

华为智简园区交换机 QoS

技术白皮书



摘 要

随着网络技术的飞速发展，新业务对带宽、时延、抖动等传输性能有着特殊的需求，因此 QoS、HQoS 技术应运而生，从而为新业务提供端到端的服务质量保证。

目 录

摘 要.....	i
1 QoS 特性简介	4
1.1 什么是 QoS.....	4
1.2 QoS 度量指标.....	5
1.2.1 带宽/吞吐量.....	5
1.2.2 时延.....	6
1.2.3 时延变化（抖动）.....	6
1.2.4 丢包率.....	7
1.3 常见 QoS 业务指标.....	8
2 QoS 原理描述.....	10
2.1 QoS 服务模型.....	10
2.1.1 Best-Effort 服务模型.....	10
2.1.2 IntServ 服务模型.....	10
2.1.3 DiffServ 服务模型.....	11
2.1.4 DiffServ 模型与 IntServ 模型比较.....	15
2.1.5 基于 DiffServ 模型的 QoS 组成.....	16
2.2 流分类和标记.....	16
2.2.1 简单流分类.....	17
2.2.2 复杂流分类.....	23
2.2.3 流标记.....	27
2.3 流量监管和流量整形.....	31
3 流量监管概述	32
3.1 令牌桶工作原理.....	32
3.2 CAR.....	37
3.3 流量整形概述.....	45
3.3.2 流量监管和整形的比较.....	52

3.4 拥塞管理和拥塞避免	52
3.4.1 拥塞概述	52
3.4.2 拥塞管理	54
3.4.3 拥塞避免	61
4 HQoS 特性简介	67
4.1 产生背景	67
4.2 技术优势	67
5 HQoS 技术原理	68
5.1 基本原理	68
5.2 业务模型	69
6 应用场景	70
6.1 园区用户接入场景 HQoS 调度	70

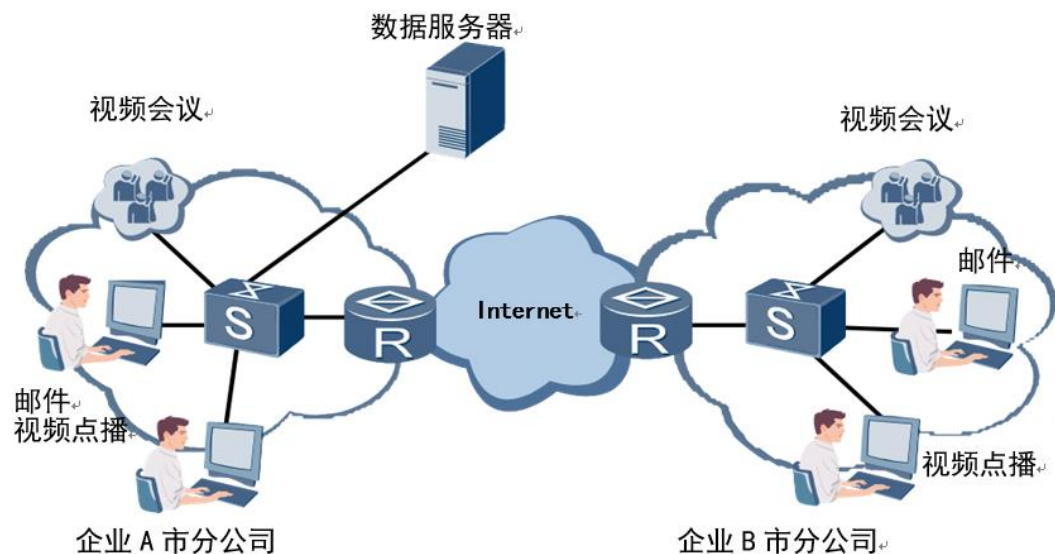
1 QoS 特性简介

1.1 什么是 QoS

随着网络技术的飞速发展，互联网中的业务越来越多样化。除了传统的 WWW、E-Mail、FTP 应用外，用户还尝试在 Internet 上拓展新业务，比如 IP 电话、电子商务、多媒体游戏、远程教学、远程医疗、可视电话、电视会议、视频点播、在线电影等。新兴的企业用户也有类似的诉求，除了基本的网页浏览外，在较集中的工作时间内还需要保证内部员工和外部访客的身份验证、异地的视频会议，大量的工作邮件，还有视频播放、FTP 文件上传下载，Telnet 特殊设备等业务。

这些新业务有一个共同特点，即对带宽、延迟、延迟抖动等传输性能有着特殊的需求，比如电视会议、视频点播等业务需要高带宽、低延迟和低延迟抖动。事务处理、Telnet 等关键任务虽然不一定要求高带宽，但要求低延迟，在拥塞发生时要求优先获得处理。

图1-1 企业网业务



网络的普及，业务的多样化，使互联网流量激增，产生网络拥塞，转发时延增加，严重时还会产生丢包，导致业务质量下降甚至不可用。所以，要在 IP 网络上开展这些实时性业务，就必须解决网络拥塞问题。

解决网络拥塞的最好的办法是增加网络的带宽。但从运营、维护的成本考虑，这是不现实的，同时增加带宽无法从根本上解决此问题，治标不治本。因此最有效的解决方案就是应用一个“有保证”的策略对网络拥塞进行管理。

QoS 技术就是在这种背景下发展起来的。QoS 是 Quality of Service (服务质量) 的简称，其目的是针对各种业务的不同需求，为其提供端到端的服务质量保证。QoS 技术在当今的互联网中应用越来越多，其作用越来越重要，如果没有 QoS 技术，业务的服务质量就无法保证。

1.2 QoS 度量指标

QoS 采用如下参数来度量，为关键业务提供服务质量保证，使其获得可预期的服务水平。

- 带宽/吞吐量
- 时延
- 时延变化（抖动）
- 丢包率

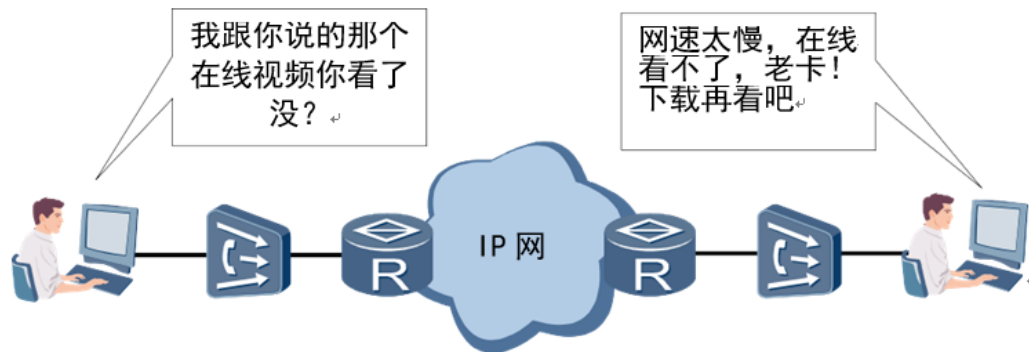
1.2.1 带宽/吞吐量

带宽 (bandwidth) 也称为吞吐量 (throughput)，是指在一个固定的时间内 (1 秒)，从网络一端传输到另一端的最大数据位数，也可以理解为网络的两个节点之间特定数据流的平均速率。带宽的单位是比特/秒 (bit/s，简称为 bps)。

带宽可以用城市的供水网做比喻来帮助理解它的含义：供水管道的直径可以衡量运水的能力。水管的直径好比是带宽，水就好比是网络传输的数据。使用粗管子就意味着拥有更宽的带宽，也就是有更大的数据传输能力。

在网络通信中，人们在使用网络时总是希望带宽越宽越好，特别是互联网功能日益强大，人们对互联网的需求不再是单一地浏览网页、查看新闻。新一代多媒体、影像传输、数据库、网络电视的信息量猛增使得带宽成为了严重的瓶颈。因此，带宽成为网络设计主要的设计点，也是分析网络运行情况的重要要素之一。

图1-2 带宽对网络的影响

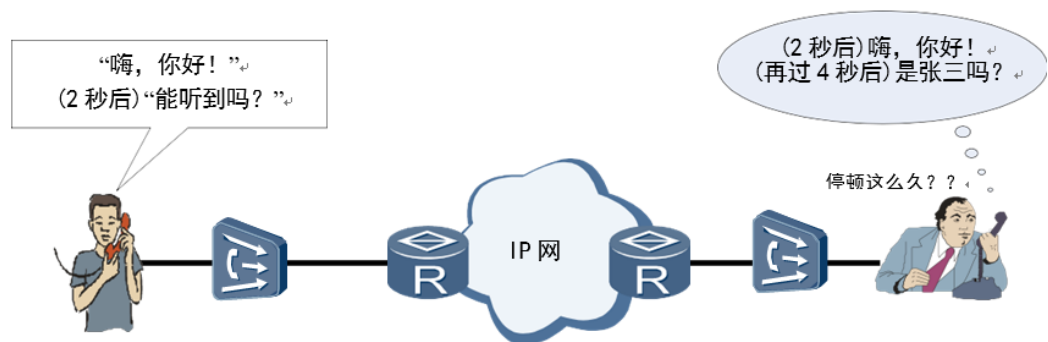


1.2.2 时延

时延（Latency）是指一个报文或分组从网络的一端传送到另一端所需要的时间。以语音传输为例，时延是指从说话者开始说话到对方听到所说内容的时间。若时延太大，会引起通话声音不清晰、不连贯或破碎。

大多数用户察觉不到小于 100 毫秒的延迟，当延迟在 100 毫秒和 300 毫秒之间时，说话者可以察觉到对方回复的轻微停顿，这种停顿可能会使通话双方都感觉到不舒服。超过 300 毫秒，延迟就会很明显，用户开始互相等待对方的回复，当通话的一方不能及时接收到期望的回复时，说话者可能会重复所说的话，这样会与远端延迟的回复碰撞，导致重复。

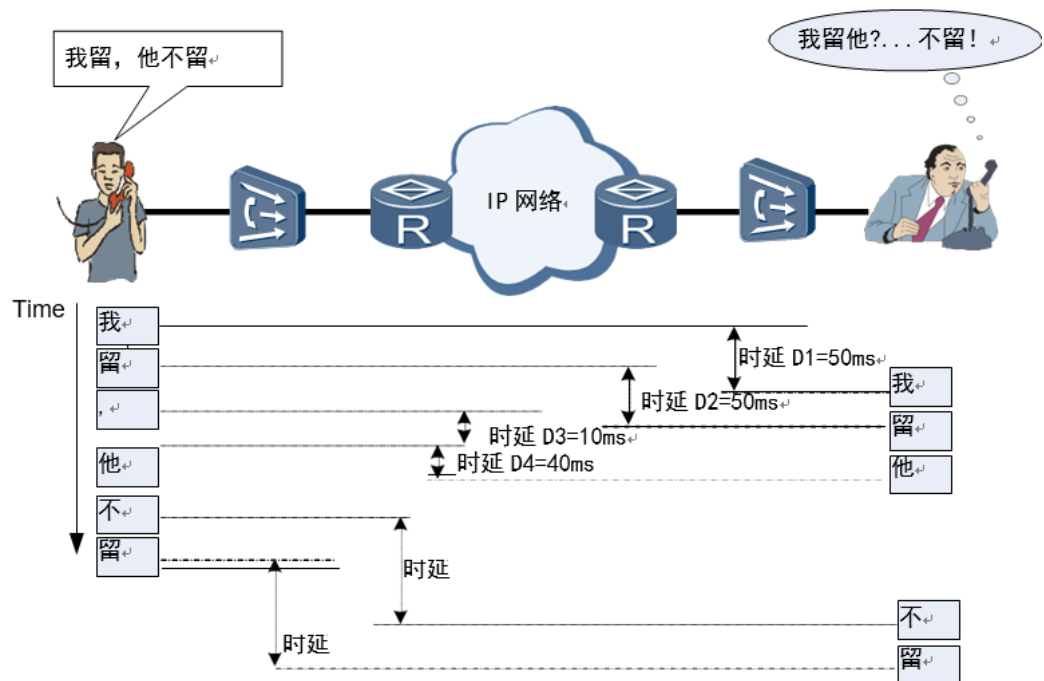
图1-3 时延对网络的影响



1.2.3 时延变化（抖动）

时延变化，也称为抖动（Jitter），是指同一业务流中不同分组所呈现的时延不同。抖动主要是由于业务流中相继分组的排队等候时间不同引起的，是对服务质量影响最大的一个问题。

某些业务类型，特别是语音和视频等实时业务是极不容忍抖动的。分组到达时间的差异将在语音或视频中造成断续。



抖动也会影响一些网络协议的处理，有些协议是按固定的时间间隔发送交互性报文，抖动过大会导致协议震荡。

所有传输系统都有抖动，只要抖动在规定容差之内就不会影响服务质量。利用缓存可以克服过量的抖动，但是这将会增加时延。

1.2.4 丢包率

少量的丢包（Loss）对业务的影响并不大，例如，在语音传输中，丢失一个比特或一个分组的信息，通话双方往往注意不到。在视频图像广播期间，丢失一个比特或一个分组可能造成在屏幕上瞬间的波形干扰，但视像很快恢复正常。使用传输控制协议（TCP）传送数据也能处理少量的丢包，因为传输控制协议允许丢失的信息重发。但大量的丢包会影响传输效率。所以，QoS 更关注的是丢包的统计数据——丢包率。丢包率是指在网络传输过程中丢失报文占传输报文的百分比。

图1-4 丢包率对网络的影响



1.3 常见 Qos 业务指标

在 IP 网络上不同的业务对带宽、时延、时延抖动和丢包率等都有不同的需求。表 1-1 和表 1-2 分别列出了几种常见业务对 QoS 从定性到定量的指标要求。表 1-3 载自 MEF 论坛，其按照业务重要性的分类，定量的分析了主要包括可用性、时延、抖动、丢包率和故障恢复时间 5 个方面的指标。

表1-1 几种常见业务的 QoS 指标

企业业务类型	带宽/吞吐量	时延	抖动	丢包率
视频电话会议	带宽需求高	对时延非常敏感	对抖动非常敏感	要求可预计的时延和丢包率
电子商务	带宽需求适当	对时延敏感	对抖动敏感	对丢包率敏感，必须可靠传输
流媒体	带宽需求高	对时延比较敏感	对抖动比较敏感	要求可预计的时延和丢包率
电子邮件、文件传输	带宽需求低	容许时延	容许抖动	尽力而为传送
HTML 网页浏览	带宽需求不定	容许适当时延	容许适当抖动	尽力而为传送
客户端/服务器 (FTP)	带宽需求适当	对时延敏感	对抖动敏感	对丢包率敏感，必须可靠传输

表1-2 几种常见业务的 QoS 定量指标

企业业务类别	时延	抖动	丢包
视频电话会议	≤50ms	≤10ms	≤0.1%
电子商务	≤200ms	≤100ms	TCP 保证

流媒体	≤1s	≤200ms	≤0.1%
电子邮件、文件传输	NA	NA	TCP 保证
HTML 网页浏览	NA	NA	NA
客户端/服务器（FTP）	NA	NA	TCP 保证

表1-3 MEF 论坛关于 QoS 定量指标

Service Class	Service Characteristics	Service Performance
Premium	Real-time IP telephony or IP video applications	Availability>99.99 % Delay<40ms Jitter<1ms Loss<0.1% Restoration time: 50ms
Silver	Bursty mission critical data applications requiring low loss and delay (eg.,Storage)	Availability>99.99 % Delay<50ms Jitter=N/A Loss<0.1% Restoration time:200ms
Bronze	Bursty data applications requiring bandwidth assurances	Availability>99.90 % Delay<500ms Jitter=N/A Loss NA Restoration time:2s
Standard	Best effort service	Availability>97.00 % Delay=N/A Jitter=N/A Loss NA Restoration time:5s

2 QoS 原理描述

2.1 QoS 服务模型

网络应用都是端到端的通信，两个主机进行通信，中间可能要跨越多个物理网络，经过多个交换机，因此要实现端到端的 QoS，就必须从全局考虑。QoS 的服务模型就是研究采用什么模式实现全局的服务质量保证。

QoS 有如下三种服务模型：

- 尽力而为（Best-Effort）服务模型
- 综合服务（Integrated Service）模型，简称 IntServ 模型
- 差分服务（Differentiated Service）模型，简称 DiffServ 模型

2.1.1 Best-Effort 服务模型

Best-Effort 是最简单的 QoS 服务模型，应用程序可以在任何时候，发出任意数量的报文，而且不需要通知网络。对 Best-Effort 服务，网络尽最大的可能性来发送报文，但对时延、可靠性等性能不提供任何保证。

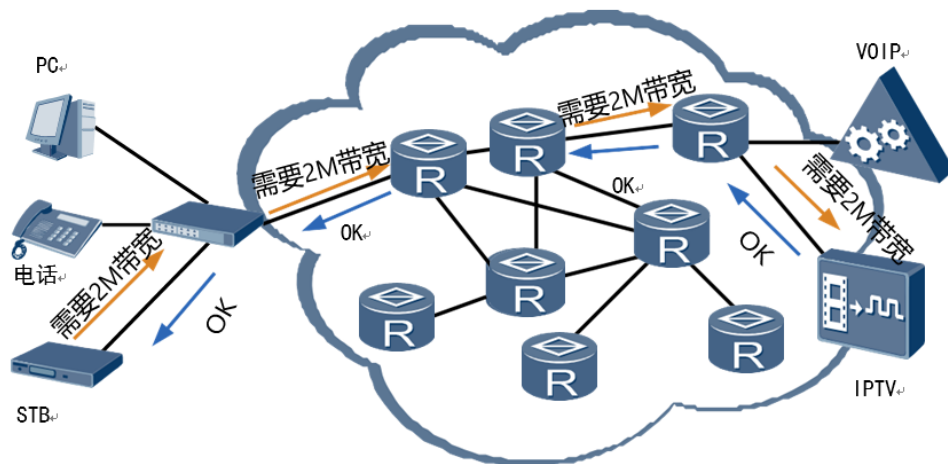
Best-Effort 服务模型适用于对时延、可靠性等性能要求不高的业务进行质量保证，是现在 Internet 的缺省服务模型，它适用于绝大多数网络应用，如 FTP、E-Mail 等。

2.1.2 IntServ 服务模型

IntServ 模型是指应用程序在发送报文前，需要通过信令（signaling）向网络描述它的流量参数，申请特定的 QoS 服务。网络根据流量参数的描述，预留资源以承诺满足该请求。在收到确认信息，确定网络已经为这个应用程序的报文预留了资源后，应用程序才开始发送报文。应用程序发送的报文应该控制在流量参数描述的范围内。网络节点需要为每个流维护一个状态，并基于这个状态执行相应的 QoS 动作，来满足对应用程序的承诺。

IntServ 模型原理类似于 MPLS-TE 技术（或者说 MPLS-TE 技术参考了 IntServ 模型更加合适），都使用了 RSVP（Resource Reservation Protocol）协议作为信令，在一条已知路径的网络拓扑上预先预留带宽、优先级等资源，路径沿途的各网元必须为每个要求服务质量保证的数据流预留想要的资源，这种资源预留的状态称为“软状态”。“软状态”是一种临时性状态，被定期的 RSVP 信息更新。通过 RSVP 信息的预留，各网元可以判断是否有足够的资源可以预留。只有所有的网元都给 RSVP 提供了足够的资源，“路径”方可建立。

图2-1 IntServ 服务模型



IntServ 模型为业务提供了一套端到端的保障制度，其优点显而易见，但是其局限性一样明显：

- MPLS-TE 的可行是因为其部署在核心的网路中，网络规模可控；而 IntServ 模型的对象是具体的端到端业务，其涉及的网络包含了核心层、汇聚层和接入层，包含的网元与 MPLS-TE 相比更是要多得多，复杂的网络限制了其发展。
- IntServ 模型要求端到端所有网络节点支持，而核心层、汇聚层和接入层的设备功能参差不齐，很难要求在这方面做到统一。

因此 IntServ 模型在 Internet 骨干网上无法得到广泛应用。

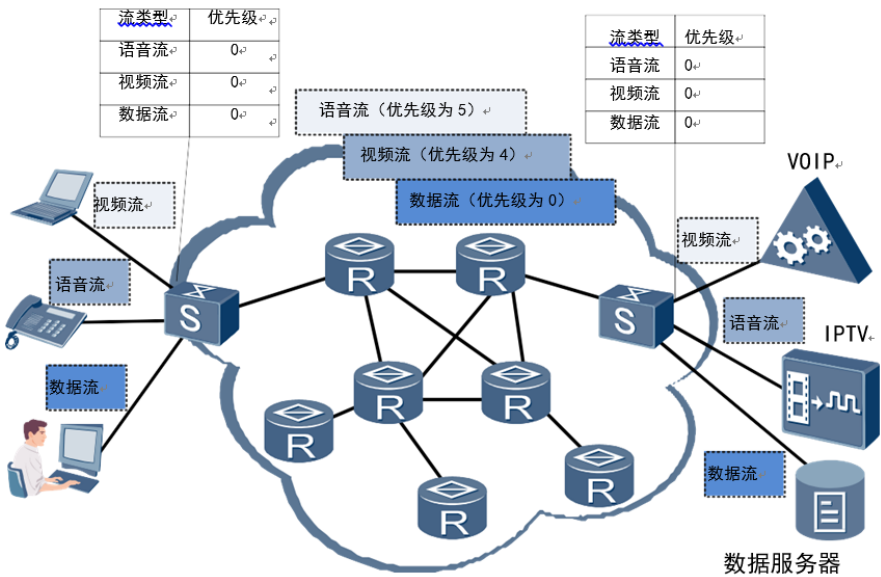
2.1.3 DiffServ 服务模型

DiffServ 模型的基本原理是将网络中的流量分成多个类，每个类享受不同的处理，尤其是网络出现拥塞时不同的类会享受不同的优先级处理，从而得到不同的丢包率、时延以及时延抖动。同一类的业务在网络中会被聚合起来统一发送，保证相同的延迟、抖动、丢包率等 QoS 指标。

Diffserv 模型中，业务流分类和汇聚工作在网络边缘由边缘节点完成。边缘节点可以通过多种条件（比如报文的源地址和目的地址、ToS 域中的优先级、协议类型等）灵活地对报文进行分类，对不同的报文设置不同的标记字段，而其他节点只需要简单地识别报

文中的这些标记，就可以进行资源分配和流量控制。因此，DiffServ 是一种基于报文流的 QoS 模型。

图2-2 DiffServ 服务模型



与 Intserv 模型相比，DiffServ 模型不需要信令。在 DiffServ 模型中，应用程序发出报文前，不需要预先向网络提出资源申请，而是通过设置报文的 QoS 参数信息，来告知网络节点它的 QoS 需求。网络不需要为每个流维护状态，而是根据每个报文流指定的 QoS 参数信息来提供服务，对报文的服务等级划分，有差别地进行流量控制和转发，提供端到端的 QoS 保证。

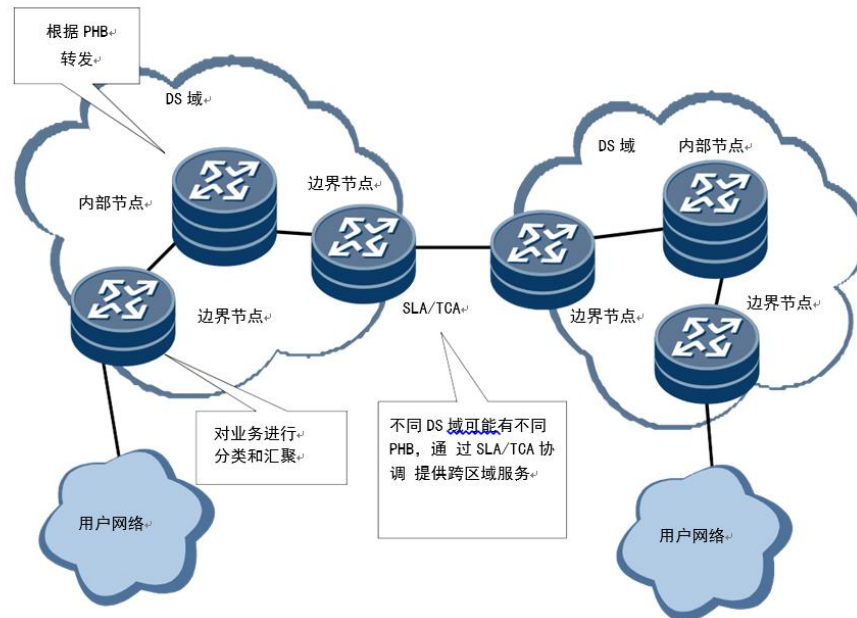
DiffServ 的基本思想是在网络边缘将进入网络的流分成各种不同的类型，将同种类型的流合并起来进行统一管理，保证相同的传输速率、延迟、抖动等服务质量参数，并对每一种类型的流在网络中分别进行处理。

业务分类和汇聚工作在网络边缘节点进行，首先数据包被标识为一定的服务类型，并记录在包头字段里，然后将数据包按一定的流量控制策略送入网络。网络中心节点通过检查包头来确定对包进行何种处理。

DiffServ 模型充分考虑了 IP 网络本身灵活性、可扩展性强的特点，将复杂的服务质量保证通过报文自身携带的信息转换为单跳行为，从而大大减少了信令的工作。因此，DiffServ 模型不但适合运营商环境使用，而且也大大加快了 QoS 在实际网络中应用的进程，是当前网络中的主流服务模型。

模型角色概念介绍

图2-3 DiffServ 服务模型概念



- DS 节点：DiffServ 功能的网络节点称为 DS 节点，图 2-3 中能看到的网元都是 DS 节点。
- DS 域（DS Domain）：一组对相同的流分类采用相同的服务提供策略和实现了相同 PHB（Per Hop Behaviors）的相连 DS 节点组成。一个 DS 域由相同管理部门的一个或多个网络组成，如一个 DS 域可以是一个 ISP，也可以是一个企业的内部网。
- DS 边界节点：负责连接另一个 DS 域或者连接一个没有 DS 功能的域。DS 边界节点负责将进入此 DS 域的业务流进行分类和可能的流量调整。
- DS 内部节点：用于在同一个 DS 域中连接 DS 边界节点和其他内部节点。DS 内部节点仅需基于 DSCP 值进行简单的流分类以及对相应的流实施流量控制。
- SLA/TCA：SLA 指用户（个人、企业、有业务往来的相邻 ISP 等）和服务提供商签署的关于业务流在网络中传递时所应当获得的待遇。SLA 包括很多方面，例如付费协议，其中的技术说明部分称为服务等级规范 SLS（Service Level Specification）。SLS 的研究重点是流量控制说明 TCS（Traffic Conditioning Specification），它描述了每个服务层次的详细性能参数，如平均速率、峰值速率、承诺突发尺寸、最大突发尺寸等，是 DiffServ 网络进行流控的主要依据。
- DS 区：一个或多个邻接的 DS 域统称为 DS 区。同一 DS 区中的不同 DS 域可有不同的 PHB，以实现不同的服务提供策略，它们之间通过 SLA（Service Level Agreements）和 TCA（Traffic Conditioning Agreement）协调提供跨区域服务。

SLA/TCA 指明了如何在 DS 域边界节点调整从一个 DS 域传向另一个 DS 域的业务流。

PHB（Per Hop Behaviors）

在每一个 DS 节点上对分组的处理称为每跳行为 PHB（Per-Hop Behavior）。PHB 描述了 DS 节点对具有明确流分类的分组采用的外部可见的转发行为。可以用优先级来定义 PHB，也可以用一些可见的服务特征如分组延迟、抖动或丢包率来定义。PHB 只定义了一些外部可见的转发行为，没有指定特定的实现方式。

RFC 定义了四种标准的 PHB：类选择码 CS（Class Selector），加速转发 EF（Expedited Forwarding），确保转发 AF（Assured Forwarding）和尽力而为 BE（Best-Effort）。其中，BE 是缺省的 PHB。

在 RFC2597 中 AF 又被划分为四个等级，即为 AF₁~AF₄；在 RFC2474 中 CS 又被划分为两个等级，即 CS₆ 和 CS₇。至此，PHB 有了 8 个细分级别，每个 PHB 在设备内部对应不同内部服务等级（Class of Service，简称 CoS），不同的服务等级将决定不同流的拥塞管理策略；同时每个 PHB 又再被划分为三个丢弃优先级（也叫做颜色 Color），分别用 Green、Yellow 和 Red 表示，不同的丢弃优先级将决定不同流的拥塞避免策略。

关于 CoS 和 Color 的详细说明请参加 2.2.1 节，关于颜色拥塞管理和拥塞避免的详细说明请参见 2.4 节。

关于各 PHB 的含义及细分级别对应的用途如表 2-1 所示。

表2-1 标准 PHB 的含义和用途

PHB	含义	细分 PHB	用途
CS(RFC 2474)	CS 表示类选择码, 代表的服务等级与在现有网络中使用的 IPPrecedence 相同。在所有标准 PHB 中, CS 的优先级最高。	CS ₇	CS ₆ 和 CS ₇ 默认用于协议报文, 比如说企业内部各个交换机之间的 STP 报文、LLDP 报文, LACP 报文等。如果这些报文无法接收会引起协议中断。
		CS ₆	

EF(RFC2598)	EF 被定义为这样的一种转发处理：从任何 DS 节点发出的信息流速率在任何情况下必须获得等于或大于设定的速率。EFPHB 在 DS 域内不能被重新标记，仅允许在边界节点重新标记。 EF 流要求低时延、低抖动、低丢包率，对应于实际应用中的视	-	EF 用于承载员工 VoIP 语音的流量，或者企业内部视频会议的数据流，因为语音要求低延迟，低抖动，低丢包率，是仅次于协议报文的最重要的报文。
AF(RFC2597)	AF 的推出是为了满足这样的需求：用户在与 ISP 订购带宽服务时，允许业务量超出所订购的规格。对不超出所订购规格的流量要求确保转发的质量；对超出规格的流量将降低服务待遇继续转发，而不只是简单地被丢弃。 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	AF4	AF4 用来承载语音的信令流量。即 VoIP 业务的协议报文。
		AF3	AF3 可以用作远端设备的 Telnet，FTP 等服务。这些业务对带宽要求适当，但对网络时延、抖动都非常敏感，同时要求完全可靠的传输（不能丢包）。
		AF2	AF2 可以用来承载企业内部 IPTV 的直播流量，可以保证员工在线视频业务流畅。直播的实时性强，需要有连续性和大吞吐量的保证，但是允许小规模丢包。
		AF1	作企业内部普通数据流业务，例如 E-Mail。普通数据对实时性和抖动等因素要求都不高，只要保证不丢包的传达即可。
BE(RFC2474)	对应于传统的 IP 分组投递服务，只关注可达性，其他方面不做任何要求。任何交换机必须支持 BE PHB。	-	企业内部尽力而为的服务，用作那些不紧急，不重要，不需要负责的业务，比如员工 HTTP 网页浏览业务。

2.1.4 DiffServ 模型与 IntServ 模型比较

表2-2 DiffServ 模型和 IntServ 模型比较

场景	DiffServ 服务模型	IntServ 服务模型
----	---------------	--------------

端到端保障	DiffServ 模型通过多个 DS 域之间接力，间接实现端到端的 QoS 保障。	直接实现端到端 QoS 保障。
网络规模	对网络规模不敏感，对于大规模网络可以通过多个 DS 域划分实现。	对网络规模敏感，过于庞大的网络规模不利于部署。
网络开销	通过报文标志位通知其他设备报文优先级，对网络无额外开销。	通过 RSVP 信令通知其他设备，并且定期刷新网络资源情况，带外信令对网络有额外开销。
网元开销	网元无需预留资源，开销小。	网元需要提前预留资源，开销大。

2.1.5 基于 DiffServ 模型的 QoS 组成

本文介绍的 QoS 都是基于 Diffserv 服务模型的，那么基于 DiffServ 服务模型的 QoS 业务可以分为以下几大类：

- **流分类和标记 (Traffic classification and marking)**：要实现差分服务，需要首先将数据包分为不同的类别或者设置为不同的优先级。将数据包分为不同的类别，称为流分类，流分类并不修改原来的数据包。将数据包设置为不同的优先级称为标记，而标记会修改原来的数据包。
- **流量监管和整形 (Traffic Policing and Shaping)**：是指将业务流量限制在特定的带宽，当业务流量超过额定带宽时，超过的流量将被丢弃或缓存。其中，将超过的流量丢弃的技术称为流量监管，将超过的流量缓存的技术称为流量整形。
- **拥塞管理和避免 (Congestion Management and Avoidance)**：拥塞管理在网络发生拥塞时，将报文放入队列中缓存，并采取某种调度算法安排报文的转发次序。而拥塞避免可以监督网络资源的使用情况，当发现拥塞有加剧的趋势时采取主动丢弃报文的策略，通过调整流量来解除网络的过载。
- **端口镜像和流镜像 (Port Mirror and Traffic Mirror)**：镜像是将指定端口的指定报文复制一份到镜像目的端口，镜像目的端口会与数据监测设备相连，用户利用这些数据监测设备来分析复制到目的端口的报文，进行网络监控和故障排除。

其中，流分类和标记是实现差分服务的前提和基础；流量监管、流量整形、拥塞管理和拥塞避免从不同方面对网络流量及其分配的资源实施控制，是提供差分服务的具体体现。

2.2 流分类和标记

流分类是对进入 DiffServ 域的业务进行分类，以便在网络中得到相应的适当处理。流分类主要目的是让其他处理此报文的应用系统或设备知道该报文的类别，并根据这种类别对报文进行一些事先约定了的处理。

业务流进入 DiffServ 域时，可以有多种方法对它进行分类，例如根据报文所携带的 QoS 优先级位，识别出不同优先级特征的流量；或根据源地址、目的地址、MAC 地址、IP 协议或应用程序的端口号等信息对流进行分类，也可以根据业务等级协议 SLA 规定的一些策略给每个数据包加上标记，从而对数据包进行分类。

当报文在 DiffServ 域边界被分类之后，网络的中间节点就可以根据分类，对不同类别的流量给予差别服务。下游（downstream）节点可以选择使用上游（upstream）节点的分类结果，也可以按照自己的分类标准对数据流重新进行分类。

根据不同的方法实现流分类的技术可以被分成“简单流分类（Behavior Aggregate Classifier）”与“复杂流分类（Multi-Field Classifier）”，本文随后的两个章节将针对这两种流分类方式进行详细的说明。

2.2.1 简单流分类

简单流分类是指采用简单的规则，如只根据 IP 报文的 VLAN 报文的 802.1p 值，IP 报文的 ToS 值、IPv6 报文的 TC 值、MPLS 报文的 EXP 域值，对报文进行粗略的分类，以识别出具有不同优先级或服务等级特征的流量，实现外部优先级和内部优先级之间的映射。

简单流分类过程实际上就是信任端口的上行报文携带的优先级标记，并进行优先级映射，即根据优先级映射表，将上行报文携带的 QoS 优先级统一映射到设备内部的服务等级和颜色，将下行报文的内部的服务等级和颜色映射成为 QoS 优先级。

简单流分类主要在网络的 DS 域内部节点部署。

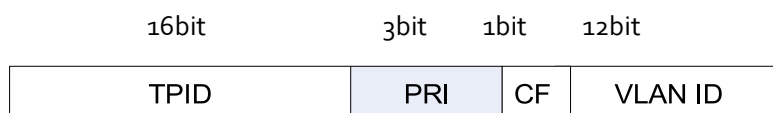
QoS 优先级分类

DiffServ 模型根据报文头中某些字段记录的 QoS 信息提供有差别的服务质量。与 QoS 相关的报文字段主要包括：

- 802.1p 字段

对于以太帧，根据 VLAN 帧头中的 802.1p（PRI）字段进行流分类，PRI 字段长为 3bit，可以表示 8 个传输优先级，按照优先级从高到低顺序取值为 7、6、.....、1 和 0，不同的优先级标识了不同等级的服务质量需求。

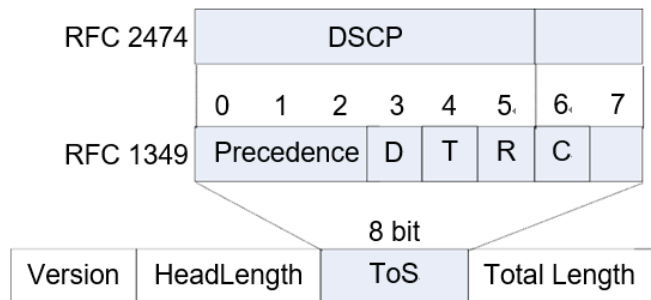
图2-4 以太报文的 802.1p 域



- IP 报文 ToS 字段

对于 IP 报文，基于 IP 包首部中的 ToS 域的前三位（即 IP Precedence）或前 6 位（即 DSCP 域）来标记报文。使用 IP 优先级可以将报文最多分成 8 类；使用 DSCP 域可将报文最多分成 64 类。

图2-5 IP 报文的 ToS 域



RFC1349 中定义的 ToS 域各比特位的含义：

- 比特 0~2 表示 Precedence 字段。代表报文传输的 8 个优先级，按照优先级从高到低顺序取值为 7、6、.....、1 和 0，与 802.1p 字段对应的报文优先级一一对应。
- D 比特表示延迟要求（Delay，0 代表正常延迟，1 代表低延迟）。
- T 比特表示吞吐量（Throughput，0 代表正常吞吐量，1 代表高吞吐量）。
- R 比特表示可靠性（Reliability，0 代表正常可靠性，1 代表高可靠性）。
- C 比特表示传输开销（Monetary Cost，0 代表正常传输开销，1 代表低传输开销）。
- 比特 6 和 7 保留。

RFC2474 则定义比特 0~6 表示 DSCP 域，其中前 3 比特是类选择代码点 CSCP（Class Selector Code Point），通过这 3 个比特位可以将 DSCP 域划分为 8 个优先级，按照优先级从高到低顺序取值为 7、6、.....、1 和 0，与 802.1p 字段对应的报文优先级一一对应。后 3 比特的含义在后文介绍中作用不大，此不详细描述。

● MPLS 的 EXP 字段

对于 MPLS 报文，则一般是根据 MPLS 报文中的 Exp 域进行流分类。Exp 域包括 3 位，通常作为 MPLS 报文的 CoS 域，表示报文传输的 8 个优先级，按照优先级从高到低顺序取值为 7、6、.....、1 和 0，与 IP 网络的 ToS 或 DSCP 字段对应的报文优先级一一对应。

图2-6 MPLS 报文的 Exp 域



802.1p、MPLS EXP、IP Precedence 字段与 DSCP 字段对应关系见表 2-3。

表2-3 802.1p、MPLS EXP、IP Precedence 字段与 DSCP 字段对应关系

IP Precedence	MPLS EXP	802.1p	DSCP
0	0	0	0
1	1	1	8
2	2	2	16
3	3	3	24
4	4	4	32
5	5	5	40
6	6	6	48
7	7	7	56

DSCP 字段与 802.1p、MPLS EXP、IP Precedence 字段对应关系见表 2-4。

表2-4 DSCP 字段与 802.1p、MPLS EXP、IP Precedence 字段对应关系

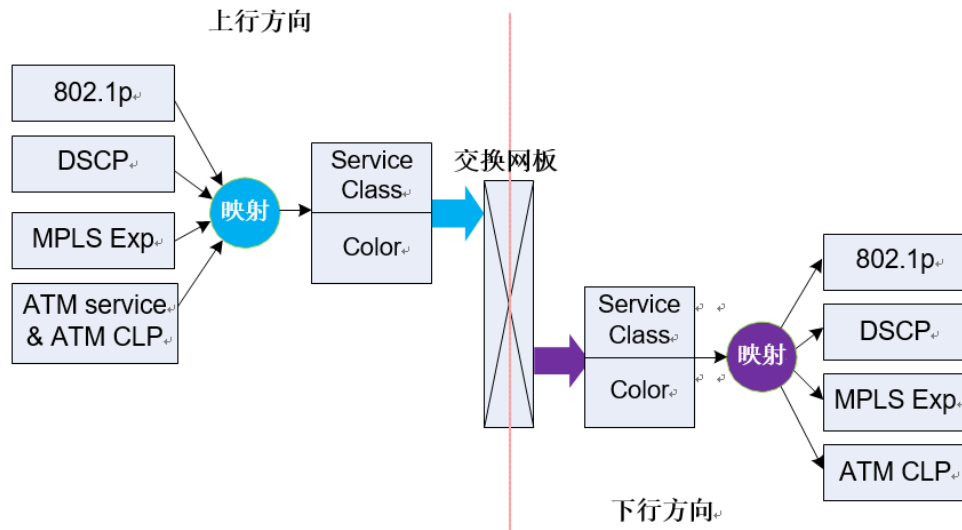
DSCP	IP Precedence	MPLS EXP	802.1p
0~7	0	0	0
8~15	1	1	1
16~23	2	2	2
24~31	3	3	3
32~39	4	4	4
40~47	5	5	5
48~55	6	6	6
56~63	7	7	7

QoS 优先级映射

不同的报文使用不同的 QoS 优先级，例如以太网报文使用 802.1p，IP 报文使用 DSCP，MPLS 报文使用 EXP。为了保证不同报文的服务质量，在报文进入设备时，需要将报文携带的 QoS 优先级统一映射到设备内部的服务等级 Class of Service（CoS，也叫做调度优先级 PHB）和丢弃优先级（也叫颜色 Color），在设备内部，根据报文的服务等级进行拥塞管理，根据报文的颜色进行拥塞避免；在报文出设备时，需要将内部的服务等级和颜色映射为 QoS 优先级，以便后续网络设备能够根据 QoS 优先级提供相应的服务质量。

将 QoS 优先级映射到服务等级和颜色是在报文上行方向进行，而服务等级和颜色映射为 QoS 优先级则是在下行方向进行，如图 2-7 所示。

图2-7 QoS 优先级映射



- CoS (Class of Service)

CoS 是指报文在设备内部的服务等级，支持 8 种取值，前文我们已经提到过了，优先级从高到低依次为 CS7、CS6、EF、AF₄、AF₃、AF₂、AF₁、BE。Class of Service 决定了报文在设备内部所属的队列类型。

服务等级的高低取决于具体的队列调度算法配置：

- 如果 8 种类型的队列都配置为 PQ 调度，则 CS7>CS6>EF>AF₄>AF₃>AF₂>AF₁>BE；
- 如果 BE 配置为 PQ 调度（当然一般不会这么配置），其余 7 种类型的队列配置为 WFQ 调度，则 BE 的优先级比其余 7 个都高；
- 如果 8 种类型的队列都配置成 WFQ 调度，则相互之间无优先级高低之分。

- Color

Color 是指报文在设备内部的丢弃优先级，用于实现同一个队列内部，当队列发生拥塞时报文丢弃顺序。Color 支持 3 种颜色划分（取值），前文已经提到过，IEEE 定义的优先级从低到高依次为 Green、Yellow、Red。

丢弃优先级的高低实际取决于对应参数的配置，例如：配置 Green 最大只能使用 50% 缓存，Red 最大可以使用 100% 缓存，则 Green 的丢弃优先级比 Red 高。所以并不是标记为 Red 的报文一定比标记为 Green 的报文丢弃优先级就高，优先级的高低完全取决于配置。

- 端口信任

在前文中介绍流分类时提到：“当报文在 DiffServ 域边界被分类之后，网络的中间节点可以根据分类结果对不同类别的流量给予差别服务。下游节点可以选择使用上游节点的分类结果，也可以按照自己的分类标准对数据流重新进行分类”。那么，如果选择使用上游节点的分类结果，则表示该节点信任上游节点的分类结果，即信任（trust）从连接上游节点的端口接收的报文所携带的 QoS 标记。因此，设备在实现 QoS 优先级映射时，可以选择信任端口的上行报文携带的优先级标记（包括 DSCP、IP Precedence、802.1p、MPLS EXP），这种模式称为端口信任模式。

目前交换机设备支持两种优先级信任模式：

➤ 信任报文的 802.1p 优先级

配置为信任 802.1p 优先级时，设备根据报文的 802.1p 优先级（对于 Untag 报文，设备使用端口优先级）对报文进行分类，并查找 802.1p 优先级到服务等级的映射表，为报文标记服务等级，以提供不同的服务质量。

➤ 信任报文的 DSCP 优先级

配置为信任 DSCP 优先级时，设备根据报文的 DSCP 优先级对报文进行分类，并查找 DSCP 优先级到服务等级的映射表，为报文标记服务等级，以提供不同的服务质量。

交换机设备根据优先级映射表实现 QoS 优先级映射。而在 DiffServ 模型中，不同 DS 域允许有不同的 PHB 映射关系，以实现不同的服务提供策略，因此设备需要允许管理员定义 DS 域并针对不同的 DS 域设定不同的优先级关系。

使用 DiffServ 域来实现 QoS 优先级映射时，具体的映射关系为包括：

- 以太报文 802.1p 优先级到 PHB 行为/颜色的映射关系请参见表 2-5。
- IP 报文 DSCP 优先级到 PHB 行为/颜色的映射关系请参见表 2-6。
- IP 报文 Precedence 优先级到 PHB 行为/颜色的映射关系请参见表 2-7。
- MPLS 报文的 EXP 优先级到 PHB 行为/颜色的映射关系请参见表 2-8。

表2-5 以太报文 802.1p 优先级到 PHB 行为/颜色的映射关系

802.1p 优先级	PHB	Color
0	BE	Green
1	AF1	Green
2	AF2	Green
3	AF3	Green
4	AF4	Green
5	EF	Green
6	CS6	Green

7	CS7	Green
---	-----	-------

表2-6 IP 报文 DSCP 优先级到 PHB 行为/颜色的映射关系

DSCP	PHB	Color	DSCP	PHB	Color
0~7	BE	Green	28	AF3	Yellow
8	AF1		29	BE	Green
9	BE		30	AF3	Red
10	AF1		31	BE	Green
11	BE		32	AF4	
12	AF1	Yellow	33	BE	
13	BE	Green	34	AF4	
14	AF1	Red	35	BE	
15	BE	Green	36	AF4	Yellow
16	AF2		37	BE	Green
17	BE		38	AF4	Red
18	AF2		39	BE	Green
19	BE		40	EF	
20	AF2	Yellow	41~45	BE	
21	BE	Green	46	EF	
22	AF2	Red	47	BE	
23	BE	Green	48	CS6	
24	AF3		49~55	BE	
25	BE		56	CS7	
26	AF3		57~63	BE	
27	BE				

表2-7 IP 报文 Precedence 优先级到 PHB 行为/颜色的映射关系

IP Precedence	PHB	Color
0	BE	Green
1	AF1	Green

2	AF2	Green
3	AF3	Green
4	AF4	Green
5	EF	Green
6	CS6	Green
7	CS7	Green

表2-8 MPLS 报文 EXP 优先级到 PHB 行为/颜色的映射关系

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

2.2.2 复杂流分类

随着网络的普及，网络中的业务越来越多样化，多种业务流共享同一网络资源，简单的流分类措施很难满足要求。这样，就要求网络具备很强的业务感知能力，能进行深度报文分析，对报文任意层次和字段的全面解析。复杂流分类可以在一定程度上满足此要求。

复杂流分类是指采用复杂的规则，如由报文的源 MAC、目的 MAC、内外层 Tag、源 IP 地址、源端口号、目的 IP 地址、目的端口号等对报文进行精细的分类。复杂流分类主要部署在网络的边缘节点。

华为交换机已实现多种复杂流分类方法和丰富的流行为，将这些流分类和对应可实施的流行为关联，形成流策略，并将流策略与指定接口、VLAN 或者全局绑定，可实现丰富的 QoS 流策略，这就是基于复杂流分类的 QoS 流策略（通常被称为“基于类的 QoS”）。

基于复杂流分类的 QoS 策略是对 QoS 策略配置的抽象，是“模板化”的 QoS 配置方式。“模板化”的最大优点是可以节省配置，支持批量修改。

流策略“模板”分为三部分：

- 流分类（Classifier）模板：定义流量类型。用 if-match 语句设定流分类的匹配规则。
- 流行为（Behavior）模板：定义针对该类流量可实施的流行为。
- 流策略（Policy）模板：在流策略模板中将流分类和流行为关联。当 Policy 模板设置完毕之后，需要将 Policy 模板应用到接口、VLAN 或者全局。

流分类（Classifier）

配置流分类可以将符合一定规则的报文分为一类，区分出用户流量，是实现差分服务的前提和基础。流分类各规则之间属于并列关系，只要匹配规则不冲突，都可以在同一流分类中配置。用户使用时，可以根据需要进行配置。

如果流分类有多个匹配规则，则这些规则之间有 And 和 Or 两种逻辑关系：

- Or 逻辑：数据包只要匹配该流分类下的任何一条 if-match 子句定义的规则就属于该类。
- And 逻辑：当流分类中有 ACL 规则时，数据包必须匹配其中一条 ACL 规则以及所有非 ACL 规则才属于该类；当流分类中没有 ACL 规则时，则报文必须匹配所有非 ACL 规则才属于该类。

可以选择一条或多条如下规则进行匹配以实现流分类，缺省的逻辑关系为 Or。

- 外层 VLAN ID
- QinQ 报文内外层 VLAN ID
- VLAN 报文 802.1p 优先级
- QinQ 报文内层 VLAN 的 802.1p 优先级
- 外层 VLAN ID 或基于 QinQ 报文内外两层 Tag 的 VLAN ID
- QinQ 报文双层 Tag
- 目的 MAC 地址
- 源 MAC 地址
- 以太网帧头中协议类型字段
- 所有报文
- IP 报文的 DSCP 优先级
- IP 报文的 IP 优先级
- 报文三层协议类型

- 入接口
- 出接口
- ACL 规则
- ACL 匹配顺序

流分类中，ACL 规则匹配过程为：

查找用户是否配置了该 ACL（因为流分类允许引用不存在的 ACL）。

规则显示顺序决定匹配顺序。规则匹配时，从 ACL 中显示的第一条规则开始查找，当找到一条符合匹配条件的规则就通知给业务，不再继续查找后续的规则。

访问控制规则可能会包含多个 rule 语句，而每个语句都指定不同的报文范围。这样，在匹配报文和访问控制规则时就会出现匹配顺序的问题。有两种匹配模式，在创建 ACL 时可以选择配置：

配置（config）模式：根据配置顺序匹配 ACL 规则。

自动（auto）模式：根据“深度优先”规则排序。

流行为（Behavior）

配置流行为即为符合流分类规则的流量指定后续动作，是配置流策略的前提条件。设备支持报文过滤、重标记、重定向、流量监管、流量统计等流行为，请根据实际需要选择配置。

表2-9 流行为在企业网中实际应用

流行为	含义	企业网实际应用
标记	设置/改写报文的优先级字段，如 VLAN 报文的 802.1p 优先级、IP 报文的 DSCP 和内部优先级，用于向下一台设备传递差分服务的 QoS 信息。其中，改写报文的优先级字段也称为“重标记”。	语音、视频和数据这些不同类型的业务，根据重要性不同在交换机上对其优先级做重标记，语音优先级最高，视频次之，数据最低。
流量监管	流量监管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量控制动作。通过配置流量监管，设备对符合流分类规则的报文的流量进行监督，对于超过规格的流量，可以采取丢弃、重标记颜色、重标记服务等级等动作。	企业网络中汇聚层交换机会连接多个接入交换机，当流量过大时会超过端口带宽，可以基于业务类型在入端口进行流量监管，对于超过规格的流量做响应处理。

流量统计	配置流量统计后，设备将对符合流分类规则的报文进行流量统计，可以帮助用户了解应用流策略后报文通过和被丢弃的情况，由此分析和判断流策略的应用是否合理，也有助于进行相关的故障诊断与排查。	企业网管常用功能，可以基于业务、基于用户监控网络中的流量，作为网络模型的分析依据。
报文过滤	最基本的安全手段。通过流分类，决定报文是被直接丢弃还是可以继续后面的转发处理。	企业网中主要有如下两种用法： 限制某些用户访问某些资源，可以通过过滤报文实现。 对于一些已知的黑名单中报文进行过滤，保护企业网络。
重定向	根据流分类决定报文的转发路径。通过配置重定向，设备将符合流分类规则的报文重定向到 CPU、指定的下一跳地址或指定接口。 包含重定向动作的流策略只能在全局、接口或 VLAN 的入方向上应用。	在出方向有链路备份的网络，可以使用指定下一跳的方式将高优先级的业务（如语音、视频等）重定向到带宽更宽或更稳定的链路上。
流镜像	设备复制一份和被观察流的原始报文一模一样的报文，并传送到指定的观察端口上。	维护手段的一种，可以通过流镜像收集某个端口某个时段进出的报文，用作问题故障分析。

流策略（Policy）

将绑定了流行为与流分类的完整流策略应用到全局、接口或者 VLAN 上，才能最终实现针对不同业务的差分服务。支持在创建流策略时指定流策略中流分类的匹配顺序，包括自动顺序（auto）和配置顺序（config）两种：

如果选择自动顺序，匹配顺序由系统预先指定的流分类优先级决定，该优先级由高到低依次为：基于二层和三层信息流分类>基于二层信息流分类>基于三层信息流分类。规则优先匹配优先级高的流分类。

如果选择配置顺序，匹配顺序由流分类在流策略中绑定的先后顺序决定。规则优先匹配绑定在先的流分类。

- 全局使用流策略

全局或 slot（框式或者盒式堆叠）的每个方向上能且只能应用一个流策略，如果在全局某方向应用了流策略，则不能在 slot 的该方向上再次应用流策略；指定 slot 在某方向应用流策略后，也不能在全局的该方向上再次应用流策略。

- 接口上使用流策略

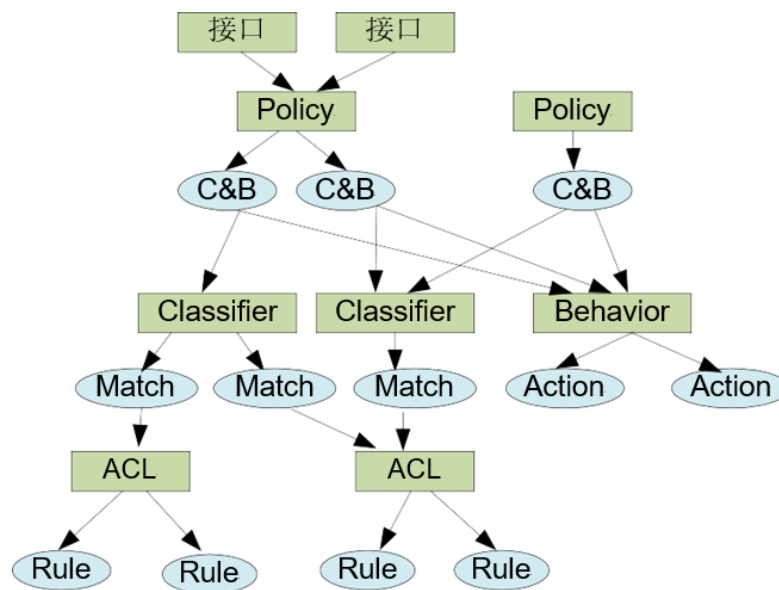
交换机每个接口的每个方向上能且只能应用一个流策略，但同一个流策略可以同时应用在不同接口的不同方向。应用后，系统对流经该接口并匹配流分类中规则的入方向或出方向报文实施策略控制。

- VLAN 上使用流策略

交换机每个 VLAN 的每个方向能且只能应用一个流策略。

以接口为例，流策略、流行为、流分类、ACL 之间的关系如图 2-8。

图2-8 接口、流策略、流行为、流分类、ACL 之间的关系



上图的关系可以总结为：

- 不同的接口可以应用相同的 Policy 模板。
- 一个 Policy 模板中可以配置一个或多个 Classifier&Behavior 对。不同的 Policy 模板可以应用相同的 Classifier&Behavior 对。
- 一个 Classifier 模板中可以配置一条或多条 if-match 语句，if-match 语句中可以引用 ACL 规则。不同的 Classifier 模板可以应用相同的 ACL 规则。一个 ACL 规则可以配置一个或多个 Rule 语句。
- 一个 Behavior 模板中可以配置一个或多个流行为。

2.2.3 流标记

设置/改写报文的优先级字段，用于向下一台设备传递差分服务的 QoS 信息的动作称为“流标记”，也称为“重标记”。2.2.1 节中详细说明了不同设备实现的报文原有优先级向内部优先级的映射关系，其流程是在端口入方向上实现的；本节关注的流标记动作则

是将报文的内部优先级映射到报文的优先级字段（修改报文的优先级字段），其流程实在端口出方向上实现的。

使用 DiffServ 域来实现流标记的优先级映射时，具体的映射关系为包括：

PHB 行为/颜色到以太报文 802.1p 优先级的映射关系请参见表 2-10。

PHB 行为/颜色到 IP 报文 DSCP 优先级的映射关系请参见表 2-11。

PHB 行为/颜色到 IP 报文 Precedence 优先级的映射关系请参见表 2-12。

PHB 行为/颜色到 MPLS 报文的 EXP 优先级的映射关系请参见表 2-13。

表2-10 PHB 行为&颜色到以太报文 802.1p 优先级的映射关系

PHB	Color	802.1p
BE	Green、Yellow、Red	0
AF1	Green、Yellow、Red	1
AF2	Green、Yellow、Red	2
AF3	Green、Yellow、Red	3
AF4	Green、Yellow、Red	4
EF	Green、Yellow、Red	5
CS6	Green、Yellow、Red	6
CS7	Green、Yellow、Red	7

表2-11 PHB 行为&颜色到 IP 报文 DSCP 优先级的映射关系

PHB	Color	DSCP
BE	Green、Yellow、Red	0
AF1	Green	10
AF1	Yellow	12
AF1	Red	14
AF2	Green	18
AF2	Yellow	20
AF2	Red	22

AF ₃	Green	26
AF ₃	Yellow	28
AF ₃	Red	30
AF ₄	Green	34
AF ₄	Yellow	36
AF ₄	Red	38
EF	Green、Yellow、Red	46
CS ₆	Green、Yellow、Red	48
CS ₇	Green、Yellow、Red	56

表2-12 PHB 行为&颜色到 IP 报文 Precedence 优先级的映射关系

PHB	Color	IP Precedence
BE	Green、Yellow、Red	0
AF ₁	Green、Yellow、Red	1
AF ₂	Green、Yellow、Red	2
AF ₃	Green、Yellow、Red	3
AF ₄	Green、Yellow、Red	4
EF	Green、Yellow、Red	5
CS ₆	Green、Yellow、Red	6
CS ₇	Green、Yellow、Red	7

表2-13 PHB 行为&颜色到 MPLS 报文的 EXP 优先级的映射关系

PHB	Color	MPLS EXP
BE	Green、Yellow、Red	0
AF ₁	Green、Yellow、Red	1
AF ₂	Green、Yellow、Red	2

AF ₃	Green、Yellow、Red	3
AF ₄	Green、Yellow、Red	4
EF	Green、Yellow、Red	5
CS ₆	Green、Yellow、Red	6
CS ₇	Green、Yellow、Red	7

使用 Map-Table 来实现流标记时，IP 报文 DSCP 字段到以太报文 802.1P 字段、IP 报文 DSCP 字段到 DSCP 字段优先级和丢弃优先级、以及 IP 报文 Precedence 字段到以太报文 802.1P 字段、IP 报文 Precedence 字段到 Precedence 字段优先级的映射关系，如表 2-14 和表 2-15 所示。

表2-14 DSCP 到 802.1p 和丢弃优先级的映射关系

Input DSCP	Output 802.1p	Output DP
0~7	0	0
8~15	1	0
16~23	2	0
24~31	3	0
32~39	4	0
40~47	5	0
48~55	6	0
56~63	7	0

表2-15 IP Precedence 到 802.1p 的映射关系

Input IP Precedence	Output 802.1p	Output Precedence
0	0	0
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5

6	6	6
7	7	7

2.3 流量监管和流量整形

流量监管和流量整形通过监督进入网络的流量速率，进而限制流量及其资源的使用，保证更好的为用户提供服务。

如果报文的发送速率大于接收速率，或者下游设备的接口速率小于上游设备的接口速率，就会引起网络的拥塞。如果不限制用户发送的业务流量，大量用户不断突发的业务数据会使网络更加拥挤。为了使有限的网络资源能够更好地发挥效用，更好地为更多的用户服务，必须对用户的业务流量加以限制。

流量监管和流量整形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策略。

3 流量监管概述

流量监管 TP（Traffic Policing）是指对进入设备的流量进行监控，确保其没有滥用网络资源。通过监控进入网络的某一流量的规格，限制它在一个允许的范围之内，若某个连接的报文流量过大，就对流量进行惩罚，比如丢弃报文，或重新设置该报文的优先级（比如限制 HTTP 不能占用超过 50% 的网络带宽），以保护网络资源和运营商的利益不受损害。

运营商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 CIR（Committed Information Rate）、峰值速率 PIR（Peak Information Rate）、承诺突发尺寸 CBS（Committed Burst Size）、峰值突发尺寸 PBS（Peak Burst Size）等流量参数，对超出 SLA 约定的流量报文可指定给予 pass（通过）、drop（直接丢弃）或 markdown（降级）等处理，此处降级是指降低服务等级（Class of Service），或者是提高丢弃等级（Color），即报文在网络拥塞时将被优先丢弃，从而保证在 SLA 约定范围内的报文享受到 SLA 预定的服务。

流量监管采用承诺访问速率 CAR（Committed Access Rate）来对流量进行控制。CAR 使用令牌桶算法进行流量速率的评估，依据评估结果，实施预先设定好的监管动作。对应于 SLA 预定的处理动作，流量监管动作包括：

- 转发（pass）：对测量结果不超过承诺速率（CIR）的报文通常处理为继续正常转发。
- 丢弃（discard）：对测量结果超过峰值速率（PIR）的报文通常进行丢弃。
- 重标记（remark）：对处于 CIR 与 PIR 之间的流量通常执行 Remark 动作，此时的报文不丢弃，而是通过 Remark 降低优先级进行尽力而为转发。

3.1 令牌桶工作原理

要实现流量的控制，必须有一种机制可以对通过设备的流量进行度量。令牌桶（Token-Bucket）是目前最常采用的一种流量测量方法，用来评估流量速率是否超过了

规定值。这里的令牌桶是指网络设备的内部存储池，而令牌则是指以给定速率填充令牌桶的虚拟信息包。

如图 3-1 所示，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。系统按设定的速度向桶中放置令牌，当桶中令牌满时，多余的令牌溢出。令牌桶只是一种流量测量方法，并不能对流量进行过滤或采取某种措施，比如说丢弃数据包等，这些操作由其他功能完成，而且令牌桶中装的是令牌而不是报文分组。

图3-1 令牌桶示意图

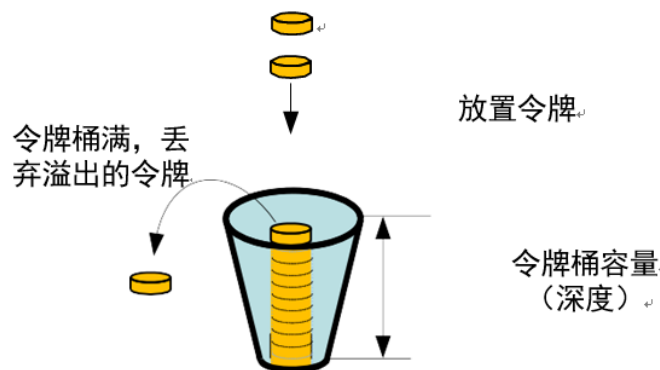
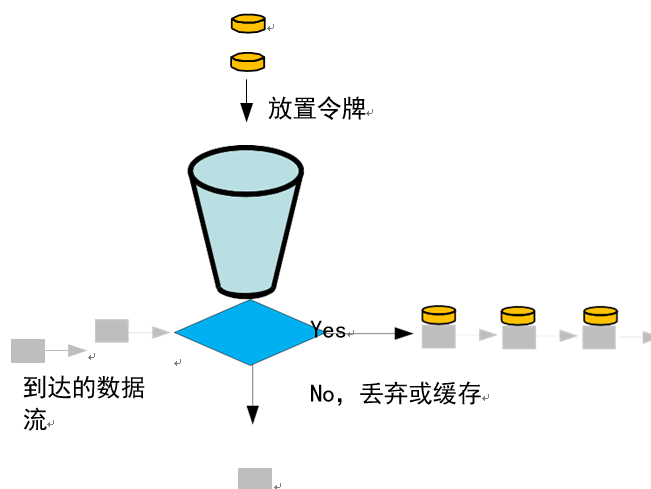


图3-2 使用令牌桶处理报文示意图



- 单速率三色标记（single rate three color marker, srTCM, RFC2697 定义，或称为单速双桶算法）算法，主要关注报文尺寸的突发。
- 双速率三色标记（two rate three color marker, trTCM, RFC2698 定义，或称为双速双桶算法）算法，主要关注速率的突发。

两种算法的评估结果都是为报文打上红、黄、绿三种颜色的标记，所以称为“三色标记”。QoS 会根据报文的颜色，设置报文的丢弃优先级，两种算法都可工作于色盲模式和非色盲模式。

单速双桶算法（srTCM）

- 单速双桶令牌桶参数

CIR（Committed Information Rate）：承诺信息速率，单位是 bit/s，表示向令牌桶中投放令牌的速率。

CBS（Committed Burst Size）：承诺突发尺寸，单位是 bit，用来定义在部分流量超过 CIR 之前的最大突发流量，即为令牌桶的容量（深度）。承诺突发尺寸必须大于报文的最大长度（最大时一个分组可以领取桶中的全部令牌）。CBS 越大，表示所允许的突发量越大。

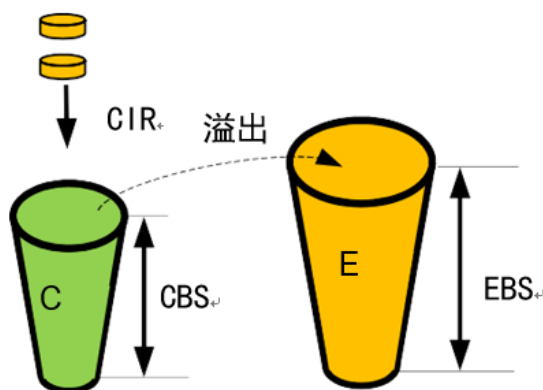
EBS（Extended burst size）：超额突发尺寸，用来定义在所有流量超过 CIR 之前的最大突发量。

- 单速双桶令牌桶结构

双桶结构由两个桶实现，为方便将两个令牌桶称为 C 桶和 E 桶。C 桶容量为 CBS，E 桶容量为 EBS，总容量是 CBS+EBS。如果不允许有突发流量，EBS 则设置成 0。

当 $EBS \neq 0$ 时，称为单速双桶。当 $EBS = 0$ ，E 桶的令牌数始终为 0，相当于只使用了一个令牌桶——C 桶，这种情况也称为单速单桶。

图3-3 单速双令牌桶示意图



- 单速双桶令牌添加方式

单速双桶令牌添加方式比较简单，先以 CIR 的速率往 C 桶中添加令牌，当 C 桶容量到达 CBS 后（C 桶满了），再以相同的速率往 E 桶中添加令牌（E 桶的令牌用做以后临时超过 CIR 的突发流量），当 E 桶容量到达 EBS 后（E 桶也满了），则新产生的令牌将会被丢弃。

初始状态下，C 桶和 E 桶都是满的。

- 单速双桶流量评估规则

当报文到来后，直接与桶中的令牌数相比较，如果有足够的令牌就转发（通常用一个令牌关联一个比特的转发权限），如果没有足够的令牌则丢弃或缓存。

为方便表示，用 T_c 和 T_e 表示桶中的令牌数量， T_c 和 T_e 初始化等于 CBS 和 EBS。色盲模式下，在对到达报文（假设报文大小为 B ）进行评估时，遵循以下规则：

- 对于单速单桶（ $EBS=0$ ）：

- 如果报文长度不超过 C 桶中的令牌数 T_c ，则报文被标记为绿色，且 $T_c=T_c-B$ ；
- 如果报文长度超过 C 桶中的令牌数 T_c ，报文被标记为红色， T_c 值不变。

- 对于单速双桶（ $EBS \neq 0$ ）：

- 如果报文长度不超过 C 桶中的令牌数 T_c ，则报文被标记为绿色，且 $T_c=T_c-B$ ；
- 如果报文长度超过 C 桶中的令牌数 T_c 但不超过 E 桶中的令牌数 T_e ，则报文被标记为黄色，且 $T_e=T_e-B$ ；

如果报文长度超过 E 桶中的令牌数 T_e ，报文被标记为红色，但 T_c 和 T_e 不变。色敏模式下，在对到达报文（假设报文大小为 B ）进行评估时，遵循以下规则：

- 对于单速单桶（ $EBS=0$ ）：

- 如果报文已被标记为绿色但报文长度不超过 C 桶中的令牌数 T_c ，则报文被标记为绿色，且 $T_c=T_c-B$ ；
- 如果报文已被标记为绿色且报文长度超过 C 桶中的令牌数 T_c ，则报文被标记为红色， T_c 保持不变；
- 如果报文已被标记为黄色或红色，都直接将报文标记为红色， T_c 保持不变。

- 对于单速双桶（ $EBS \neq 0$ ）：

- 如果报文已被标记为绿色且报文长度不超过 C 桶中的令牌数 T_c ，则报文被标记为绿色，且 $T_c=T_c-B$ ；
- 如果报文已被标记为绿色且报文长度超过 C 桶中的令牌数 T_c 但不超过 E 桶中的令牌数 T_e ，则报文被标记为黄色，且 $T_e=T_e-B$ ；

- 如果报文已被标记为黄色但报文长度不超过 E 桶中的令牌数 T_e ，则报文被标记为黄色，且 $T_e = T_e - B$ ；
- 如果报文已被标记为黄色且报文长度超过 E 桶中的令牌数 T_e ，则报文被标记为红色，且 T_e 保持不变；
- 如果报文已被标记为红色，直接将报文标记为红色， T_c 和 T_e 不变。

双速双桶算法（trTCM）

● 双速双桶令牌桶参数

CIR（Committed Information Rate）：承诺信息速率，表示端口允许的信息流平均速率，单位是 bit/s。

CBS（Committed Burst Size）：承诺突发尺寸，用来定义在部分流量超过 CIR 之前的最大突发流量，单位为 bit。承诺突发尺寸必须不小于报文的最大长度。

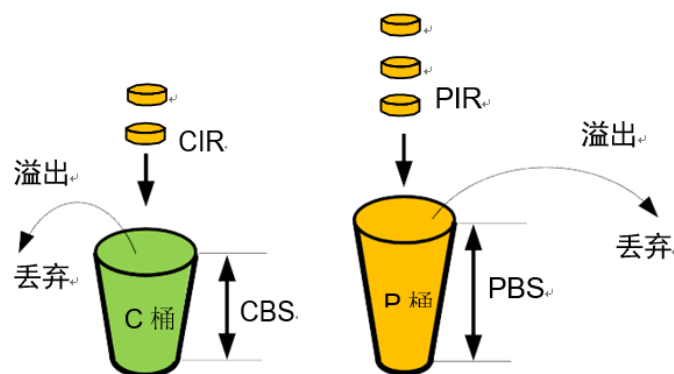
PIR（Peak Information Rate）：表示峰值信息速率，表示端口允许的突发流量的最大速率，单位是 bit/s。该值必须不小于 CIR 的设置值。

PBS（Peak Burst Size）：表示峰值突发尺寸，用来定义每次突发所允许的最大的流量尺寸。

● 双速双桶令牌桶结构

双速率三色标记算法业界都使用两个令牌桶，但它关注的是速率的突发，所以不像单速率三色标记算法那样把第一个桶中溢出的令牌放到第二个桶中，而是使用两个独立的令牌桶，存在两个令牌填充速率。为方便将两个令牌桶称为 C 桶和 P 桶，C 桶容量为 CBS，令牌填充速率为 CIR，P 桶容量为 PBS，令牌填充速率为 PIR。

图3-4 双速双桶示意图



● 双速双桶令牌添加方式

初始状态时 C 桶和 P 桶都是满的。往 C 桶和 P 桶分别以 CIR 和 PIR 的速率填充令牌。因这两个令牌桶是相互独立的，当其中一个桶被填满时，这个桶新产生的令牌将会被丢弃，而另一个桶则不受影响，继续填充令牌。

- 双速双桶流量评估规则

双速率三色标记算法关注的是速率的突发，首先评估的是数据流的速率是否符合规定的突发要求，其规则是先比较 P 桶，再比较 C 桶。

双速率三色标记算法也有色盲模式和色敏模式两种。为方便用 T_c 和 T_p 表示桶中的令牌数量， T_c 和 T_p 初始化等于 CBS 和 PBS。

色盲模式下，在对到达报文（假设数据包大小为 B）进行评估时，遵循以下规则：

- 如果报文长度超过 P 桶中的令牌数 T_p ，则报文被标记为红色，且 T_c 和 T_p 保持不变
- 如果报文长度不超过 P 桶中的令牌数 T_p 但超过 C 桶中的令牌数 T_c ，则报文被标记为黄色，且 $T_p = T_p - B$ ，
- 如果报文长度不超过 C 桶中的令牌数 T_c ，报文被标记为绿色，且 $T_p = T_p - B$ ， $T_c = T_c - B$ 。
- 色敏模式下，在对到达报文（假设报文大小为 B）进行评估时，遵循以下规则：
- 如果报文已被标记为绿色且报文长度超过 P 桶中的令牌数 T_p ，则报文被标记为红色，且 T_p 和 T_c 不变。
- 如果报文已被标记为绿色且报文长度不超过 P 桶中的令牌数 T_p 但超过 C 桶中的令牌数 T_c ，则报文被标记为黄色，且 $T_p = T_p - B$ ， T_c 不变。
- 如果报文已被标记为绿色且报文长度不超过 C 桶中的令牌数 T_c ，则报文被标记为绿色，且 $T_p = T_p - B$ ， $T_c = T_c - B$ 。
- 如果报文已被标记为黄色，则只比较 P 桶，如果报文长度超过 P 桶中的令牌数 T_p ，则报文被标记为红色，且 T_p 和 T_c 不变。
- 如果报文已被标记为黄色，且报文长度不超过 P 桶的令牌数，则报文被标记为黄色，且 $T_p = T_p - B$ ， T_c 不变。
- 如果报文已被标记为红色，直接将报文标记为红色， T_c 和 T_p 不变。

3.2 CAR

流量监管采用承诺访问速率 CAR（Committed Access Rate）来对流量进行控制。CAR 利用令牌桶来衡量每个数据报文是超过还是遵守所规定的报文速率。

CAR 主要有两个功能：

- 流量速率限制：通过使用令牌桶对流经端口的报文进行度量，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。
- 流分类：通过令牌桶算法对流量进行测量，根据测量结果给报文打上不同的流分类内部标记（包括服务等级与丢弃优先级）。

在华为交换机上，CAR 功能有两种实现方式：

- 基于接口的 CAR：如果不限制用户发送的流量，大量用户不断突发的数据会使网络更拥挤。为了使资源能够更好地发挥效用，可以通过配置流量监管对用户的流量加以限制，流量被限制在一个合理的范围之内，从而保护网络资源和用户的利益。
- 基于流的 CAR：若需要对接口出/入方向某类流量进行控制时，可以配置基于流的流量监管。相同的流策略可以在不同的接口下应用，当匹配流分类规则的报文的接收或发送速率超过限制速率时，直接被丢弃。基于流的流量监管，可以通过流分类，为不同业务提供更细致的差分服务。

华为交换机依据 RFC2697、RFC2698 实现 CAR 功能。CAR 中令牌添加方式是报文触发，添加令牌的数量是 $CIR \times$ 当前时间与上次添加令牌的时间之差。向桶内注入令牌后，再判断桶内的令牌数是否满足传送该报文的要求。

CAR 支持单速单桶、单速双桶、双速双桶的标记方式。本文举例介绍这三种标记的色盲模式如何处理报文（色敏模式的处理与之类似）。

单速单桶场景

假设设备端口的 CIR 设置为 1Mbps，CBS 为 2000bytes，初始状态时 C 桶满（单速单桶的 EBS 为 0，只有一个 C 桶）。

- 假设第 1 个到达的报文是 1500bytes 时，检查 C 桶发现令牌数大于数据包的长度，所以数据包被标为绿色，C 桶减少 1500bytes，还剩 500bytes。
- 假设 1ms 之后到达第 2 个报文 1500bytes，先填充令牌，新增令牌 = $CIR \times$ 时间间隔 = $1Mbps \times 1ms = 1000bit = 125bytes$ ，加上 C 桶原来剩余的令牌 500bytes，此时 C 桶共有 625bytes，令牌不够，报文标记为红色。
- 假设又过 1ms 后到达第 3 个报文 1000bytes，但 C 桶只有 625bytes，小于报文长度，因此新增令牌 $CIR \times 1ms = 1000bit = 125bytes$ ，此时 C 桶共有 750bytes，依然不够，因此报文被标记为红色。
- 假设又过 20ms 后到达第 4 个报文 1500bytes，但 C 桶只有 750bytes，小于报文长度，因此 C 桶新增令牌 $CIR \times 20ms = 2000bit = 250bytes$ ，C 桶此时令牌数 3250bytes，而 CBS = 2000bytes，因此溢出 1250bytes 令牌被丢弃，此时 C 桶大于报文长度，报文标记为绿色，C 桶减少 1500bytes，剩 500bytes。

以上过程汇总后请参见表 3-1。

表3-1 单速单桶的处理结果

包序号	时刻 (ms)	包长 (bytes)	与上次添加令牌的 间隔	本轮增加令牌	令 牌 增 加 后 C 桶令牌	报文处理 后 C 桶剩 余令牌	报文标 记结果
-	-	-	-	-	2000	2000	-
1	0	1500	0	0	2000	500	绿色
2	1	1500	1	125	625	625	红色
3	2	1000	1	125	750	750	红色
4	22	1500	20	2500	2000	500	绿色

单速双桶场景

假设设备端口的 CIR 设置为 1Mbps，CBS 为 2000bytes，EBS 为 2000bytes，初始状态时 C 桶和 E 桶满。

- 假设第 1 个到达的报文是 1500bytes 时，检查 C 桶发现令牌数大于数据包的长度，所以数据包被标为绿色，C 桶减少 1500bytes，还剩 500bytes，E 桶保持不变。
- 假设 1ms 之后到达第 2 个报文 1500bytes，但 C 桶只有 500bytes，小于报文长度，因此新增令牌 $CIR \times 1ms = 1000bit = 125bytes$ ，此时 C 桶共有 625bytes，依然不够。检查 E 桶有足够令牌，因此报文标记为黄色，E 桶减少 1500bytes，剩 500bytes，C 桶不变。
- 假设又过 1ms 后到达第 3 个报文 1000bytes，但 C 桶只有 625bytes，小于报文长度，因此新增令牌 $CIR \times 1ms = 1000bit = 125bytes$ ，此时 C 桶共有 750bytes，依然不够，检查 E 桶也不够，因此报文被标记为红色，C 桶、E 桶令牌数不变。
- 假设又过 20ms 后到达第 4 个报文 1500bytes，但 C 桶只有 750bytes，小于报文长度，因此 C 桶新增令牌 $CIR \times 20ms = 20000bit = 2500bytes$ ，C 桶此时令牌数 3250bytes，而 CBS=2000bytes，因此溢出 1250bytes 添加到 E 桶，此时 E 桶有 1750bytes。由于此时 C 桶大于报文长度，报文标记为绿色，C 桶减少 1500bytes，剩 500bytes，E 桶不变。

以上过程汇总后请参见表 3-2。

表3-2 单速双桶的处理结果

包序号	时刻 (ms)	包长 (byte s)	时间 间隔	本轮 增加 令牌	令牌增加 后各桶令 牌(C桶)	令牌增加 后各桶令 牌(E桶)	报文处理 后各桶剩 余令牌 (C桶)	报文处理 后各桶剩 余令牌 (E桶)	报 文 标 记 结 果
-	-	-	-	-	2000	2000	2000	2000	-

1	0	1500	0	0	2000	2000	500	2000	绿色
2	1	1500	1	125	625	2000	625	500	黄色
3	2	1000	1	125	750	500	750	500	红色
4	22	1500	20	2500	2000	1750	500	1750	绿色

双速双桶场景

假设设备端口的 CIR 设置为 1Mbps，PIR 设置为 2Mbps，CBS 为 2000bytes，PBS 为 2000bytes，初始状态时 C 桶和 P 桶满。

- 第 1 个到达的报文假设是 1500bytes 时，检查发现报文长度不超过 P 桶也不超过 C 桶，所以数据包被标为绿色，C 桶和 P 桶都减少 1500bytes，C 桶还剩 500bytes，P 桶还剩 500bytes。
- 假设 1ms 后到达第 2 个报文 1500bytes，超过 P 桶，因此 P 桶新增令牌 $PIR \times 1ms = 2000bit = 250bytes$ ，此时 P 桶共有 750bytes，依然小于报文长度。因此报文标记为红色，P 桶、C 桶令牌数不变。
- 假设又过 1ms 后到达第 3 个报文 1000bytes，超过 P 桶，因此 P 桶新增令牌 $PIR \times 1ms = 2000bit = 250bytes$ ，此时 P 桶共有 1000bytes，等于报文长度，再检查 C 桶，此时 C 桶 500bytes，小于令牌数，因此 C 桶新增令牌 $CIR \times 1ms = 1000bit = 125bytes$ ，此时 C 桶 625bytes，仍然小于报文长度，因此报文被标记为黄色，P 桶减少 1000bytes，剩 0bytes，C 桶不变。
- 假设又过 20ms 之后到达报文 1500bytes，但 P 桶没有令牌，不够发送报文，因此 P 桶新增令牌 $PIR \times 20ms = 40000bit = 5000bytes$ ，超过 P 桶容量 PBS，因此 $PBS = 2000bytes$ ，溢出的令牌丢弃；这样 P 桶有 2000bytes，大于报文长度，因此比较 C 桶，C 桶此时令牌数 625bytes，小于报文长度，因此 C 桶新增令牌 $CIR \times 20ms = 20000bit = 2500bytes$ ，大于 CBS 的 2000bytes，因此溢出的令牌丢弃，C 桶此时令牌数 2000bytes，大于报文长度，报文被标记为绿色，C 桶减少 1500bytes 还剩 500bytes，P 桶减少 1500bytes 还剩 500bytes。

以上过程汇总后请参见表 3-3。

表3-3 双速双桶的处理结果

包序号	时刻 (ms)	包长 (bytes)	时间间隔	本轮增加令牌	令牌增加后各桶	令牌增加后各桶令牌 (P 桶)	报文处理后各桶剩余令牌 (C 桶)	报文处理后各桶剩余令牌 (P 桶)	报文标记结果
-	-	-	-	-	2000	2000	2000	2000	-
1	0	1500	0	0	2000	2000	500	500	绿色
2	1	1500	1	125	500	750	500	750	红色

3	2	1000	1	125	625	1000	625	0	黄色
4	22	1500	20	2500	2000	2000	500	500	绿色

三种令牌桶的使用场景比较

单速单桶和单速双桶关注报文尺上的突发，其令牌添加方式和报文处理流程比较简单；双速双桶关注速率上的突发，其令牌添加方式和报文处理流程相对复杂。

单速和双速各有优点，不同的实现方式决定了其具有一定的性能差异（丢包率、突发流量处理性能、大小包混合转发性能、数据转发平缓程度等），在实际应用中，应针对不同的流量特征选择恰当的标记方式。

- 如果只是为了限制带宽，使用单速单桶。
- 如果在限制带宽的基础上，还要对输入流量的突发情况进行区分，做不同的标记处理，则使用单速双桶。注意：标记为 yellow 的动作一定要同标记为 green 的配置的不一样，否则限速效果与单速单桶一样。
- 如果在限制带宽的基础上，还要对输入流量的带宽情况进行区分，区分出带宽是小于 CIR 还是在 CIR~PIR 之间，则使用双速双桶。注意：标记为 yellow 的动作也要同标记为 green 的动作配置得不一样，否则限速效果与单速单桶一样。

表3-4 三种场景的比较

整形方	优点	缺点	适用场景
单速单桶	仅用于限制带宽，思路和配置简单	对超过单桶容量的突发流量没有任何的宽容余地。	对于优先级较低的业务，（比如访问外网的 Http 流量），对于超过额度的流量直接丢弃保证其他业务，不考虑突发。
单速双桶	除了带宽限制，还可以容许一部分流量突发，并且可以区分突发业务和正常业务。	思路较单速单桶复杂，需要考虑 E 桶的容量。	较为重要的业务，或者理解为容许有突发的业务（如公司邮件数据，邮件是 IT 较为重要的业务之一），对于突发流量有宽容。
双速双桶	最为细致的流量带宽划分，可以区别带宽小于 CIR 还是在 CIR~PIR 之间。	方案部署前需要充分考虑 CIR，CBS，PIR 和 PBS 的取值，并且根据不同业务做区分。	重要业务建议使用，可以更好的监控流量的突发程度，对流量分析起到指导作用。

CAR 参数设置

在令牌桶算法中，CIR 设置越大，令牌产生的速率越大，则分组获得令牌就越多，流向网络的流量也就越大，因此，CIR 的大小是控制流入网络中流量多少的关键。另外，令牌桶桶深 CBS 也是一个重要参数，随着令牌桶桶深的增加，C 桶中积累令牌的数目也将越多，流入网络的流量就越大。

对于承诺突发流量 CBS，不应该小于报文的最大长度。例如，当 CIR=100Mbps，CBS=200bytes，那么对于 1500byte/packet 的流量而言，每次与令牌桶比较，包长都大于 CBS，即使数据流速率小于 100Mbps，也全部被标记为红色或黄色，导致 CAR 效果不准。

对于超额突发流量 EBS，单位为 bit。在华为设备上，CBS 与 EBS 是由两个单独的令牌桶承载的，CBS 与 EBS 的大小没有什么必然联系，如果不允许有突发流量的话，EBS 只要设置成 0。要使令牌桶有扩展突发能力只要 EBS 设置的大于 0。

令牌桶算法的带宽参数设置取决于实际网络业务的限速需要。桶深是一个重要参数，具体该如何设置，则取决于具体的业务流量情况。原则上，桶深需要满足如下条件：

- 桶深 \geq MTU
- 桶深 \geq 业务流量的正常突发

对于条件 1，比较直观，容易操作；但对于条件 2，实际操作比较困难，因此出现了一些经验性的计算公式，对华为交换机总结的经验性公式为：

- 带宽 \leq 100Mbps 时，桶深(Bytes)=带宽(bps)*1000/8
- 带宽 $>$ 100Mbps 时，桶深(Bytes)=100,000(bps)*1000/8

假设一个对于语音流量占用带宽较小（最大 G.711 编码的按照 100Kbps 计算），下行接入 100 路 VoIP 电话的端口同时使用占用带宽为 10Mbps，根据以上计算公式，预留的桶深为 1250MBytes；对于占用带宽较大的视频带宽，以高清低交互视频占用带宽 2Mbps 计算，下行接入 10 路视频业务的端口同时占用带宽为 20Mbps，根据以上计算公式，预留的桶深为 2500MBytes。

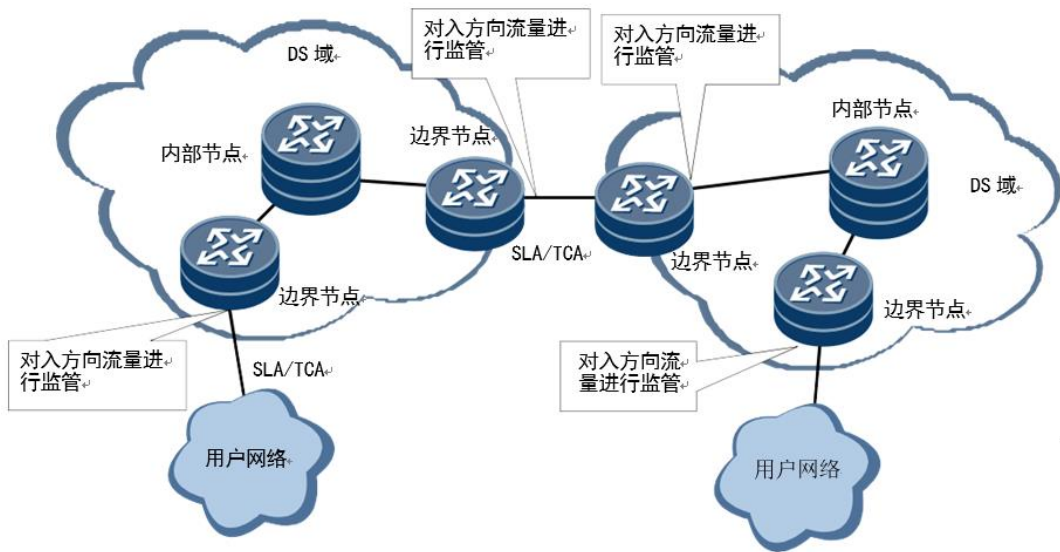
除了桶深，突发流量速率 PIR 的设置也是方案使用和命令配置中不可避免的，在工程上 PIR 的速率一般定义为 CIR 的 1.5 倍，过大的 PIR 会导致设备过高的负荷，反而适得其反。

如上例分析的那样，语音 CIR 一般设置为 100Kbps，那么对应的 PIR 为 150Kbps；视频业务 CIR 一般为 2Mbps，那么对应的 PIR 为 3Mbps。

流量监管的应用

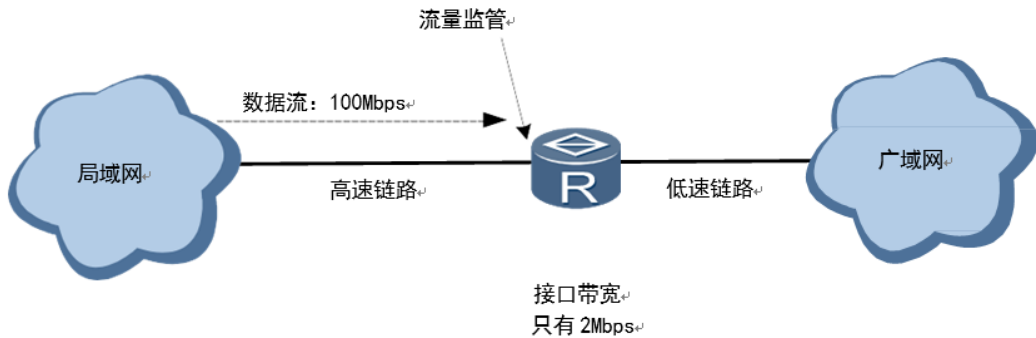
流量监管主要应用于网络边缘入口处，对超出 SLA 约定的流量报文给予通过、直接丢弃或降级处理，从而保证在 SLA 约定范围之内的报文享受到 SLA 预定的服务，同时保证核心设备的正常数据处理，如图 3-5 所示。

图3-5 流量监管的应用一



企业用户通过接入交换机连接广域网和企业内部局域网，局域网的带宽（100Mbps）通常比广域网（2Mbps 或更低）高。当局域网用户试图通过广域网发送大量数据时，在网络边缘就会发生拥塞。这种情况下，可以在网络边缘交换机的入口处进行流量监管，限制大流量数据的速率。如图 3-6 所示。

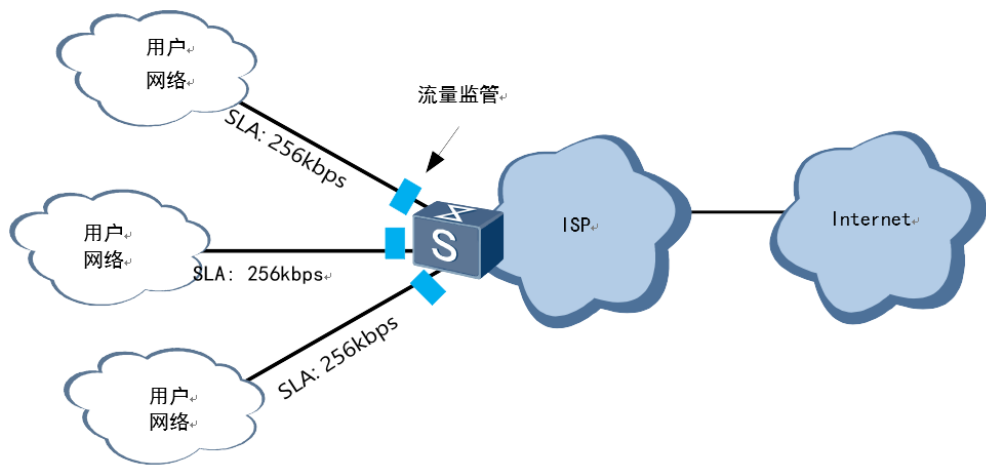
图3-6 流量监管的应用二



● 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。例如图 3-7，企业用户某台边缘交换机接入了三个部门网络。每个部门用户不能发送超过 256kbps 的流量，但有时他们可能会随意发送。此时，可以在边缘交换机的入接口上进行流量监管，将部门用户流量限制在 256Kbps 以下，超过 256Kbps 的流量将被丢弃。

图3-7 企业网络基于接口的流量监管



● 基于类的流量监管

基于类的 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。

例如图 3-8，假设某企业有 3 个部门用户（1.1.1.1、1.1.1.2、1.1.1.3）集合到一台交换机。根据规定，每个部门不能发送超过 256Kbps 的数据流，但有时他们可能会随意发送。当一个部门发送了大量的数据流时，可能会干涉到其他部门的数据流，即使这些部门按照 256kbps 甚至更低速率发送。此时，可以在交换机的入接口上基于源地址进行流分类和流量监管，监测进入路由器的速率，如果某个部门的速率超过 256Kbps，则丢弃该部门超过速率的分组。

图3-8 企业网络基于流的流量监管

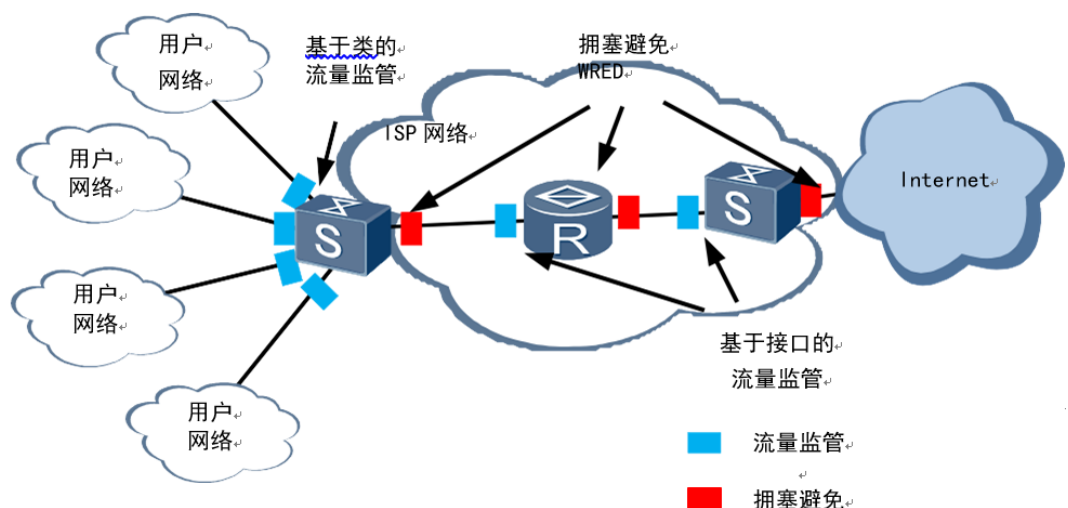


流量监管和其他 QoS 策略的配合

流量监管通常和拥塞避免、重标记等其他的 QoS 策略配合使用，共同提供全网 QoS 保障。

图 3-9 是流量监管和拥塞避免配合使用的场景。假设有 4 个用户网络集合到一台 ISP 网络的边缘交换机。根据 SLA 规定，每个用户网络都不能发送超过 256Kbps 的 FTP 数据流，但有时他们可随意发送，甚至会发送超过 1Mbps 的 FTP 数据流。当一个用户网络发送了大量的 FTP 数据流时，可能会干涉到某个其他用户网络的 FTP 数据流，即使这些用户网络按照 256kbps 甚至更低速率在发送。此种情况中，可以在每个入接口上配置基于类的流量监管，监测入接口的 FTP 的速率，并重标记报文的 DSCP 值。如果速率小于等于 256Kbps，分组被标记为 AF11，如果速率在 256Kbps 到 1Mbps 之间，分组被标记为 AF12，如果速率超过 1Mbps 的分组标记为 AF13。在 ISP 网络的其他节点出接口上对 AF 类流量配置 WRED（Weighted Random Early Detection）丢弃策略，避免网络拥塞。WRED 根据分组的 DSCP 标记丢弃该类分组。首先丢弃 AF13 分组，AF12 分组和 AF11 分组就会在最后被丢弃。

图3-9 流量监管和拥塞避免配合使用

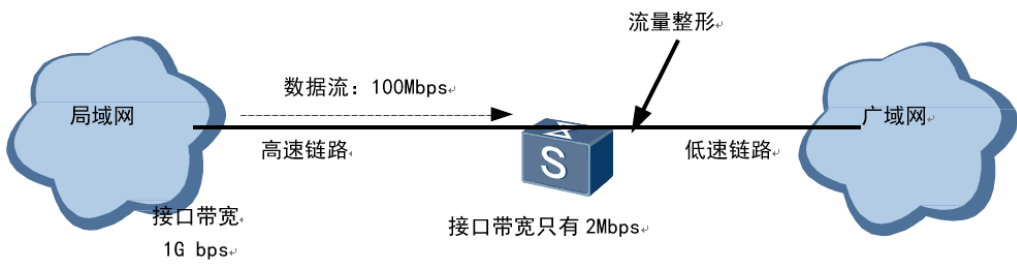


3.3 流量整形概述

流量整形是对输出报文的速率进行控制，使报文以均匀的速率发送出去。

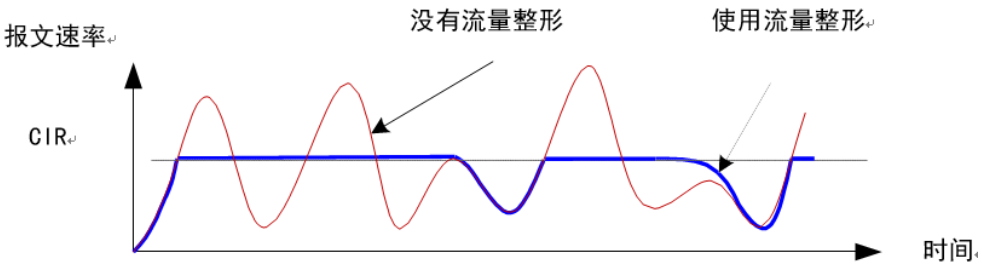
流量整形通常是为了使报文速率与下游设备相匹配。当从高速链路向低速链路传输数据，或发生突发流量时，带宽会在低速链路出口处出现瓶颈，导致数据丢失严重。这种情况下，需要在进入低速链路的设备出口处进行流量整形，如图 3-10。

图3-10 从高速链路向低速链路传输数据



可以通过在上游设备的接口出方向配置流量整形，将上游不规整的流量进行削峰填谷，输出一条比较平整的流量（如图 3-11），从而解决下游设备的瞬时拥塞问题。

图3-11 使用流量整形的效果



流量整形通常使用缓冲区和令牌桶来完成，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地发送这些被缓冲的报文。流量整形是在队列调度之后，数据包在出队列的过程中进行调度。

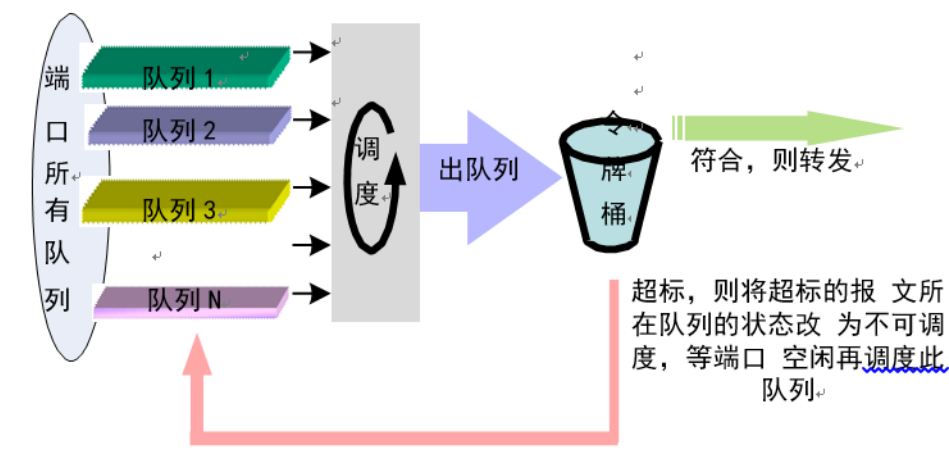
分类及比较

流量整形分为两种：

接口流量整形：也叫接口限速 LR（Line rate），限制接口发送的所有报文（包括紧急报文）的总速率，是对整个出接口进行流量整形，不区分优先级。如图 3-12 所示。

- 队列调度之后在出队的时候，对所有队列都进行令牌桶评估。
- 在令牌桶评估后，如果数据包速率符合要求，则转发；如果数据包速率超标，则将该数据包所在队列的状态改为不可调度，等空闲时再调度此队列。

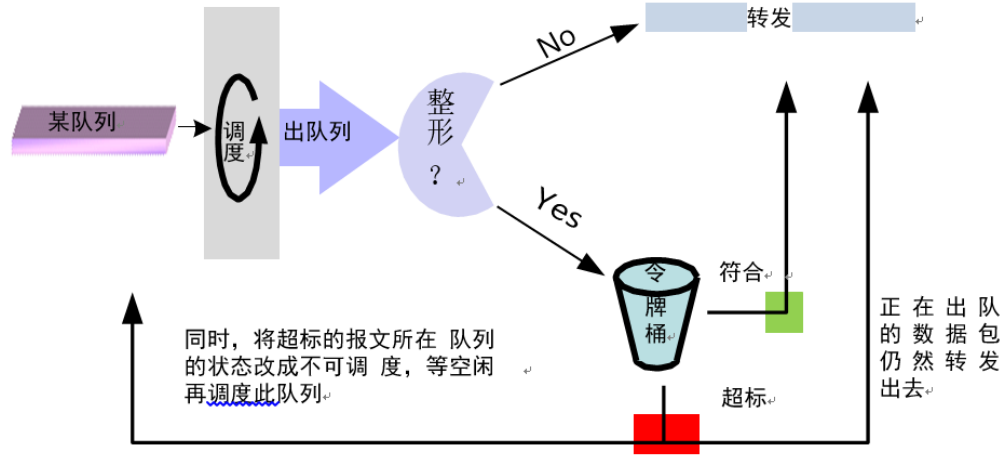
图3-12 端口流量整形过程



队列流量整形：也就是所谓的 Shaping，对出接口的每个队列进行流量整形，区分优先级。如图 3-13 所示。

- 队列调度之后在出队的时候，对于不需要整形的报文，直接转发；对于需要进行整形的报文，则先进行令牌桶评估。
- 在令牌桶评估后，如果数据包速率符合要求，则被标记为绿色并转发；如果数据包速率超标，则当前正在出队的数据包仍然转发出去，同时，将该数据包所在队列的状态改为不可调度，等令牌桶填充了新的令牌时再调度此队列。队列的状态改为不可调度后，该队列允许报文继续入队，但入队满了的时候会丢弃报文。因此，虽然流量整形使超额的数据能够从接口平滑地输出，但并不表示流量整形永远不会丢包。

图3-13 队列流量整形过程



两种流量整形方式各有优势，可以针对不同的需求场景应用。用户可以根据自己的场景和需求进行选择。

表3-5 两种流量整形方式的比较

流量整形方式	优点	缺点	适用场景
基于端口	思路和配置简单	这个端口使用统一的模型，不能够基于业务区别对待。	如果端口下挂的业务模型比较单一（例如银行企业交易网，端口下都为财务交易数据；例如 110 警务平台，端口下都为语音业务），可以考虑使用端口流量整形。
基于队列	不同的业务使用不同队列，可以基于业务进行流量整形。	方案复杂，需要分别考虑各种业务模型；配置上较端口复杂。	如果端口下挂的是混合业务模型（数据、语音和视频业务混合），需要针对不同业务进行整形（指定具体业务整形），可以考虑使用队列流量整形。
基于端口 & 队列（层次化整形）	既可以基于业务整形，同时又可以考虑整个带宽的设置，实现层次化的管理。	如果该接口上同时配置接口队列整形，则接口整形的 CIR 必须大于等于接口队列整形的 CIR 之和；否则，流量整形会出现异常现象。	如果 PIR 之和已经超过端口的最大带宽，或者超过端口下接网络允许的最大带宽时，需要基于端口做限速，保证总流量不超过规定值。一般场景下也可以使用，体现层次化的流量管理。（推荐）

流量整形参数设置

流量整形分为基于端口和基于队列两种，基于端口只能支持 C 桶，而基于队列的整形可以同时支持 C 桶和 P 桶。

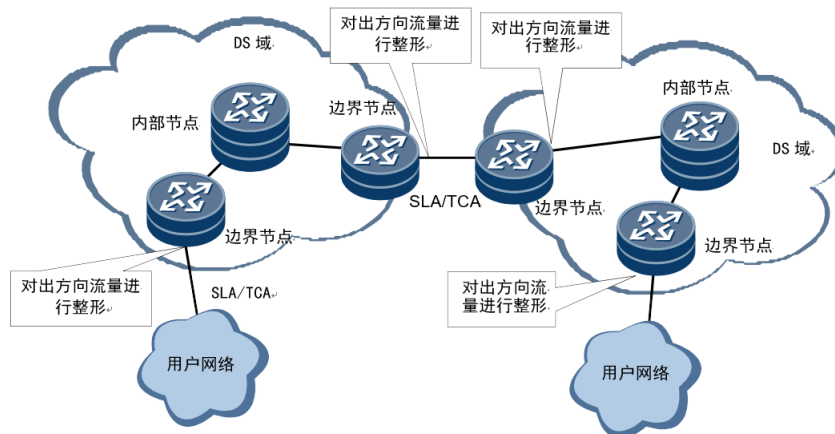
整形的参数需要结合入端口的 C 桶和 P 桶的设置，需要考虑端口的缓存能力，以及业务上 SLA 的合同规定。

企业内部的局域网带宽足够业务部署，一般不用流量整形，流量整形常用在局域网接 Internet 的出接口上，需要根据 SLA 会各个业务做合理分配。例如某企业租用了 ISP 的带宽为 5Mbps，合理分配业务带宽为视频流量 2.5Mbps，语音业务 1Mbps，数据流量 1.5Mbps，那么 CIR、CBS、PIR、PBS 的数据可以根据计算而得。

流量整形的应用

流量整形的典型应用是基于下游网络节点的流量监管的 SLA 指标来控制本地流量的输出，减少报文丢失。

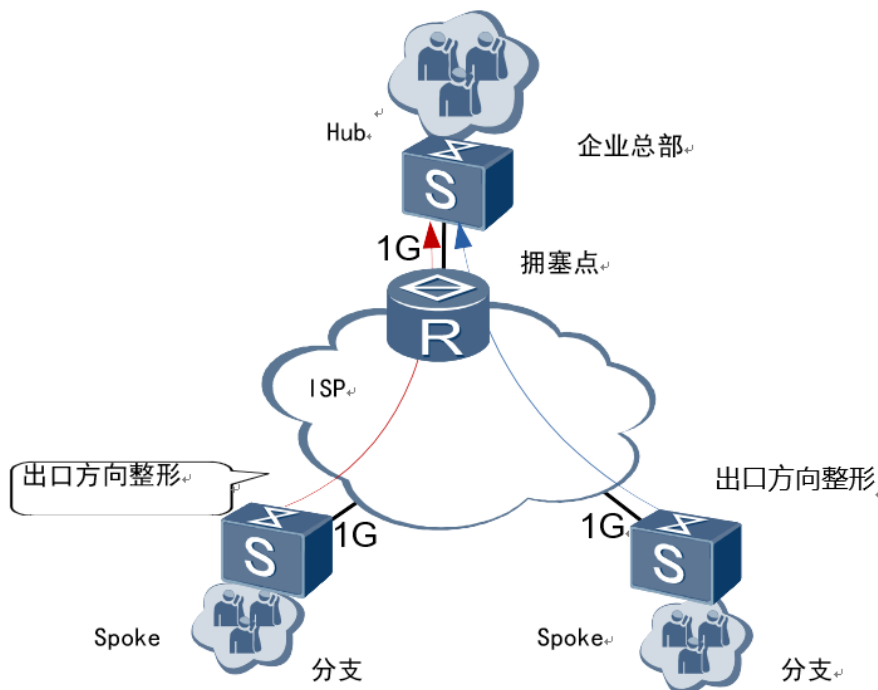
图3-14 流量整形的应用



- 基于端口的流量整形对接口上所有通过的报文进行流量整形。

如图 3-15 所示，假设一个 Hub-Spoke 组网模式的企业网络，总部与分支机构通过 ISP 使用专线相连，其链路带宽为 1Gbps。当所有分支同时向总部发送数据时，可能会在连接总部的 ISP 边缘节点造成拥塞。为了避免在 ISP 丢包，可以在各分支的边缘设备出口处进行端口流量整形。

图3-15 Hub-Spoke 企业网基于端口的流量整形



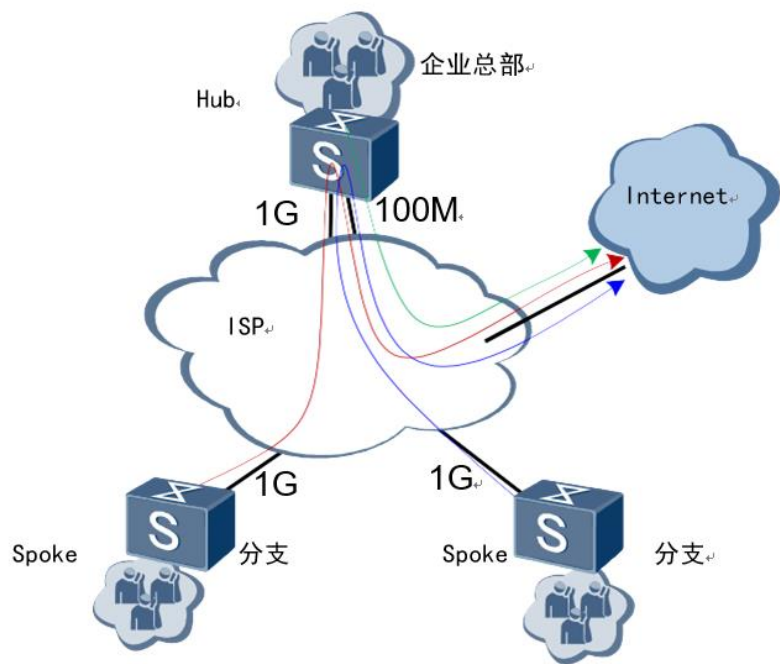
- 基于队列的流量整形

对接口上通过的某类报文（基于简单流分类）分别进行流量整形，从而可实现针对业务（如语音、数据、视频）的流量整形。

如图 3-16，假设一个 Hub-Spoke 组网模式的企业，总部与分支机构通过 ISP 使用专线相连，其链路带宽为 1Gbps。分支机构访问 Internet 是经过总部的，而总部到 Internet 的链路只有 100Mbps。如果所有分支机构以高速率同时访问 Internet，则可能使分支机构及总部访问 Internet 的总 web 流量超过 100Mbps，导致 web 流量在 ISP 被丢弃。

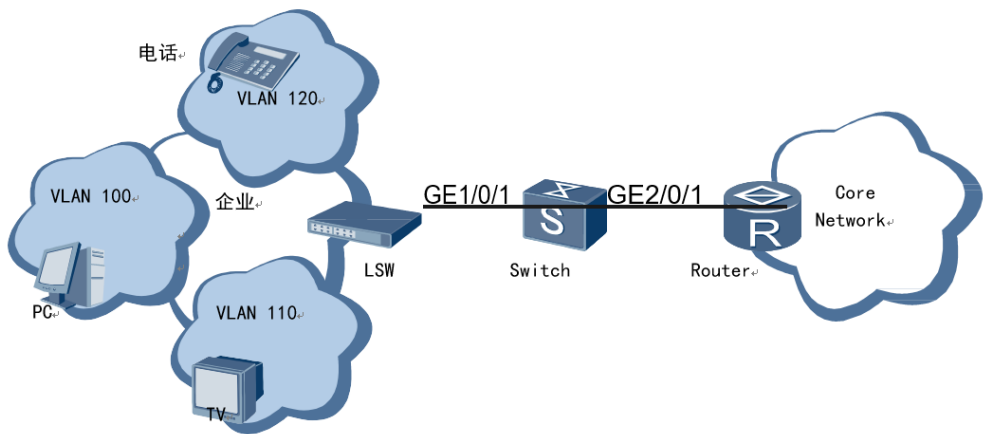
如果在分支机构的出口设备及总部访问 Internet 的出口设备上配置基于队列对 web 流量进行整形，可防止 web 流量的丢失。

图3-16 Hub-Spoke 企业网基于队列的流量整形



- 基于队列&端口的流量整形（层次化的流量整形）

图3-17 层次化的流量整形



Switch 通过接口 GE0/0/2 与路由器互连，来自企业局域网的业务有语音、视频、数据，这些业务可经由 Switch 和路由器到达互联网，如图 3-17 所示。由于来自企业局域网的流量速率大于互联网接口 GE0/0/2 的速率，那么在出接口 GE0/0/2 处可能会发生带宽抖动。为减少带宽抖动，同时保证各类业务带宽要求，可以在 GE0/0/2 处做层次化流量整形。先基于端口做流量整形，保证出端口流量不会超过允许的带宽；然后基于不同的业务类型（语音、视频和数据）的优先级基于队列做流量整形。

3.3.2 流量监管和整形的比较

- 相同点：
 - 作用都是监控网络流量
 - 都是用令牌桶算法评估流量速率
 - 主要都用于网络边缘
- 差异点如下表所示：

表3-6 流量监管和流量整形差异点

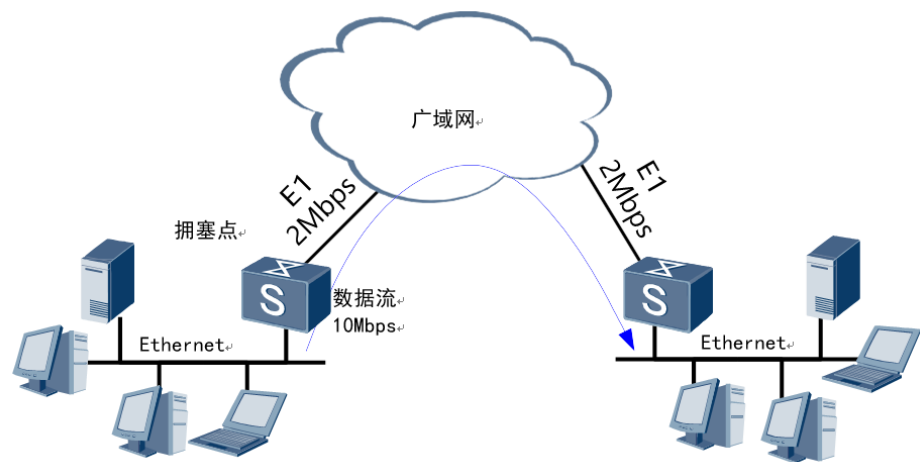
流量监管	流量整形
应用于端口入方向	应用于端口出方向
丢弃超额流量或将超额流量重标记为低优先级	缓存超出策略/协定 SLA 规定的超额流量
不需要额外的内存资源，不会带来延迟和抖动	需要缓存超额流量，可能会带来延迟和抖动
丢包可能引发重传（在应用层面看也可能引起延迟和抖动）	较少的丢包，因而较少导致重传
可以重标记流量（着色）	不能重标记

3.4 拥塞管理和拥塞避免

3.4.1 拥塞概述

拥塞是在共享网络上多个用户竞争相同的资源（带宽、缓冲区等）时发生的问题。例如，由于广域网的带宽通常要比局域网的带宽小，当一个局域网的用户向另一个局域网的用户发送数据时，由于广域网的带宽小于局域网的带宽，数据将不可能按局域网发送的速度在广域网上传输。此时，处在局域网和广域网之间的边缘交换机上将发生拥塞，如图 3-18 所示。

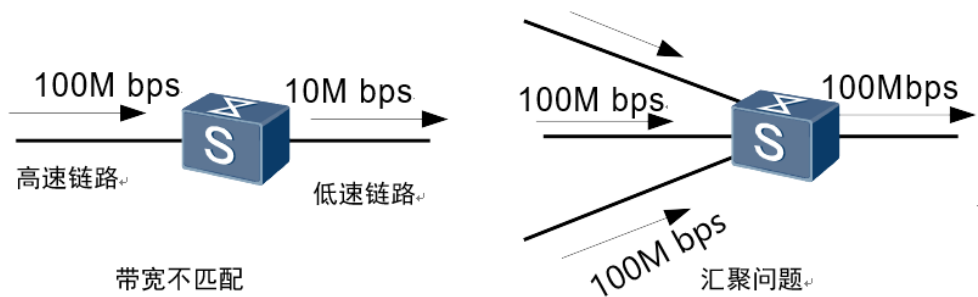
图3-18 网络拥塞



拥塞经常发生于图 3-19 所示的情况：

- 速率不匹配：分组从高速链路进入设备，再由低速链路转发出去。
- 汇聚问题：分组从多个接口同时进入设备，由一个没有足够带宽的接口转发出去。

图3-19 链路带宽瓶颈



不仅仅是链路带宽的瓶颈会导致拥塞，任何用于正常转发处理的资源的不足（如可分配的处理器时间、缓冲区、内存资源的不足）都会造成拥塞。此外，在某段时间内对所到达的流量控制不力，使之超出了可分配的网络资源，也是引发网络拥塞的一个因素。

拥塞有可能会引发一系列的负面影响：

- 拥塞增加了报文传输的延迟和延迟抖动。
- 过高的延迟会引起报文重传。
- 拥塞使网络的有效吞吐率降低，造成网络资源的损害。

- 拥塞加剧会耗费大量的网络资源（特别是存储资源），不合理的资源分配甚至可能导致系统陷入资源死锁而崩溃。

可见，拥塞使流量不能及时获得资源，是造成服务性能下降的源头。然而在分组交换以及多用户业务并存的复杂环境下，拥塞又是常见的。因此采取有效的避免拥塞以及防止拥塞加剧的方法是必需的。

任何一个实用的网络都需要解决网络拥塞的管理问题，也就是解决有限的网络资源与用户需求间的矛盾，在满足用户对服务质量要求的前提下尽可能地充分利用网络资源。

通常用如下两种策略来缓解网络拥塞：

- 拥塞管理（Congestion Management）：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技术，将从一个接口发出的所有报文放入多个队列，按照各个队列的优先级进行处理。不同的队列调度算法用来解决不同的问题，并产生不同的效果。
- 拥塞避免（Congestion Avoidance）：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量控制机制。拥塞避免用于防止因为线路拥塞而使设备的队列溢出。

3.4.2 拥塞管理

队列概述

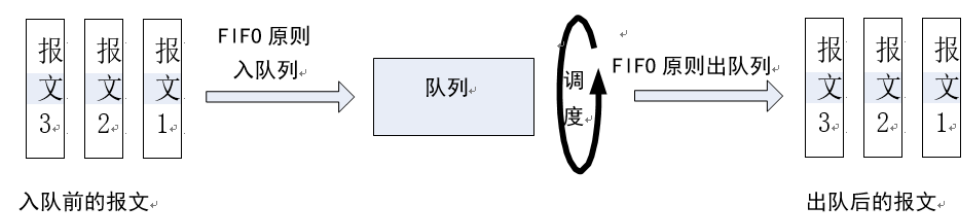
拥塞管理的中心内容是当拥塞发生时如何制定一个策略，用于决定报文转发的处理次序和丢弃原则，一般采用队列技术。

队列指的是在缓存中对报文进行排序的逻辑。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暂存在缓存中。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由队列调度算法决定。

华为交换机设备的每个端口上都有 8 个下行队列，称为 CQ（Class Queue）队列，也叫端口队列（Port-queue），在交换机内部与前文提到的 8 个 PHB 一一对应，分别为 BE、AF1、AF2、AF3、AF4、EF、CS6 和 CS7。

单个队列的报文采用 FIFO（First In First Out）原则入队和出队。

图3-20 FIFO 处理流程



队列调度

QoS 支持多种形式的队列调度，本节中——介绍。

● PQ（Priority Queuing）调度

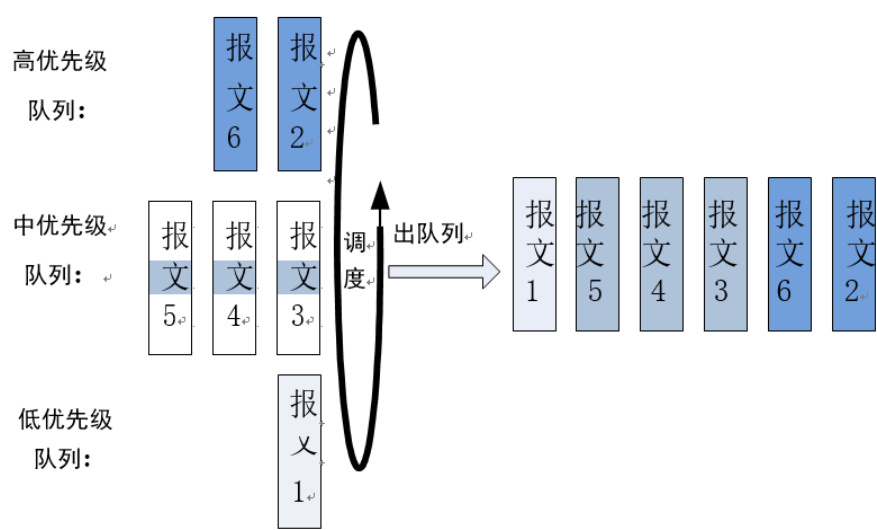
PQ（Priority Queuing）调度，就是严格按照队列优先级的高低顺序进行调度。只有高优先级队列中的报文全部调度完毕后，低优先级队列才有调度机会。

采用 PQ 调度方式，将延迟敏感的关键业务放入高优先级队列，将非关键业务放入低优先级队列，从而确保关键业务被优先发送。

PQ 调度的缺点是：拥塞发生时，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报文就会由于得不到服务而“饿死”。

假设端口有 3 个采用 PQ 调度的队列，分别为高优先（High）队列、中优先（Medium）队列、和低优先（Low）队列，它们的优先级依次降低。如图 3-21，其中报文编号表示报文到达顺序。

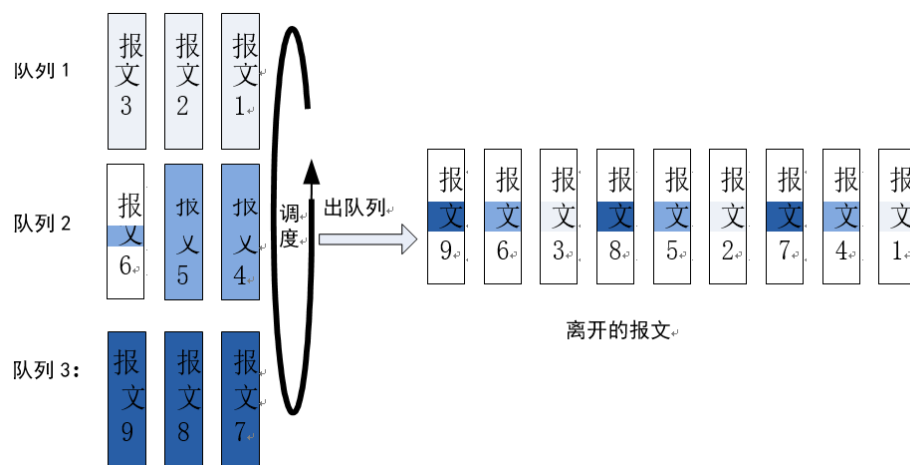
图3-21 PQ 调度



如图 3-21 所示，在报文出队的时候，首先让高优先队列中的报文出队并发送，直到高优先队列中的报文发送完，然后发送中优先队列中的报文，直到发送完，接着是低优先队列。在调度低优先级队列时，如果高优先级队列又有报文到来，则会优先调度高优先级队列。这样，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后发送。

● RR（Round Robin）调度

RR 调度采用轮询的方式，对多个队列进行调度。RR 以环形的方式轮询多个队列。如果轮询的队列不为空，则从该队列取走一个报文；如果该队列为空，则直接跳过该队列，调度器不等待。



如上图所示，RR 调度各个队列之间没有优先级之分，都能够有相等的概率得到调度。

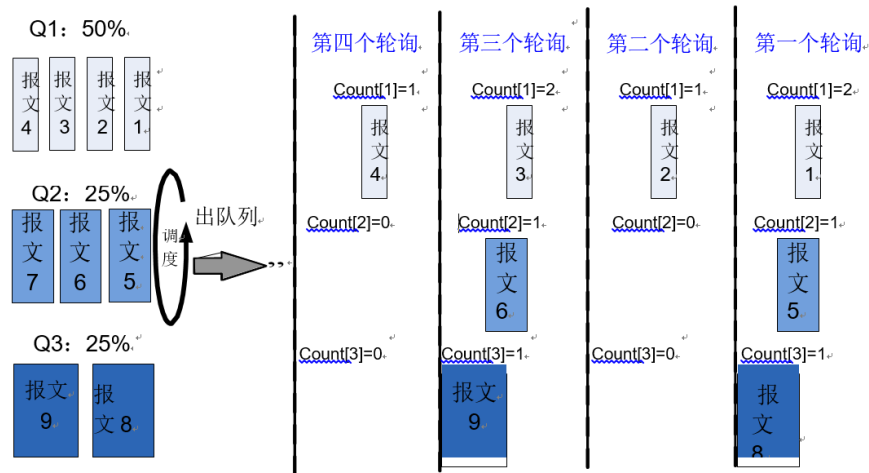
RR 调度的缺点是：所有队列无法体现优先级，对于延迟敏感的关键业务和非关键业务无法得到区别对待，使得关键业务无法及时得到处理。

● WRR（Weighted Round Robin）调度

加权轮询 WRR（Weighted Round Robin）调度主要解决 RR 不能设置权重的不足。在轮询的时候，WRR 每个队列享受的调度机会和该队列的权重成比例。RR 调度相当于权值为 1 的 WRR 调度。WRR 的实现方法是为每个队列设置一个计数器 Count，根据权重进行初始化。每次轮询到一个队列时，该队列输出一个报文且计数器减一。当计数器为 0 时停止调度

该队列，但继续调度其他计数器不为 0 的队列。当所有队列的计数器都为 0 时，所有计数器重新根据权重初始化，开始新一轮调度。在一个循环中，权重大的队列被多次调度。

图3-22 WRR 调度



假设某端口有 3 个队列采用 WRR 调度，为每个队列配置一个权值，依次为 50%、25%、25%，详细的调度过程如下：

首先计数器初始化：Count[1]=2，Count[2]=1，Count[3]=1。

➤ 第 1 个轮询：

从队列 1 取出报文 1 发送，Count[1]=1；从队列 2 取出报文 5 发送，Count[2]=0；从队列 3 取出报文 8 发送，Count[3]=0。

➤ 第 2 个轮询：

从队列 1 取出报文 2 发送，Count[1]=0；由于 Count[2]=0，Count[3]=0，队列 2 和队列 3 不参与此轮调度。此时，Count[1]=0，Count[2]=0，Count[3]=0，将计数器重新初始化：Count[1]=2，Count[2]=1，Count[3]=1。

➤ 第 3 个轮询：

从队列 1 取出报文 3 发送，Count[1]=1；从队列 2 取出报文 6 发送，Count[2]=0；从队列 3 取出报文 9 发送，Count[3]=0。

➤ 第 4 个轮询：

从队列 1 取出报文 4 发送，Count[1]=0；由于 Count[2]=0，Count[3]=0，队列 2 和队列 3 不参与此轮调度。此时，Count[1]=0，Count[2]=0，Count[3]=0，将计数器重新初始化：Count[1]=2，Count[2]=1，Count[3]=1。

从统计上看，各队列中的报文流被调度的次数与该队列的权值成正比，权值越大被调度的次数相对越多。如果该端口为 100Mbps，则可以保证最低权重的队列至少获得 25Mbit/s 带宽，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。

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。

WRR 调度有两个缺点：

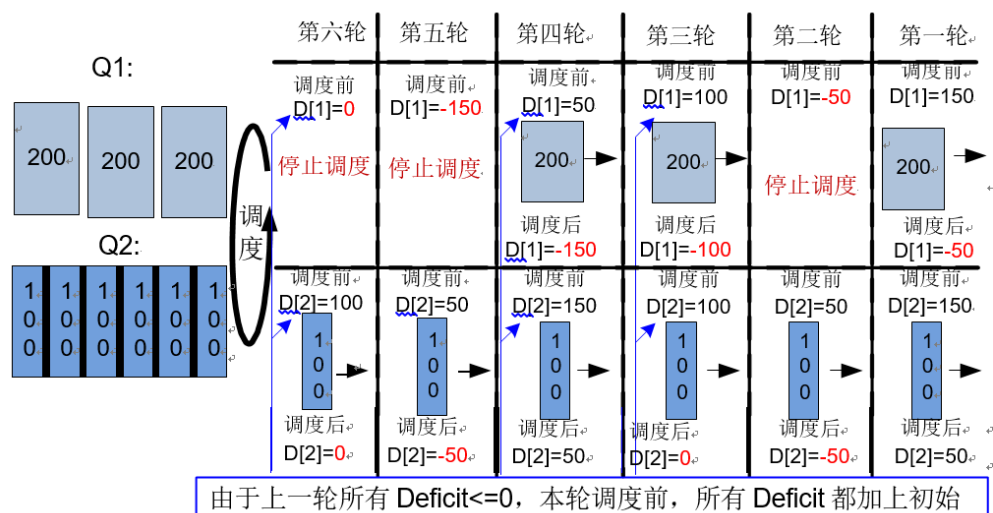
- WRR 调度按照报文个数进行调度，因此每个队列没有固定的带宽，同等调度机会下大尺寸报文获得的实际带宽要大于小尺寸报文获得的带宽。而用户一般关心的是带宽。当每个队列的平均报文长度相等或已知时，通过配置 WRR 权重，用户能够获得想要的带宽；但是，当队列的平均报文长度变化时，用户就不能通过配置 WRR 权重获取想要的带宽。
- 低延时需求业务（如语音）得不到及时调度。
- DRR（Deficit Round Robin）调度

差分轮询 DRR（Deficit Round Robin）调度实现原理与 RR 调度基本相同。

DRR 与 RR 的区别是：RR 调度是按照报文个数进行调度，而 DRR 是按照报文长度进行调度。DRR 为每个队列设置一个计数器 Deficit，Deficit 初始化为一次调度允许的最大字节数，一般为接口 MTU。每次轮询到一个队列时，该队列输出一个报文且计数器 Deficit 减去报文长度。如果报文长度超过了队列的调度能力，DRR 调度允许 Deficit 出现负值，以保证长报文也能够得到调度。但下次轮循调度时该队列将不会被调度。当计数器为 0 或负数时停止调度该队列，但继续调度其他计数器为正数的队列。当所有队列的 Deficit 都为 0 或负数时，将所有队列的 Deficit 计数器加上初始值，开始新一轮调度。

假设某端口 MTU=150Bytes，有 2 个队列 Q1 和 Q2 采用 DRR 调度，Q1 队列中有多个 200Bytes 的长报文，Q2 队列中有多个 100Bytes 的端报文，则调度过程如图 3-23 所示。

图3-23 DRR 调度



由图 3-23 可以看出，经过第 1~6 轮 DRR 调度，Q1 队列被调出了 3 个 200Bytes 的报文，Q2 队列被调出了 6 个 100Bytes 的报文。从长期的统计看，Q1 和 Q2 的实际输出

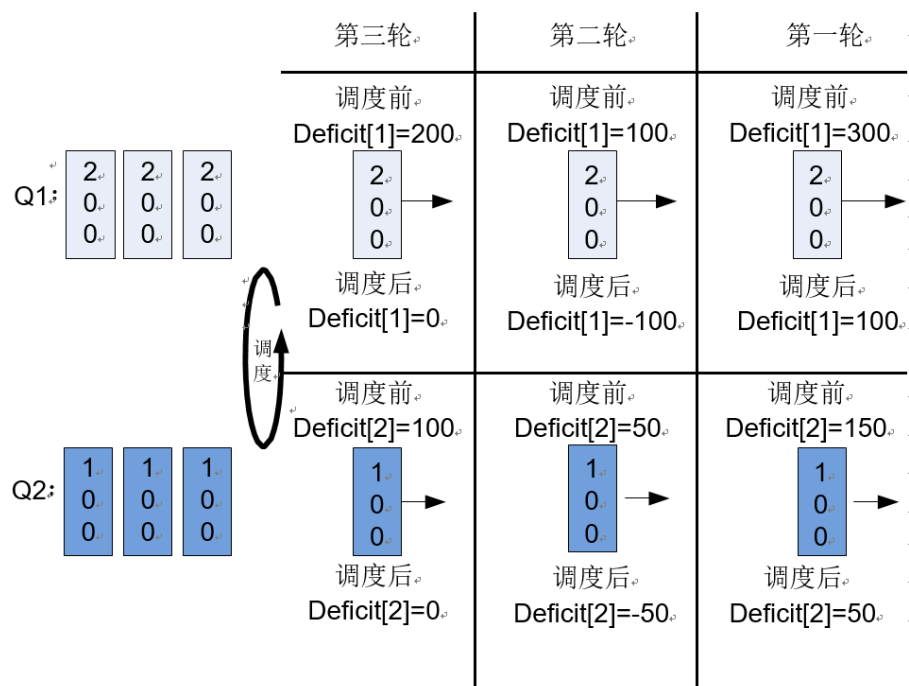
带宽比是 1:1，为公平的比例。DRR 调度避免了采用 PQ 调度时低优先级队列中的报文可能长时间得不到服务的缺点。但是，DRR 调度不能设置权重，且也具有低延时需求业务（如语音）得不到及时调度的缺点。

● DWRR（Deficit Weighted Round Robin）调度

差分加权轮询 DWRR（Deficit Weighted Round Robin）调度主要解决 DRR 不能设置权重的不足。DRR 调度相当于权值为 1 的 DWRR 调度。

DWRR 为每个队列设置一个计数器 Deficit，Deficit 初始化为 $\text{Weight} \times \text{MTU}$ 。每次轮询到一个队列时，该队列输出一个报文且计数器 Deficit 减去报文长度。当计数器为 0 时停止调度该队列，但继续调度其他计数器不为 0 的队列。当所有队列的计数器都为 0 时，所有计数器的 Deficit 都加上 $\text{Weight} \times \text{MTU}$ ，开始新一轮调度。

图3-24 DWRR 调度



假设某端口 MTU=150Bytes，有 2 个队列 Q1 和 Q2 采用 DRR 调度，Q1 队列中有多个 200Bytes 的长报文，Q2 队列中有多个 100Bytes 的端报文，Q1 和 Q2 配置权重比为 $\text{weight}_1:\text{weight}_2=2:1$ 。则 DWRR 调度过程如图 3-24。

➤ 第一轮调度

$\text{Deficit}[1]=\text{weight}_1 \times \text{MTU}=300$ ， $\text{Deficit}[2]=\text{weight}_2 \times \text{MTU}=150\text{Bytes}$ ，从 Q1 队列取出 200Bytes 报文发送，从 Q2 队列取出 100Bytes 发送；发送后， $\text{Deficit}[1]=100$ ， $\text{Deficit}[2]=50$ 。

➤ 第二轮调度

从 Q1 队列取出 200Bytes 报文发送，从 Q2 队列取出 100Bytes 发送；发送后， $\text{Deficit}[1]=-100$ ， $\text{Deficit}[2]=-50$ 。

➤ 第三轮调度

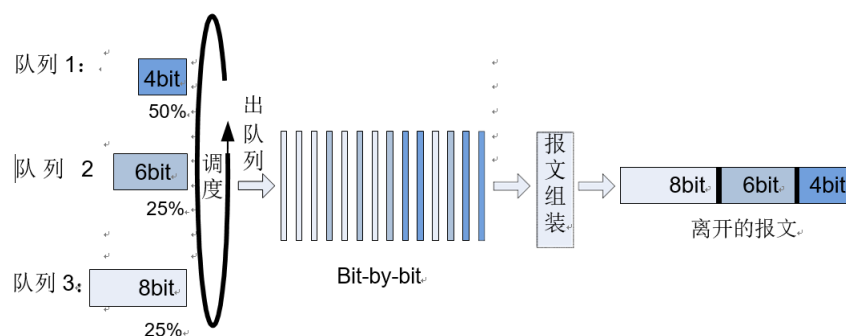
此时两个队列都为负，因此， $\text{Deficit}[1]=\text{Deficit}[1]+\text{weight1}*\text{MTU}=-100+2*150=200$ ， $\text{Deficit}[2]=\text{Deficit}[2]+\text{weight2}*\text{MTU}=-50+1*150=100$ 。从 Q1 队列取出 200Bytes 报文发送，从 Q2 队列取出 100Bytes 发送；发送后， $\text{Deficit}[1]=0$ ， $\text{Deficit}[2]=0$ 。

由上图可以看出，经过第 1~3 轮 DWRR 调度，Q1 队列被调出了 3 个 200Bytes 的报文，Q2 队列被调出了 3 个 100Bytes 的报文。从长期的统计看，Q1 和 Q2 的实际输出带宽比是 2:1，与权重比相符。DWRR 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，也避免了各队列报文长度不等或变化较大时，WRR 调度不能按配置比例分配带宽资源的缺点。但是，DWRR 调度也具有低延时需求业务（如语音）得不到及时调度的缺点。

● WFQ（Weighted Fair Queuing）调度

加权公平队列 WFQ（Weighted Fair Queuing）调度是按队列权重来分配每个流应占有出口的带宽。同时，为了使得带宽分配更加“公平”，WFQ 以 bit 为单位进行调度，类似于图 3-25 的 bit-by-bit 调度模型。

图3-25 WFQ 调度



Bit-by-bit 调度模型可以完全按照权重分配带宽，防止长报文比短报文获得更多带宽，从而减少大小报文共存时的时延抖动。

但 Bit-by-bit 调度模型只是理想化的模型，实际上，华为交换机实现的 WFQ 是按照一定的粒度，例如 256B、1KB，或其他粒度，具体按何种粒度，与单板类型相关。

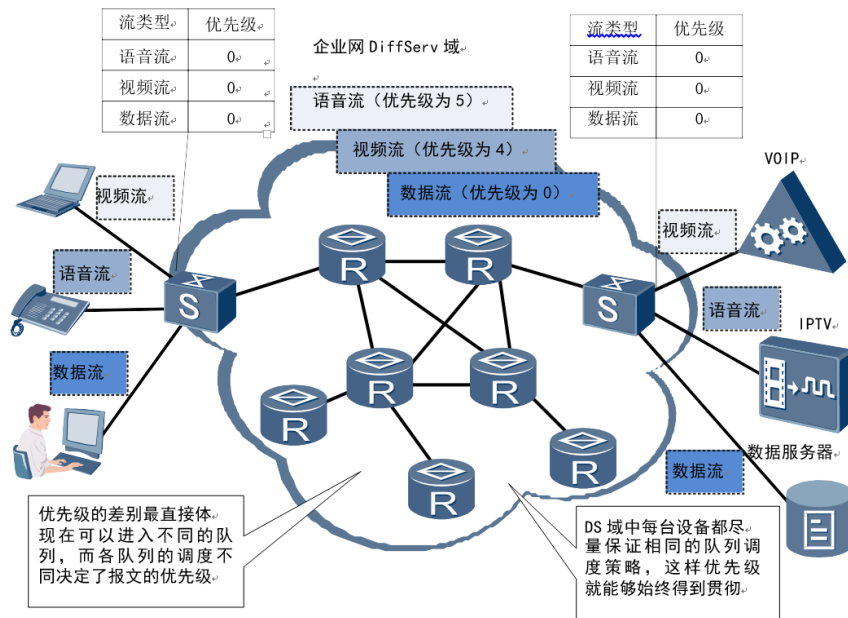
WFQ 的优点主要有以下几点：

- 不同的队列获得公平的调度机会，从总体上均衡各个流的延迟。
- 短报文和长报文获得公平的调度：如果不同队列间同时存在多个长报文和短报文等待发送，让短报文优先获得调度，从而在总体上减少各个流的报文间的抖动。
- 从统计上看，权重越小，所分得的带宽越少。权重越大，所分得的带宽越多。

拥塞管理的应用

拥塞管理的应用就是我们常说的队列调度，在 QoS 方案部署中比较常用。

图3-26 队列调度



在 DS 域中，客户针对不同业务打上不同的内部和外部优先级，而不同优先级的报文之所以可以被区别对待就是因为其可以安排在不同的队列中，不同队列享受不同的调度策略。因此，队列调度策略是报文优先级划分的目的，是报文能够被区别对待的手段之一。

一般情况下优先级较高的队列使用 PQ 调度，保证业务时刻可以得到调度而不受其他低优先级队列的影响；而低优先级队列之间使用轮训调度，避免低优先级队列出现饿死的情况。华为交换机当前没有配置 wrr、dwrr 权重时，缺省是按照 WRR 算法调度，队列权重均为 1。

3.4.3 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。

支持两种丢弃策略：

- 尾丢弃
- WRED

尾丢弃

传统的丢包策略采用尾部丢弃（Tail Drop）的方法，同等的对待所有的报文，不对服务等级进行区分。在拥塞发生期间，队列尾部的数据报文将被全部丢弃，直到拥塞解除。

这种丢弃策略会引发 TCP 全局同步现象。所谓 TCP 全局同步现象，是指当多个队列同时丢弃多个 TCP 连接报文时，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；而后这几个 TCP 连接又会在某个时刻同时出现流量高峰。如此反复，使网络流量忽大忽小，影响链路利用率。

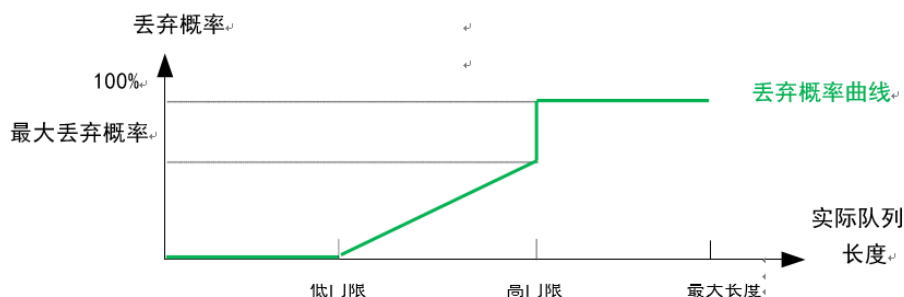
WRED

加权随机早期检测 WRED（Weighted Random Early Detection）是在队列拥塞前进行报文丢弃的一种拥塞避免机制。WRED 通过随机丢弃报文避免了 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，无论何时总有 TCP 连接在进行较快的发送，提高了线路带宽的利用率。

WRED 为每个队列都设定一对低门限和高门限值，并规定：

- 当队列长度小于低限时，不丢弃报文，丢弃概率为 0%。
- 当队列长度超过高限时，丢弃所有新到来的报文，即进行尾丢弃，丢弃概率为 100%。
- 当队列长度在低限和高限之间时，开始随机丢弃新到来的报文，且设定了一个最大丢弃概率，队列越长，丢弃概率越大。如果以报文长度为横坐标，丢弃概率为纵坐标，则丢弃概率曲线如图 3-27。

图3-27 WRED 丢弃概率曲线

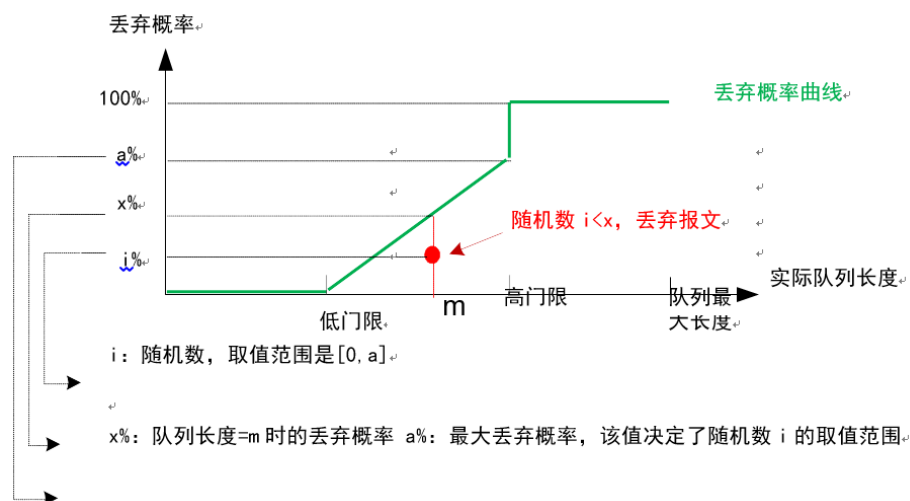


WRED 详细实现过程如图 3-28，WRED 为每个到来的报文赋予一随机数 i （ $0 < i \% < \text{最大丢弃概率}$ ），并用该随机数与当前队列的丢弃概率比较，小于丢弃概率则新到的报文

被丢弃。假设最大丢弃概率为 $a\%$ ，当前队列长度为 m ，对应的丢弃概率为 $x\%$ ，如果随机数不大于 x ，则丢弃新到的报文；如果随机数 $x < i < a$ ，则不丢弃新到的报文。

假设队列最低门限为 40% ，最高门限为 80% ，丢弃最高概率为 20% ，当前队列长度为总队列的 50% ，当前对应丢弃概率为 5% 。若此时一个报文进入队列，会给他赋予一个随机值，范围为 $[0, 20]$ ，若随机值落在 $[0, 5]$ 之间则丢弃报文，否则报文进入队列。由于当前队列长度离最大门限较远，所以报文被丢弃的概率不大；若当前队列长度为总队列的 75% ，当前对应丢弃概率为 18% ，那么随机数落在 $[0, 18]$ 之间的概率会大大增加，也就是说被丢弃的概率会大大增加。

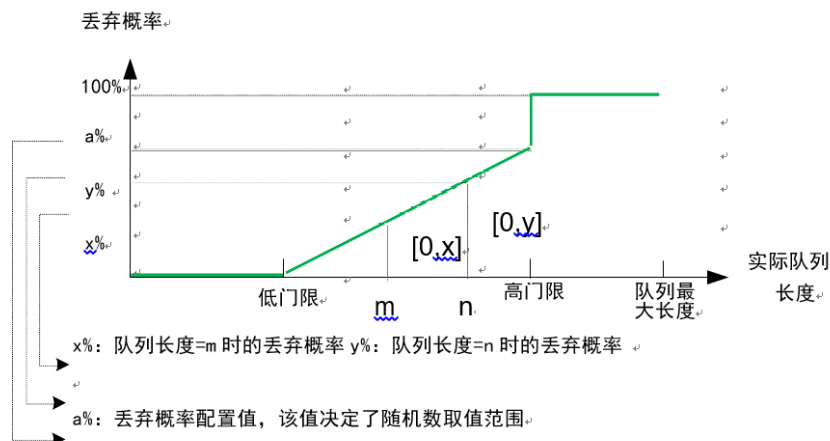
图3-28 WRED 实现过程



如图 3-28，假设队列长度为 m （低门限 $< m <$ 高门限）时丢弃概率为 $x\%$ ，则当随机数落入区间 $[0, x]$ 时新到的报文被丢弃；长度为 n （ $m < n <$ 高门限）时丢弃概率为 $y\%$ ，则当随机数落入区间 $[0, y]$ 时报文被丢弃。区间 $[0, y]$ 比 $[0, x]$ 的范围大，随机数落入区间 $[0, y]$ 比落入区间 $[0, x]$ 的可能性大，因此，队列越长，新到的报文被丢弃的可能性越高。

假设当前队列长度为总队列长度的 50% 时，丢弃概率为 12% ，那么随机数落入 $[0, 12]$ 之间时新到的报文将被丢弃；当队列长度为 60% 时丢弃概率为 15% （队列越长，丢弃概率越高），那么随机数落入 $[0, 15]$ 之间时新到的报文将被丢弃，很显然随机数落入 $[0, 15]$ 之间个概率要大于 $[0, 12]$ ，因此也就说明了队列越长，新到的报文被丢弃的可能性越高。

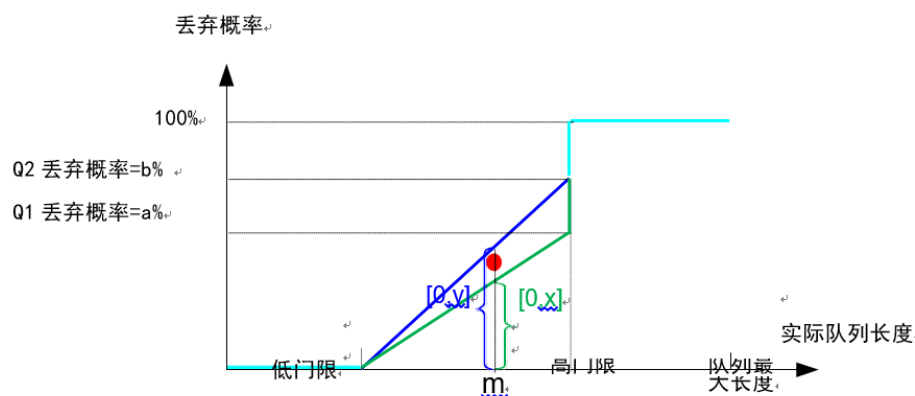
图3-29 队列越长，新到的报文被丢弃的可能性越高



假设有两个队列 Q_1 和 Q_2 ，对应的最低门限和最高门限一致，最大丢弃概率分别为 $a\%$ 和 $b\%$ ，则两个队列的丢弃概率曲线图如图 3-29。假设当队列长度为 m 时， Q_1 的丢弃概率为 $x\%$ ， Q_2 的丢弃概率为 $y\%$ ，则当随机数落入区间 $[0, x]$ 时 Q_1 新到的报文被丢弃；

当随机数落入区间 $[0, y]$ 时 Q_2 新到的报文被丢弃。区间 $[0, y]$ 比 $[0, x]$ 的范围大，随机数落入区间 $[0, y]$ 比落入区间 $[0, x]$ 的可能性大，因此，相同队列长度时，最大丢弃概率配置得越大，丢弃可能性越高。

图3-30 相同队列长度时，最大丢弃概率配置得越大，丢弃可能性越高



WRED 参数设置

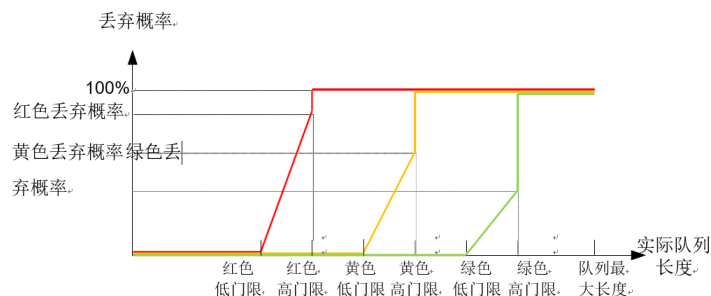
对于实时性要求比较高的业务一般使用的是尾丢弃。因为这种报文要提供最大限度的带宽保证。采用尾丢弃是只有当报文队列达到最大长度时才会丢弃，由于使用 PQ 调度，抢占其他业务的带宽，所以当发生拥塞时，实时性的业务带宽能够得到最大的保证。

对于 WFQ 队列，一般采用 WRED。由于 WFQ 队列是按权重分享带宽，容易发生拥塞，采用 WRED 策略有效的避免了 TCP 全局同步现象。

- 高低门限及丢弃概率的设置

实际应用时，WRED 低门限百分比建议从 50%开始取值，根据不同颜色的丢弃优先级逐级调整。一般推荐绿色报文设置的丢弃概率比较小，高、低门限值比较大；黄色报文次之；红色报文设置的丢弃概率最大，高、低门限值最小，如图 3-31。这样，在网络趋近拥塞时，红色报文由于设置的低门限值比较小，丢弃概率比较大，红色报文最先开始被丢弃；随着队列的长度逐渐增长，最后才开始丢弃绿色报文。如果队列长度达到相应颜色的最大门限，这种颜色的报文开始作尾丢弃。

图3-31 三种丢弃优先级的 WRED 丢弃概率曲线



- 队列最大长度的设置

华为交换机允许在配置 WRED 时修改队列最大长度（`qos queue queue-index length length-value`，配置接口优先级队列的长度）。上一节中介绍过，当网络拥塞时，报文在缓存中产生堆积，被延迟处理，延迟时间的大小主要取决于队列的缓存长度以及该队列获得的输出带宽。因此，在队列输出带宽一定的情况下，队列越短，时延越小；队列越长，时延越大。

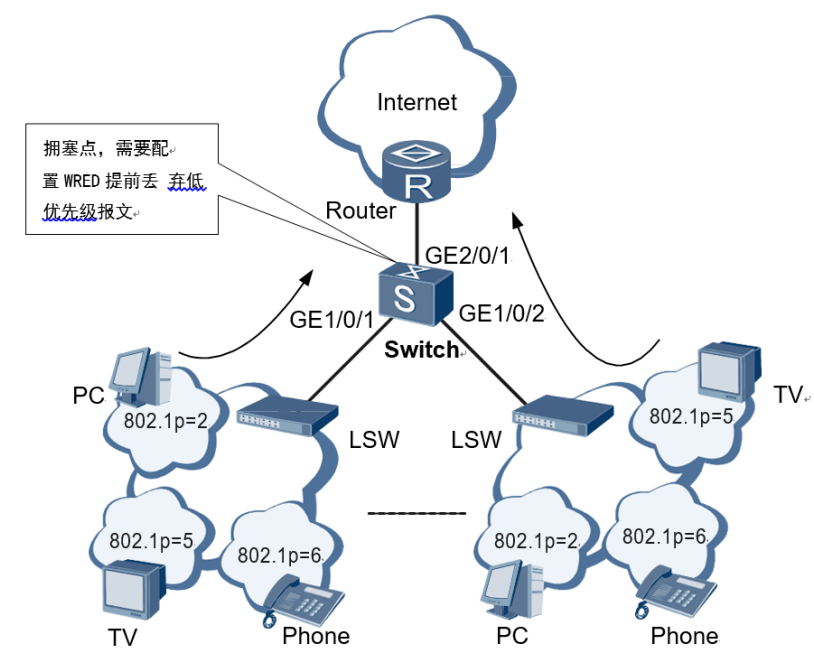
队列不能设置过短。如果某队列过短，缓存空间过小，容易造成丢包，流量即使速率不大也容纳不下。队列越短，容忍的突发量越小。队列越大，容忍的突发量越大。

队列也不能设置过长。如果队列太长，时延过大。尤其对于 TCP 应用，通信两端会预测网络拥塞情况，在发送完一个报文后，启动超时定期器等待对方应答，如果在定时器超时前没有等到应答，发送方会重发报文。如果报文在网络中被缓存的时间过长，不丢弃与丢弃没有区别。

拥塞避免的应用

在 2.4.1 节中已经描述了拥塞出现的场景，一般 WRED 配置在接入层交换机上行链接汇聚层的端口，或者在汇聚层类似汇聚功能的交换机或者路由器链接上行的端口，如图 3-32 所示 WRED 应用在 Switch 的 GE2/0/1 端口。

图3-32 拥塞避免的应用场景



WRED 配置在接口的出方向，基于模板配置，模板中可以区分颜色配置，最终模板在队列上应用。因此可以这样理解，WRED 可以基于不同的队列，不同的颜色分配报文的丢弃参数。越重要的报文其最低门限和最高门限都较不重要的报文要高，而最高丢弃概率则较低。推荐使用的各种颜色报文 WRED 参数设置如下表。

表3-7 WRED 基于颜色的推荐使用模板

队列（PHB）	阈值下限（%）	阈值上限（%）	丢弃概率
Green	80	100	10
Yellow	60	80	20
Red	40	60	30

由于 WRED 在队列调度之后生效，不同队列的优先级已经通过队列调度体现，因此推荐各队列都统一使用如上的 WRED 模板，不再基于队列区分，避免配置过于复杂却效果不大。

表3-8 WRED 基于队列的推荐使用模板

队列（PHB）	阈值下限（%）	阈值上限（%）	丢弃概率
高优先级（CS7、CS6）	80	100	10
中优先级（EF、AF1-AF4）	60	80	20
低优先级（BE）	40	60	30

4 HQoS 特性简介

4.1 产生背景

传统 QoS 技术可通过对视频、语音、数据等业务流的分类，为不同的业务提供差异化的服务。但是随着网络设备性能的快速提升，允许接入的用户数及其业务量不断增加，传统的 QoS 在应用中遇到了新的挑战：

- 传统 QoS 技术是基于端口进行调度的，这样导致流量管理对用户不敏感，无法实现基于用户的调度。
- 传统 QoS 技术很难做到同时对多个用户的多个业务进行控制。

为了解决上述问题，HQoS 应运而生。HQoS 通过多级队列区分用户及其业务，实现对多个用户的多种业务提供精细化的服务质量保证。

4.2 技术优势

华为敏捷交换机实现 HQoS 调度，具备以下优点：

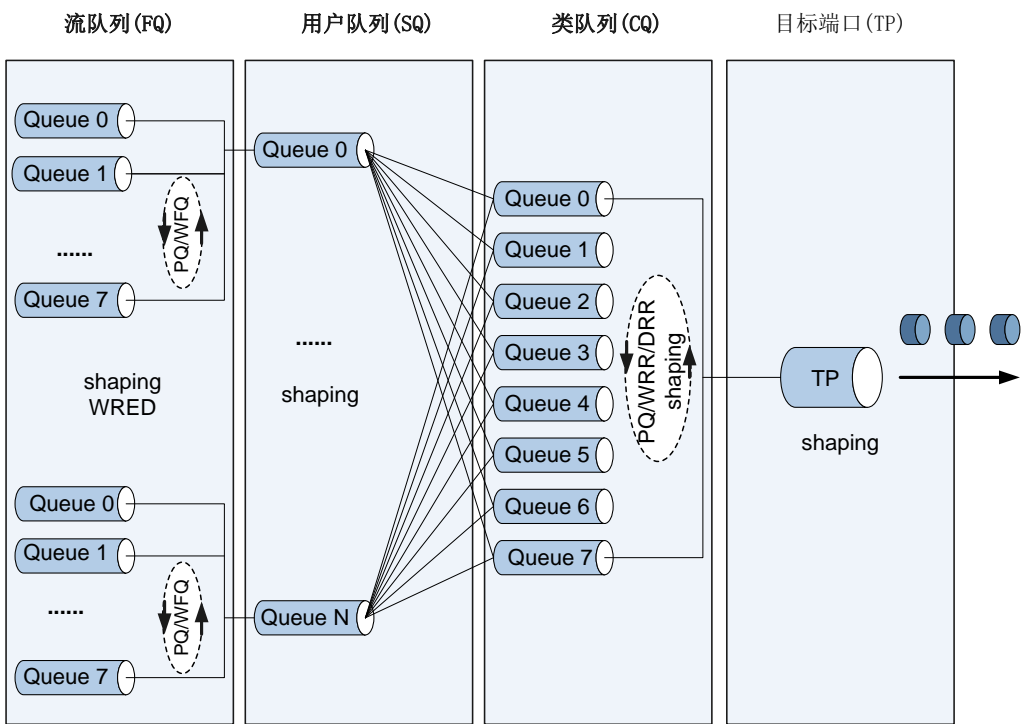
- 精细化标识业务,提升用户体验：可以精细地标识用户及其业务，提供基于业务、用户、端口队列、端口等层次化的多级调度，保证高级用户的高优先级业务体验。
- 融合 DiffServ 模型，增强整网 QoS 部署能力：HQoS 基于队列实现层次化调度，可和 DiffServ 模型很好融合，在网络关键节点（例如核心层设备）上提供多用户多业务带宽保证，提升整网 QoS 部署能力。

5 HQoS 技术原理

5.1 基本原理

HQoS 本质上是一种多级队列调度机制，目前在设备上支持流队列（FQ）、用户队列（SQ）、类队列（CQ）、目标端口队列（TP）等层级，如图 5-1 所示。多级队列以树状结构汇聚，流队列为叶子节点，目标端口队列为根节点。报文做层次化调度时，首先进入叶子节点 FQ，按序经过 FQ>SQ>CQ>TP 逐级调度后，从根结点 TP 发送出去。

图5-1 HQoS 队列调度示意图



- 流队列 FQ (Flow Queue)

HQoS 可以针对每个用户的业务流进行队列调度，每个用户都有 8 个流队列，分别对应 8 个业务优先级（BE、AF1、AF2、AF3、AF4、EF、CS6、CS7）。每个流队列可以配置 WRED 丢弃、流量整形或者队列调度，其中队列调度支持 PQ（优先级队列）和 WFQ（加权公平队列）两种机制。

● 用户队列 SQ（Subscriber Queue）

用户队列主要用来区分不同的用户。一个 SQ 对应一个用户，用户队列可以配置基于用户的流量整形，限制每个用户的总带宽。

● 类队列 CQ（Class Queue）

HQoS 下行每个物理端口对应 8 个 CQ 队列，用来根据业务优先级区分下行所有用户的业务流。每个 CQ 队列可配置 WRED 丢弃、流量整形或者队列调度，其中队列调度支持 PQ、WRR、DRR 等机制。

设备支持配置流队列到类队列的映射，通过建立 FQ CQ 的队列映射，可以灵活的控制流队列某一服务等级队列中的业务流量进入端口队列的某一服务等级队列。

● 目标端口队列 TP（Target Port）

目标端口 TP 即设备的物理接口，数据经过 TP 的调度后从相应的端口转发出去。设备支持在每个 TP 端口上配置流量整形。

5.2 业务模型

在网络部署中，调度队列和实际业务一般可基于表 5-1 所示的对应关系规划组网。

表5-1 多级队列和实际业务对应关系

队列名称	业务名称	场景说明
流队列（FQ）	用户业务流	每个流队列对应某用户业务，例如语音、视频、数据等业务。
用户队列（SQ）	用户	用户队列和实际用户对应。
类队列（CQ）	端口队列	类队列对应物理端口下的端口队列。
目标端口（TP）	端口	TP 队列对应实体物理端口。

需要说明，在具体网络规划中，如果用户不启用 HQoS 调度，则数据报文可旁路 FQ 和 SQ 队列，直接经过 CQ>TP 队列发送出去。

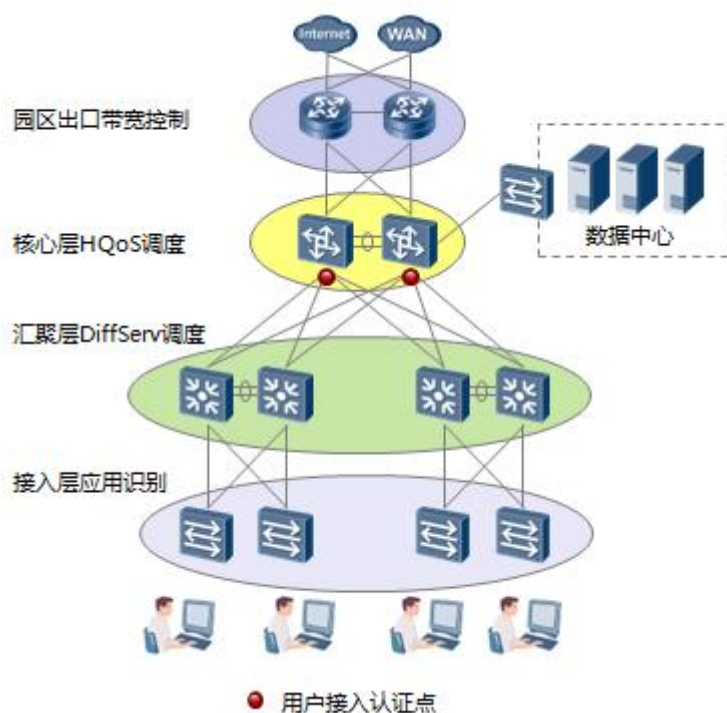
6 应用场景

6.1 园区用户接入场景 HQoS 调度

在企业园区中，HQoS 调度一般部署在用户认证设备上。在用户接入方案中，认证点一般部署在汇聚层或者核心层交换机上，HQoS 调度也部署在相应位置。

如图 6-1 所示，本图以用户认证点在核心层交换机为例，说明 HQoS 的规划。

图6-1 企业园区接入场景 HQoS 调度方案



在园区网中，HQoS 调度一般结合整网的 DiffServ 模型实施：

- 接入层应用识别

接入交换机需要担负数据流的识别、分类以及流标记的工作。在实际部署的时候，接入交换机上不同的端口接入不同的终端，在接入层可以给这些不同的业务分配不同的优先级（802.1P 或 DSCP 等）。

- 汇聚层 DiffServ 调度

汇聚层设备端口信任 DSCP（或者 802.1P），基于接入层标识的 QoS 参数，通过队列调度、流量整形、拥塞避免等技术实施 QoS 策略，保证高优先级业务优先获得调度。

- 核心层 HQoS 调度

HQoS 调度一般部署在网络下行方向。

在网络下行方向，用户需要访问服务器区域等存在大数据的资源，存在拥塞的可能，故在核心层交换机用户侧端口出方向上部署 HQoS 调度。

在具体实施中，用户级调度进入 SQ 队列，实施流量整形；业务级调度进入 FQ 队列，实施 WRED 和队列调度。



在核心层上行方向，信任接入层标识 DSCP（或者 802.1P，实施 DiffServ 调度。

- 出口路由器带宽控制

对于出口路由器，同样作为 DiffServ 域，信任设备标识的 DSCP/802.1P 参数，实施 QoS 策略。需要说明的是，在路由器的 WAN 口上，由于受限于出口带宽，需要进行 WAN 口出方向带宽控制。

华为技术有限公司
深圳龙岗区坂田华为基地
电话：+86 755 28780808
邮编：518129
www.huawei.com

商标声明

 HUAWEI, HUAWEI,  是华为技术有限公司商标或者注册商标，在本手册中以及本手册描述的产品中，出现的其它商标，产品名称，服务名称以及公司名称，由其各自的所有人拥有。

免责声明

本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺，华为不对您在本文档基础上做出的任何行为承担责任。华为可能不经通知修改上述信息，恕不另行通知。

版权所有©华为技术有限公司 2019。保留一切权利。

非经华为技术有限公司书面同意，任何单位和个人不得擅自摘抄、复制本手册内容的部分或全部，并不得以任何形式传播。