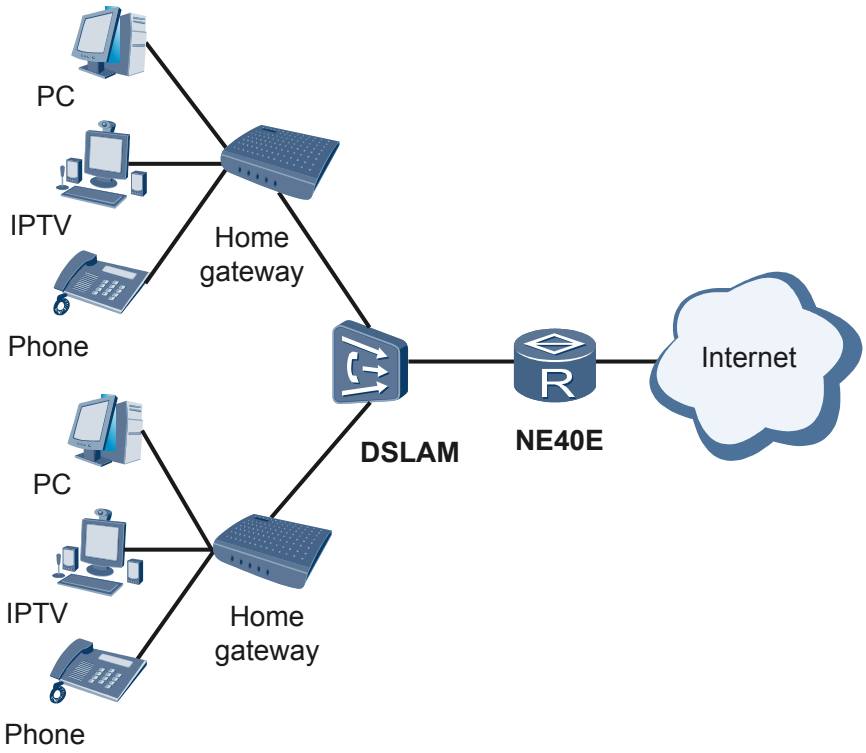
7.4 应用

7.4.1 联合流量整形的典型应用

用户有 HSI、IPTV 和 VoIP 三种业务，其中 IPTV 业务包含 IPTV UC 和 IPTV MC 两种，用户想在保证 IPTV 业务的同时又不影响其它业务，可以采用联合流量整形功能来实现。即把 IPTV MC 和 IPTV UC 业务加入联合流量整形中，用户流量会先对 IPTV MC 和 IPTV UC 做整形，然后再参与用户的 HQoS 调度。

在 NE20E-X6 上配置 HQoS 和联合流量整形，对下行的流量做联合流量整形。

图 7-2 联合流量整形典型组网



8 最后一公里 QoS

关于本章

8.1 介绍

定义

最后一公里 QoS(Last Mile QoS)，又称链路级 QoS。最后一公里是指用户和 DSLAM 之间的距离，NE20E-X6 能够基于 DSLAM 和用户之间运行的链路层协议来调整下行的流量。

目的

由于 BRAS 设备上的报文封装与 DSLAM 上报文封装不同，如果按照 BRAS 上的封装做流量整形，可能由于未包含额外的链路开销，造成 DSLAM 流量超过线路实际能力而发生拥塞。DSLAM 不支持 QoS，拥塞后不能按照 BRAS 上的调度模型丢包，导致用户的带宽无法保证。因此需要在 BRAS 设备上进行 QoS 整形时能够模拟 DSLAM 的报文封装，扣除 ATM 用户信元头或以太帧头等多余开销，使 QoS 整形更加精确。

8.2 参考标准和协议

不涉及

8.3 原理描述

8.3.1 最后一公里 QoS 的基本原理

8.3.1 最后一公里 QoS 的基本原理

由于 ME60 和 DSLAM 设备对报文的封装不同，如果按照 ME60 上的封装做流量整形，可能超出 DSLAM 的线路能力，在 DSLAM 拥塞后不能按照配置的调度模型进行丢包，导致用户的带宽无法保证。最后一公里 QoS 保证了 ME60 在进行 QoS 整形时能够扣除 ATM 用户信元头或以太帧头等多余开销，使 QoS 整形更加精确，防止由于未扣除额外的链路开销，造成 DSLAM 流量超过线路实际能力而发生拥塞。

最后一公里 QoS 调度包括以下两种模式：

- ATM 信元模式链路级 QoS 调度模式

  当 DSLAM 和用户之间使用 ATM 传输时，NE20E-X6 和 DSLAM 之间是以太封装，为了使链路层的传输速率保持一致，需要根据 ATM 和以太封装类型，调整报文的字节。

  调整的公式为：调整字节=ATM 信元封装开销－以太帧封装开销。

  说明

  调整字节有可能为负值。

  常用的调整字节数如**表 8-1**。

表 8-1 最后一公里 QoS 常用调整字节表

报文映射关系	以太封装开销	ATM 封装开销	调整字节
PPPOE->PPPOEOA(LLC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：4 Ethernet Header：14 QinQ：8 PPPOE Header：6 ATM Tail：8	12
PPPOE->PPPOA(LLC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：4 ATM Tail：8	-16
PPPOE->IPOEOA(LLC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：4 Ethernet Header：14 QinQ：8 ATM Tail：8	6
PPPOE->IPOA(LLC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：4 ATM Tail：8	-16
IPOE->IPOA(LLC)	Ethernet Header：14 QinQ：8	ATM Header：4 ATM Tail：8	-10
PPPOE->PPPOEOA(VC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：0 Ethernet Header：14 QinQ：8 ATM Tail：8 PPPOE Header：6	8
PPPOE->PPPOA(VC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：0 ATM Tail：8	-20

报文映射关系	以太封装开销	ATM 封装开销	调整字节
PPPOE->IPOEOA(VC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：0 Ethernet Header：14 QinQ：8 PPPOE Header：6 ATM Tail：8	8
PPPOE->IPOA(VC)	Ethernet Header：14 QinQ：8 PPPOE Header：6	ATM Header：0 ATM Tail：8	-20
IPOE->IPOA(VC)	Ethernet Header：14 QinQ：8	ATM Header：0 ATM Tail：8	-14

说明

- LLC 封装：允许在一条 ATM 虚电路上复用多个协议，这时需要在传统的 PDU 前加上 IEEE 802.2 逻辑链路控制（LLC）信头，以此来标识所传送的 PDU 的协议。
- VC 复用：一个高层协议由一条 ATM 虚电路来承载。

- 以太网帧链路级 QoS 调度模式

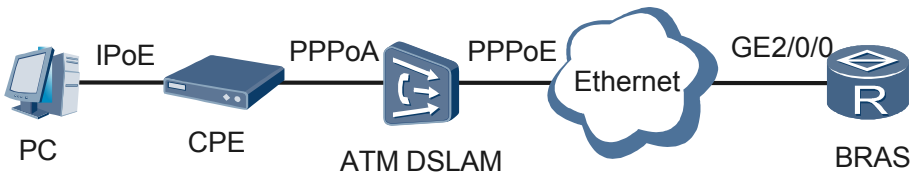
  用户和 DSLAM 之间通过以太传输，DSLAM 和 NE20E-X6 之间通过 VLAN 或 QinQ 传输。配置链路级 QoS 后，NE20E-X6 可以去掉 BRAS 和 DSLAM 间的 VLAN 头开销。

8.4 应用

8.4.1 最后一公里 QoS 的 ATM DSLAM 应用

路由器作为 BRAS 设备，用户通过以 PPPoA 方式接入网络，路由器和 DSLAM 之间链路类型为以太类型。为了使链路层的传输速率保持一致，要求在路由器上配置最后一公里 QoS，防止 DSLAM 发生拥塞。

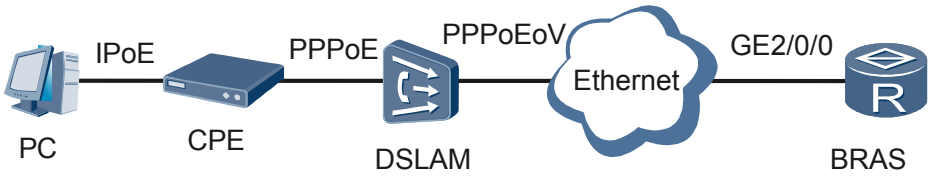
图 8-1 最后一公里 QoS 的 ATM DSLAM 典型组网



8.4.2 最后一公里 QoS 的 Ethernet DSLAM 应用

当 DSLAM 与用户之间使用以太传输时，BRAS 与 DSLAM 之间通过 VLAN 或者 QinQ 传输，配置最后一公里，可以去掉 BRAS 和 DSLAM 之间的 VLAN 头开销。

图 8-2 最后一公里 QoS 的 Ethernet DSLAM 典型组网



8.5 术语与缩略语

术语

术语	解释
ATM	异步传输模式。是一种面向连接的网络技术，使用固定大小（53 字节）的信元传送多种服务类型的数据（如文本、音频和视频数据）。信元大小固定使得对信元的处理可以通过硬件进行，从而缩短转发延时，ATM 主要设计用来充分利用高速的传输介质，如 E3、SONET、T3 等

缩略语

缩略语	英文全称	中文全称
DSLAM	Digital Subscriber Line Access Multiplexer	数字用户线接入复接器
ATM	Asynchronous Transfer Mode	异步传输模式
VC	Virtual Circuit	虚电路

9 组播用户虚拟调度

关于本章

9.1 介绍

定义

组播虚拟调度是指用户加入组播，需要按照组播 VLAN 复制或组播复制，而复制点在下游设备上，就需要对该用户的单播带宽进行相应的调整，实现用户的组播与单播带宽联动。

目的

组播虚拟调度是用户级的调度，在保证用户总带宽有限的情况下，实现用户单播带宽与组播带宽的联动，保证 BTV 业务的质量。

9.2 参考标准和协议

不涉及

9.3 原理描述

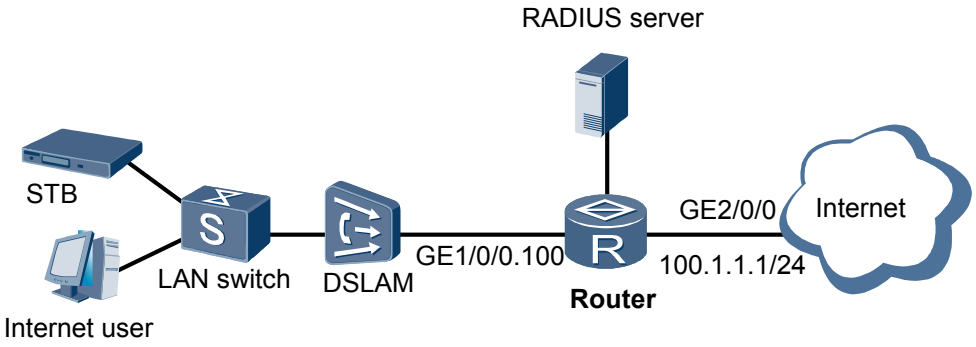
9.3.1 组播虚拟调度实现的基本原理

9.3.1 组播虚拟调度实现的基本原理

组播虚拟调度是用户级的调度。当组播复制点在下游设备上时，BRAS 设备上只复制一份流量到下游，而不是按用户复制该组播流量，这就使得组播流量在 BRAS 设备上不能进入用户队列与用户的单播流量一起调度，单播流量仍按照用户的最大带宽进行转发，这就导致在下游设备复制组播流量时，复制后的祖播流量带宽无法得到保证。为了使用户组播流量带宽得到保证，当用户加入某组播组时，就需要在 BRAS 设备上将该用户的单播带宽减掉该组播组的带宽，使得总带宽不变，从而保证组播流量宽。当用户离开该组播组时，再将该用户的单播带宽还原，从而最终实现用户单播带宽与组播带宽的联动，保证 BTV 业务的质量和用户的体验。

如图所示 DSLAM 到用户的最大带宽受距离限制，假设最大只能 3M。正常上网时使用了全部的 3M 带宽，此时用户又点播了 2M 的组播节目，单播流量加上组播流量共 5M 大于用户的 3M 带宽，在 DSLAM 到 LAN Switch 之间发生拥塞造成丢包。由于 DSLAM 不具备 QoS 能力会随机丢包，组播流量被丢弃导致组播节目质量没保障。因此在 BRAS 设备上根据组播流量带宽动态调节单播流量。DSLAM 将用户的 IGMP Report 报文通过用户 VLAN 发送到 BRAS 设备，BRAS 收到 IGMP Report 消息后知道用户点播了组播节目，则会将用户的单播流量减少到 1M，此时满足组播流量为 2M 且用户总带宽为 3M，保证了组播节目的质量。

图 9-1 组播虚拟调度原理示意图



9.4 应用

当组播流量的复制点不在 BRAS 设备上应用组播虚拟调度，其中包括如下两个典型场景。

9.4.1 单边缘组播虚拟调度典型组网

9.4.2 双边缘组播虚拟调度典型组网

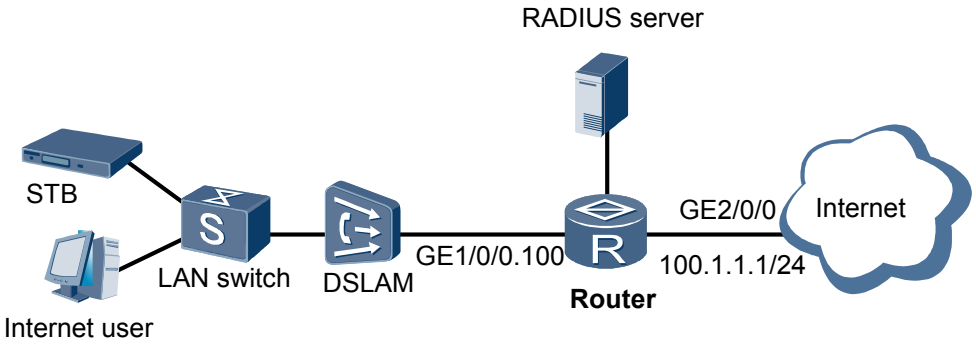
9.4.1 单边缘组播虚拟调度典型组网

在单边缘的这种组网中（一台设备同时承担用户接入和组播流量转发），BRAS 设备需要同时具有虚拟调度和组播数据复制两个功能。

如图 9-2 所示，用户终端上线时向路由器发送上线请求报文，这些报文携带了 Option 82 信息或外层 VLAN 信息。路由器根据 Option 82 信息或外层 VLAN 信息识别一个家庭的所有业务流，按照家庭对用户业务带宽进行调度。

当路由器检测到用户点播的组播节目需要进行虚拟调度时，则根据该节目带宽和用户总带宽调整用户单播业务带宽。同时，路由器通过组播 VLAN 下发组播流量，由下层设备（如 DSLAM）复制组播流量到用户。

图 9-2 单边缘组播虚拟调度典型组网



9.4.2 双边缘组播虚拟调度典型组网

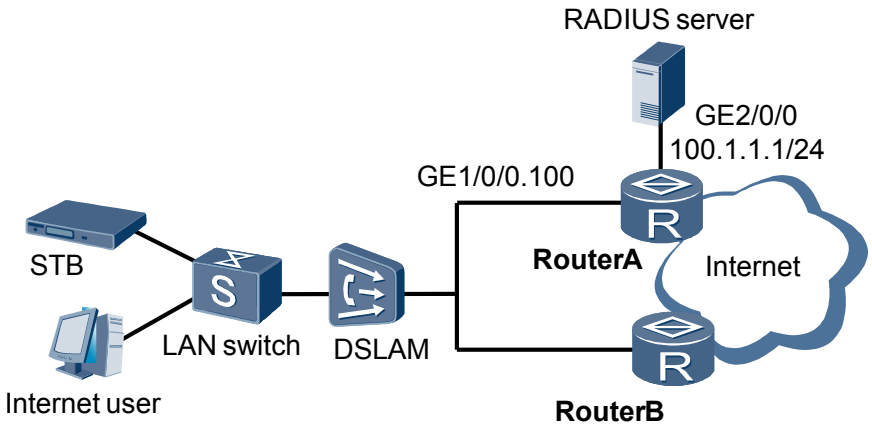
如图 9-3 所示，用户可以通过两台 BRAS 设备接入网络（称为双边缘组网），RouterA 仅完成用户的虚拟调度，而 RouterB 复制一份组播数据到下游设备。

和单边缘组网类似，RouterA 根据 Option 82 信息或外层 VLAN 信息识别一个家庭的所有业务流，按照家庭对用户业务带宽进行虚拟调度。

当 RouterA 检测到用户点播的组播节目需要进行虚拟调度时，则根据该节目带宽和用户总带宽调整用户单播业务带宽。同时，RouterB 通过组播 VLAN 下发组播流量，由下层设备复制组播流量到用户。

RouterB 在向用户转发组播数据的同时也把组播数据转发到 RouterA。RouterA 根据接收到的组播数据，测量组播流量带宽，进行组播虚拟调度。

图 9-3 双边缘组播虚拟调度典型组网



9.5 术语与缩略语

术语

术语	解释
IGMP	Internet Group Management Protocol，称为因特网组管理协议，是 IP 组播在末端网络上使用的主机对路由器的信令机制。 主机通过 IGMP 加入或者离开组播组；路由器通过 IGMP 了解下游网段是否存在组播组成员。
PIM	Protocol Independent Multicast，协议无关组播，属于组播路由协议。 网络中单播路由畅通是 PIM 转发的基础。PIM 利用现有的单播路由信息，对组播报文执行 RPF 检查，从而创建组播路由表项，构建组播分发树。

术语	解释
（S，G）	属于组播路由表项，S 表示组播源，G 表示组播组。 源地址为 S、组地址为 G 的组播报文，到达路由器后，从（S，G）表项中的下游接口转发出去。 通常，将源地址为 S，组地址为 G 的组播报文表示为（S，G）报文。
(*,G)	属于 PIM 路由表项，*表示任意源，G 表示组播组。 (*,G)表项适用于所有组地址为 G 的组播报文。不论是哪个组播源发出的，只要是发往组播组 G 的组播报文，都应该从(*,G)表项中的下游接口转发出去。

缩略语

缩略语	英文全称	中文全称
IGMP	Internet Group Management Protocol	因特网组管理协议
RP	Rendezvous Point	汇聚点
AS	Autonomous System	自治系统

10 CRTP 和 ECRTP

关于本章

10.1 介绍

定义

Compressed RTP (CRTP)是一种在报文传输前对报文头进行压缩的链路效率机制。它能够有效的降低网络负担，加速报文的传输速率并节省占用带宽。

目的

在 NGN（Next Generation Network）承载网络中，很多运营商都存在着传输资源缺乏的情况。在 IP NGN 的业务中，IP/UDP/RTP 报文头有近 40 字节，而实际每个报文中的语音数据（采用好的语音压缩算法）一般都小于 30 字节。这样，报文开销字节太大，传输效率较低。

为此，IETF 定义了一系列报文头压缩技术来解决报文传输效率问题。其中，RFC2508 定义的 CRTP 可以将 IP、UDP 和 RTP 头部由 40 字节压缩到 2 或 4 字节（不带 UDP 校验和为 2 字节，带 UDP 校验和则为 4 字节），极大地降低了报文头的冗余度，提高了链路的带宽利用率和传输效率。这对于低速链路具有显著效果。很大程度上解决了传输资源问题。

受益

运营商受益：

提高链路带宽利用率

用户受益：

无

10.2 参考标准和协议

文档编号	描述
RFC2508	Compressing IP/UDP/RTP Headers for Low-Speed Serial Links
RFC2733	An RTP Payload Format for Generic Forward Error Correction
RFC3096	Requirements for robust IP/UDP/RTP header compression
RFC3409	Lower Layer Guidelines for Robust RTP/UDP/IP Header Compression
RFC3545	Enhanced Compressed RTP (CRTP) for Links with High Delay, Packet Loss and Reordering
RFC3545	A Transport Protocol for Real-Time Applications

10.3 特性增强

无

10.4 原理描述

10.4.1 报文头压缩原理

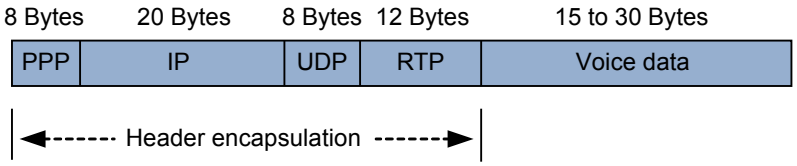
一条链路在通信过程中，报文头的变化情况如下

- IP、UDP、RTP 报文头部有约一半的字段保持不变，如源 IP 地址、目的 IP 地址、源端口号、目的端口号、RTP 的负载类型等。
- 有些字段是冗余的，如 UDP 首部长度字段可以从 IP 首部长度信息中计算得到；IP 首部长度可以从链路层帧头长度信息计算得到。另外，链路层可以实现良好的错误检测（如 PPP 的 CRC），IP 首部校验和也可以忽略。
- 有些字段虽然在传输过程中发生了变化，但报文分组之间的区别是恒定的，二次差分为 0。

CRTP 的压缩算法正是基于承载语音业务报文首部的变化规律进行报文头压缩的。

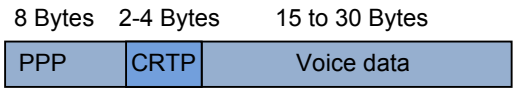
传统网络中，IP 协议通过 RTP 协议承载话音，假设 RTP 数据使用 PPP 链路传输，则帧格式如**图 10-1** 所示。

图 10-1 封装 RTP 的数据帧格式



压缩后的 CRTP 报文封装格式如**图 10-2** 所示。

图 10-2 封装 CRTP 报文的数据帧格式



CRTP 在链路层定义了除 IPv4 和 IPv6 外的四种新的报文格式：

- FULL_HEADER：传送未压缩报文首部和数据。FULL_HEADER 与 IPv4 或 IPv6 报文的区别在于它必须携带 CID 和 4 位的顺序号用来建立压缩方和解压缩方的同步。为了避免增加首部长度，CID 和顺序号被插入到 IP 和 UDP 报文头的长度字段中。

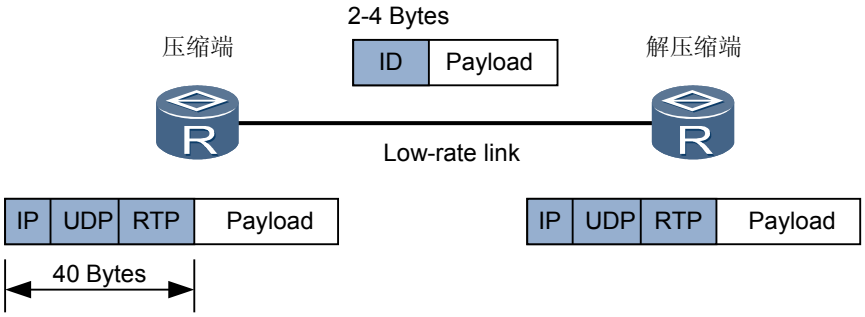
说明

为了检测数据包传输过程的丢失情况，每个数据包（包括压缩和未压缩的）都必须携带 CID 和顺序号。

- COMPRESSED_UDP：携带压缩的 IP 和 UDP 报文首部，后面是未压缩的 RTP 报文首部和数据。当 RTP 首部中应为常量的字段发生变化时使用此类型。
- COMPRESSED_RTP：携带压缩的 RTP、UDP 和 IP 报文首部。当 RTP 报文首部变化字段的二次差分值为 0 时使用此包类型。解压端在前一个报文分组的未压缩首部加上一次差分值就可以重建该包。
- CONTEXT_STATE：由解压端发送给压缩端的特殊报文，用来传输已经或者可能已经失去同步的 CID，请求更新上下文。该报文仅通过点到点链路发送，不携带 IP 报文头。

RTP 报文头压缩的处理过程如图 10-3 所示。

图 10-3 RTP 报文头压缩处理过程



1. 压缩端发送一个 FULL_HEADER 报文给解压缩端。后续压缩数据报文时，将报文首部中保持不变的字段和冗余字段剔除，并对变化字段进行差分编码。压缩后的报文携带 CID 来指示该压缩报文属于哪个上下文。

   说明

   CRTP 使用下文标识符 CID（Context Identifier）来标识 IP 源地址、IP 目的地址、UDP 源端口、UDP 目的端口以及 RTP 的 SSRC 字段组合。这些字段组合称为上下文。

2. 解压端收到压缩报文后，根据 CID 在上下文列表中进行检索，便可知道压缩报文的上下文信息。如果不考虑任何信息丢失，解压端通过将差分结果叠加到前一个报文的未压缩报文首部上便可以重建报文首部。
3. 如果链路上发生压缩报文丢弃或压缩报文被损坏，解压端会丢弃损坏的压缩报文，同时发送 CONTEXT_STATE 报文，请求压缩端以 FULL_HEADER 形式重发此报文。

10.4.2 ECRTP

增强的 RTP 报文头压缩即 ECRTP（Enhanced Compression RTP）。

对时延大、丢包率高、存在报文乱序的链路，CRTP 需要频繁的发送 FULL_HEADER 同步报文，压缩效率大大降低。RFC3545 提出了 Enhanced CRTP，用于增强 CRTP 功能，降低了链路质量对压缩效率的影响。

ECRTP 主要通过改变压缩端向解压缩端更新会话环境的方式，来增强 CRTP 对链路质量的适应性，具体的增强主要体现在如下几个方面：

- 压缩方周期性发送扩展 COMPRESSED_UDP 报文（报文格式被扩展，以携带更多的报文头变化信息），以刷新解压缩方的上下文信息，避免两端上下文不同步。
- 在不携带 UDP 校验和的情况下，增加 CRTP 头校验和字段。解压缩方可根据 CRTP 头校验和判断是否存在解压错误，从而进行“二次尝试”，减少两端不同步造成的丢包。
- 压缩方连续发送 N+1 个同步报文，避免同步报文丢失造成上下文不同步。N 的取值可以根据链路质量情况进行合理的配置。

CRTP 适用于低时延的可靠的点对点链路。ECRTP 适用于时延较长、丢包率高、存在报文乱序的质量较差的点对点链路中。MPLS 网络中推荐使用 ECRTP。

10.4.3 CRTP/ECRTP 在产品的支持情况

目前 NE20E-X6 支持对 RTP/UDP/IP 报文头进行压缩。支持两种压缩方式：CRTP 和 ECRTP。

CRTP/ECRTP 支持如下功能：

- 支持报文头压缩的接口类型包括：
  - PPP 链路上的 RTP 报文头压缩。
  - 支持 CE1、CT1、CT3、CPOS 接口通道化出来的 Serial 接口和 MP-Group 接口。
- 支持 CRTP 链路上的简单流分类。不同优先级的报文入不同的队列进行调度，并支持业务优先级的映射。
- 支持 CRTP 链路上的复杂流分类。复杂流分类在报文进行 CRTP 压缩之前、解压缩之后进行。分类后的报文，可以配置业务优先级映射，或配置流量监管（针对 CRTP 压缩之前和解压缩之后的报文长度进行）。

无论是简单流分类还是复杂流分类，都需要注意：IP 报文头的优先级字段发生变化，会造成 CRTP 无法正常压缩和解压缩，而发送 FULL_HEADER 报文，这将造成链路效率的下降。因此，在实际网络中应该保证同一链路上发送的 IP 报文优先级尽量不改变。

10.5 应用

无。

10.6 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
CRTP	Compressed RTP	压缩实时协议
RTP	Real-Time Transport Protocol	实时传输协议

缩略语	英文全称	中文全称
ECRTP	Enhanced Compressed RTP	增强压缩实时协议
UDP	User Datagram Protocol	用户数据协议
NGN	Next Generation Network	下一代网络

11 QPPB

关于本章

11.1 介绍

定义

QPPB 是通过 BGP 传播 QOS 策略（QoS Policy Propagation Through the Border Gateway Protocol）的简称。

目的

在部署大型复杂网络时，需要执行大量的复杂流分类，而且无法按照团体属性、ACL、Prefix 或 AS-Path 对报文进行分类。如果网络结构不稳定，需经常变化网络结构时，配置修改的工作量非常大甚至难以实施，可以通过部署 QPPB 减少配置修改的工作量。

QPPB 的最大优点是可由 BGP 路由发送者通过设置 BGP 属性预先对路由进行分类，BGP 路由接收者可以依据 BGP 路由发送者设置属性对 BGP 路由应用不同的本地 QoS 策略。

在复杂组网环境中，在经常需要动态修改路由分类策略的情况下，应用 QPPB 可以简化路由接收者上的策略修改，只需要修改 BGP 路由发送者上的路由策略就可以满足需求。

11.2 参考标准和协议

无

11.3 特性增强

无

11.4 原理描述

11.4.1 QPPB 原理描述

QPPB 技术主要通过 BGP 传播的路由属性设置 QoS 参数、制定 QoS 流量策略、实现 QoS 保障，可以分为对路由发送者和对路由接收者的设置。

BGP 路由发送者在向邻居发送路由时，先匹配路由策略，为发送的不同路由信息设置不同的 BGP 路由属性。这些路由属性作为路由分类的标识。路由属性可以设置为：

- 路由信息的 AS 路径列表
- 路由信息的团体属性列表
- 路由信息的路由权值
- 地址前缀列表

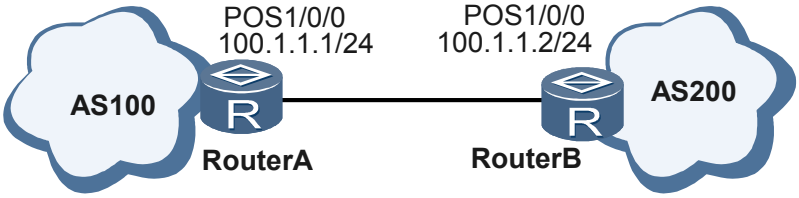
BGP 邻居在接收到路由后，执行如下动作：

- 基于 BGP 团体列表、BGP AS-Paths list 和 ACL、Prefix list 等过滤器匹配路由策略，为接收到的 BGP 路由设置 IP 优先级和 Traffic behavior name 等参数将 BGP 路由信息及相关联的 QoS 参数一起下发到 FIB 表。
- 对分类后的数据流配置 QoS 流量策略，在数据转发过程中，对发送到目的网段的数据包可以依据从 FIB 中获取的 IP 优先级和 Traffic behavior name 等参数使用不同的 QoS 策略，从而实现 QoS 保证。

QPPB 技术实际只是在路由发送方通过路由分类设置路由属性，在接收方根据目的网段的路由属性设置 QoS 策略，不是在 BGP 路由信息中发送 QoS 策略。因此 QPPB 技术但需要整网统一协调的路由及 QoS 策略，并且不同的节点之间能够相互信任。

11.4.2 QPPB 实现机制

图 11-1 QPPB 原理图



如图 11-1 所示，RouterA 和 RouterB 是 BGP 邻居，RouterA 是 BGP 路由的发送方，RouterB 是 BGP 路由的接收方。

RouterA 发送给 RouterB 的 BGP 路由设置了团体属性，RouterB 收到 RouterA 发送的 BGP 路由后，通过路由策略匹配团体属性为接收的 BGP 路由设置 QoS 参数。

具体配置思路如下：

1.在 RouterA 和 RouterB 上配置 BGP 的基本功能、配置 BGP 发布的本地网络路由、配置建立 BGP 连接所使用的接口等。

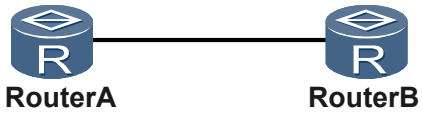
2.在路由发送方 RouterA 上通过路由策略为发送的 BGP 路由设置路由属性：首先定义路由策略的过滤器，可以是 ip-prefix、ACL 列表等。然后通过匹配过滤规则，设置 BGP 的路由属性，可以是 as-path、community、cost、preference、traffic-index 等。

3.在路由接收方 RouterB 上，通过匹配路由策略为接收到的满足某些属性的特定 BGP 路由设置 QoS 策略。目前，产品支持为 BGP 路由设置 Traffic behavior name 和 IP 优先级。

4.在路由接收方 RouterB 与 RouterA 相连的接口上，配置根据 Traffic behavior name 和 IP 优先级应用策略。

11.5 应用

图 11-2 QPPB 应用示意图



如图 11-2 所示,RouterA 向 RouterB 通告带有属性的 BGP 路由，RouterB 收到 RouterA 通告的路由后，通过匹配 BGP 团体列表、ACL、BGP AS path list，为 BGP 路由设置 IP 优先级、Trafficbehavior name。流量从 RouterA 发往 RouterB，在 RouterB 上与 RouterA 相连的接口上使能 QPPB 策略后，从 RouterA 发送到 RouterB 的数据包就会应用相应的 QoS 策略。

12 ATM QoS

关于本章

12.1 介绍

定义

ATM 网络中，用户在进行数据传输的过程中，为了保证数据传输的服务质量，需要在每个虚连接上携带 ATM QoS 的参数。通过在虚连接及其 ATM 接口上执行 ATM QoS 的策略，保证服务质量。

目的

ATM 网络本身具备丰富的 QoS 能力。随着 ATM 网络与 PSN 网络的结合，需要在此过程中保持 ATM 网络的 QoS 能力，建立 IP 优先级、MPLS 优先级、VLAN 优先级与 ATM 优先级的对应关系，使 ATM 网络中优先级较高的报文在 IP 网络中也以较高的优先级进行转发，同时使 IP 网络中优先级较高的报文在 ATM 网络中也能以较高的优先级进行转发。

ATM 网络与 PSN 网络相互结合使用主要有两种情况：

- ATM 透明传输：传统的 ATM 网络向 PSN 网络迁移，利用 MPLS 隧道充当 PW，连接两端的 ATM 网络。PW 中使用 MPLS 报文封装并透传 AAL5 数据帧或 ATM 信元。
- 1483B 封装的 IPoEoA 和 1483R 封装的 IPoA：路由器位于 ATM 网络边缘，承担 IP 网络接入功能。数据报文在 ATM 网络上传输时使用 AAL5 帧封装——IPoA、IPoEoA 等。在路由器处作 ATM 终结，将 IP 报文转发至其它类型的接口，或将以太二层报文转发至以太接口。

12.2 参考标准和协议

无

12.3 特性增强

不涉及

12.4 原理描述

12.4.1 ATM QoS 实现机制
当 ATM 网络向 PSN 网络迁移或作为 IP 网络的承载层时，为提供端到端的 QoS，需要将 IP 网络的 QoS 机制与 ATM 网络的 QoS 机制结合起来。

12.4.1 ATM QoS 实现机制

当 ATM 网络向 PSN 网络迁移或作为 IP 网络的承载层时，为提供端到端的 QoS，需要将 IP 网络的 QoS 机制与 ATM 网络的 QoS 机制结合起来。

NE20E-X6 支持通过单 PVC 和 PVC-Group 两种方案，将 IP 网络的 QoS 机制与 ATM 网络的 QoS 机制结合起来。

ATM QoS 实现机制包括以下方面：

- ATM 简单流分类
- ATM 强制流分类
- ATM 复杂流分类
- ATM 流量整形
- ATM PVC 的拥塞管理

ATM 简单流分类

ATM QoS 支持 ATM 透明传输、1483R 和 1483B 三种方式下的简单流分类。支持在 ATM 子接口、VE 接口或 PVC、PVP 使能 ATM 简单流分类和配置 ATM 简单流分类映射关系。

ATM 透明传输包括 ATM 信元透明传输和 ATM 帧格式透明传输。

- VCC、VPC 方式的透明传输是 ATM 信元透明传输。透明传输的基本单位是 ATM 信元，大小固定 53 个字节，与标准 ATM 链路上传输的基本单位是一致的。
- SDU 方式的透明传输是帧格式的透明传输。透明传输的基本单位是帧，大小由上行 PE 上接收到的报文和用户配置的 MTU 值共同决定。

1483R 协议封装 IP 报文，实现 IPoA 业务。1483B 协议封装以太报文，实现 IPoEoA 业务。

- ATM 透明传输方式简单流分类原理

  可以在 MPLS 网络入 PE 的 AC 侧，根据 ATM 网络的服务类型和 CLP 值映射到路由器的内部优先级；在 MPLS 网络入 PE 的 PW 侧，根据路由器内部优先级映射到 MPLS 网络的 EXP 值，使 ATM 网络中的 QoS 参数在 MPLS 网络中传递。（对于 SDU 透明传输方式，在 MPLS 网络入 PE 的 AC 侧，只要该 SDU 的任何一个信元的 CLP 为 1，则整个 SDU 的 CLP 值为 1，否则为 0。该 CLP 值结合 PVC 的服务类型映射到路由器内部优先级。在 MPLS 网络入 PE 的 PW 侧，与其它方式的信元透明传输一致。）

  在 MPLS 网络出 PE 的 PW 侧，根据 MPLS 外层封装的优先级（EXP 值）进行转发处理。在 MPLS 网络出 PE 的 AC 侧，使用被封装的信元原有的优先级，进行转发处理。（对于 SDU 透明传输方式，在 MPLS 网络出 PE 的 PW 侧，与其它方式的信元透明传输一致。在 MPLS 网络出 PE 的 AC 侧，如果 SDU 的 CLP 为 1，则分段为信元时设置所有的信元 CLP 为 1，否则设置为 0）。

  通过上述的简单流分类，实现 ATM 网络的 QoS 参数从一个 ATM 网络经 PSN 网络透明传输到另一个 ATM 网络。

- 1483R 和 1483B 方式简单流分类原理

  在 ATM 网络边缘，承担 IP 网络接入功能的路由器上使能简单流分类，进行 DSCP 到 ATM 业务优先级的映射。

  在接入路由器的 PVC 上行，1483R/1483B 报文的业务优先级由 IP 报文的 DSCP 值来决定，进行 QoS 处理。

  在接入路由器的 PVC 下行，根据路由器内部优先级映射到 ATM 网络中的 CLP，实现了 IP 网络中 DSCP 到 ATM 业务优先级的映射。

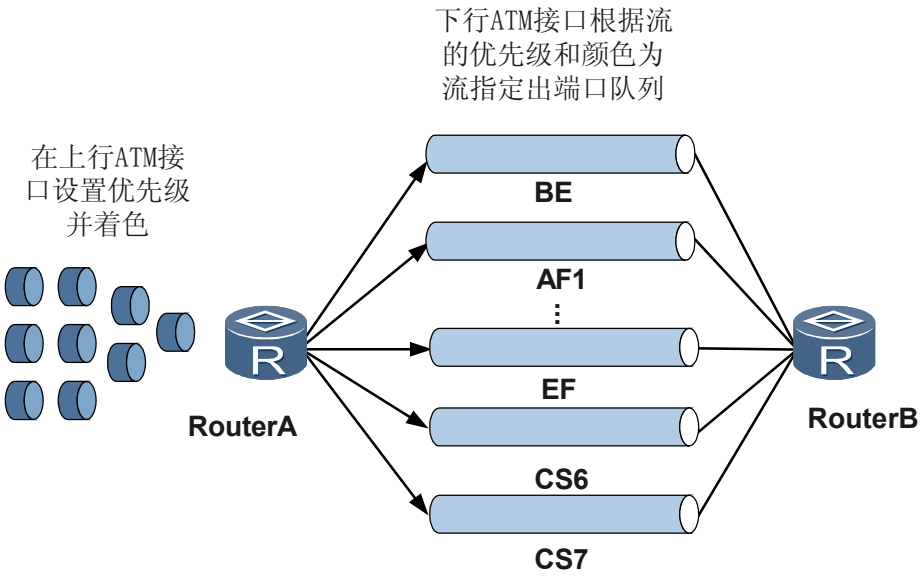
ATM 强制流分类

在 ATM 网络中，虽然 ATM 信元本身携带有优先级的信息，但是根据 ATM 信元的优先级来实现 IPoA、信元透明传输、IWF 简单流分类功能，比较复杂。可以采用强制流分

类的方案，并不关注 ATM 信元的服务类型和优先级，在 ATM 网络的边缘路由器的上行接口，通过配置强制流分类，强制为某个 PVC、某个接口（包括主接口和子接口）或者某个 PVP的流量指定 IP 报文的业务优先级和颜色，并根据指定的业务优先级和报文颜色在 ATM 网络边缘路由器的下行接口应用 QoS 策略。

如图 12-1 所示，在 RouterA 的上行 ATM 接口强制对某个流指定优先级和颜色，下行 ATM 接口会根据为这个流指定的优先级和颜色为流指定队列和调度方式，通过这种方式实现 ATM 的 QoS 功能。

图 12-1 ATM 强制流分类原理图



ATM 主接口、ATM 子接口、ATM PVC 模式、ATM PVP 模式和 VE 接口支持强制流分类。

ATM 复杂流分类

在部署 IPoA 和 IPoEoA 业务中，如果需要对进入 ATM 网络的流量或 ATM 网络中的流量进行分类管理和限制（例如对语音、视频、数据等业务分别对待，分配不同的带宽、保证不同的时延；对来自不同用户的流量分别对待，保证不同的带宽和优先级），则需要根据报文的 DSCP 值、协议类型、IP 地址、端口号等参数对不同分类的业务提供差别服务，配置 ATM 复杂流分类的 QoS 流量策略。

ATM 复杂流分类的实现，体现在 QoS 流量策略的应用中。一个 QoS 策略中关联了已经定义的流分类和流行为，将 QoS 策略应用在 ATM 接口上，就可以在该 ATM 接口上提供需要执行的服务质量保证。

具体实现步骤如下：

1. 定义流分类
2. 定义流行为
3. 定义流量策略，关联流分类和流行为
4. 在 ATM（子）接口或 Virtual-Ethernet 接口上应用流量策略

说明

由于 ATM 不支持 IPv6 协议和 IPv4 组播协议，因此 NE20E-X6 不支持 IPv6 报文和 IPv4 组播报文的 ATM 复杂流分类功能。

ATM 流量整形

在 ATM 网络流量很大的时候，超出规格的报文将被直接丢弃。如果不希望下游网络因为上游发送数据流量过大而造成拥塞或大量报文的直接丢弃，可以通过在上游路由器的出接口配置 ATM 流量整形，限制流出 ATM 网络的某一连接的流量与突发，使这类报文以比较均匀的速度向外发送，以利于网络上下游之间的带宽匹配。

ATM 流量整形的配置分为两步骤：

- 在系统视图配置 ATM 的业务类型和整形参数。可以配置业务类型为确定速率 CBR（Constant Bit Rate）、非实时可变速率 NRT-VBR（Non Real Time-Variable Bit Rate）或实时可变速率 RT-VBR（Real Time-Variable Bit Rate）。
- 在 ATM PVC 或 PVP 上指定 ATM 业务类型，应用流量整形参数。

ATM PVC 的拥塞管理

在 ATM 网络中，对超过 PVC 带宽的报文不丢弃，而是缓存起来，在网络有空闲带宽的时候再发送出去。通过配置 ATM PVC 的拥塞管理，可以将此类报文按照某种算法入 PVC 的八个队列，之后按照队列调度机制进行发送。ATM PVC 队列包括 PQ 和 WFQ 队列。

13 L2TP QoS

关于本章

13.1 介绍

定义

L2TP QOS 是基于 L2TP 场景的 QOS 业务，包括限速、调度和流映射。

目的

在 L2TP 业务批发的应用场景中，LAC 只做业务批发，实际的业务控制点在 LNS 上，LAC 和 LNS 之间是一 L2TP 隧道，这样有必要在 LNS 上对进入 L2TP 隧道的流量以及隧道内的业务流量做精细化控制，降低 LAC 和 LNS 之间因业务流的无序竞争造成的业务质量下降。同时，运营商也可以控制不同业务管道接入骨干网的流量，限制隧道内不同用户的突发流量。

L2TP HQoS 主要针对 LNS 侧用户做 QoS 调度，以便对 L2TP 隧道及隧道内的业务流量进行精细化的控制。

13.2 参考标准和协议

无

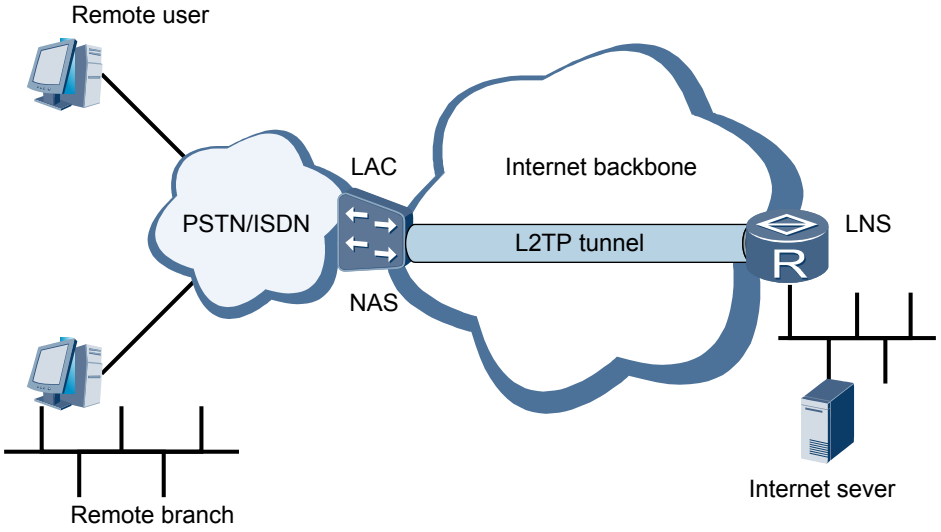
13.3 特性增强

不涉及

13.4 原理描述

13.4.1 L2TP QoS 实现机制

图 13-1 L2TP 组网模型



在图 13-1 中，用户需要通过二层网络登录到私网。

L2TP 访问集中器 LAC（L2TP Access Concentrator）是二层网络上有 PPP 端系统和 L2TP 处理能力的设备，一般是本地 ISP 的接入设备。LAC 位于 L2TP 网络服务器 LNS（L2TP Network Server）和远端系统（远程用户和远程分支机构）之间。

在 LAC 上可以对用户流量进行限速，流映射，调度等。

LNS（L2TP Network Server）是接受 PPP 会话的一端，通过 LNS 验证，用户就可以登录到私网上，访问私网资源。

在 LNS 上可以对用户流量进行限速、流映射和调度等。

LNS 上的 QoS 调度模式分为两种：

- 按隧道调度

  在这种调度模式下，不区分每个用户包含哪些业务，因此不为每个用户分配用户队列，而为隧道分配用户队列。
- 按会话调度

  在这种调度方模式下，为每个用户分配用户队列，为隧道分配组队列，每个用户包含 1 ～ 8 个优先级队列，各队列间可以做 SP 或 WFQ 调度。