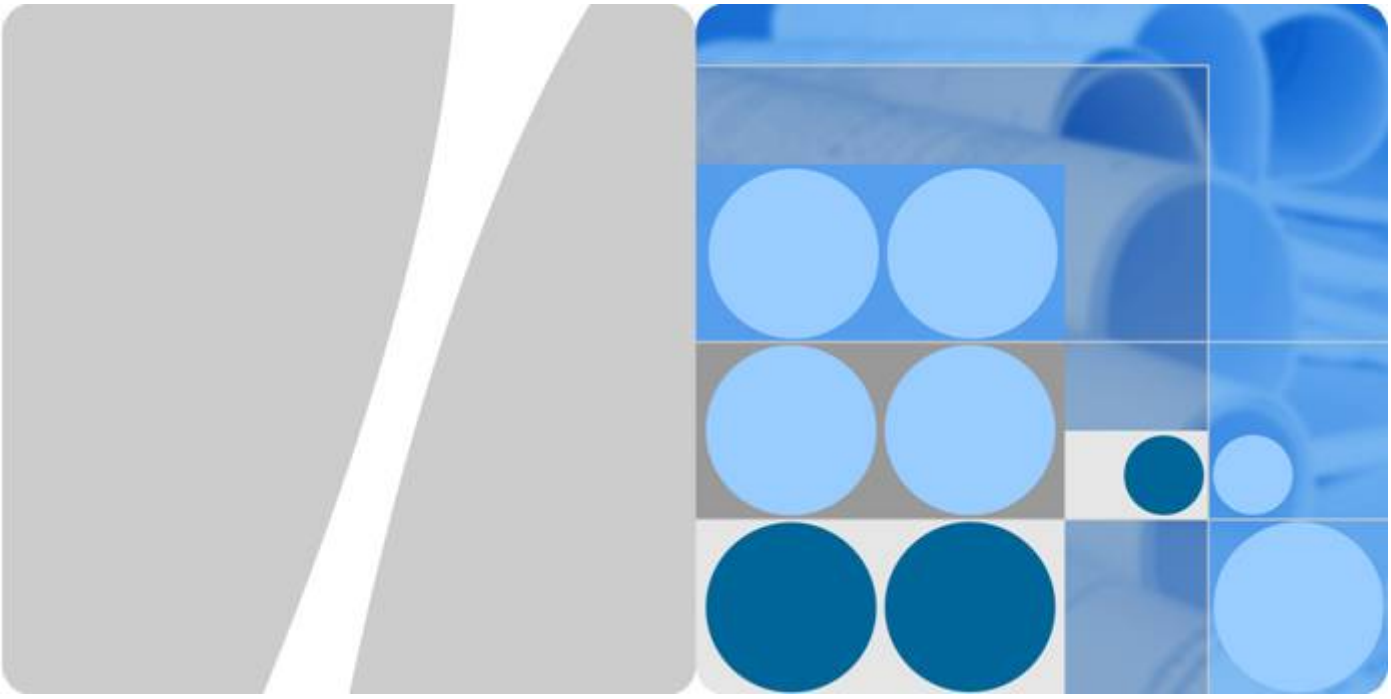


S9700 核心路由交换机

V200R001C00

特性描述-QoS

文档版本　01

发布日期　2012-03-15

华为技术有限公司

HUAWEI

版权所有 © 华为技术有限公司 2012。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

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： 4008302118

前 言

读者对象

本文档针对 QoS 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。

华为专有和保密信息
版权所有 © 华为技术有限公司

命令行格式约定

格式	意义
粗体	命令行关键字（命令中保持不变、必须照输的部分）采用**加粗**字体表示。
斜体	命令行参数（命令中必须由实际值进行替代的部分）采用*斜体*表示。
[ ]	表示用“[ ]”括起来的部分在命令配置时是可选的。
{ x | y | ... }	表示从两个或多个选项中选取一个。
[x | y | ...]	表示从两个或多个选项中选取一个或者不选。
{ x | y | ... }*	表示从两个或多个选项中选取多个，最少选取一个，最多选取所有选项。
[x | y | ...]*	表示从两个或多个选项中选取多个或者不选。
&<1-n>	表示符号&前面的参数可以重复 1 ～ n 次。
#	由“#”开始的行表示为注释行。

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 (2012-03-15)

第一次正式发布。

版权所有 © 华为技术有限公司

目 录

版权所有 © 华为技术有限公司

1 QoS

关于本章

版权所有 © 华为技术有限公司

1.1 介绍

定义

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，在 Internet 中，QoS 用于评估网络传送分组的服务能力。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，是对分组投递过程中可为延迟、延迟抖动、丢包率等核心需求提供支持的服务能力的评估。

- 带宽

  又可称为吞吐量，表示一定时间内业务流的平均速率，单位通常是 kbit/s。

- 时延

  表示业务流穿过网络时需要的平均时间。对于网络中的一个设备来说，一般将时延的需求理解为几种等级。例如分为两种时延等级，通过优先队列的调度方法使得高优先级的业务尽可能快地获得服务，而低优先级的业务则需要等待没有高优先级业务时才能获得服务。

- 时延抖动

  表示业务流穿过网络的时间的变化。

- 丢包率

  表示业务流在传送过程中的丢失比率。由于现代的传输系统具有很高的可靠性，信息的丢失往往发生在网络出现拥塞时。最常见的情况是队列溢出导致分组丢失。

目的

- 传统的分组投递业务

  传统的 IP 网络无区别地对待所有的报文，网络设备处理报文采用的策略是先入先出 FIFO（First In First Out），它依照报文到达时间的先后顺序分配转发所需要的资源。所有报文共享网络和交换设备的带宽等资源，至于得到资源的多少完全取决于报文到达的时机。这种服务策略称作 BE（Best-Effort），它尽最大的努力将报文送到目的地，但对分组投递的延迟、延迟抖动、丢包率和可靠性等需求不提供任何承诺和保证。

  传统的 BE 服务策略只适用于对带宽、延迟性能不敏感的 WWW（World Wide Web）、文件传输、E-Mail 等业务。

- 新业务引发的新需求

  随着计算机网络的高速发展，越来越多的网络接入 Internet。Internet 无论从规模、覆盖范围和用户数量都拓展得非常快。越来越多的用户使用 Internet 作为数据传输的平台，开展各种应用。同样地，服务提供商也希望通过新业务的开展来增加收益。除了传统的 WWW、E-Mail、FTP（File Transfer Protocol）应用外，用户还尝试在 Internet 上拓展新业务，比如远程教学、远程医疗、可视电话、电视会议、视频点播等。企业用户也希望通过 VPN（Virtual Private Network）技术，将分布在各地的分支机构连接起来，开展一些事务性应用，比如访问公司的数据库或通过 Telnet 管理远程设备。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、视频点播需要高带宽、低延迟和低延迟抖动的保证。事务处理、Telnet 等关键任务虽然不一定要求高带宽，但非常注重低延迟，在拥塞发生时要求优先获得处理。

  新业务的不断涌现对 IP 网络的服务能力提出了更高的要求，用户已不再满足于能够简单地将报文送达目的地，而是还希望在投递过程中得到更好的服务，诸如支持

版权所有 © 华为技术有限公司

为用户提供专用带宽、减少报文的丢失率、管理和避免网络拥塞、调控网络的流量、设置报文的优先级等。

所有这些，都要求网络应当具备更为完善的服务能力。

1.2 参考标准和协议

与 QoS 特性相关的参考资料清单如下：

文档	描述	备注
RFC 2474	Differentiated Services Field	-
RFC 2475	Architecture for Differentiated Services	-
RFC 2597	A Single Rate Three Color Marker	-
RFC 2598	An Expedited Forwarding PHB	-

1.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

表 1-1 QoS 特性的版本支持

产品	版本
S9700	V200R001

1.4 原理描述

1.4.1 拥塞的产生、影响和对策

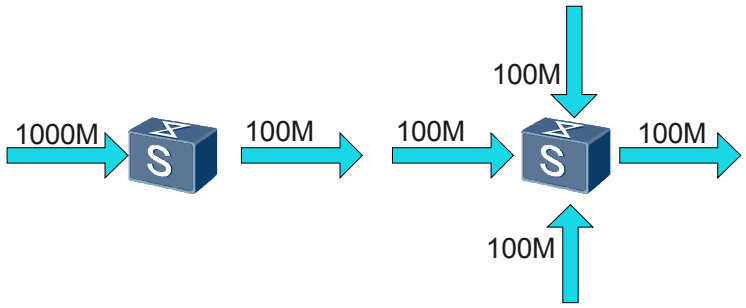
传统网络所面临的服务质量问题，主要是由网络拥塞引起的。拥塞是指由于供给资源的相对不足而造成服务速率下降（引入了额外的延迟）的一种现象。

拥塞的产生

在 Internet 分组交换的复杂环境下，拥塞极为常见。

版权所有 © 华为技术有限公司

图 1-1 拥塞产生示意图



（1）不同速率接口流量拥塞　（2）相同速率接口流量的拥塞

以所示两种情况为例：

- 分组流从高速链路进入 S9700，由低速链路转发出去。
- 分组流从相同速率的多个接口同时进入 S9700，由一个相同速率接口转发出去。如果流量以线速到达，那么就会遭遇资源的瓶颈而导致拥塞。

不仅仅是链路带宽的瓶颈会导致拥塞，任何用以正常转发处理的资源（如处理器时间、缓冲区、内存资源）的不足，都会造成拥塞。此外，在某个时间内 S9700 对所到达的流量控制不力，导致超出可分配的网络资源，也是引发网络拥塞的一个因素。

拥塞的影响

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。

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。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，必须慎重加以对待。

对策

增加网络带宽是解决资源不足的一个直接途径，但增加的资源很快就会被海量的网络流量所吞没，所以增加资源并不能解决所有导致网络拥塞的问题。解决网络拥塞问题的一个更有效的办法是增加网络设备在流量控制和资源分配上的功能，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。

1.4.2 服务模型

服务模型，是指一组实现端到端 QoS 保证的方式，包括 Best Effort、IntServ 和 DiffServ 三种服务模型。

版权所有 © 华为技术有限公司

Best Effort 模型

Best Effort 模型（即尽力而为模型）是一个单一的服务模型，也是最简单的服务模型。应用程序可以在任何时候，发出任意数量的报文，而且不需要事先获得批准，也不需要通知网络。Best Effort 模型中，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 模型是 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，它通过先入先出（FIFO）调度方式来实现。

IntServ 模型

IntServ（Integrated Service）模型是一个综合服务模型，它的特点是在发送报文前要先向网络提出申请。这个请求是通过信令来完成的，一个实例是资源预留协议 RSVP（Resource Reservation Protocol）。应用程序首先通过 RSVP 信令通知网络它的 QoS 需求（如时延、带宽、丢包率等指标）。在收到资源预留请求后，传送路径上的网络节点实施许可控制（Admission control），验证用户的合法性并检查资源的可用性，决定是否为应用程序预留资源。

一旦认可并为应用程序的报文分配了资源，则只要应用程序的报文控制在流量参数描述的范围内，网络节点将承诺满足应用程序的 QoS 需求。预留路径上的网络节点可以通过执行报文的分类、流量监管、低延迟的排队调度等行为，来满足对应用程序的承诺。IntServ 模型常与组播应用结合，适用于需要保证带宽、低延迟的实时多媒体应用，如电视会议、视频点播等。

当前，采用 RSVP 协议的 IntServ 模型定义了两种业务类型：

- 保证型服务（Guaranteed Service）提供保证的带宽和时延限制来满足应用程序的要求。如 VoIP（Voice over IP）应用可以预留 10M 带宽和要求不超过 1 秒的时延。
- 负载控制型服务（Controlled-Load Service）保证即使在网络过载（overload）的情况下，仍能对报文提供类似 Best Effort 模型在未过载时的服务质量——即在网络拥塞的情况下，保证某些应用程序报文的低时延和低丢包率需求。

可以提供端到端的 QoS 投递服务是 IntServ 模型的最大优点。IntServ 模型的最大缺点是可扩展性不好。网络节点需要为每个资源预留维护一些必要的软状态（Soft State）信息；在与组播应用相结合时，还要定期地向网络发资源请求和路径刷新信息，以支持组播成员的动态加入和退出。

上述操作要耗费网络节点较多的处理时间和内存资源。在网络规模扩大时，维护的开销会大幅度增加，对网络节点特别是核心节点线速处理报文的性能造成严重影响。因此，IntServ 模型不适宜于在流量汇集的骨干网上大量应用。

DiffServ 模型

为了在 Internet 上针对不同的业务提供有差别的服务质量，IETF 定义了 DiffServ（Differentiated Service）模型。

DiffServ 模型是一种多服务模型，它可以满足不同的 QoS 需求。与 IntServ 模型不同，应用程序在发出报文前，通过设置报文头部的优先级字段，向网络中各设备通告自己的 QoS 需求，而不需要通知途经的网络设备为其预留资源。

DiffServ 模型中，网络不需要为每个流维护状态，它根据每个报文携带的优先级来提供特定的服务。可以用不同的方法来指定报文的 QoS，如 IP 报文的优先级（IP Precedence），报文的源地址和目的地址等。网络通过这些信息来进行报文的分类、流量整形、流量监管和队列调度。

版权所有 © 华为技术有限公司

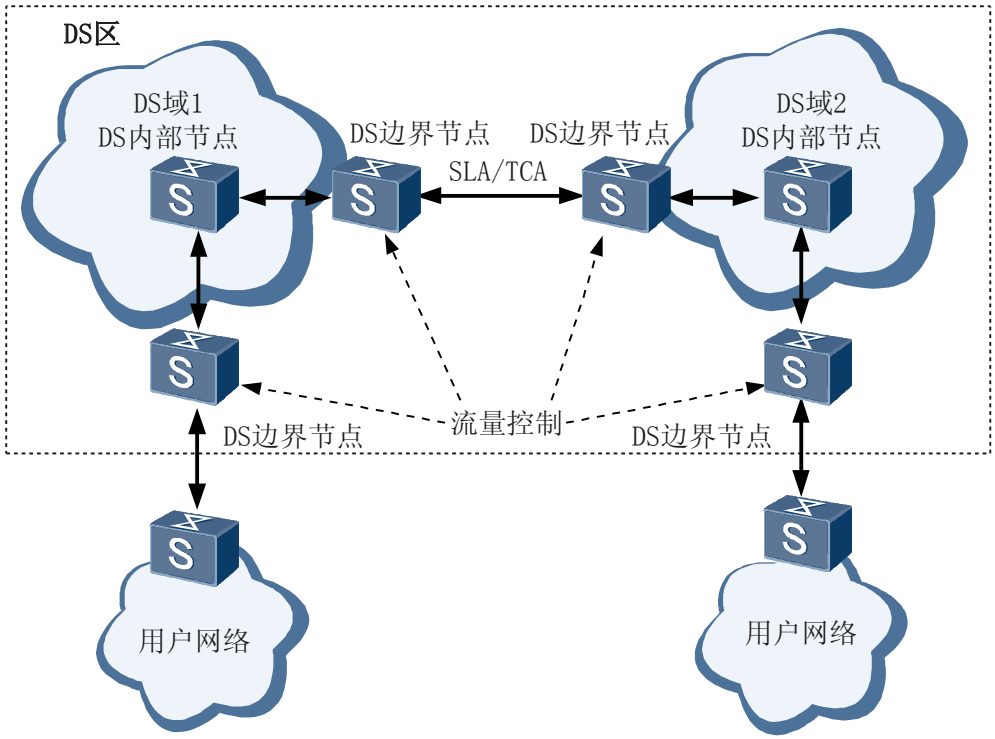
DiffServ 模型一般用来为一些重要的应用提供端到端的 QoS。通常在配置 DiffServ 模型后，边界设备通过报文的源地址和目的地址等信息对报文进行分类，对不同的报文设置不同的优先级，并标记在报文头部。而其他设备只需要根据设置的优先级来进行报文的调度。

1.4.3 DiffServ 实现技术

DiffServ 体系结构

DiffServ 体系结构定义了实现差分服务的系统模型和基本功能组件。在一个网络节点上，实现差分服务的基本功能组件包括逐跳行为 PHB（Per-Hop Behavior）、业务流的分类以及流量调整（包括流量监管与流量整形、拥塞避免与拥塞管理）等功能。差分服务建立在一种 DS 域模型之上，并规定了一个 DS 域的边界节点和内部节点。在边界节点上，对进入网络的业务流进行分类、流量调整和优先级标记，并按照 DS 域所支持的 PHB 组中的一个 PHB 转发。在内部节点上，将根据边界节点标记的 DSCP/802.1p 优先级所定义的 PHB 来选择该业务流的转发行为，为业务流分配带宽资源。

图 1-2 Diff-Serv 体系结构示意图



DiffServ 体系结构如所示，其中：

- DS 节点

  DS 节点指实现 DiffServ 功能的网络节点。DS 节点可分为 DS 边界节点和 DS 内部节点。

- DS 边界节点

  DS 边界节点负责连接另一个 DS 域或者连接一个没有 DS 功能的域的节点。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，以保证穿过此 DS 域的业务流被适当标记，并按照 DS 域所支持的 PHB 组中的一个 PHB 转发。

版权所有 © 华为技术有限公司

对于不同方向的业务流，DS 边界节点既可以是 DS 域的输入（Ingress）节点，又可以是 DS 域的输出（Egress）节点。业务流在 Ingress 节点处进入 DS 域，在 Egress 节点处离开 DS 域。Ingress 节点负责保证进入 DS 域的业务流符合本域和此节点直连的另一个域之间的服务等级协定 SLA（Service Level Agreements）或流量控制协定 TCA（Traffic Conditioning Agreement）。Egress 节点依据两个域之间的 TCA 细节，对转发到其直连的对等域的业务流执行流量调整功能。

- DS 内部节点

  DS 内部节点负责连接同一 DS 域中的其他 DS 内部节点或 DS 边界节点。DS 内部节点负责根据 IP 报头中的 DS 字段或 VLAN 报文的 802.1p 字段所定义的 PHB 来为该业务流选择转发行为。无论是 DS 边界节点还是 DS 内部节点都必须能够根据业务流的 DSCP 或者 802.1p 选择相应的 PHB 进行转发操作。

- DS 域

  DiffServ 模型的实现基于 DS 域，DS 域由一组采用相同的服务提供策略和实现了相同 PHB 组集合的相连 DS 节点组成。一个 DS 域由 DS 边界节点和 DS 内部节点组成，边界节点构成了 DS 域的边界，内部节点构成了 DS 域的核心。

- SLA

  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）。

- TCA

  TCA 指用户与服务提供商签署的关于业务分类准则、业务模型及相应处理的协定。去掉了商业条款的 TCA 称为 TCS（Traffic Conditioning Specification）一个 SLA 中可以包含 TCA。对于业务的处理而言，SLA 或 SLS 指明的是比较一般的内容，例如采用什么样的机制。而 TCA 或 TCS 则比较具体，例如具体的带宽要求。

- DS 区

  一个或多个邻接的 DS 域统称为 DS 区。DS 区可以支持贯穿区内多个 DS 域的分类业务。DS 区中的 DS 域可能支持不同的 PHB 组，和 QoS 字段到 PHB 的映射规则。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 和 TCA 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。

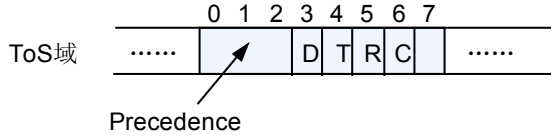
记录 QoS 信息的字段

为了在 Internet 上针对不同的业务提供有差别的 QoS 服务质量，人们根据报文头中的某些字段记录 QoS 信息，从而让网络中的各设备根据此信息提供有差别的服务质量。这些和 QoS 相关的报文字段包括：

- IP 报文头中的 Precedence 字段

  根据 RFC791 定义，IP 报文头 ToS（Type of Service）域中的 Precedence 字段标识了报文的优先级，IP 报文中的 Precedence 字段位置如图 1-3 所示。

图 1-3 IP 报文中的 Predecence 字段



版权所有 © 华为技术有限公司

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。最高优先级是 7 或 6，经常是为路由选择或更新网络控制通信保留的，用户级应用仅能使用 0 ～ 5 级。

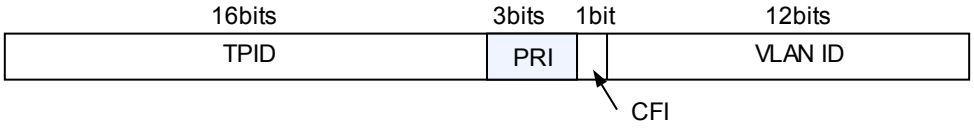
除了 Predecence 字段外，ToS 域中还包括 D、T、R 三个比特：

- 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）。
- ToS 域中的比特 6 和 7 保留。

- VLAN 帧头中的 802.1p 优先级

  通常二层交换机之间交互 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，VLAN 帧中的 PRI 字段位置如图 1-4 所示。

图 1-4 VLAN 帧中的 802.1p 优先级



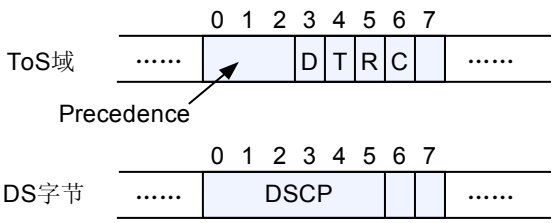
在 802.1Q 头部中包含 3 比特长的 PRI 字段。PRI 字段表示 8 个传输优先级，按照优先级从高到低顺序取值为 7、6、……、1 和 0。此外，802.1Q 头部中还包含 TPID（Tag Protocol Identifier）、CFI（Canonical Format Indicator）和 VLAN ID 字段。

- DSCP 字段

  在 DiffServ 方案中，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 DSCP（Differentiated Service Code Point）。在网络中根据实施好的 QoS 机制来区分每一类通信（依据每个报文的 DSCP 值），并为之提供服务（包括资源分配、分组丢弃策略等）。

  RFC1349 重新定义了 IP 报文中的 ToS 域，增加了 C 比特，表示传输开销（Monetary Cost）。之后，IETF DiffServ 工作组在 RFC2474 中将 IP 报文头 ToS 域中的比特 0 ～ 5 重新定义为 DSCP，并将 ToS 域改名为 DS（Differentiated Service）字节。DS 字节格式如图 1-5 所示。

图 1-5 DSCP 字段格式



DS 字节的低 6 位（比特 0 ～ 5）用作区分服务代码点 DSCP，高 2 位（比特 6、7）是保留位。DSCP 中的低 3 位（比特 0 ～ 2）是类选择代码点 CSCP（Class Selector

版权所有 © 华为技术有限公司

Code Point），它表示了一类 DSCP。网络中支持 DiffServ 技术的各设备根据 DSCP 值选择相应的转发行为。

Diff-Serv 模型

- DiffServ 模型的基本思想

  为了在 Internet 上针对不同的业务提供有差别的服务质量，IETF 定义了 DiffServ 模型。

  在采用 DiffServ 模型的应用中，设备在发送报文前通过设置 IP 报文头部 ToS 域中的优先级字段，向网络中各设备通告自己的 QoS 需求。报文传播路径上的各设备通过分析 IP 报文头来获知报文的服务需求类别。在实施 DiffServ 时，接入设备需要首先对报文进行分类，并在 IP 报文头部标记服务类别。下游的设备只需简单地识别报文中的这些服务类别，并按照要求转发报文。因此，DiffServ 模型是一种基于报文流的 QoS 解决方案。

- 标准的 PHB 行为

  IETF Diff-Serv 工作组将网络节点对报文实施调度、监管等转发行为定义为 PHB（Per-Hop Behaviors）。网络中各设备根据 DSCP 值选择相应的 PHB 行为。

  目前，IETF 定义了四种标准的 PHB：CS（Class Selector）、EF（Expedited Forwarding）、AF（Assured Forwarding）和 BE（Best-Effort），并将 BE 作为缺省 PHB。

  - CS

    CS 表示类选择码，代表的服务等级与 IP Precedence 相同，DSCP 取值为“XXX000”，X 为 0 或 1。

  - EF

    EF 表示加速转发行为，代表 DiffServ 网络中最高的服务质量。应用于低丢包率、低时延、高带宽的业务，信息流的在任何情况下都能获得等于或大于设定的速率。DSCP 取值为“101110”。

  - AF

    AF 表示确保转发行为，应用于带宽保证、低时延的关键数据业务。对未超出带宽限度的流量提供转发质量保证，对超出限度的流量降低服务等级后继续转发，而不是直接丢弃。

    根据 RFC 2597 的描述，目前定义了四类 AF，每类 AF 用“AFi”表示，其中 1<=i<=4，即这四类 AF 是：AF1、AF2、AF3、AF4。并且在每类 AF 中，又定义了 3 种丢弃优先级，每种丢弃优先级用“AFij”表示，其中 1<=j<=3，“j”值越大，表明丢弃优先级越高。各类 AF 业务对应的 DSCP 取值见**表 1-2**。

表 1-2 各类 AF 业务对应的 DSCP 值

丢弃优先级	AF1	AF2	AF3	AF4
低	AF11 001010	AF21 010010	AF31 011010	AF41 100010
中	AF12 001100	AF22 001100	AF32 011100	AF42 100100
高	AF13 001110	AF23 010110	AF33 011110	AF43 100110

版权所有 © 华为技术有限公司

- BE

  BE 表示尽力而为转发行为，应用于不需要严格 QoS 保证的尽力发送业务，只关注可达性，其他方面不做任何要求，如传统的 IP 分组投递服务。DSCP 取值为“000000”。

DiffServ 功能组件

流分类、流量监管、流量整形、拥塞管理和拥塞避免是构造有区别地实施服务的基石，它们主要完成如下功能：

- 流分类：依据一定的匹配规则识别出对象。流分类是有区别地实施服务的前提。
- 流量监管：对进入交换机的特定流量的规格进行监管。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。
- 流量整形：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交换机可供给的网络资源，避免不必要的报文丢弃和拥塞。
- 拥塞管理：网络拥塞时必须采取的解决资源竞争的措施。通常是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。
- 拥塞避免：过度的拥塞会对网络资源造成损害。拥塞避免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。

在这些功能组件中：流分类是基础，它依据一定的匹配规则识别出报文，是有区别地实施服务的前提；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，是有区别地提供服务具体体现。

1.4.4 基于类的 QoS

基于类的 QoS 是指通过对流量按照某种规则进行分类，并对同种类型的流量关联某种动作，形成某种策略，将该策略应用后，可实现基于类的流量监管、重新标记优先级、重定向等功能。

流分类

流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。流分类包括简单流分类和复杂流分类。

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 DSCP（DiffServ Code Point）字段对报文进行粗略的分类，以识别出具有不同优先级特征的流量。

- 简单流分类的分类规则

  S9700 可以依据以下信息对报文进行简单流分类：

  - IP 报文的 DSCP 优先级
  - VLAN 报文的 802.1p 优先级
  - IP 报文中的 ip-precedence
  - MPLS 报文的 EXP 优先级
- 简单流分类的部署思路

版权所有 © 华为技术有限公司

DS 内部节点负责根据 IP 报头中的 ToS 字段或 VLAN 报文的 802.1p 字段所定义的 PHB 来为该业务流选择转发行为。因此，要求 DS 内部节点能支持简单流分类以选择业务流的转发行为。

按照接口入方向和出方向的不同，简单流分类可以分为上行和下行：

- 上行简单流分类配置在接口的入方向上，根据 DiffServ 域中定义的映射关系将报文的优先级映射到 PHB 行为并标记颜色。

  通过上行配置简单流分类可以区分不同的业务（如语音、视频、数据等）。拥塞管理、队列调度时，不同的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队列，保证低延时。

  上行若不配置简单流分类，Tag 报文按照各自的 802.1p 优先级入队列，Untag 报文按照接口配置的缺省 802.1p 优先级入队列。

- 下行简单流分类配置在接口的出方向上，根据 DiffServ 域中定义的映射关系将报文的 PHB 行为和被标记的颜色映射到优先级。

  通过下行配置简单流分类，可以实现重标记报文优先级的功能，下游设备将能够根据重标记的优先级为报文提供不同的 QoS 保证。

  下行若不配置简单流分类，系统按照缺省的优先级映射关系重标记报文优先级。

复杂流分类

复杂流分类是指根据报文携带的二三层信息或者借助 ACL（Access Control List）规则，根据 IP 五元组（源 IP 地址、目的 IP 地址、源端口号、目的端口号、报文类型）、TCP SYN 等信息对报文进行分类，依据该分类为报文提供相应的服务质量。

- 复杂流分类的分类规则

  S9700 可以根据以下二层信息对报文进行复杂流分类：

  - VLAN 报文外层 Tag 的 ID 信息
  - VLAN 报文内层 Tag 的 ID 信息
  - VLAN 报文外层 Tag 的 802.1p 优先级
  - VLAN 报文内层 Tag 的 802.1p 优先级
  - VLAN 报文的双层 Tag 信息
  - 源 MAC 地址
  - 目的 MAC 地址
  - 出接口
  - 入接口
  - 基于二层封装的协议字段

  S9700 也可以根据以下三层信息对报文进行复杂流分类：

  - IP 报文的 DSCP 优先级
  - IP 报文的 IP 优先级
  - MPLS 报文的 EXP 优先级
  - TCP 报文的 TCP SYN 标志
  - IP 协议类型（即 IPv4 协议或 IPv6 协议）

  S9700 还可以借助 ACL 依据以下信息对报文进行复杂流分类：

  - IP 优先级或 DSCP 优先级
  - 源 IP 地址前缀

版权所有 © 华为技术有限公司

- 目的 IP 地址前缀
- IP 报文承载的协议号
- 分片标志
- TCP SYN 标志
- TCP 或 UDP 源端口号或端口范围
- TCP 或 UDP 目的端口号或端口范围

- 复杂流分类的部署思路

  DS 边界节点担负着流量监管、防止窃取和拒绝服务（Theft and Denial of Service）的攻击、实施流过滤等访问控制功能。因此，要求 DS 边界节点能支持复杂流分类以识别出更具体的流。

优先级映射

优先级映射用来实现 QoS 优先级与设备内部的服务等级（又称为内部优先级，包括 PHB 行为/颜色）之间的转换。

不同的报文使用不同的 QoS 优先级，例如 VLAN 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，在报文进设备时，需要将报文携带的 QoS 优先级映射到设备内部的 PHB 行为和颜色；在报文出设备时，需要将内部的 PHB 行为和颜色映射为 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。

优先级映射基于简单流分类实现从 QoS 优先级到服务等级（包括 PHB 行为、颜色）或从服务等级到 QoS 优先级的映射，并利用 DS 域来管理和记录 QoS 优先级和服务等级之间的映射关系。

S9700 提供两种优先级映射模式：

- IP 报文的 DSCP 优先级映射

  在业务流进入设备时，S9700 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。

  在业务流流出设备时，S9700 根据报文的服务等级，查找服务等级到 DSCP 优先级的映射表，将设备服务等级转换为对接网络的 QoS 优先级，以便对接的网络设备能够根据 QoS 优先级提供相应的服务质量。

- VLAN 报文的 802.1p 优先级映射

  在业务流进入设备时，S9700 根据报文的 802.1p 优先级（对于 Untag 报文，S9700 使用接口上配置的缺省优先级）对报文进行分类，并查找 802.1p 优先级到服务等级映射表，为报文标记服务等级，以提供不同的服务质量。

  在业务流流出设备时，S9700 根据报文的服务等级，查找服务等级到 802.1p 优先级的映射表，将设备服务等级转换为对接网络的 QoS 优先级，以便对接的网络设备能够根据 QoS 优先级提供相应的服务质量。

系统中存在一个已建立的 DS 域：**default** 域，用户可以选择采用系统缺省的 DS 域、也可以重新定义 DS 域，S9700 最多允许创建 7 个新的 DS 域。应用时，将选择的 DS 域绑定在相应接口上，系统对出入该接口的流进行相应的优先级映射。

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。

版权所有 © 华为技术有限公司

在 S9700 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、流镜像、安全和流量统计。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。S9700 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 S9700 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接口板 CPU、指定接口、指定的下一跳地址或下一跳标签 LSP（Label Distribution Path）。

  S9700 支持指定多个下一跳。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。

- 安全

  安全动作是指对报文实施逆向安全检查 URPF（Unicast Reverse Path Forwarding）。安全动作本身不是 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。

- 流量统计

  流量统计用于统计指定业务流的数据报文，它统计的是匹配流分类的报文中通过和丢弃的报文数量和字节数。

  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。

流策略

流策略是将复杂流分类和流行为关联后形成的完整的 QoS 策略。流策略可以应用到接口、全局、接口板或者 VLAN 中，从而将流策略中绑定的流分类和流行为应用到这些地方。

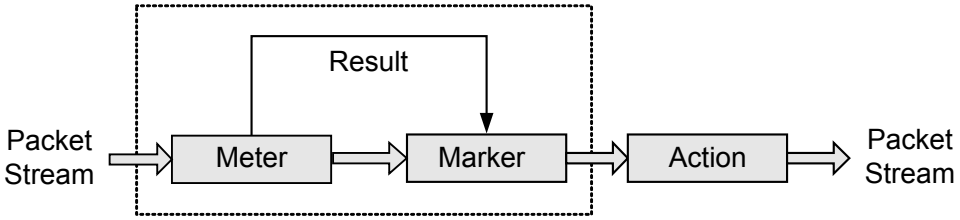
1.4.5 流量监管

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运营商的利益。S9700 使用 CAR 来进行流量监管，既可以通过配置和应用 QoS CAR 模板来实现，也可以通过配置和应用流策略来实现。

版权所有 © 华为技术有限公司

流量监管的原理

图 1-6 流量监管组件

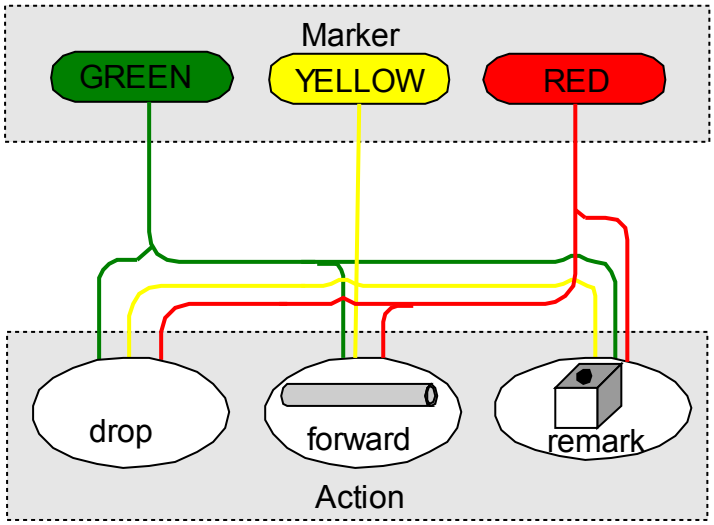


如图 1-6 所示，S9700 的流量监管由三部分组成：

- 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：
  - forward：对测量结果为“符合”的报文继续转发。
  - remark：修改报文内部优先级后再转发。
  - drop：对测量结果为“不符合”的报文进行丢弃。

  默认情况下，green、yellow 进行转发，red 报文丢弃。

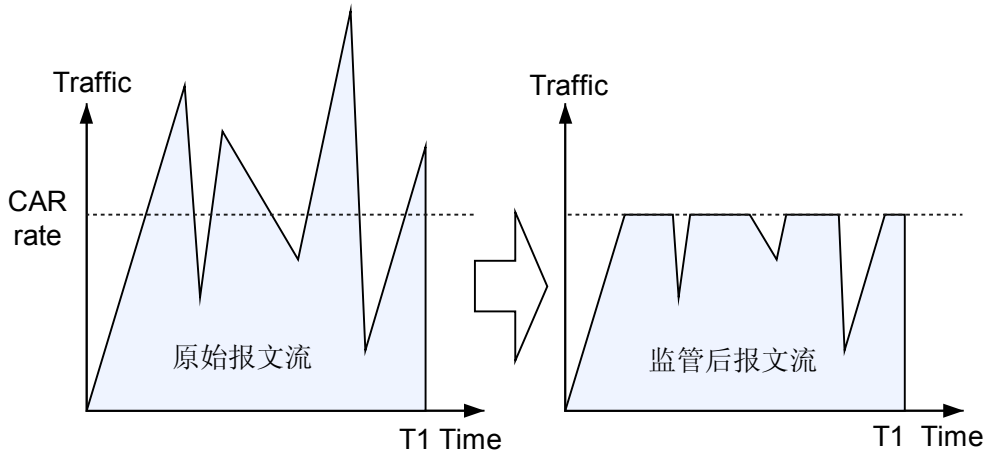
图 1-7 流量监管动作



经过流量监管，如果某流量速率超过标准，S9700 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，报文被丢弃。如图 1-8 显示了流量监管时网络流量被限制在规定的速率范围内的速率曲线图，超过速率的部分被完全削除。

版权所有 © 华为技术有限公司

图 1-8 流量监管的报文流曲线图



1.4.6 流量整形

概述

当下游设备的接口速率小于上游设备的接口速率或发生突发流量，在下游设备接口处可能出现流量拥塞的情况，此时用户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，从而解决下游设备的拥塞问题。

S9700 同时支持队列整形和端口整形。

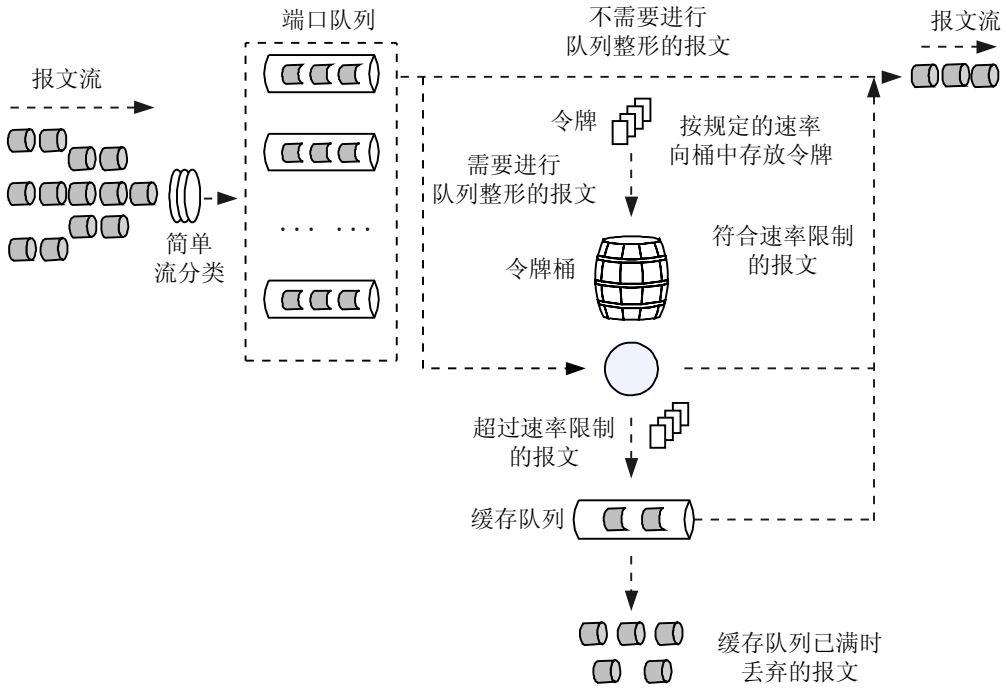
处理流程

流量整形是一种应用于端口或端口队列的流量控制技术，可以对从接口上经过的所有报文或对从接口上经过的某类（基于简单流分类）报文进行速率限制。

流量整形也是通过令牌桶进行流量控制。下面以端口队列整形为例介绍流量整形的处理流程，其处理流程如图 1-9 所示。

版权所有 © 华为技术有限公司

图 1-9 流量整形处理流程图



具体处理流程如下：

1. 当报文到来的时候，首先对报文进行分类，使报文进入不同的端口队列。
2. 若报文进入的端口队列没有配置队列整形功能，则直接发送该端口队列的报文；否则，进入下一步处理。
3. 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。
4. 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。

端口队列整形后，如果该端口同时配置了端口整形，则系统还要按照端口整形速率对报文流进行速率控制。其处理流程与端口队列整形相似，但不需要步骤 **1** 和步骤 **2**。

流量整形与流量监管区别

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控制时，直接丢弃不符合速率要求的报文。而流量整形则会将不符合速率要求的报文先行缓存，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。

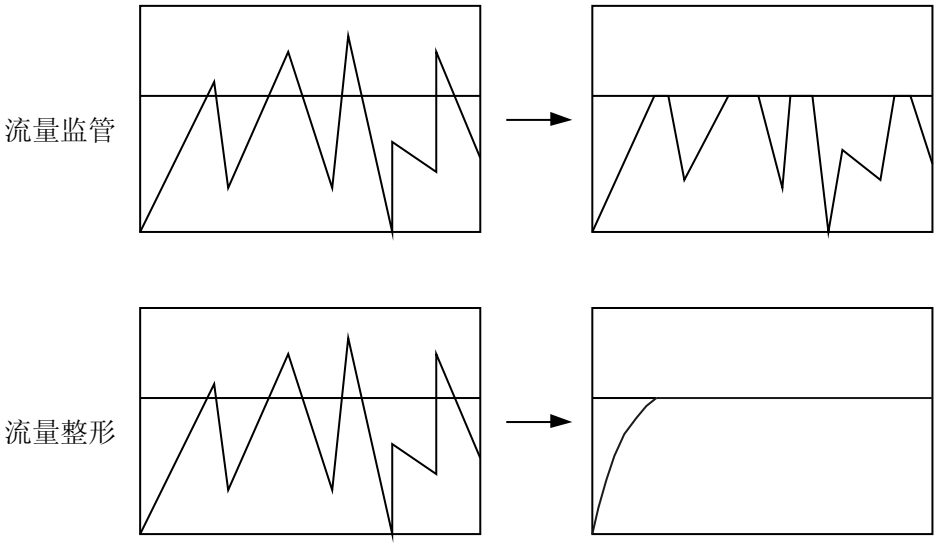
版权所有 © 华为技术有限公司

表 1-3 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，需要较多的缓冲资源缓存报文。
流量监管	支持重标记，不需使用额外的缓冲。	较多丢弃报文，可能引发重传。

图 1-10 说明了流量监管与流量整形的区别。

图 1-10 流量监管与流量整形的区别



1.4.7 拥塞管理

当时延敏感业务要求得到比非时延敏感业务更高质量的 QoS 服务时，而且网络中间歇性的出现拥塞，此时需要进行拥塞管理；如果任何时候都出现拥塞，则需要增加带宽。拥塞管理一般采用排队技术，使用不同的调度算法来发送队列中的报文流。

根据排队和调度策略的不同，S9700 上的拥塞管理技术分为 PQ、DRR、PQ+DRR、WRR 和 PQ+WRR。每种调度算法都是为了解决特定网络流量的问题，并对带宽资源的分配、延迟、抖动等有着十分重要的影响。

在 S9700 上，每个接口出方向上都拥有 8 个队列，以队列索引号进行标识，队列索引号分别为 7、6、……、1 和 0（PQ 调度时，7 队列优先级最高，0 队列优先级最低）。S9700 根据本地优先级和队列之间的映射关系，自动将分类后的报文流送入各队列，然后按照各种队列调度机制进行调度。

- PQ 调度

  PQ 调度，针对于关键业务类型应用设计，PQ 调度算法维护一个优先级递减的队列系列并且只有当更高优先级的所有队列为空时才服务低优先级的队列。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（如 E-Mail）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。

版权所有 © 华为技术有限公司

如图 1-11 所示，Queue 7 比 Queue 6 具有更高的优先权，Queue 6 比 Queue 5 具有更高的优先权，依次类推。只要链路能够传输分组，Queue 7 尽可能快地被服务。只有当 Queue 7 为空，调度器才考虑 Queue 6。当 Queue 6 有分组等待传输且 Queue 7 为空时，Queue 6 以链路速率接受类似地服务。当 Queue 7 和 Queue 6 为空时，Queue 5 以链路速率接收服务，以此类推。

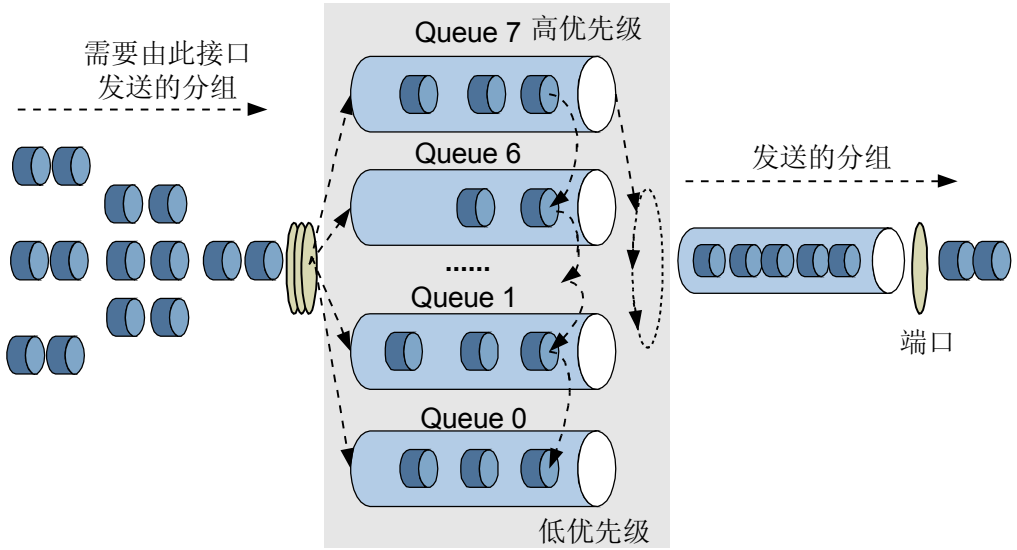
PQ 调度算法对低时延业务非常有用。假定数据流 X 在每一个节点都被映射到最高优先级队列，那么当数据流 X 的分组到达时，则分组将得到优先服务。

然而 PQ 调度机制会使低优先级队列中的报文由于得不到服务而“饿死”。例如，如果映射到 Queue 7 的数据流在一段时间内以 100%的输出链路速率到达，调度器将从不为 Queue 6 及以下的队列服务。

避免队列饥饿需要上游设备精心规定数据流的业务特性以确保映射到 Queue 7 的业务流不超出输出链路容量的一定比例，这样可以使 Queue 7 常常为空，允许调度器为低优先级队列服务。

缺省情况下，S9700 的调度模式为 PQ。

图 1-11 PQ 调度示意图



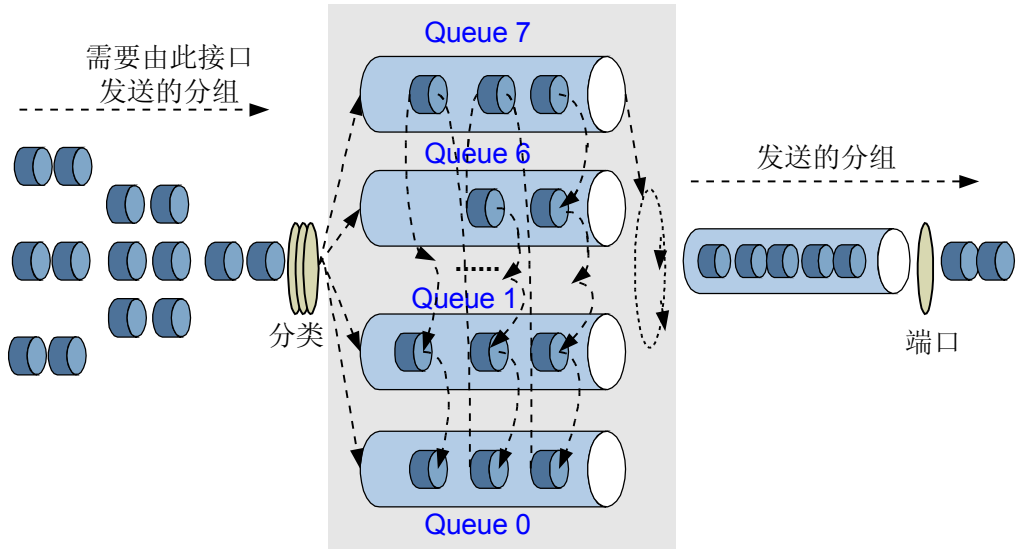
- WRR 调度

  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。实际上，RR 调度相当于权值为 1 的 WRR 调度。

  WRR 队列示意图如图 1-12 所示。

版权所有 © 华为技术有限公司

图 1-12 WRR 调度示意图



在进行 WRR 调度时，S9700 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。例如，用户根据需要为接口上 8 个队列指定的权值分别为 4、2、5、3、6、4、2 和 1，按照 WRR 方式进行调度的结果请参见**表 1-4** 所示。

表 1-4 WRR 调度的结果

队列索引	Q7	Q6	Q5	Q4	Q3	Q2	Q1	Q0
队列权值	4	2	5	3	6	4	2	1
参加第 1 轮调度的队列	Q7	Q6	Q5	Q4	Q3	Q2	Q1	Q0
参加第 2 轮调度的队列	Q7	Q6	Q5	Q4	Q3	Q2	Q1	-
参加第 3 轮调度的队列	Q7	-	Q5	Q4	Q3	Q2	-	-

版权所有 © 华为技术有限公司

队列索引	Q7	Q6	Q5	Q4	Q3	Q2	Q1	Q0
参加第 4 轮调度的队列	Q7	-	Q5	-	Q3	Q2	-	-
参加第 5 轮调度的队列	-	-	Q5	-	Q3	-	-	-
参加第 6 轮调度的队列	-	-	-	-	Q3	-	-	-
参加第 7 轮调度的队列	Q7	Q6	Q5	Q4	Q3	Q2	Q1	Q0
参加第 8 轮调度的队列	Q7	Q6	Q5	Q4	Q3	Q2	Q1	-
参加第 9 轮调度的队列	Q7	-	Q5	Q4	Q3	Q2	-	-
参加第 10 轮调度的队列	Q7	-	Q5	-	Q3	Q2	-	-
参加第 11 轮调度的队列	-	-	Q5	-	Q3	-	-	-

版权所有 © 华为技术有限公司

队列索引	Q7	Q6	Q5	Q4	Q3	Q2	Q1	Q0
参加第 12 轮调度的队列	-	-	-	-	Q3	-	-	-

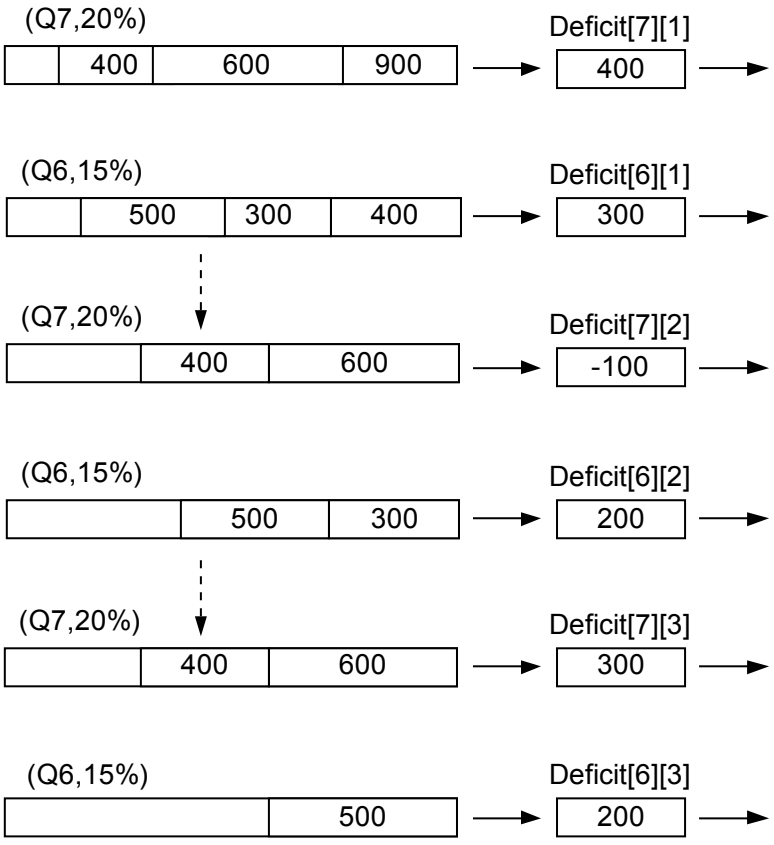
从统计上看，各队列中的报文流被调度的次数与该队列的权值成正比，权值越大被调度的次数相对越多。由于 WRR 调度的以报文为单位，因此每个队列没有固定的带宽，同等调度机会下大尺寸报文获得的实际带宽要大于小尺寸报文获得的带宽。

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。WRR 队列还有一个优点是，虽然多个队列的调度是轮询进行的，但对每个队列不是固定地分配服务时间片——如果某个队列为空，那么马上换到下一个队列调度，这样带宽资源可以得到充分的利用。但 WRR 调度无法使低延时需求业务得到及时调度。

- DRR 调度

  DRR（Deficit Round Robin）调度同样也是 RR 的扩展，相对于 WRR 来言，解决了 WRR 只关心报文，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的问题，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 队列调度如图 1-13 所示。

图 1-13 DRR 调度示意图



版权所有 © 华为技术有限公司

假设用户配置各队列权重为 40、30、20、10、40、30、20、10（依次对应 Q7、Q6、Q5、Q4、Q3、Q2、Q1、Q0），如图 1-13 所示，队列 Q7、Q6 的能够分别获取 20%、15%的带宽，当前 Q7 队列中有 400bytes、600bytes、900bytes 的报文，Q6 队列中有 500bytes、300bytes、400bytes 的报文。每次调度时，系统按权重为各队列分配带宽，假设 Q7 队列为 400bytes/s，Q6 队列为 300bytes/s。Deficit 表示每次调度时各队列的带宽赤字。

- 第一次调度

  Deficit[7][1] = 400，Deficit[6][1] = 300，从 Q7 队列取出 900bytes 报文发送，从 Q6 队列取出 400bytes 发送；发送后，Deficit[7][1] = -500，Deficit[6][1] = -100。

- 第二次调度

  Deficit[7][2] = -500+400 = -100，Deficit[6][2] = -100+300 = 200，由于 Q7 队列 Deficit 值为负，Q7 队列不会被调度；从 Q6 队列取出 300bytes 发送；发送后，Deficit[6][2]= -100。

- 第三次调度

  系统设置 Deficit[7][3] = -100+400 = 300，Deficit[6][3] = -100+300 = 200，由于 Q7 队列 Deficit 值变为正，Q7 队列再次被调度，从 Q7 队列取出 600bytes 报文发送，从 Q6 队列取出 500bytes 发送；发送后，Deficit[7][1] = -300，Deficit[6][1] = -200。如此这样循环调度，最终 Q7、Q6 队列获取的带宽将分别占总带宽的 20%、15%，因此，用户能够通过设置权重获取想要的带宽。

但 DRR 调度仍然没有解决 WRR 调度中低延时需求业务得不到及时调度的问题。

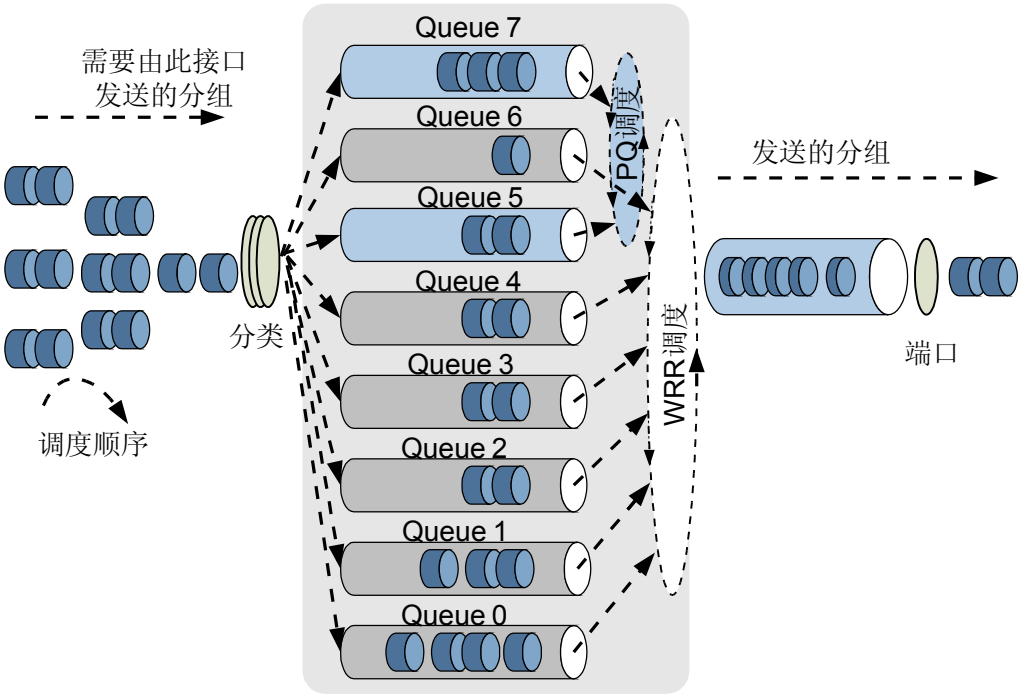
- PQ+WRR 调度

  PQ 调度和 WRR 调度各有优缺点，为了克服单纯采用 PQ 调度或 WRR 调度时的缺点，PQ+WRR 调度以发挥两种调度的各自优势，不仅可以通过 WRR 调度可以让低优先级队列中的报文也能及时获得带宽，而且可以通过 PQ 调度可以保证了低延时需求的业务能优先得到调度。

  在 S9700 上，用户可以配置队列的 WRR 参数，根据配置将接口上的 8 个队列分为两组，一组（例如 Queue-7、Queue-5）采用 PQ 调度，另一组（例如 Queue-6、Queue-4、Queue-3、Queue-2、Queue-1 和 Queue0 队列）采用 WRR 调度。PQ +WRR 调度示意图如图 1-14 所示。

版权所有 © 华为技术有限公司

图 1-14 PQ+WRR 混合调度示意图



在调度时，S9700 首先按照 PQ 方式调度 Queue 7、Queue 5 队列中的报文流，只有这些队列中的报文流全部调度完毕后，才开始以 WRR 方式循环调度其他队列中的报文流。Queue 6、Queue 4、Queue 3、Queue 2、Queue 1 和 Queue 0 队列包含自己的权值。重要的协议报文和有低延时需求的业务报文应放入采用 PQ 调度的队列中，得到优先调度的机会，其余报文放入以 WRR 方式调度的各队列中。

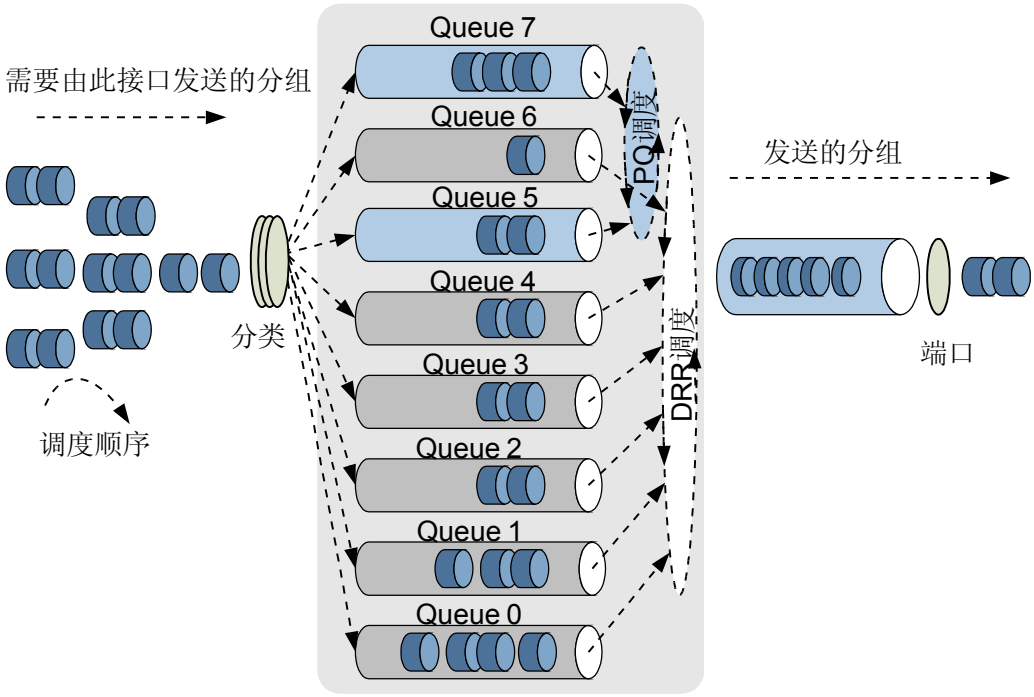
- PQ+DRR 调度

  与 PQ+WRR 相似，其集合了 PQ 调度和 DRR 调度各有优缺点。单纯采用 PQ 调度时，低优先级队列中的报文流长期得不到带宽，而单纯采用 DRR 调度时低延时需求业务（如语音）得不到优先调度，如果将两种调度方式结合起来形成 PQ+DRR 调度，不仅能发挥两种调度的优势，而且能克服两种调度各自的缺点。

  S9700 接口上的 8 个队列被分为两组，用户可以指定其中的某几组队列进行 PQ 调度，其他队列进行 DRR 调度。

版权所有 © 华为技术有限公司

图 1-15 PQ+DRR 调度示意图



如图 1-15 所示，在调度时，S9700 首先按照 PQ 方式优先调度 Queue 7 和 Queue 5 队列中的报文流，只有这些队列中的报文流全部调度完毕后，才开始以 DRR 方式调度 Queue 6、Queue 4、Queue 3、Queue 2、Queue 1 和 Queue 0 队列中的报文流。其中，Queue 6、Queue 4、Queue 3、Queue 2、Queue 1 和 Queue 0 队列拥有自己的保证带宽和峰值带宽。

重要的协议报文以及有低延时需求的业务报文应放入需要进行 PQ 调度的队列中，得到优先调度的机会，其他报文放入以 DRR 方式调度的各队列中。

1.4.8 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。

本节介绍拥塞避免的基本知识，具体包括：

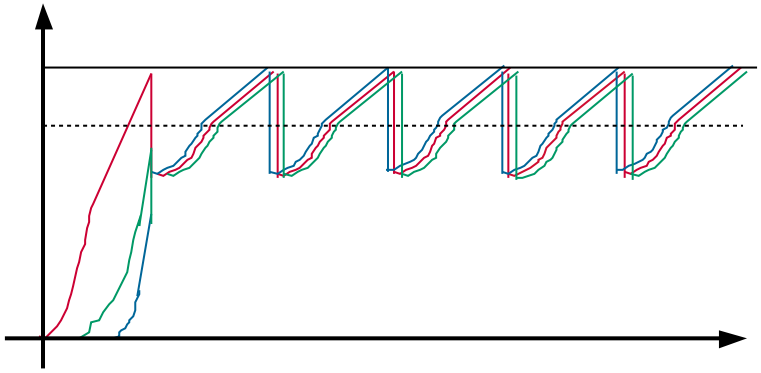
- 传统的尾部丢包策略

  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，导致 TCP 连接始终无法建立。所谓 TCP 全局同步现象如图，三种颜色表示三条 TCP 连接，当同时丢弃多个 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而导致流量降低，之后又会在某个时间同时出现流量高峰，如此反复，使网络流量忽大忽小。

版权所有 © 华为技术有限公司

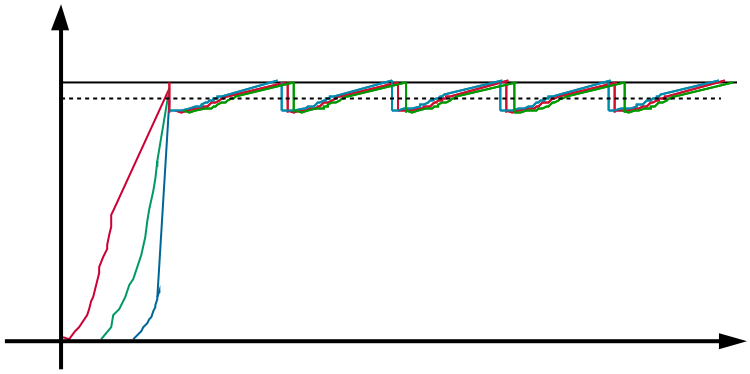
图 1-16 尾部丢包示意图



- WRED

  为避免 TCP 全局同步现象，出现了 RED（Random Early Detection）技术。RED 通过随机地丢弃数据报文，让多个 TCP 连接不同时降低发送速度，从而避免了 TCP 的全局同步现象。使 TCP 速率及网络流量都趋于稳定。

图 1-17 RED 算法示意图



基于 RED 技术，S9700 实现了 WRED（Weighted Random Early Detection）。在接口出队列上，WRED 根据报文的优先级及类型将报文分为四类，绿色、黄色、红色及 non-tcp，对这四种报文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，开始丢包，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，这样按照一定的丢弃概率主动丢弃队列中的报文，从而一定的程度上避免拥塞问题。

1.4.9 接口出方向限速

利用 LR（Line Rate，端口限速）可以在一个端口上限制发送报文的总速率。LR 也是采用令牌桶进行流量控制。如果在设备的某个端口上配置了 LR，所有经由该端口发送的报文首先要经过 LR 的令牌桶进行处理。如果令牌桶中有足够的令牌，则报文可以发

版权所有 © 华为技术有限公司

送；否则，报文将被丢弃。与流量监管相比，端口限速能够限制在端口上通过的所有报文。当用户只要求对所有报文限速时，使用端口限速比较简单。

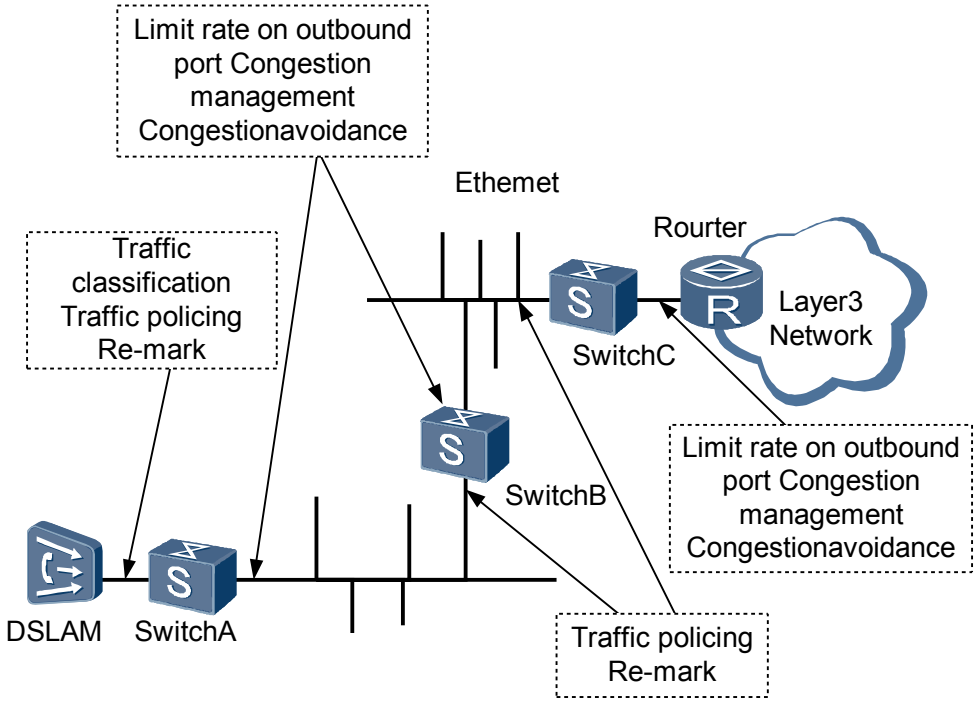
1.4.10 本地优先级与入队列索引关系

在默认的情况下，本地优先级（报文的服务等级）与端口队列的对应关系是一对一。在实际部署时，有时需要两类或多类不同的服务等级放入同一个队列中进行调度，从而有效地节约设备缓存。

1.5 应用

在以太网上，报文流根据需要可以封装为 VLAN 帧。S9700 能够从 VLAN 帧头获取 802.1p 优先级，并作为提供有差别 QoS 的参考依据。基于类的 QoS 在 Ethernet 网络上的部署如图 1-18 所示。

图 1-18 QoS 在 Ethernet 网络的部署



网络边缘交换机的入口处

在网络边缘 S9700 的入口处，根据业务流的 QoS 服务质量要求对业务流首先进行流分类；然后对业务流进行流量监管，惩罚超出速率限制的突发流量；同时根据需要重标记报文流的某些字段，以便网络内设备根据新标记值作为分类标准。

网络边缘交换机的出口处

分类后的业务流按照优先级关系，被送入 S9700 接口出方向上的不同队列中。在网络边缘 S9700 的出口处，通过对队列实施各种调度策略进行拥塞管理；在每个队列的尾部，通过 WRED 技术进行拥塞避免，从而在发现拥塞有加剧趋势时主动丢弃报文；在发出报文前对接口出方向上的流量进行限速。

版权所有 © 华为技术有限公司

网络内部交换机

在以太网内部各交换机的入口处，既可以重新对流进行分类，也可以直接根据来自上游的重标记信息作为分类依据，同时还可以再次重标记；之后，不同类别的业务流被送入各队列。在出口处，需要进行拥塞管理、拥塞避免和接口出方向限速。

1.6 术语与缩略语

术语

术语	解释
差分服务	即 Differentiated Service，简称为 DiffServ。DiffServ 是一个多服务模型，可以满足不同的 Qos 需求。应用程序在发出报文前，不需要通知通信设备，而且网络不需要为每个流维护状态。它根据每个报文指定的 QoS，来提供特定的服务，包括进行报文的分类、流量整形、流量监管和排队。主要实现技术包括 CAR 和队列技术。
承诺访问速率	英文全称是 Committed Access Rate。借助令牌桶机制，若桶中存在足够数量的令牌用来转发报文，则称流量遵守或符合速率限制，否则称为不符合或超过速率限制。CAR 支持单速单桶、srTCM、trTCM 等算法，根据预先设置的匹配规则对流量进行评估，并对流量进行度量和监管。
承诺突发尺寸	英文全称是 Committed Burst Size。表示令牌桶的容量，即每次突发所允许的最大流量尺寸。突发尺寸必须大于最大报文长度。
承诺信息速率	英文全称是 Commit Information Rate。向桶中放置令牌的平均速率，即允许的流平均速率。通常情况下，流量的速率小于承诺信息速率。

缩略语

缩略语	英文全称	中文全称
QoS	Quality of Service	服务质量
srTCM	Single Rate Three Color Marking	单速率三色标记
trTCM	Two Rate Three Color Marking	双速率三色标记
WRED	Weighted Random Early Discard	加权随机早期丢弃
WFQ	Weighted Fair Queue	加权公平队列
VLAN	Virtual Local Area Network	虚拟局域网
MQC	Modular QoS Command	模块化 QoS 命令

版权所有 © 华为技术有限公司