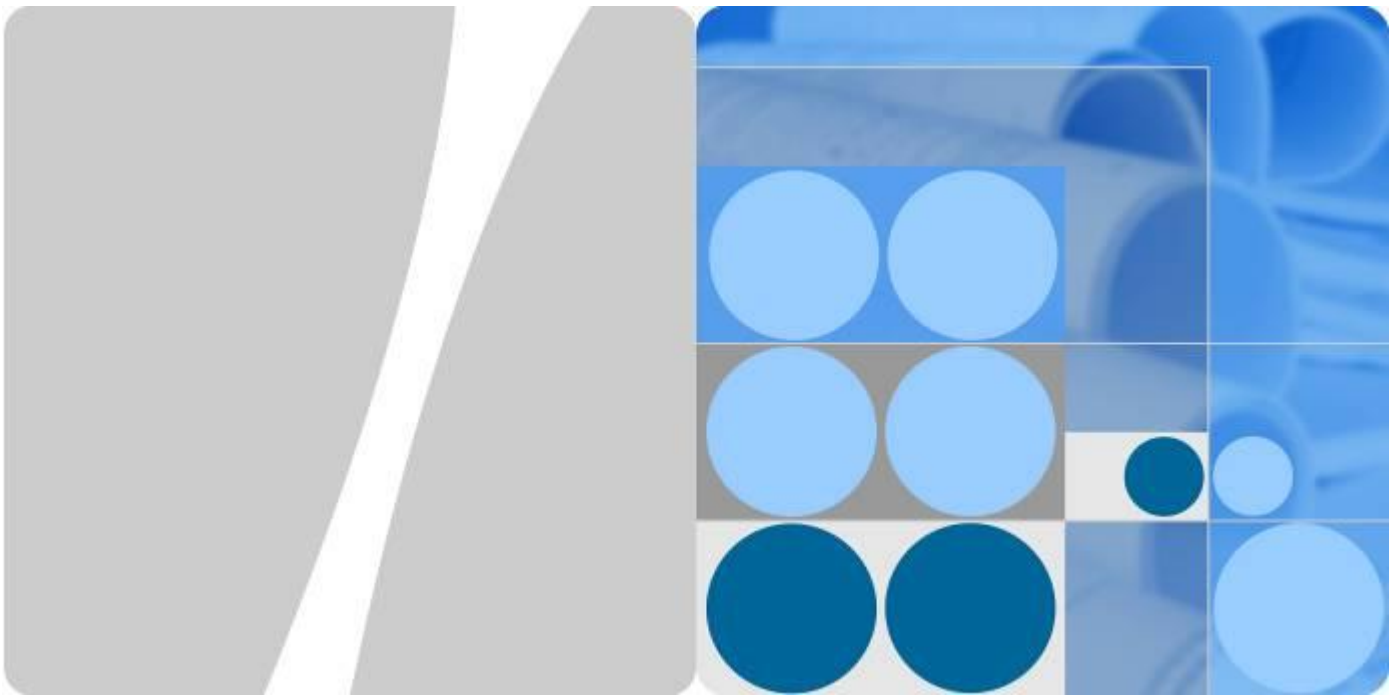


华为AR G3企业路由器

Qos 技术白皮书

文档版本	V1.1
发布日期	2011.10.31

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

目 录

1概述 5
1.1 介绍 5
1.2 QoS服务模型 6
1.3 AR QoS技术优势 7
2AR QoS技术实现 8
2.1 总体介绍 8
2.2 流量分类与标记 9
2.3 流量监管 12
2.4 流量整形 13
2.5 拥塞避免 15
2.6 拥塞管理 16
2.7 HQoS 20
2.8 链路分片与交叉 24
2.9 QoS配置模型 25
3典型应用 28
3.1 AR QoS基本应用 28
参考资料清单 33
术语与缩略语 34

插图目录

图 1 QoS硬件架构逻辑框图 8
图 2 流量监管示意图 12
图 3 srTCM算法示意图 13
图 4 trTCM算法示意图 13
图 5 队列整形示意图 14
图 6 RED丢弃示意图 15
图 7 DRR调度示意图 17
图 8 WFQ调度示意图 18
图 9 CBQ调度示意图 19
图 10 HQoS队列模型 21
图 11 HQos分层队列调度举例 23
图 12 链路分片与交叉 24
图 13 嵌套流策略 27
图 14 AR在企业网的QoS基本应用场景 30
图 15 示例HQos业务分解示意图 32

1 概述

1.1 介绍

服务质量QoS（Quality of Service）用于评估服务方满足客户服务需求的能力。由于网络提供的服务是多样的，因此可以基于不同方面进行评估。通常所说的QoS，是在分组传输过程中通过对吞吐量、时延、抖动、丢包率等因素的评估，来衡量网络的服务能力。

- 吞吐量

网络的两个节点之间在不丢包的情况下传输的最大速率，单位是字节每秒或比特每秒。

- 丢包率

指网络传输过程中的数据包丢失的比率。非拥塞状态下，丢包率应为零；拥塞时，根据服务要求对数据包有选择地丢弃。

- 时延

数据包在网络的两个节点之间传输，从发送端到接收端的时间间隔。对于网络中的一个设备来说，时延指数据包第一个比特进入到最后一个比特输出时间间隔。

- 抖动

即时延的变化。同一连接的不同数据包的时延不同，产生抖动。如，网络的两个节点之间传输报文时，从发送端到接收端第一个报文耗时10ms，第二个报文耗时15ms，则时延的变化为5ms。

1.2 QoS服务模型

QoS服务模型是一个逐步发展的过程，当前主要包括以下几种方式：

- Best-Effort service（尽力而为服务模型）一个单一的服务模型，网络尽最大的可能性来发送报文，不提供任何关于带宽、时延、抖动及丢包率的保证。尽力而为的服务是当前 Internet 提供的主要的一种服务，适用于大多数数据应用，如 FTP、WWW、E-Mail 等。不提供 Qos 相关配置时多是该种模型

- Integrated service（综合服务模型，简称 IntServ）是一个综合服务模型，通过 RSVP 信令机制提供端到端的 QoS 保证，从源端到目的端的每个网络节点都要运行 RSVP。应用在发送报文前先向网络提出带宽、时延等服务申请，每个网络节点视当前资源使用状况为其分配资源，并监视应用连接每条流的状态。IntServ 模型对路由器的要求很高，当网络中的数据流数量很大时，路由器的存储和处理能力会遇到很大的压力，部署困难且可扩展性较差。MPLS RSVP-TE 采用该种模型

- Differentiated service（差分服务模型，简称 DiffServ）是一种多服务模型，根据每个报文携带的优先级将网络上传输的业务分成不同的业务流，通过逐跳行为（PHB）为每种业务提供差分服务。DiffServ 的特点是实现简单、扩展性好 ，但由于其结构中网络和端系统之间缺乏信令通信 ，不能提供端到端的 QoS 保证。其他单节点的 Qos 配置采用的是该种模型

AR采用DiffServ模型实现QoS功能。

1.3 AR QoS技术优势

- 具有强大的流量识别能力，可以对报文进行全面识别和分类。不仅支持传统的三、四层 ACL 分类，还集成 DPI 引擎，可以解析报文的七层信息，可以识别有状态的协议。
- 支持全面的报文优先级标记，优先级映射表支持全面。
- 支持入方向/出方向的流量监管，支持单速率三色算法和双速率三色算法。
- 支持硬件流量管理器（TM），对基本转发性能零影响。
- 支持层次化 QoS，支持三级多级调度和三级整形。
- 最大支持 8k 队列，在出端口统一调度。
- 支持丰富的调度算法，包括 SP、WRR、DRR、WFQ 等多种算法，满足低速、高速各种链路应用。
- 支持 Tail-Drop、WRED 等拥塞避免算法。
- 支持统一的 QoS 策略配置模型，便于用户理解与使用。
- 支持动态修改已经生效的 QoS 策略，修改过程中不影响流量匹配。

2 AR QoS 技术实现

2.1 总体介绍

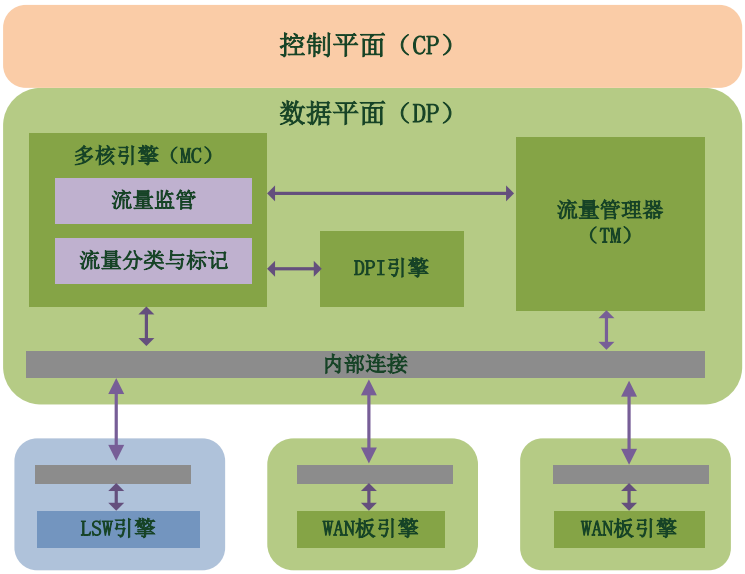


图 1　QoS硬件架构逻辑框图

AR的流量转发是集中式处理模型，如上图所示，从LAN侧转发到WAN的流量都经过数据平面进行转发，对流量集中控制，统一部署QoS功能。对于LAN接口或LAN插卡，为了充分发挥其转发能力，LAN与LAN接口之间的转发，则通过LSW硬件实现QoS功能。

AR QoS主要处理部件包括三个部分：

- 多核软件完成流量分类与标记、流量监管等功能
- 流量管理器完成队列相关处理，包括调度、整形、WRED、HQoS 等，低端设备由软件实现，高端设备通过硬件实现。软硬件实现 Qos 的功能相同，硬件实现 Qos 不会影响转发性能
- DPI 引擎可以识别七层应用与有状态的协议，提供强大的流量分类能力

2.2 流量分类与标记

2.2.1 流量分类

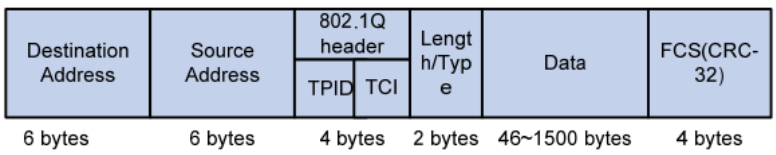
流量分类就是根据不同类型的数据报文划分优先级或多个服务类，这些优先级和服务类是区分服务模型的前提和基础。根据分类结果可以设置不同的流量策略，从而提供不同的服务。AR QoS具有强大的流量分类能力，可以根据报文L2～L7的信息进行分类。

- L2 字段：
  - 以太链路包括源 MAC 地址、目的 MAC 地址、802.1P、VLAN ID、以太网协议类型；
  - MPLS 报文的 MPLS-EXP；
  - ATM 报文的 PVC、ATM-CLP；
  - 帧中继的 DLCI、FR-DE。
- L3 字段：源 IP 地址、目的 IP 地址、ToS、DSCP、三层协议类型等
- L4 字段：TCP、UDP 协议及四层端口号
- L7 字段：通过 DPI 技术作应用识别，包括七层信息和有状态的协议。

其中VLAN的8021.p、IP的Tos DSCP、MPLS的EXP等几个重要字段称为QoS优先级字段，下面分别介绍其含义和作用：

- VLAN 802.1P 优先级

  通常二层路由器之间交互 VLAN 帧。根据 IEEE 802.1Q 定义，VLAN 帧头中的 PRI 字段（即 802.1p 优先级）标识了服务质量需求。

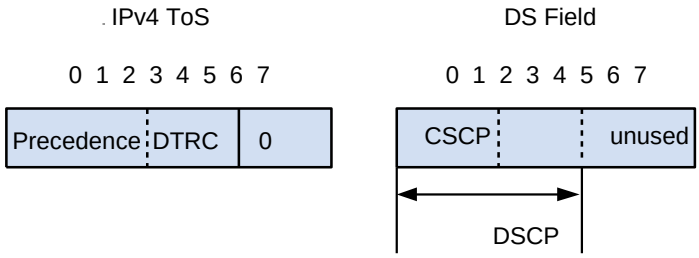


  如图所示，4 个字节的 802.1Q 标签头包含了 2 个字节的 TPID（Tag Protocol Identifier，标签协议标识，取值为 0x8100 ）和 2 个字节的 TCI （Tag Control Information，标签控制信息），其中 TCI 中的前 3 位即 802.1p 优先级。

- IP ToS 优先级和 DSCP 优先级

在 RFC791、RFC134 和 RFC1349 中定义了 IPv4 报文头的 ToS（Type of Service）字段，如图所示，ToS 字段包含 3bits 的优先级（Precedence）、D bit、T bit、R bit 和 C bit，ToS 字段的最高位 bit 必须为 0。D bit 代表延迟（Delay），T bit 代表吞吐量（Throughput），R bit 代表可靠性（Reliability），C bit 代表花费（Cost）。在实施 QoS 时，路由器会检查报文的优先级。其余的比特位未使用。

在 RFC2474 中对 IPv4 报文头的 ToS 字段进行了重新定义，称为 DS（Differentiated Services）字段。DS 字段的低 6 位（0～5 位）用作区分服务代码点 DSCP（DS CodePoint），高 2 位(6、7 位)是保留位。DS 字段的低 3 位(0～2 位)是类选择代码点 CSCP(Class Selector CodePoint)，相同的 CSCP 值代表一类 DSCP。DS 节点根据 DSCP 的值选择相应的 PHB(Per-Hop Behavior）。



- MPLS EXP 优先级

对于 MPLS 报文，通常将标签信息中的 EXP 域作为 MPLS 报文的 CoS 域，与 IP 网络的 ToS 域等效，用来区分数据流量的服务等级，以支持 MPLS 网络的 DiffServ。

2.2.2 优先级标记

标记是对流量进行分类之后，根据服务级别设置报文中的QoS字段。对于IP报文，就是对IP优先级或者DSCP进行设置。对于MPLS报文，就是对MPLS头部的EXP域进行设置。对于二层以太报文，就是对VLAN Tag报文中的8021P域进行设置。

除了各种报文的本身携带优先级之外，还有一个重要概念：本地优先级（LP）。本地优先级由报文优先级或端口默认优先级影射而来，表示报文在设备内的转发优先级，用于确定报文从出接口

发送时的队列号，值越大优先级越高。在转发出路由器时本地优先级失去意义，可以根据配置的模式影射回报文优先级，但当配置模式为直接设置报文优先级时本地优先级没有作用。

报文数据流进入设备端口之后，设备会根据端口配置的信任模式来分配报文的各类优先级。端口的信任模式如下，对于二层网络中的报文，可以选择信任802.1P模式；对于三层网络中的报文，可以选择信任DSCP模式；对于MPLS报文，可以选择信任EXP模式。 缺省模式下端口模式为不信任报文的优先级，完全信任端口配置的优先级，即所有报文在相同端口的内部优先级相同

- 信任端口的优先级

  缺省情况下，端口模式为不信任报文优先级，即信任端口优先级，按照端口的优先级，根据映射表为报文分配优先级。

- 信任 DSCP 模式

  配置为信任 DSCP 优先级时，根据报文的 DSCP 优先级作为索引，查看 DSCP 映射表，得到报文的 LP 优先级，在设备内转发的时候使用 LP 作为拥塞处理的优先级值。当报文从设备转发出去时，把映射后的优先级更新到出报文的 VLAN tag、IP、DSCP 或 MPLS 标签的 EXP 字段。

- 信任 802.1P 模式

  配置为信任 802.1P 优先级时，根据报文的 802.1P 优先级作为索引，查看 802.1P 映射表，得到报文的 LP 优先级，在设备内转发的时候使用 LP 作为拥塞处理的优先级值。当报文从设备转发出去时，把射后的优先级更新到出报文的 VLAN tag、IP、或 DSCP 字段。

- 信任 EXP 模式

  配置为信任 EXP 优先级时，根据报文的 EXP 优先级作为索引，查看 EXP 映射表，得到报文的 LP 优先级，在设备内转发的时候使用 LP 作为拥塞处理的优先级值。当报文从设备转发出去时，把映射后的优先级更新到出报文的 VLAN tag、IP、DSCP 或 MPLS 标签的 EXP 字段。

2.3 流量监管

流量监管TP（Traffic Policing）就是对流量进行控制，通过监督进入网络的流量速率，对超出部分的流量进行“惩罚”，使进入的流量被限制在一个合理的范围之内，从而保护网络资源和企业网用户的利益。

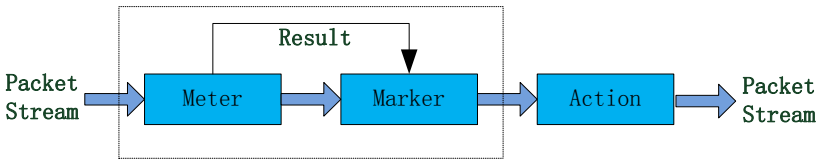


图 2　流量监管示意图

如上图所示，AR 流量监管由三部分组成：

- Meter：通过令牌桶机制对网络流量进行度量，向 Marker 输出度量结果。
- Marker：根据 Meter 的度量结果对报文进行染色，报文会被染成 green、yellow、red 三种颜色。
- Action：根据 Marker 对报文的染色结果，对报文进行一些动作，动作包括：
  - pass：对测量结果为“符合”的报文继续转发。
  - pass + remark：修改报文内部优先级后再转发。
  - discard：对测量结果为“不符合”的报文进行丢弃。

经过流量监管，如果某流量速率超过标准，AR可以选择降低报文优先级再进行转发或者直接丢弃。默认情况下，green、yellow进行转发，red报文丢弃。

AR的监管算法遵照RFC 2697和RFC 2698定义的算法，支持单速率三色标记（srTCM）和双速率三色标记（trTCM）。

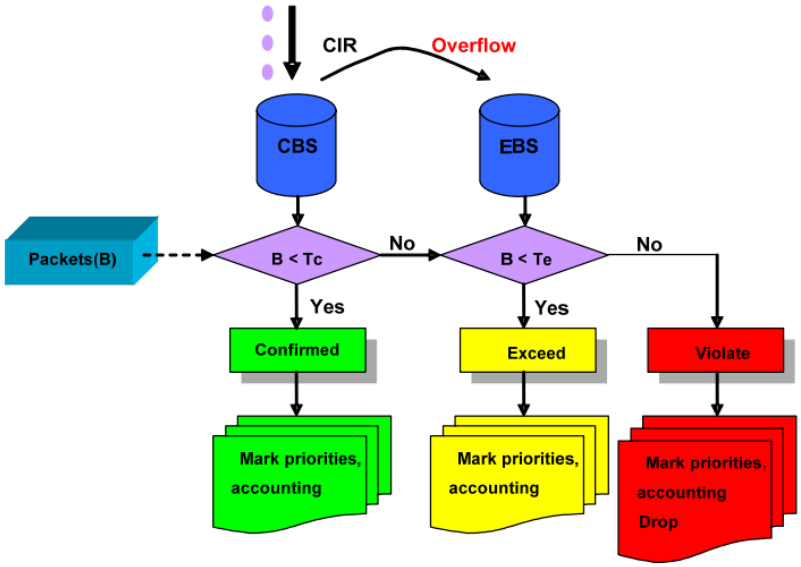


图 3　srTCM算法示意图

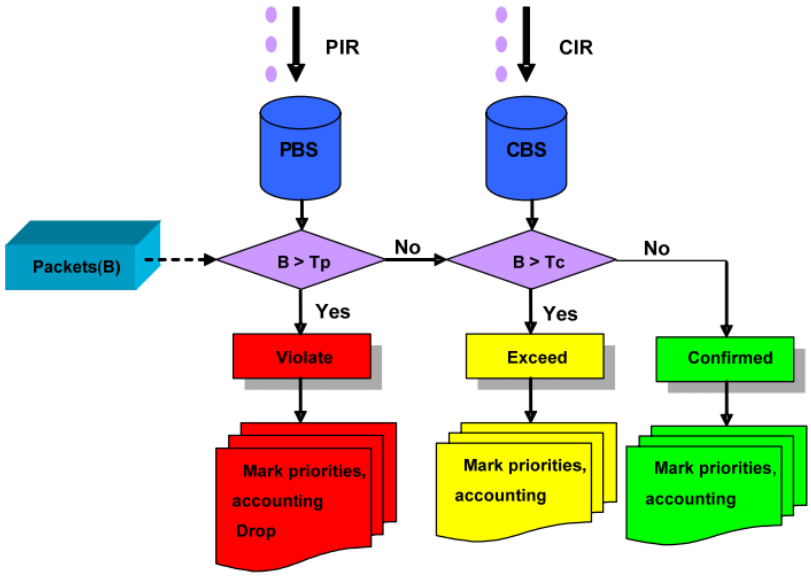


图 4　trTCM算法示意图

2.4 流量整形

通用流量整形（Generic Traffic Shaping，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，使这一流量的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。

GTS也是通过令牌桶进行流量控制。当报文的发送速度过快时，首先在缓冲区进行缓存，在令牌桶的控制下，再均匀的发送这些被缓存的数据。

GTS的基本处理过程如图所示，其中用于缓存报文的队列称为GTS队列。

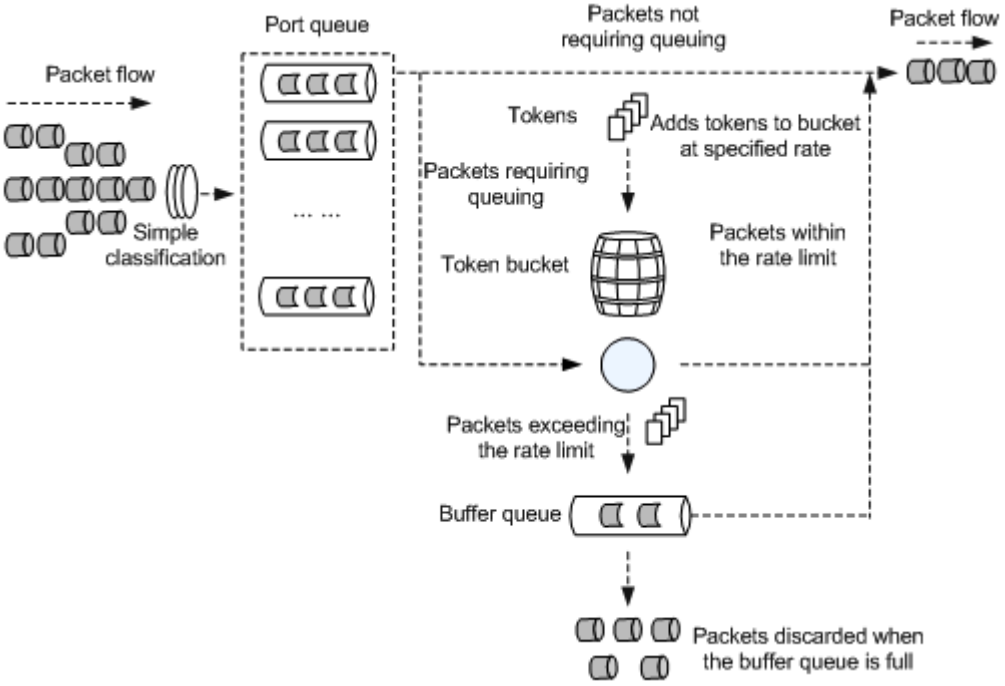


图 5　队列整形示意图

- 当报文到来的时候，首先对报文进行分类，使报文进入不同的队列。
- 若报文进入的队列没有配置队列整形功能，则直接发送该队列的报文；否则，进入下一步处理。
- 按用户设定的队列整形速率（CIR）向令牌桶中放置令牌。如果令牌桶中有足够的令牌可以用来发送报文，则报文直接被发送，在报文被发送的同时，令牌做相应的减少。如果令牌桶中没有足够的令牌，则将报文放入缓存队列，如果报文放入缓存队列时，缓存队列已满，则丢弃报文。
- 缓存队列中有报文的时候，系统按一定的周期从缓存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缓存队列中的报文不能再发送或缓存队列中的报文全部发送完毕为止。

AR支持三级流量整形，分别为队列整形、逻辑接口（如子接口、PVC等）整形、物理接口整形。三级整形以汇聚的关系，按设置的速率逐级控制。

2.5 拥塞避免

拥塞避免（Congestion Avoidance）是指通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞发生或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。

传统的丢包策略采用尾部丢弃（Tail-Drop）的方法。当队列的长度达到最大值后，所有新入队列的报文（缓存在队列尾部）都将被丢弃。这种丢弃策略会引发TCP全局同步现象：当队列同时丢弃多个TCP连接的报文时，将造成多个TCP连接同时进入拥塞避免和慢启动状态以降低并调整流量，而后又会在某个时间同时出现流量高峰，如此反复，使网络流量不停震荡。

为避免TCP全局同步现象，可使用RED或WRED（Weighted Random Early Detection）技术。RED通过对报文采取随机丢弃方式，使得不同TCP连接的报文被丢弃的概率不相同，避免了TCP的全局同步现象。当某个TCP连接的报文被丢弃，开始减速发送的时候，其他的TCP连接仍然有较高的发送速度。这样，任何时候总有TCP连接在进行较快的发送，提高了线路带宽的利用率。

AR的流队列支持基于报文IP优先级和DSCP值的WRED丢弃策略。对每一种优先级都可以独立设置报文的丢弃高门限、低门限和丢弃概率。当队列的长度小于低门限时，不丢弃报文；当队列的长度超过高门限时，丢弃所有到来的报文；当队列的长度在上限和下限之间时，开始随机丢弃到来的报文（队列越长，丢弃概率越高，但有一个最大丢弃概率）。

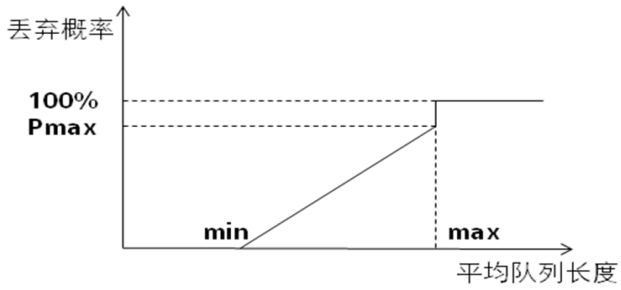


图 6　RED丢弃示意图

AR支持64个丢弃模板，每个流队列都可以使用任一模板，以设置不同的丢弃参数。

直接采用队列的绝对长度和低门限、高门限比较并进行丢弃，将会对突发性的数据流造成不公正的待遇，不利于数据流的传输。AR采用平均队列长度和设置的队列上限、下限比较来确定丢弃的概率。队列的平均长度是队列长度被低通滤波后的结果。它既反映了队列的变化趋势，又对队列长度的突发变化不敏感，避免了对突发性数据流的不公正待遇。

2.6 拥塞管理

AR的典型应用是部署在企业出口，接入到广域网(WAN)。WAN侧接口带宽通常要比LAN侧接口的带宽小，当从LAN传送数据到WAN侧时，在WAN接口就会产生丢包，即发生了拥塞。

拥塞管理是指网络在发生拥塞时，对流量的发包顺序进行管理和控制。拥塞管理一般采用队列调度技术，目前AR支持的队列调度技术有：

- LAN接口支持的调度模式：PQ、DRR、PQ+DRR、WRR、PQ+WRR
- WAN接口支持的调度模式：PQ、WFQ、PQ+WFQ、CBQ

本文描述PQ、WRR、DRR、WFQ以及CBQ的基本原理，并说明在AR中的应用情况。

2.6.1 PQ调度

PQ队列是针对关键业务应用设计的。关键业务有一个重要的特点，即在拥塞发生时要求优先获得服务以减小响应的延迟。PQ严格按照优先级从高到低的次序，优先发送较高优先级队列中的分组，当较高优先级队列为空时，再发送较低优先级队列中的分组。这样，将关键业务的分组放入较高优先级的队列，将非关键业务的分组放入较低优先级的队列，可以保证关键业务的分组被优先传送，非关键业务的分组在处理关键业务数据的空闲间隙被传送。

PQ的缺点是如果较高优先级队列中长时间有分组存在，那么低优先级队列中的分组将可能一直得不到服务。为了防止其他低优先级队列饿死，AR实现了速率可控优先级队列（RCPQ），即PQ队列支持最大带宽限制。

2.6.2 WRR调度

WRR（Weight Round Robin）加权循环调度在RR（Round Robin）调度的基础上演变而来，在队列之间进行轮流调度，根据每个队列的权重来调度各队列中的报文流。RR调度即权值为1的WRR调度。

在进行WRR调度时，AR根据每个队列的权值进行轮循调度。调度一轮权值减一，权值减到零的队列不参加调度,当所有队列的权限减到0时，开始下一轮的调度。，假设所有流量的报文长度相同，则从统计上看，各队列中的报文流所获得的带宽与权重成正比。例如，出口带宽10M，有4个队列，其权重为4、3、2、1，则每个队列得到的带宽分别为4M、3M、2M、1M。

WRR调度避免了采用PQ调度时低优先级队列中的报文可能长时间得不到服务的缺点。其优点是算法简单，效率较高，适合于硬件实现。但WRR最初是针对固定包长（ATM信元）设计的调度算法，其缺点是对于变长包调度不准确。

另一缺点是不能保证高优先级业务（如VOIP）的时延。AR的解决方法是支持PQ和WRR混合调度，高优先级报文进入PQ，先获得服务，其他报文进行WRR调度。

2.6.3 DRR调度

DRR（Deficit Round Robin）调度，解决了WRR不能支持变长包而出现的带宽分配不公平的缺点，通过调度过程中考虑了包长的因素，从而达到调度的速率公平性。

DRR算法的原理：每个队列设置一个服务配额（Quantum）和一个差额计数器（Deficit counter，简称DC），其中配额可以设置权重。调度后的带宽比例即权重的比例。处理过程如下：

1、调度器访问每个非空队列，如果队列头部的包长度大于DC，则调度器移动到下一个队列。

2、如果队列头部的包小于或等于DC，则变量DC减去包长字节数，同时把报文发送出去。调度器继续输出包和减少DC值，直到队列头部的包长度大于变量DC值。剩余的DC值将作为信用值累加到下次轮询时使用。

3、如果队列输出包直到队列为空，则设置DC为零，此时调度器将服务下一个非空队列。

如下例子，四个队列A、B、C、D，配额为500字节，权重分别为1、5、1、1，调度过程如图：

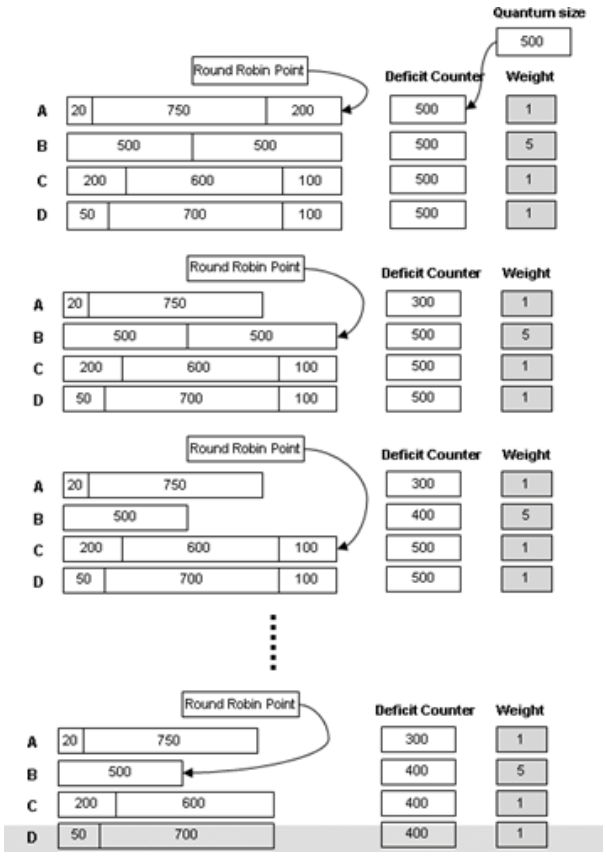


图 7　DRR调度示意图

DRR算法很好地解决变长包调度不准确的问题。其缺点是：

1、　如果遇到一个空队列，服务器立即移到下一个队列。如果队列错过了它的传输时序就只能等到下一个属于它的时序才能传输。如果每个队列都在使用，那么该队列的数据包要等到所有的队列都处理完之后才能被处理。这是轮询算法都存在的问题。

2、　带宽较大的服务分配的权重也较大，其一轮服务定额就会比较大，会导致流量突发，调度后可能丢包，低速网络尤其严重。

考虑到DRR不适用于低速链路，AR只在LAN侧接口支持DRR调度，WAN侧接口采用WFQ算法。

2.6.4 WFQ调度

WFQ算法是一种基于流的公平调度算法，是GPS（Generalized Processor Sharing）系统的近似实现。它的基本思想是通过虚拟时间机制实现，数据包在入队列时，依据包的大小、服务速率和报文优先级等信息分配一虚拟完成时间，每次调度，比较一组队列头的数据包的虚拟完成时间，选择虚拟完成时间值最小的数据包发送。虚拟完成时间的作用是为了报文出队时进行排序，真正意义上它只是一个序列号。

WFQ 在计算报文序列号时增加了权重，从统计上，WFQ 使权重大的报文获得优先调度的机会多于低优先权的报文。WFQ 能够按流的“会话”信息（协议类型、源和目的TCP 或UDP 端口号、源和目的IP 地址、ToS 域中的优先级等）自动进行流分类，并且尽可能多地提供队列，以将每个流均匀地放入不同队列中，从而在总体上均衡各个流的延迟。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，所得的带宽越少。优先级的数值越大，所得的带宽越多。

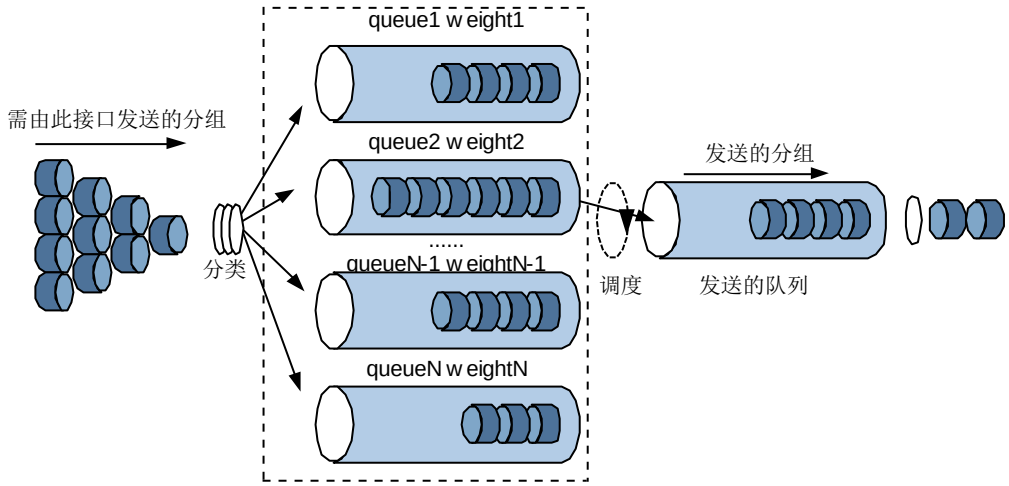


图 8　WFQ调度示意图

例如：接口中当前有8个流，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）之和，即1+2+3+4+5+6+7+8=36。每个流所占带宽比例为：（自己的优先级数+1）/（所有（流的优先级+1）之和）。即每个流可得的带宽比例分别为：1/36、2/36、3/36、4/36、5/36、6/36、7/36、8/36。

WFQ可以公平地分享网络资源，在拥塞发生时能很好地均衡各个流的延迟和延迟抖动。其缺点是算法实现复杂。AR在低端设备通过软件实现，对转发性能存在一定影响。但高端设备通过硬件实现，对转发性能没有任何影响。

另外，AR还支持PQ和WFQ的混合调度，既能保证高优先级业务的低时延，又能保证其他业务得到公平调度。

2.6.5 CBQ调度

CBQ（Class-based Queue）基于类的加权公平队列是WFQ功能的扩展，用户可以自定义流量分类，进入相应的队列作调度。CBQ首先根据IP优先级或者DSCP优先级、输入接口、IP报文的五元组等规则来对报文进行分类，然后让不同类别的报文进入不同的队列。对于不匹配任何类别的报文，送入系统定义的缺省类。

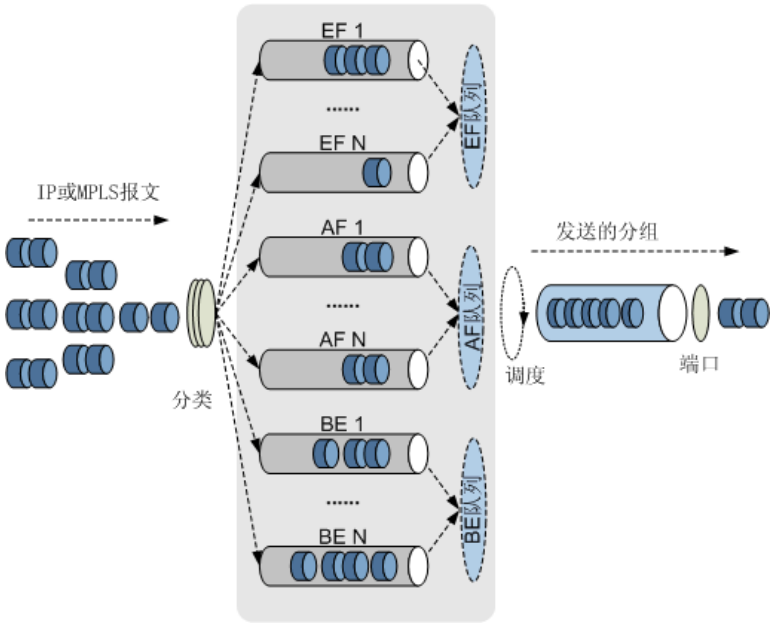


图 9　CBQ调度示意图

如图所示CBQ提供三类队列：

- EF队列：满足低时延业务
- AF队列：满足需要带宽保证的关键数据业务
- BE队列：满足不需要严格QoS保证的尽力发送业务

下面分别介绍每类队列的功能及作用：

- EF队列：

EF队列是具有高优先级的队列，一个或多个类的报文可以被设定进入EF队列，不同类别的报文可设定占用不同的带宽。在调度出队的时候，若EF队列中有报文，则总是优先发送EF队列中的报文，直到EF队列中没有报文时，或者超过为EF队列配置的最大预留带宽时才调度发送其他队列中的报文。

进入EF队列的报文，在接口没有发生拥塞时（此时所有队列中都没有报文）都可以被发送；在接口发生拥塞时（队列中有报文时）会被限速，超出规定流量的报文将被丢弃。这样，属于EF队列的报文既可以获得空闲的带宽，又不会占用超出规定的带宽，保护了其他报文的应得带宽。此外，由于只要EF队列中有报文，系统就会发送EF队列中的报文，所以EF队列中的报文被发送的延迟最多是接口发送一个最大长度报文的时间，无论是时延还是时延抖动，EF队列都可以将之降低为最低限

度。这为对时延敏感的应用（如VoIP业务）提供了良好的服务质量保证。

由于EF队列中的报文一般是语音报文（VoIP），采用的是UDP报文，所以没有必要采用WRED的丢弃策略，采用尾丢弃策略即可。

- AF队列

每个AF队列分别对应一类报文，用户可以设定每类报文占用的带宽。在系统调度报文出队的时候，按用户为各类报文设定的带宽将报文出队发送，可以实现各个类的队列的公平调度。当接口有剩余带宽时，AF队列按照权重分享剩余带宽。同时，在接口拥塞的时候，仍然能保证各类报文得到用户设定的最小带宽。

对于AF队列，当队列的长度达到队列的最大长度时，缺省采用尾丢弃的策略，但用户还可以选择用WRED丢弃策略。

- BE队列

当报文不匹配用户设定的所有类别时，报文被送入系统定义的缺省类。虽然允许为缺省类配置AF队列，并配置带宽，但是更多的情况是为缺省类配置BE队列。BE队列使用WFQ调度，使所有进入缺省类的流量按报文优先级进行基于流的队列调度。

对于BE队列，当队列的长度达到队列的最大长度时，缺省采用尾丢弃的策略，但用户还可以选择用WRED丢弃策略。

2.7 HQoS

2.7.1 介绍

传统的QoS基于接口进行流量调度，单个接口只能区分业务优先级，无法区分用户。只要属于同一优先级的流量，使用同一个接口队列，彼此之间竞争同一个队列资源。因此，传统的QoS无法对接口上多个用户的多个流量进行区分服务。

例如：有两个用户同时发送AF4的流量，用户1发送10M，用户2发送1G。但AF4的流量我们限速为10M。传统的QoS不区分用户，由于用户2发送的AF4流量大，用户2的报文有很大几率进入队列，而用户1的报文则被丢弃的概率非常大。因此用户1的流量就受到了其它用户的影响。这在运营商拓其针对企业、签约用户的业务中是非常不利的。因为运营商无法保证其所有用户的流量，也就无法吸引更多的用户来购买他的带宽和服务套餐。

目前，越来越多的企业用户通过向运营商租用专线的方式来构建自己的企业网，不同企业之间，其业务侧重点和所需要的服务质量是有差别的。这就要求运营商能够依据不同企业的业务需求提供不同的调度策略和QoS保证。传统的QoS无法区分用户，所以无法对不同的企业用户提供有差别的队列调度服务。

随着网络用户数量的持续增长和网络业务的不断丰富，用户和运营商都希望能够提供区分用户和用户业务的服务，以获得更好的服务质量和更多的利润。HQoS（Hierarchical Quality of Service）

基于多级队列实现层次化调度，不仅区分了业务，也区分了用户。既能够提供精细化的服务质量保证，又能够从整体上节约网络运行维护成本。

2.7.2 AR支持的HQoS

AR支持的HQos是在下行TM中完成，通过报文上行解析并携带的信息进行物理级别、逻辑级别、应用或业务级别等多个调度级别，每一级别可以使用不同的特征进行流量管理，达到对流量的进一步细分和控制。

2.7.3 HQos的队列划分

HQoS基于队列实现层次化调度，目前在AR上支持三级队列：Level3流队列（Flow Queue）、Level2用户队列（Subscriber Queue）、 Level1端口队列（Port Queue）。三级队列以树状结构汇聚，流队列为叶子节点，端口队列为根结构。报文作层次化调度时，首先进入叶子节点，经过多级调度后，从根结点发送出去。

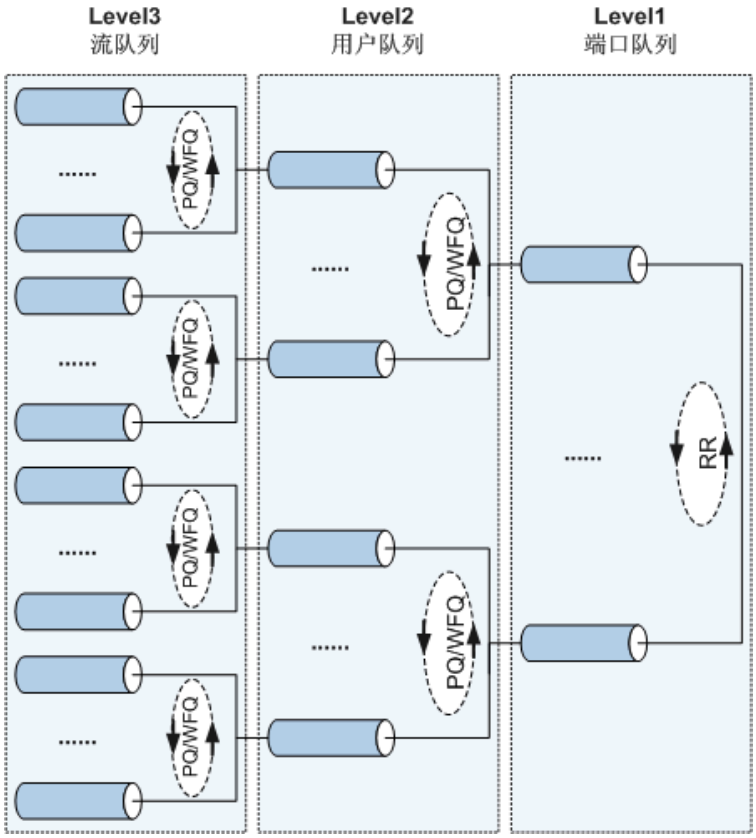


图 10　HQoS队列模型

流队列（Flow Queue）：

属于业务级别，可以用于管理某一用户各种不同业务的带宽，每个用户的同类业务可以被认为是一个业务流，HQoS能够针对每个用户的不同业务流进行队列调度。流队列一般与业务类型相对应，包括EF、AF、BE等，用户可以配置流队列的调度方式。流队列一般由用户配置

用户队列（Subscriber Queue）：

属于逻辑级别，可以用于管理接口上各用户的带宽，来自同一用户的所有业务可以被认为是一个用户队列，HQoS可以使该用户队列下的所有业务共享一个用户队列的带宽。可以基于如下信息区分HQoS用户队列：

基于“以太接口+VLAN”

基于“帧中继接口+DLCI”

基于“ATM接口+PVC”

端口队列（Port Queue）：

属于物理级别，用于管理整个物理接口的带宽，每个端口一个队列，端口队列之间进行轮询调度（RR），用户仅可以配置基于端口的流量整形，但其调度方式不可配置。

举例说明：

HQOS业务分类有三级，第一级为物理端口，第二级使用VLAN ID区分用户，第三级使用IP DSCP区分不同的业务类型，具体需要如下：有两个用户接入设备，分配给这两个用户的总带宽为500M。其中，

- 用户A的流量策略：
  - Service VLAN ID = 10
  - 总带宽150 MB
  - 分三个业务类：
  - 最优先业务: 例如语音，其IP Precedence = 5
  - VPN业务: 其 IP Precedence = 3，要求保证60%带宽
  - Internet业务:其IP Precedence = 0，要求保证20%带宽
- 用户B的流量策略：
  - Service VLAN ID = 15
  - 总带宽：CIR: 75 MB, PIR: 100 MB
  - 使用两个不同的用户VLAN承载不同业务类:
  - Customer VLAN 2
    - VPN业务: 其IP DSCP = 32 in case of congestion，要求保证70%可用带宽
    - 数据业务: 其IP DSCP = 8，要求保证20%可用带宽
  - Customer VLAN 100
    - Internet业务: 不超过10 MB

该应用的分层队列模型如下图所示。

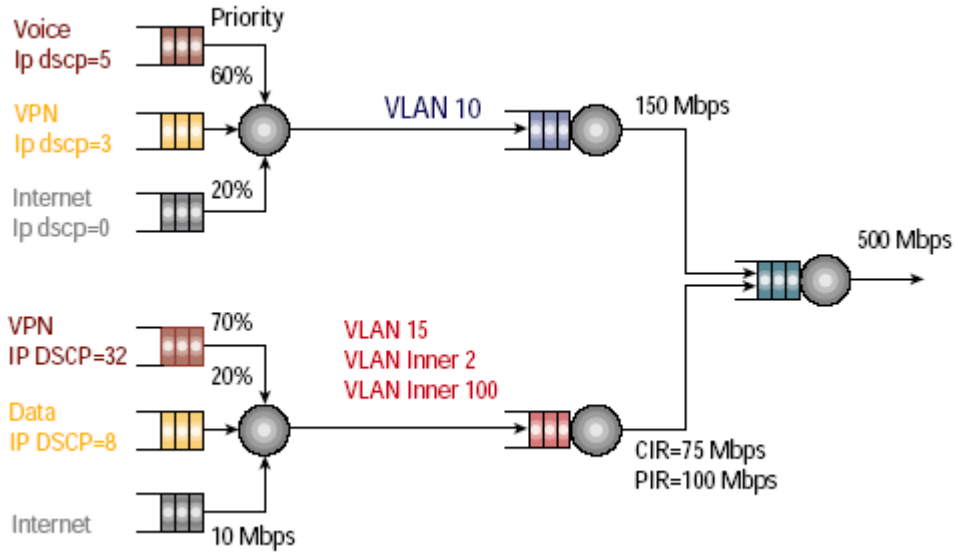


图 11　HQos分层队列调度举例

2.7.3.1　HQoS调度器

HQoS通过分级的方式，来实现更加精细化的调度，为用户QoS业务层面提供丰富的业务支撑。

AR提供了三级调度器，即流队列调度器、用户队列调度器和端口队列调度器。流队列调度器和用户队列调度器都支持PQ、WFQ、PQ+WFQ调度。端口队列调度器使用轮询调度RR（Round Robin）方式。上例中根据业务、逻辑端口的权重配置HQos调度器可以控制

单个用户A/B的某一种业务Voice/VPN/Internet的带宽

单个用户A/B的三种业务Voice/VPN/Internet的带宽分配

单个用户A/B的所有业务的带宽

多个用户A B之间的带宽分配

多个用户A B的总带宽

2.7.3.2　HQoS整形器

整形器实现报文的缓存及限速功能。AR支持三级整形器，即流队列整形器、用户队列整形器和端口队列整形器。报文进入设备后先缓存到队列，再限速从队列发送报文，整形器配合限速算法可以保证承诺速率并限制最大速率。在各级的整形器中都是通过令牌桶进行流量控制。

1.报文在令牌桶进行缓存，在令牌桶中匀速的增加令牌，

2.得到令牌的报文得以转发，保证均匀的发送这些被缓存的数据。报文转发消耗掉等同于报文长度的令牌，

3. 当剩余令牌不能是当前报文转发时，报文缓存在令牌桶中，直到令牌被匀速的注入进来达到报文长度值

4. 在令牌桶缓存报文的过程中，如果发生桶满则导致报文丢弃

2.7.3.3 HQoS丢弃器

丢弃器在报文入队列之前将根据丢弃策略丢弃报文。HQoS支持的3种队列支持不同的丢弃方式：

- 端口队列：尾部丢弃
- 用户队列：尾部丢弃
- 流队列：尾部丢弃和 WRED

尾部丢弃是在拥塞发生期间，队列尾部的数据报文将被丢弃，直到拥塞解决。

WRED是基于DSCP或IP优先级的一种丢弃策略。每一种优先级都可以独立设置报文的丢包的高门限、低门限及丢包率，报文到达低门限时，开始根据权重的配置进行加权后丢弃，权重高的丢弃的优先级低，权重低的丢弃的优先级高，到达高门限时丢弃所有的报文，随着门限的增高，丢包率不断增加，最高丢包率不超过设置的丢包率，直至到达高门限，报文全部丢弃。由于优先级信息作用在流队列，所以可以根据报文的优先级信息进行加权丢弃，优先随机丢弃优先级低的报文

2.8 链路分片与交叉

链路分片与交叉（LFI）是在PPP链路和FR链路应用的低速链路技术。

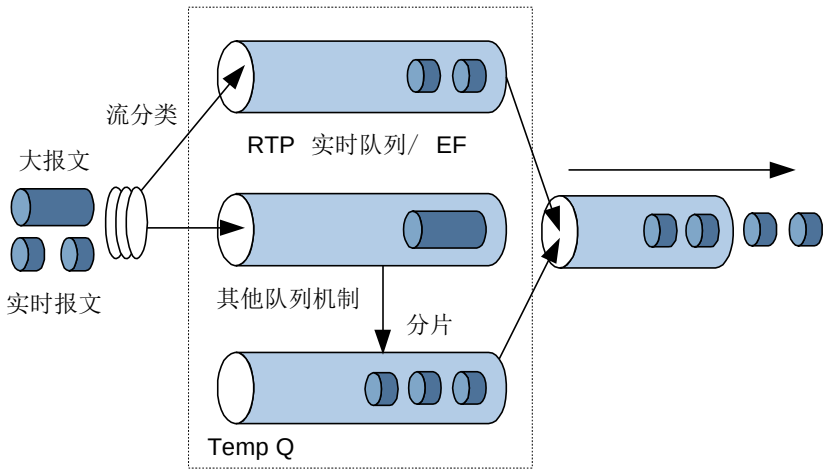


图 12 链路分片与交叉

在低速串行链路上，实时交互式通信，如Telnet和VoIP，往往会由于大型分组的发送而导致阻塞延迟，例如，正好在大报文被调度而等待发送时，语音报文到达，它需要等该大报文被传输完毕后才能被调度。对于诸如交互式语音等实时应用而言，大报文导致的这种阻塞延迟太长了，对端将听到断断续续的话音。交互式语音要求端到端的延迟不大于100～150ms。

一个1500bytes（即通常MTU的大小）的报文需要花费215ms穿过56Kbps的链路，这超过了人所

能忍受的延迟限制。为了在相对低速的链路上限制实时报文的延迟时间，例如56Kbps Frame Relay或64Kbps ISDN B通道，需要一种方法将大报文进行分片，将小报文和大报文的分片一起加入到队列。

LFI将大型数据帧分割成小型帧，与其他小片的报文一起发送，从而减少在带宽较小的链路上的延迟和抖动。

如上图所示，链路分片与交叉的处理过程：大报文和实时报文一起到达某个接口时，除了RTP实时队列和EF队列中的报文外，其他队列中的大报文将被分成若干小包放入分片队列进行发送；但如果此时RTP实时队列和EF中有缓存的报文，则优先调度RTP实时队列和EF队列，否则继续调度分片队列，这样就避免了在低速链路上传送大包对实时报文造成的时延与抖动。

2.9 QoS配置模型

QoS策略提供了一组模板化的命令行配置方式，目的是将基于ACL的QoS配置命令整合在一起，包含三个要素：流分类器、流行为、QoS策略。

- 流分类器（traffic classifier）：采用一定的规则识别出符合某类特征的报文。
- 流行为（traffic behavior）：对报文做的一些 QoS 动作集合。
- 流策略（traffic policy）：将指定的流分类器和流行为关联后形成完整的 QoS 策略。

QoS策略可以应用于接口或子接口，更方便地配置QoS功能。

2.9.1 分类器

流分类器用来定义一组流量匹配规则，来对报文进行分类。

流量分类采用一定的规则识别符合某类特征的报文，从而把具有某类共同特征的报文划分为一类，它是有区别地进行服务的前提和基础。

分类器中规则之间的关系分为：and或者or，默认关系为or。

- and：报文只有匹配了所有的规则，设备才认为报文属于这个类
- or：报文只要匹配了类中的一个规则，设备就认为报文属于这个类。

流分类器的匹配是以ACL为基础的，但是却又不同于ACL。二者之间的最主要区别在于流分类器只有分类匹配一个作用，而没有表明对符合分类的流做出什么动作，而ACL本身是为了进行访问控制，所以附带有deny和permit的动作。而且二者所匹配的范围不同，流分类器所能匹配的流范围大于ACL，可以说ACL中的匹配范围是Class中的一个子集。比如流分类器可以匹配入接口，ACL则不支

持。

2.9.2 流行为

流行为用来定义针对报文所做的QoS动作。进行复杂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行为关联起来才有意义。

在AR中针对复杂流分类可实施的流行为包括禁止/允许、重标记、重定向、流量监管、流量整形、流镜像、流量统计、队列调度。除deny外，其他流行为可以组合使用。

- 禁止/允许

  禁止/允许是最简单的流控动作。AR 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。

- 重标记

  重标记是对报文的优先级字段进行设置。在不同的网络中报文使用不同的优先级字段，例如 VLAN 网络使用 802.1p，IP 网络使用 ToS，MPLS 网络使用 EXP。因此需要 AR 可以针对不同的网络对报文的优先级进行重标记。

  通常网络的边界节点设备需要对进入的报文进行优先级重标记。网络内部的节点设备按照边界节点所标记的优先级提供相应等级的 QoS 服务，或者按自己的标准重新进行标记。

- 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。

  通过重定向可以实现策略路由。这种策略路由是静态的，当配置中的下一跳不可用时，系统将按原来的转发路径转发报文。

- 流量监管

  流量监管就是一种通过对流量规格的监督，来限制流量及其资源使用的流控动作。通过流量监管，可以控制某个流的规格，对于超过规格的流量，可以采取丢弃、重标记颜色、重标记优先级或其他的 QoS 措施。

- 流量整形

  流量整形也是通过对流量规格的监督，来限制流量及其资源使用的流控动作。它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设备可供给的网络资源，避免不必要的报文丢弃和拥塞。流量整形通过限制流出某一网络的某一连接的流量，使这类报文以比较均匀的速度向外发送。

- 流镜像

流镜像，即将指定的数据包复制到用户指定的目的地，以进行网络检测和故障排除。

- 流量统计

流量统计用于统计指定业务流的数据包，它统计的是设备中转发的数据包中匹配已定义的复杂流分类规则的数据信息。

流量统计本身不是 QoS 控制措施，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

包括 EF、AF、WFQ 队列调度模式，流量整形（TS），WRED 等与队列相关机制的配置。

2.9.3 流策略

流策略是将分类器和流行为关联后形成的完整的QoS策略。可以根据具体的行为决定是将策略应用在接口的出方向或入方向上，比如流量监管既可以应用在出方向也可以应用在入方向，而流量整形只能应用在接口的出方向上。

2.9.4 流策略嵌套

流策略嵌套是指一个QoS策略中包含另一个QoS策略，如图所示，即父策略的行为（动作）是一个子策略。使用流策略嵌套时，对于命中流分类的某一类报文，除了执行父策略中定义的行为外，还由子策略再对该类流量进行分类，执行子策略中定义的行为。

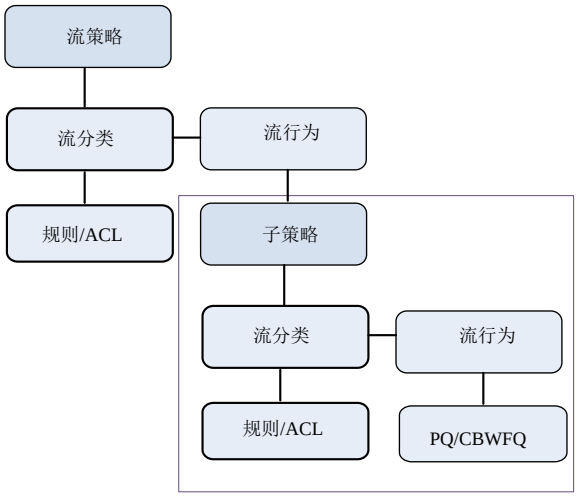


图 13　嵌套流策略

AR支持两层策略嵌套，子策略下面不能再有嵌套。

嵌套策略为HQoS提供了层次化的配置模型：在HQoS中，通过父策略区分网络中的不同用户，通过子策略区分用户的不同业务，从而提供区分用户和用户业务的精细化服务。

3 典型应用

3.1 AR QoS基本应用

AR部署在企业网的出口处，用以接入WAN侧网络。由于WAN链路带宽有限，需要对不同的业务提供差分服务，如减少语音报文的抖动和延时、保证重要业务的带宽等。企业内部LAN侧的不同业务流量进入AR设备时，先作流量分类和流量监管，再在WAN接口通过队列机制控制报文的优先发送顺序，把报文发送到WAN侧网络下游设备。

AR Qos应用的基本前提是根据组网环境、VPN/VLAN部署、应用业务分类等计算并划分出带宽模型和流量模型，根据相应的要求和服务承诺来进行相应的Qos优先级分类并标识、流量监管、流量整形、流量调度等功能。

AR Qos实现的服务模型包含传统Qos和HQos，对于传统Qos来讲，无论配置为基于物理组网环境(端口级别)的Qos，基于VPN/VLAN部署(用户级别)的Qos、基于应用业务(业务级别)的Qos都只能进行一次Qos的调度，多粒度同时配置的时候以粒度最细的Qos配置为最终调度的参数进行优先级分类并标识、流量监管、流量整形、流量调度。对于HQos来讲，可以同时配置端口级、用户级、业务级的流量调度模型，AR对流量的调度会按照各级调度的配置参数进行调度、整形、监管。

应用举例：国内宽带多业务通信网的部署，网络设计采用层次化设计。网络TOPO分为三层，总社为核心层，分社及社属企业等一级接入点为接入层，重点用户为第三层。

流量模型：

核心层：总社接入带宽为100M，各分社接入向运营商A、B接入带宽为均为30M。

接入层：接入带宽为30M，重点用户用vlan标识从运营商A接入，接入保证带宽为10M，峰值带宽为15M

运营商A承载业务：

语音业务：其IP Precedence = 5，要求保证40%带宽要求重标记 DSCP值为 40

视频业务: 其IP Precedence = 3，要求保证40%带宽，要求重标记 DSCP值为 20

VPN业务：其IP Precedence = 0，要求保证20%带宽

运营商B承载业务：

Internet业务:其IP Precedence = 0

重点用户：接入带宽为10M，峰值带宽为15M，重点用户业务分为两类：

视频业务: 其IP Precedence = 3，要求保证80%带宽 要求重标记 DSCP值为 40

Internet业务:其IP Precedence = 0，要求保证20%带宽

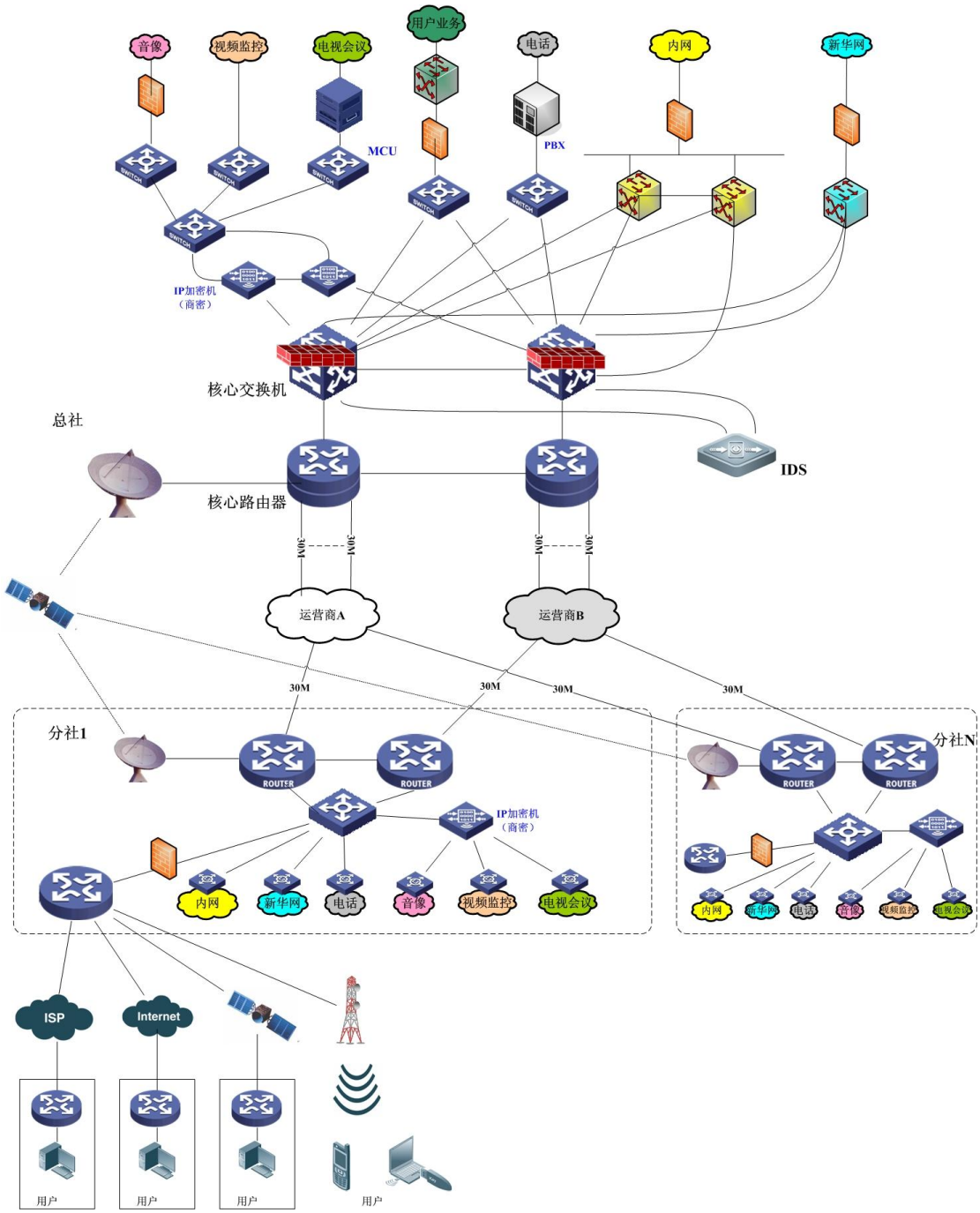


图 14　AR在企业网的QoS基本应用场景

1、组网环境分析：

由于租用了不同运营商的两条相同带宽的链路，为了充分利用链路资源，需要对经过广域网的流量进行分流。可将VPN、音像、语音三个优先级较高的业务由运营商A承载，并设置相应的优先级顺序，Internet等优先级略低的业务由运营商B承载。同样的对于租用不用的运营商端口，也可以进行相应的Qos策略配置，得到基于端口的Qos策略

核心路由器上基于端口配置：

运营商A、B对于总社接入端口配置：CIR:100M,

运营商A、B对于各分社接入端口配置：CIR:30M

2、VPN\VLAN部署分析

重点客户用vlan与普通的分社用户加以区分，使接入层能够对内部的用户进行相应的细分提供不同的带宽保证和Qos策略，得到用户级别的Qos策略。

接入层路由器上基于用户（带有vlan的接口）配置：

运营商A对于重点客户的vlan接入逻辑端口配置：CIR:10M PIR:15M

3、应用业务分类分析

各类接入业务根据不同的数据报文细分，设备可以通过DPI技术进行应用识别，为每类业务提供精细化的QoS控制，可以基于报文L4～L7内容对对流量进行智能分类，识别相关的保证业务和应用，根据识别结果结合QoS的流量标记、队列调度、整形等技术，来保障关键业务的高优先级和带宽，对非关键业务进行其他Qos限制。

接入层路由器上基于精细化Qos规则（Access Control List规则）配置：

运营商A 接入层业务：

匹配语音业务ACL规则：CIR:8M Remark DSCP:40

匹配视频业务ACL规则：CIR:8M Remark DSCP:20

匹配VPN业务ACL规则：CIR:4M

运营商A 重点客户业务：

匹配用户信息和视频业务ACL规则：CIR:8M Remark DSCP:40

匹配用户信息和Internet业务ACL规则：CIR:2M

运营商B 接入层业务：

匹配Internet业务ACL规则：CIR:30M

将抽象出来的各级调度的规则按照HQos调度模型即为

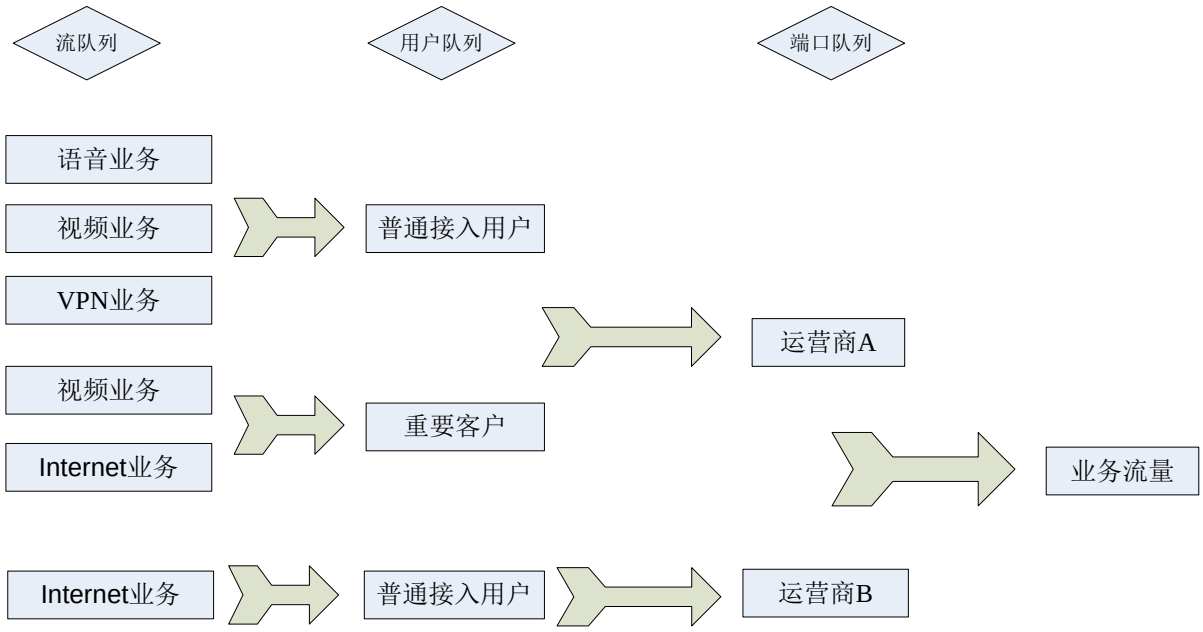


图 15　示例HQos业务分解示意图

按照以上分析，可以将举例模型中的流量，根据需求进行以上的配置，可以配置单次调度的Qos和多次调度的HQos,达到对流量规划和业务的精细化识别后为有不同服务需求的业务提供有区别的服务，正确地分配和使用资源。在进行资源分配和流量控制的过程中，尽可能地控制好那些可能引发网络拥塞的直接或间接因素，减少拥塞发生的概率；并在拥塞发生时，依据业务的性质及其需求特性权衡资源的分配，将拥塞对QoS的影响减到最小。

参考资料清单

[1] RFC 1633 Integrated Services in the Internet Architecture: an Overview
[2] RFC 2205 Resource Reservation Protocol (RSVP)-Version1 Functional Specification
[3] RFC 2210 The use of RSVP with IETF Integrated Services
[4] RFC 2211 Specification of the Controlled-Load Network Element Service
[5] RFC 2212 Specification of Guaranteed Quality of Service
[6] RFC 2215 General Characterization Parameters for Integrated Service Network Elements
[7] RFC 2474 Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6
[8] RFC 2475 An Architecture for Differentiated Services
[9] RFC 2597 Assured Forwarding PHB Group
[10]RFC 2598 An Expedited Forwarding PHB
[11]RFC 3260 New Terminology and Clarifications for Diffserv
[12]RFC 3246 An Expedited Forwarding PHB
[13]RFC 2697 A Single Rate Three Color Marker
[14]RFC 2698 A Two Rate Three Color Marker
[15]RFC 4594 Configuration Guidelines for DiffServ Service Classes

术语与缩略语

缩略语	英文全名	中文解释
ACL	Access Control List	访问控制列表
DPI	Deep Packet Inspection	深度报文解析
QoS	Quality of Service	服务质量
HQOS	Hierarchical QOS	层次化QOS
BA	Behavior Aggregate	行为聚合
PHB	Per Hop Behavior	每一跳行为
CAR	Committed Access Rate	承诺接入速率
GTS	Generic Traffic Shaping	通用流量整形
WRED	Weighted Random EarlyDetection	加权随机早期检测
FIFO	First In First Out	先进先出队列
SP	Strict Priority	严格优先级
WRR	Weighted Round Robin	加权轮循队列调度
DRR	Deficit Round Robin	赤字轮循队列调度
PQ	Priority Queue	优先级队列
WFQ	Weighted Fair Queue	加权公平队列调度
CBQ	Class-based WFQ	基于类的加权公平队列
RCPQ	Rate-Controlled Priority Queue	速率可控优先级队列