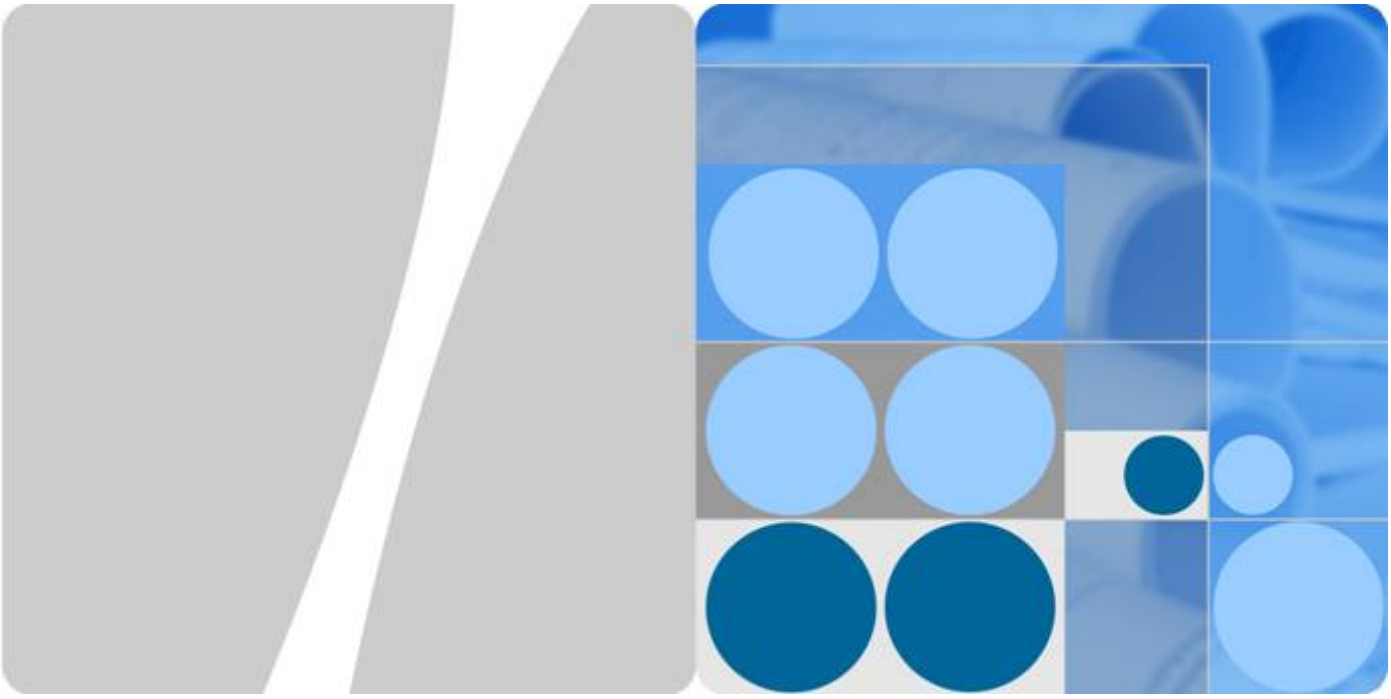


Huawei AR1200 系列企业路由器

V200R002C00

特性描述-QoS

文档版本　01

发布日期　2011-12-30

华为技术有限公司

商标声明


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

注意

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

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

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 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。

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&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 (2011-12-30)

第一次正式发布。

目 录

1 QoS

关于本章

服务质量 QoS（Quality of Service）就是指网络通信过程中，允许用户业务在带宽、时延、时延抖动和丢包率等方面获得可预期的服务水平。

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带宽、时延、时延抖动、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

对企业的网络流量进行调控，避免并管理网络拥塞，减少报文的丢失率，为企业中的用户提供专用带宽，或者为不同的业务（语音、视频、数据等）提供不同的服务质量。

受益

用户可以使用各类带宽服务，如保证时间敏感业务的低时延、多媒体业务的带宽保证等。

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	Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6	-
RFC 2475	An Architecture for Differentiated Services	-
RFC 2597	Assured Forwarding PHB Group	-
RFC 2598	An Expedited Forwarding PHB	-

文档	描述	备注
RFC 3260	New Terminology and Clarifications for Diffserv	-
RFC 3246	An Expedited Forwarding PHB	-
RFC 2697	A Single Rate Three Color Marker	-
RFC 2698	A Two Rate Three Color Marker	-
RFC 4594	Configuration Guidelines for DiffServ Service Classes	-

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

产品	最低支持版本
AR1200	V200R001C00

说明

SAC（Smart Application Control）功能最低支持版本为 V200R002C00。

依赖特性

- 基于流策略的 QoS 通过 ACL 进行复杂流分类。
- 基于流策略的镜像功能需要配置观察端口，依赖于镜像特性。

1.4 原理描述

1.4.1 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，AR1200 使用的是 DiffServ 服务模型。

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。网络设备可以在任何时候，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。

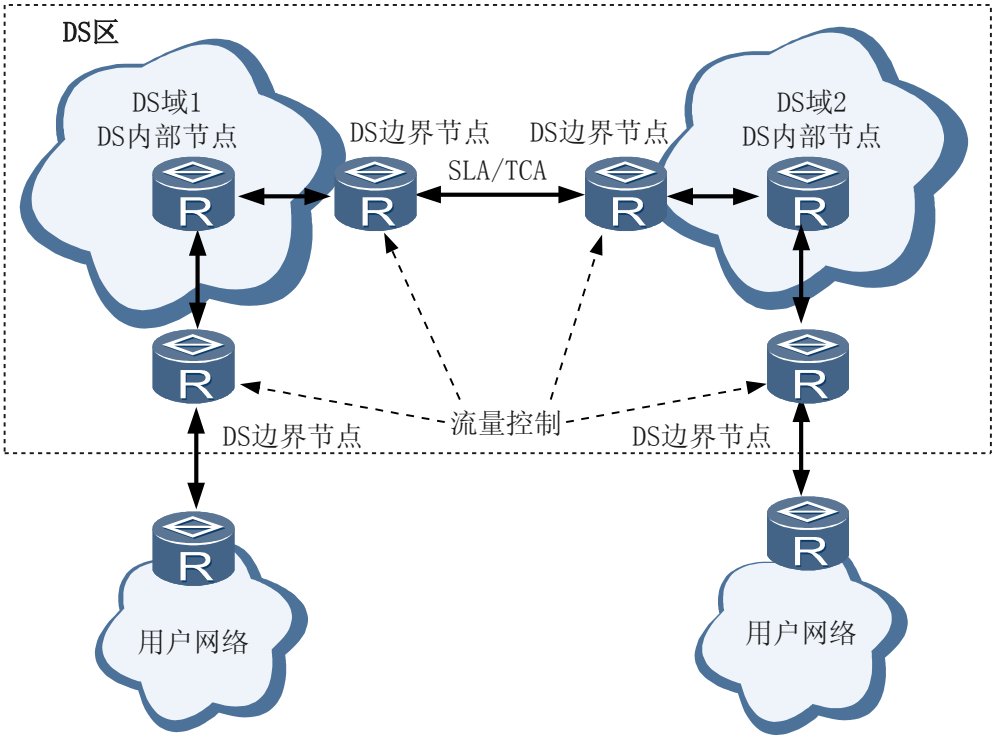
DiffServ 模型中，网络不需要为每个流维护状态，它根据每个报文携带的优先级来提供特定的服务。可以用不同的方法来指定报文的 QoS，如 IP 报文的优先级（IP Precedence），报文的源地址和目的地址等。网络通过这些信息来进行报文的分类、流量整形、流量监管和队列调度。

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。

- 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业务流的分类、逐跳行为 PHB(Per-Hop Behavior)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 1-1 Diff-Serv 体系结构示意图

  DiffServ 体系结构如上图所示，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。

    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。

- 标准的 PHB 行为

  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。

  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 1-1**。

  表 1-1 各类 AF 业务对应的 DSCP 值

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

- 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：对进入路由器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路由器可供给的网络资源，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：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。

1.4.2 优先级映射

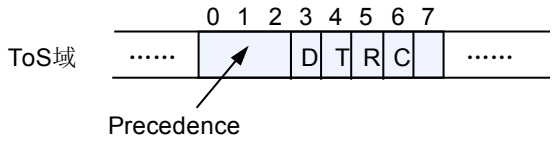
记录 QoS 信息的字段

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 1-2 所示。

  图 1-2 IP 报文中的 Predecence 字段

  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。

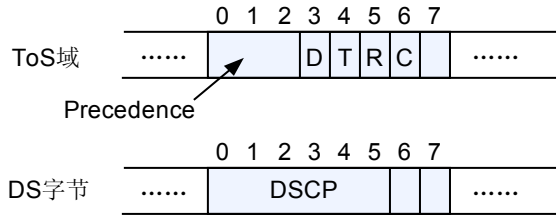
  除了 Predecence 字段外，ToS 域中还包括 D、T、R 三个比特：

  - 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。

- DSCP 字段

  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 字节格式如图 1-3 所示。

  图 1-3 DSCP 字段格式

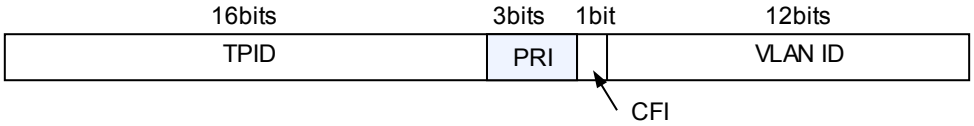
  

  DS 字段的低 6 位（0 ～ 5 位）用作区分服务代码点 DSCP（DS Code Point），高 2 位（6、7 位）是保留位。DS 字段的低 3 位（0 ～ 2 位）是类选择代码点 CSCP（Class Selector Code Point），相同的 CSCP 值代表一类 DSCP。DS 节点根据 DSCP 的值选择相应的 PHB（Per-Hop Behavior）。

- VLAN 帧头中的 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，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

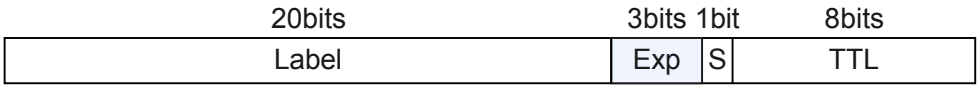


在 802.1Q 头部中包含 3 比特长的 PRI 字段。PRI 字段表示 8 个传输优先级，按照优先级从高到低顺序取值为 7、6、……、1 和 0。

- MPLS EXP 字段

  MPLS 报文与普通的 IP 报文相比增加了标签信息。标签的长度为 4 个字节，封装结构如图 1-5 所示。

图 1-5 MPLS 标签的封装格式



标签共有 4 个域：

- Label：20 比特，标签值字段，用于转发的指针。
- Exp：3 比特，保留字段，用于试验，现在通常用做 CoS（Class of Service）。
- S：1 比特，栈底标识。MPLS 支持标签的分层结构，即多重标签，S 值为 1 时表明为最底层标签
- TTL：8 比特，和 IP 分组中的 TTL（Time To Live）意义相同。

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。Exp 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。

在 IP 网络，由 IP 报文的 IP 优先级或 DSCP 标识服务等级。但是对于 MPLS 网络，由于报文的 IP 头对 LSR 设备是不可见的，所以需要在 MPLS 网络的边缘对 MPLS 报文的 EXP 域进行标记。

缺省的情况下，在 MPLS 网络的边缘，将 IP 报文的 IP 优先级直接拷贝到 MPLS 报文的 EXP 域；但是在某些情况下，如 ISP 不信任用户网络、或者 ISP 定义的差别服务类别不同于用户网络，则可以根据一定的简单分类策略，依据内部的服务等级重新设置 MPLS 报文的 EXP 域，而在 MPLS 网络转发的过程中保持 IP 报文的 ToS 域不变。

在 MPLS 网络的中间节点，根据 MPLS 报文的 EXP 域对报文进行分类，并实现拥塞管理，流量监管或者流量整形等 PHB。

优先级映射

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。当报文经过不同网络时，为了保持报文的优先级，需要在连接不同网络的网关处配置这些优先级标记的映射关系。

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 优先级映射到本地优先级 LP（即设备为报文分配的一种具有本地意义的优先级，对应出端口队列序

号）；在报文出设备时，可以将本地优先级映射为 QoS 优先级或直接重标记报文优先级，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。

AR1200 的本地优先级与 802.1P 优先级采用固定的映射关系，无需用户配置，因此本地优先级对用户不可见。具体映射关系如表 1-2 和表 1-3：

表 1-2 AR1200 主控板上 FE 接口 802.1P 到 LP 的映射关系表

802.1P	LP
0	0
1	0
2	1
3	1
4	2
5	2
6	3
7	3

表 1-3 AR1200 其余接口 802.1P 到 LP 的映射关系表

Input 802.1P	LP
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7

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。

- 使用端口的优先级

  缺省情况下，端口模式为不信任报文优先级，即使用端口优先级，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 802.1P、DSCP、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 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 802.1P、DSCP、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。

设备提供多张优先级映射表，分别对应相应的优先级映射关系。缺省情况下，

- DSCP 到本地优先级、到 802.1P 的映射关系如**表 1-4**、**表 1-5**，DSCP 到 DSCP 的优先级映射保持不变。
- 802.1P 到本地优先级、到 DSCP 的映射关系如**表 1-6**、**表 1-7**，802.1P 到 802.1P 的优先级映射保持不变。
- 端口优先级到本地优先级、到 DSCP 的映射关系与 802.1P 优先级到本地优先级、到 DSCP 的映射关系一致，端口优先级到 802.1P 的优先级映射保持不变。

表 1-4 AR1200 主控板上 FE 接口 DSCP 到 LP 和 802.1P 的映射关系表

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

表 1-5 AR1200 其余接口 DSCP 到 LP 和 802.1P 的映射关系表

Input DSCP	LP	Output 802.1P
0 ～ 7	0	0
8 ～ 15	1	1
16 ～ 23	2	2
24 ～ 31	3	3
32 ～ 39	4	4
40 ～ 47	5	5

Input DSCP	LP	Output 802.1P
48 ～ 55	6	6
56 ～ 63	7	7

表 1-6 AR1200 主控板上 FE 接口 802.1P 到 LP 和 DSCP 的映射关系表

Input 802.1P	LP	Output DSCP
0	0	0
1	0	8
2	1	16
3	1	24
4	2	32
5	2	40
6	3	48
7	3	56

表 1-7 AR1200 其余接口 802.1P 到 LP 和 DSCP 的映射关系表

Input 802.1P	LP	Output DSCP
0	0	0
1	1	8
2	2	16
3	3	24
4	4	32
5	5	40
6	6	48
7	7	56

1.4.3 QoS 策略

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。

1.4.3.1 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。

分类规则见表 1-8：

表 1-8 复杂流分类的分类规则

层级	分类规则
二层	● 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 ● MPLS 报文的 EXP 优先级 ● 源 MAC 地址 ● 目的 MAC 地址 ● 基于二层封装的协议字段 ● FR DE ● FR DLCI ● ATM PVC ● ACL 4000 ～ 4999
三层	● IP 报文的 DSCP 优先级 ● IP 报文的 IP 优先级 ● IP 协议类型（即 IPv4 协议） ● ACL 2000 ～ 3999
四层	● RTP 端口号 ● TCP 报文的 TCP SYN 标志

层级	分类规则
其他	● 入接口

1.4.3.2 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 AR1200 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。AR1200 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 AR1200 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。请参见**拥塞管理**中的“CBQ”。

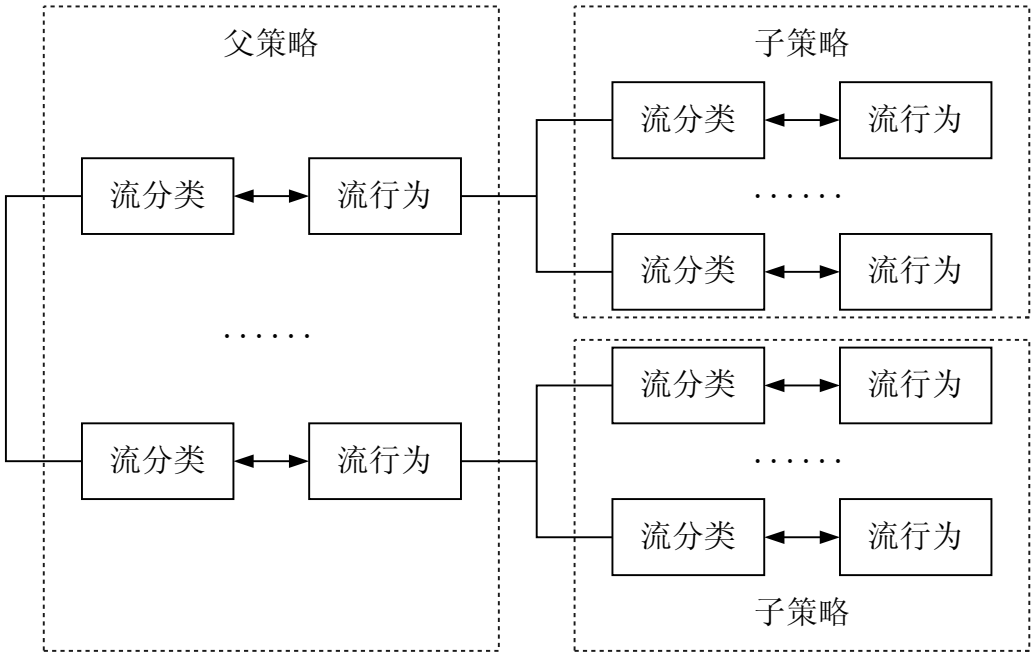
1.4.3.3 流策略

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。

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 1-6** 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 1-6 流策略嵌套示意图



AR1200 支持两层策略嵌套，子策略下面不能再有嵌套。

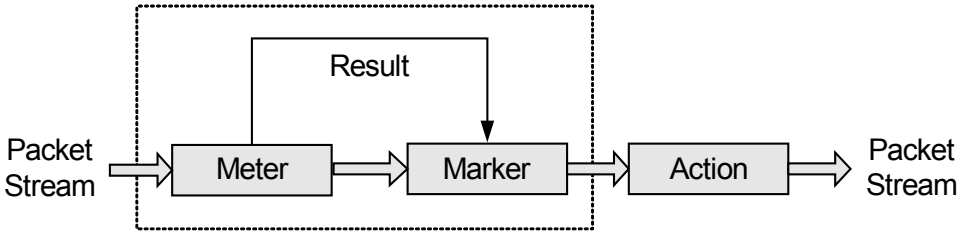
嵌套策略为 HQoS 提供了层次化的配置模型：在 HQoS 中，通过父策略区分网络中的不同用户，通过子策略区分用户的不同业务，从而提供区分用户和用户业务的精细化服务。

1.4.4 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。AR1200 使用 CAR 来进行流量监管。

流量监管的原理

图 1-7 流量监管组件

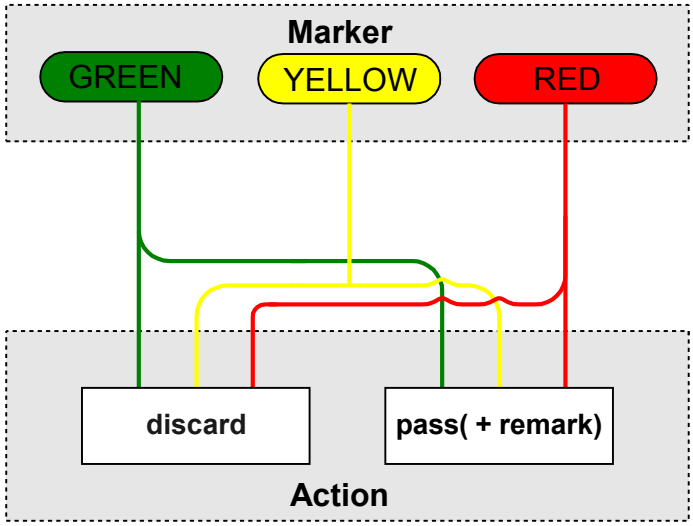


如图 1-7 所示，AR1200 的流量监管由三部分组成：

- 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：
  - 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。

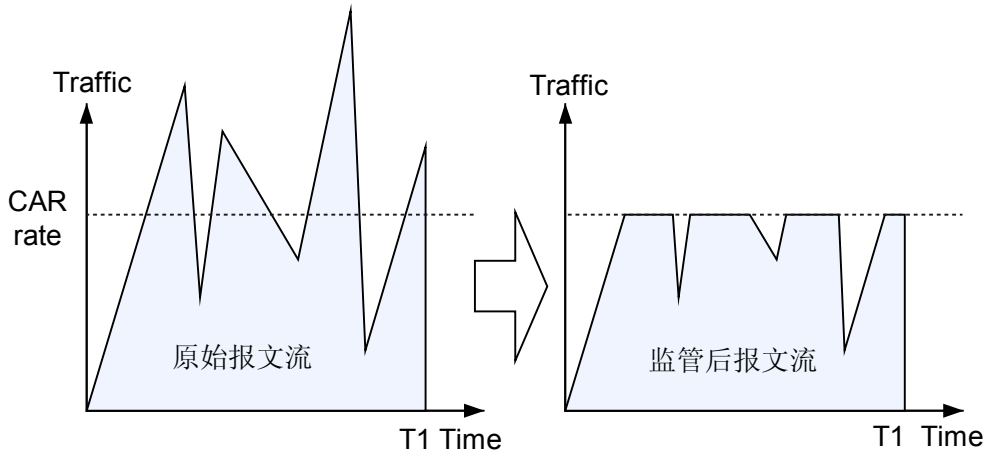
  默认情况下，green、yellow 进行转发，red 报文丢弃。

图 1-8 流量监管动作



经过流量监管，如果某流量速率超过标准，AR1200 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 1-9 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。

图 1-9 流量监管的报文流曲线图



1.4.5 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。

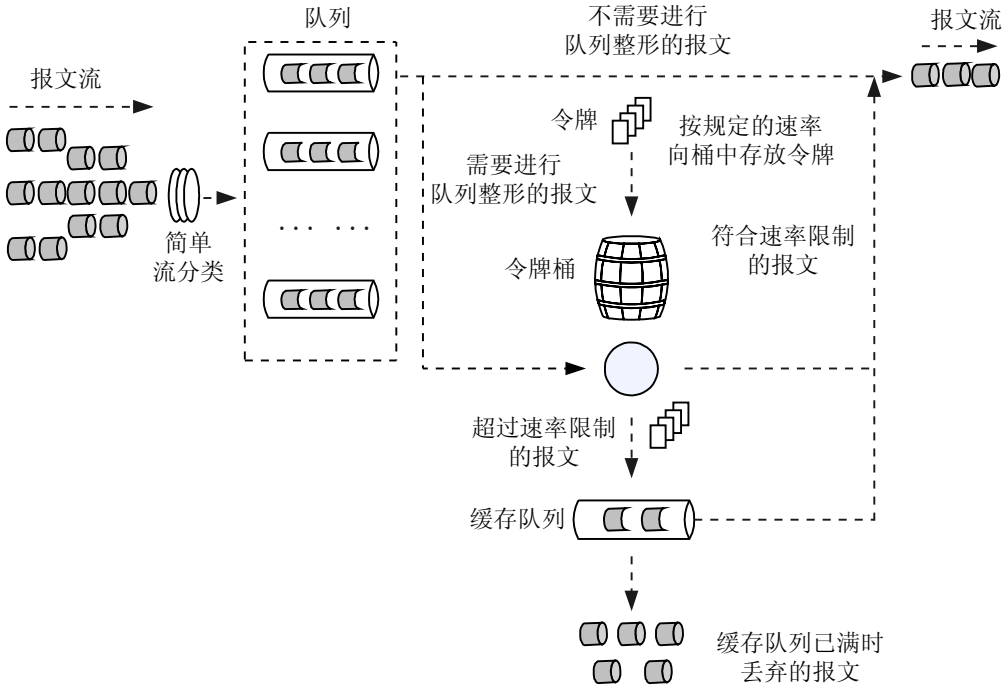
AR 支持三级流量整形。

处理流程

流量整形是一种应用于接口、子接口或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。下面以接口或子接口下基于简单流分类的队列整形为例介绍流量整形的处理流程，其处理流程如**图 1-10** 所示。

图 1-10 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，使报文进入不同的队列。
2. 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。
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。

队列整形后，如果该接口或子接口同时配置了端口整形，则系统还要逐级按照子接口整形速率、接口整形速率对报文流进行速率控制。其处理流程与队列整形相似，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：

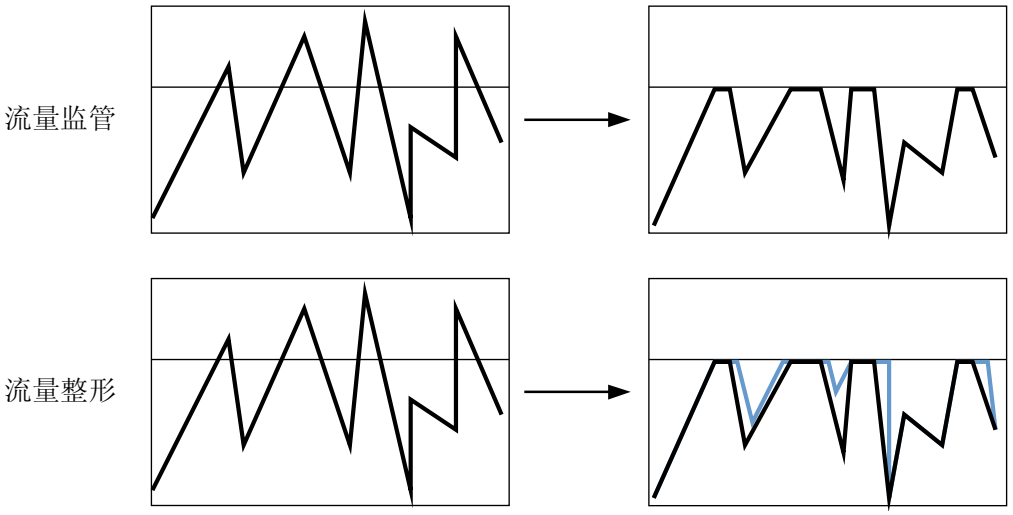
- 利用流量监管进行报文控制时，直接丢弃不符合速率要求的报文。而流量整形则会将不符合速率要求的报文先行缓存，当令牌桶有足够的令牌时，再均匀的向外发送这些被缓存的报文。
- 流量整形可能会增加延迟，而流量监管几乎不引入额外的延迟。

表 1-9 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 1-11 说明了流量监管与流量整形的区别。

图 1-11 流量监管与流量整形的区别



1.4.6 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，AR1200 LAN 接口上的拥塞管理技术分为 PQ、DRR、PQ+DRR、WRR、PQ+WRR，WAN 接口上的拥塞管理技术分为 PQ、WFQ 和 PQ+WFQ。

在 AR1200 上，缺省情况下，每个接口出方向上都拥有 4 个或 8 个队列（AR1200 主控板上 FE 接口为 4 个队列，其余均为 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。缺省情况下，LAN 侧所有队列均采用 WRR 调度模式，WAN 侧所有队列均采用 WFQ 调度模式。AR1200 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。

如图 1-12 所示，Queue7 比 Queue6 具有更高的优先权，Queue6 比 Queue5 具有更高的优先权，依次类推。只要链路能够传输分组，Queue7 尽可能快地被服务。只有当 Queue7 为空，调度器才考虑 Queue6。当 Queue6 有分组等待传输且 Queue7 为空时，Queue6 以链路速率接受类似地服务。当 Queue7 和 Queue6 为空时，Queue5 以链路速率接收服务，以此类推。

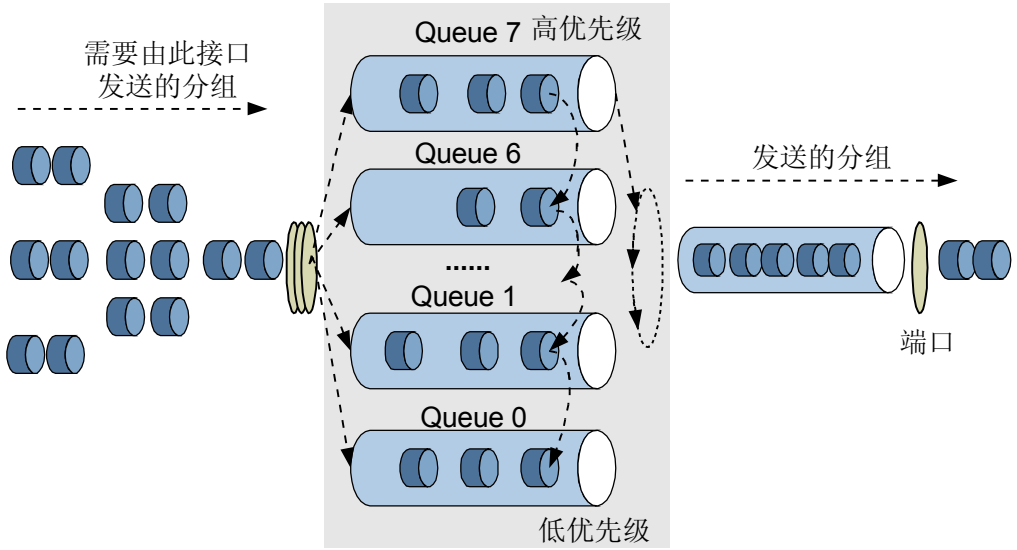
PQ 调度算法对低时延业务非常有用。假定数据流 X 在每一个节点都被映射到最高优先级队列，那么当数据流 X 的分组到达时，则分组将得到优先服务。

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 Queue7 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 Queue6 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 Queue7 的业务流不超出输出链路容量的一定比例，这样可以使 Queue7 常常为空，允许调度器为低优先级队列服务。

缺省情况下，AR1200 的调度模式为 PQ。

图 1-12 PQ 调度示意图

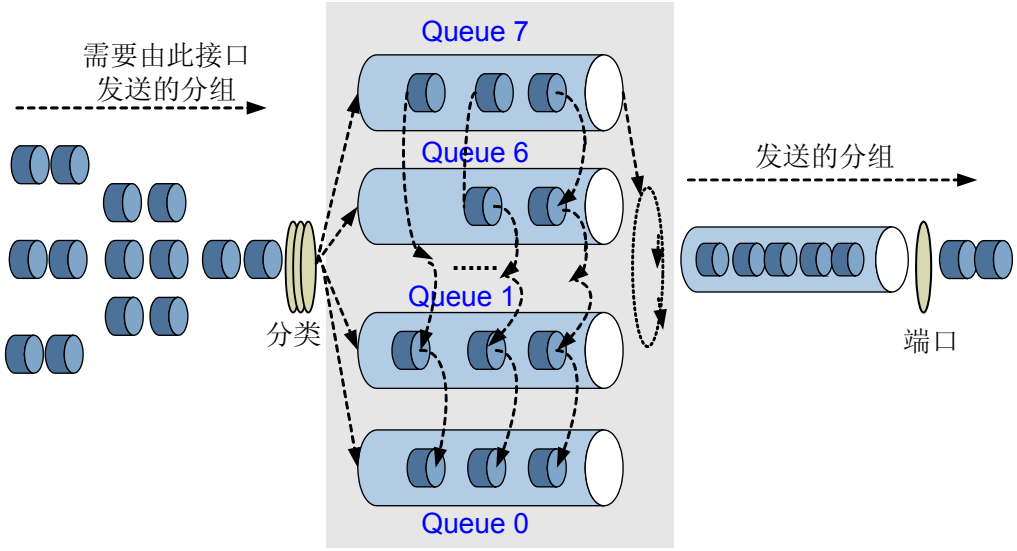


- WRR 调度

  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 1-13 所示。

图 1-13 WRR 调度示意图



在进行 WRR 调度时，AR1200 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 1-10** 所示。

表 1-10 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	-	-

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	-	-	Q4	Q3	Q2	-	-
参加第 11 轮调度的队列	-	-	Q5	-	Q3	-	-	-

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

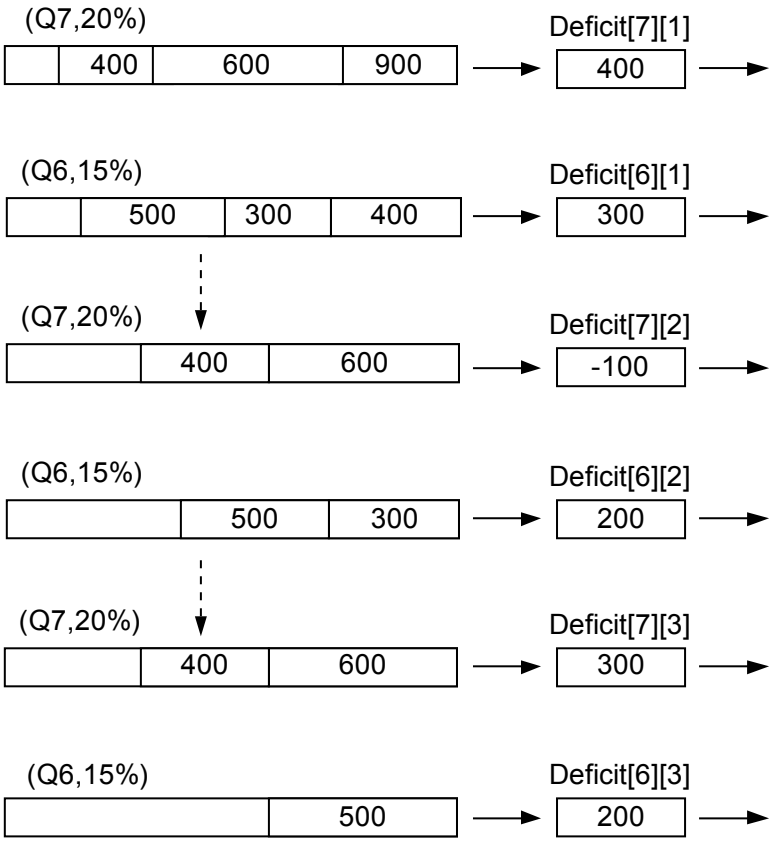
从统计上看，各队列中的报文流被调度的次数与该队列的权值成正比，权值越大被调度的次数相对越多。由于 WRR 调度的以报文为单位，因此每个队列没有固定的带宽，同等调度机会下大尺寸报文获得的实际带宽要大于小尺寸报文获得的带宽。

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 1-14 所示。

图 1-14 DRR 调度示意图



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 1-14 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] = -300。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。

- WFQ 调度

  公平队列 FQ（Fair Queuing）的目的是尽可能公平地分享网络资源，使所有流的延迟和抖动达到最优：

  - 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。

  与 FQ 相比，WFQ（Weighted Fair Queue）在计算报文调度次序时增加了优先权方面的考虑。从统计上，WFQ 使高优先权的报文获得优先调度的机会多于低优先权的报文。

  WFQ 调度在报文入队列之前，先对流量进行分类，有两种分类方式：

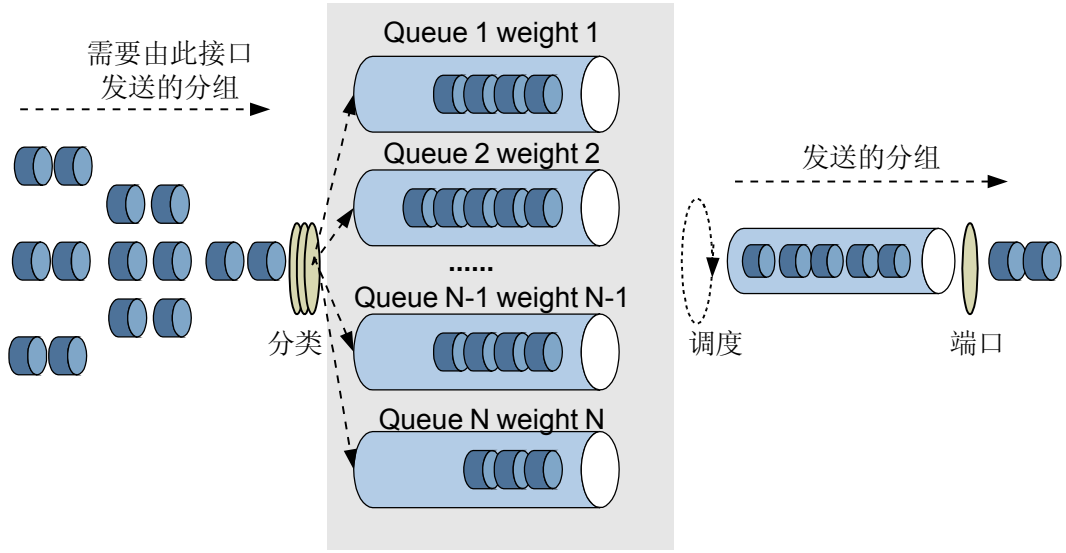
  - 按流的“会话”信息分类：

    根据报文的协议类型、源和目的 TCP 或 UDP 端口号、源和目的 IP 地址、ToS 域中的优先级位等自动进行流分类，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。这种方式只有 CBQ 的 default-class 支持。

  - 按优先级分类：

    通过优先级映射把流量标记为本地优先级，每个本地优先级对应一个队列号。每个接口预分配 4 个或 8 个队列，报文根据队列号进入队列。默认情况，队列的 WFQ 权重相同，流量平均分配接口带宽。用户可以通过配置修改权重，高优先权和低优先权按权重比例分配带宽。

图 1-15 WFQ 调度示意图

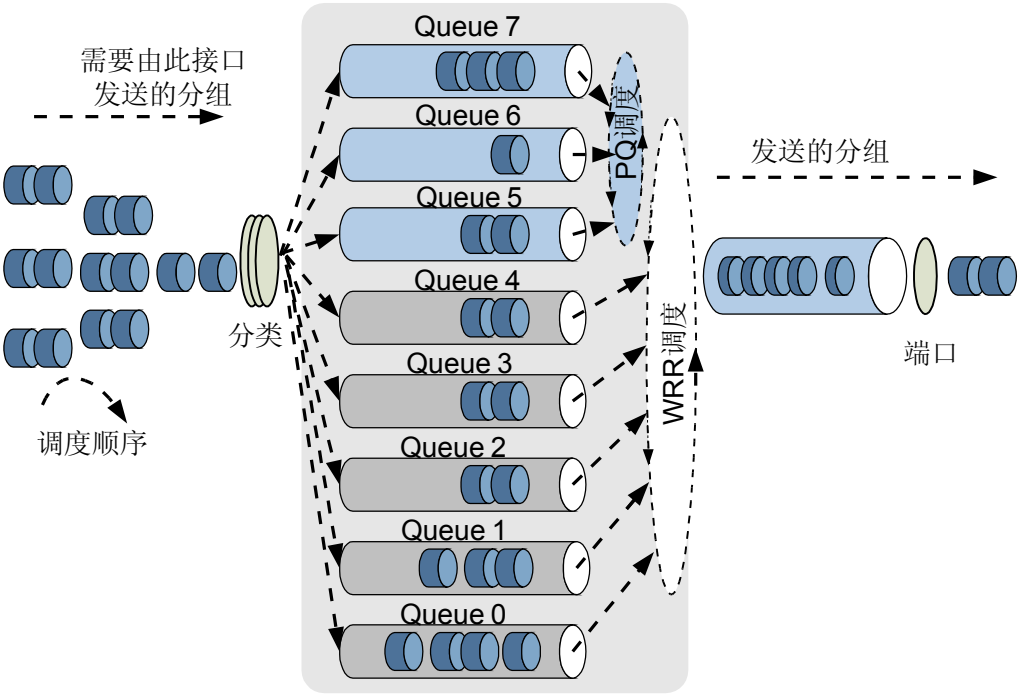


- PQ+WRR 调度

  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 AR1200 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 Queue7、Queue6、Queue5）采用 PQ 调度，另一组（例如 Queue4、Queue3、Queue2、Queue1 和 Queue0 队列）采用 WRR 调度。AR1200 上只有 LAN 侧接口支持 PQ+WRR 调度。PQ+WRR 调度示意图如图 1-16 所示。

图 1-16 PQ+WRR 混合调度示意图



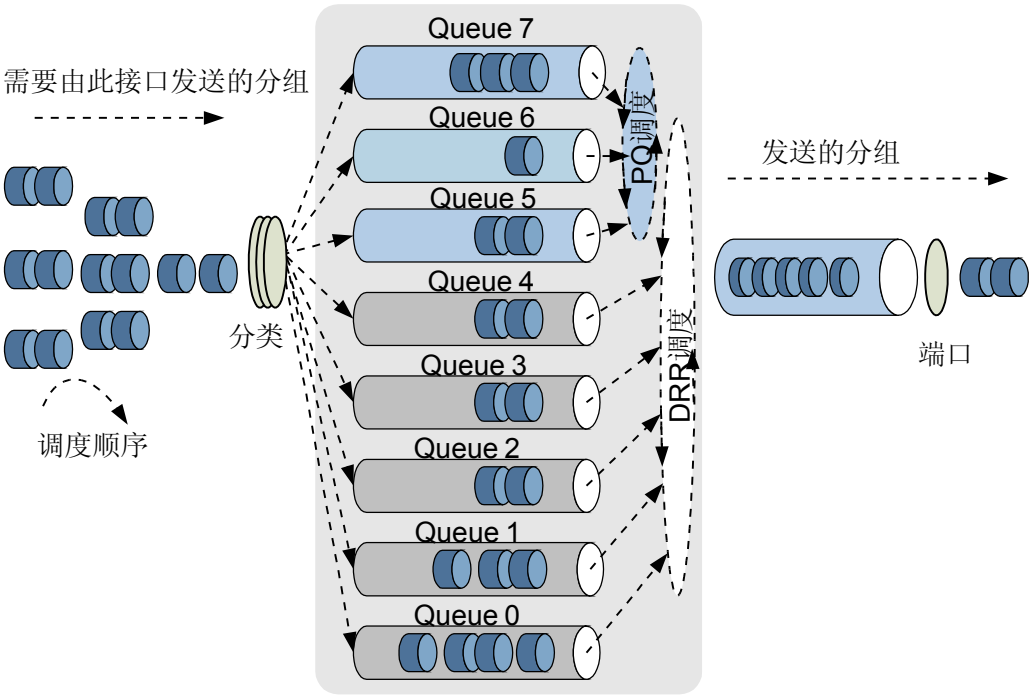
在调度时，AR1200 首先按照 PQ 方式调度 Queue7、Queue6、Queue5 队列中的报文流，只有这些队列中的报文流全部调度完毕后，才开始以 WRR 方式循环调度其他队列中的报文流。Queue4、Queue3、Queue2、Queue1 和 Queue0 队列包含自己的权值。重要的协议报文和有低延时需求的业务报文应放入采用 PQ 调度的队列中，得到优先调度的机会，其余报文放入以 WRR 方式调度的各队列中。

- 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。

  AR1200 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。AR1200 的 WAN 接口和主控板上 FE 接口不支持 PQ+DRR 调度模式。

图 1-17 PQ+DRR 调度示意图



如图 1-17 所示，在调度时，AR1200 首先按照 PQ 方式优先调度 Queue7、Queue6 和 Queue5 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 Queue4、Queue3、Queue2、Queue1 和 Queue0 队列中的报文流。其中，Queue4、Queue3、Queue2、Queue1 和 Queue0 队列包含自己的权值。

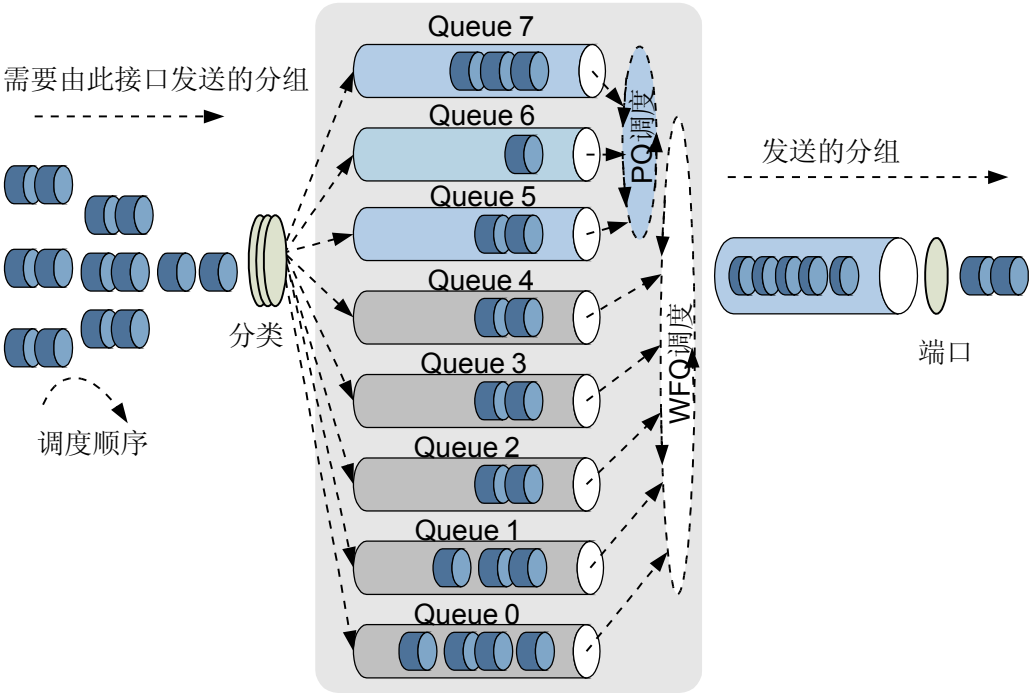
重要的协议报文以及有低延时需求的业务报文应放入需要进行 PQ 调度的队列中，得到优先调度的机会，其他报文放入以 DRR 方式调度的各队列中。

- PQ+WFQ 调度

  与 PQ+WRR 相似，其集合了 PQ 调度和 WFQ 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 WFQ 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+WFQ 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。

AR1200 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 WFQ 调度。只有 WAN 侧接口支持 PQ+WFQ 调度。

图 1-18 PQ+WFQ 调度示意图



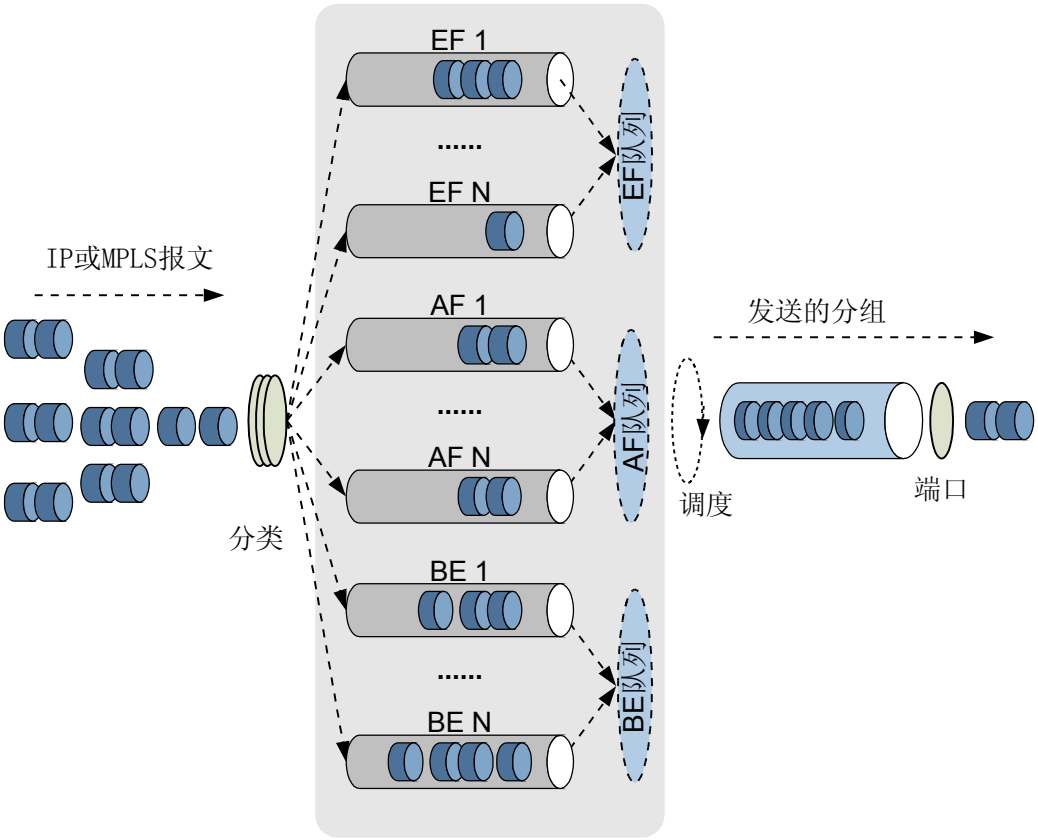
如图 1-18 所示，在调度时，AR1200 首先按照 PQ 方式优先调度 Queue7、Queue6 和 Queue5 队列中的报文流，只有这些队列中的报文流全部调度完毕后，才开始以 WFQ 方式调度 Queue4、Queue3、Queue2、Queue1 和 Queue0 队列中的报文流。其中，Queue4、Queue3、Queue2、Queue1 和 Queue0 队列包含自己的权值。

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 WFQ 方式调度的各队列中。

- CBQ 调度

  CBQ（Class-based Queueing）基于类的加权公平队列是对 WFQ 功能的扩展，为用户提供了定义类的支持。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 1-19 CBQ 调度示意图



如图 1-19 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
- 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。
- 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报文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。

1.4.7 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。

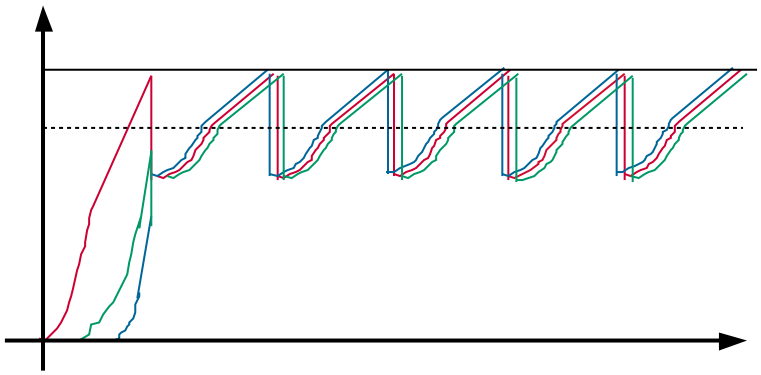
本节介绍拥塞避免的基本知识，具体包括：

- 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。

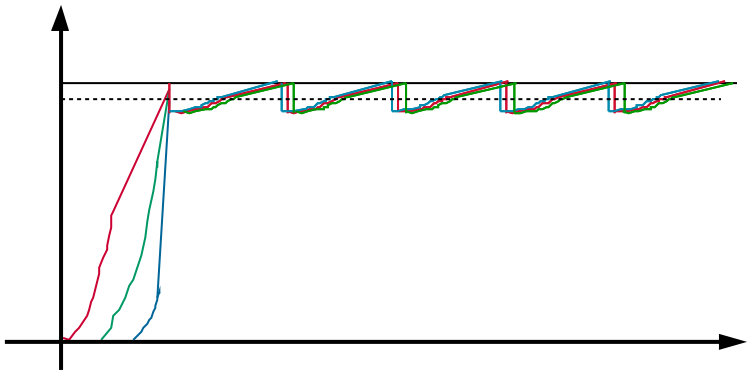
  图 1-20 尾部丢包示意图

- 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 1-21 RED 算法示意图



基于 RED 技术，AR1200 实现了 WRED（Weighted Random Early Detection）。流队列支持基于 DSCP 或 IP 优先级进行 WRED 丢弃。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。

1.4.8 HQoS

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。

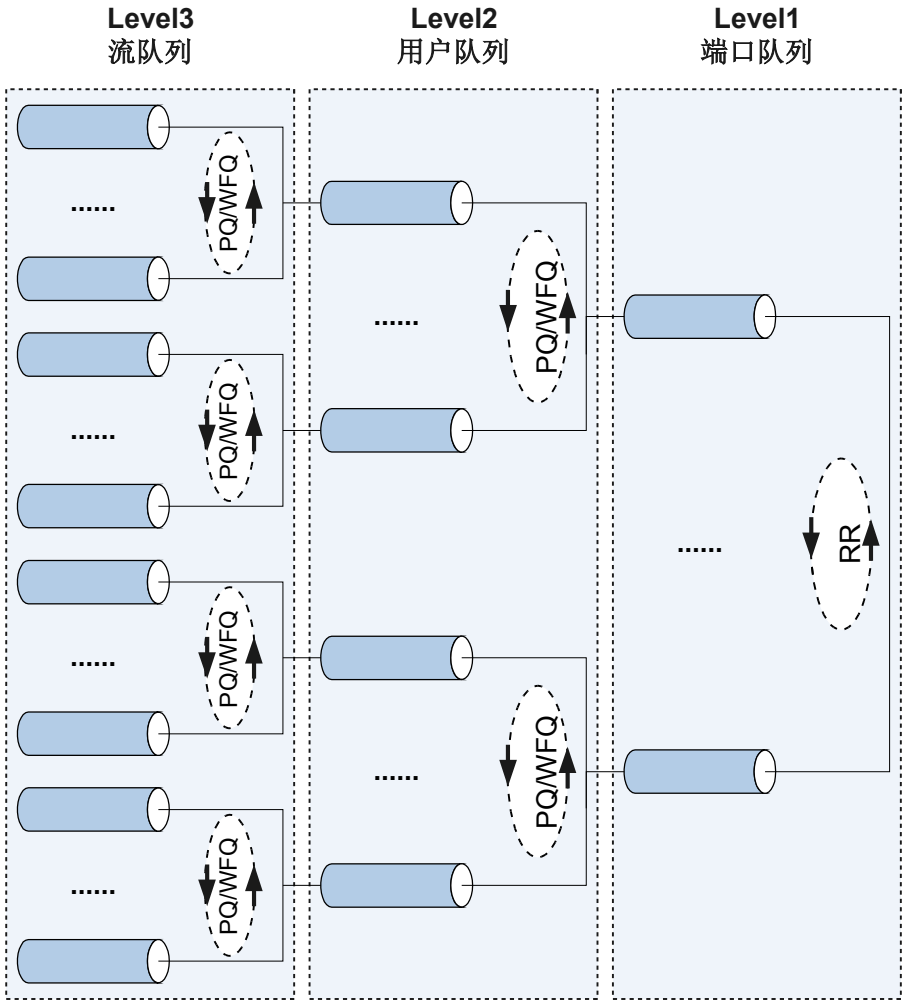
目前，越来越多的企业用户通过向运营商租用专线的方式来构建自己的企业网，不同企业之间，其业务侧重点和所需要的服务质量是有差别的。这就要求运营商能够依据不同企业的业务需求提供不同的调度策略和 QoS 保证。传统的 QoS 无法区分用户，所以无法对不同的企业用户提供有差别的队列调度服务。

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。

HQoS 支持的队列

如图 1-22 所示，HQoS 基于队列实现层次化调度，目前在 AR1200 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 1-22 HQoS 队列调度示意图



- 流队列

  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。

- 用户队列

  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。

- 端口队列

  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。

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。

AR1200 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。

以企业用户的 HQoS 部署为例，企业用户主要有三种业务：语音通讯（VoIP）、视频会议（VC）和数据业务（DATA），每个用户队列对应一个企业用户，每个流队列对应一种业务。通过部署 HQoS，可以实现：

- 控制单个企业用户三种业务之间的流量调度
- 控制单个企业用户三种业务的总带宽
- 控制多个企业用户之间的带宽分配
- 控制多个企业用户的总带宽

HQoS 整形器

整形器实现报文的缓存及限速功能。AR1200 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。

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。

1.4.9 SAC

智能应用控制 SAC（Smart Application Control）是一个智能的应用识别与分类引擎，利用 DPI（Deep Packet Inspection）深度报文检测技术，对报文中的第 4 ～ 7 层内容和一些动态协议(如 HTTP、FTP、RTP)进行检测和识别，根据分类结果实施精细化 QoS 策略控制。

DPI 技术

传统流量分类技术只能检测 IP 报文的 4 层以下的内容，包括源地址、目的地址、源端口、目的端口以及业务类型等。而 DPI 在分析报文头的基础上，增加了对应用层的分析，是一种基于应用层的流量检测和控制技术。

典型的 DPI 识别技术包括：特征识别、关联识别、行为识别等。这三类识别技术分别适用于不同类型的协议，相互之间无法替代，通过综合的运用这三大技术，能够有效的识别网络上的各类应用，从而实现对网络数据流的深度控制。

- 特征识别

  特征识别是 DPI 的最基本技术。不同的应用通常会采用不同的协议，而不同的应用协议具有各自的特征。这些特征可能是特定的端口、特定的字符串或者特定的比特

序列，能标识该协议的特征称为特征码。特征识别技术，即通过匹配数据报文中的特征码来确定应用。协议的特征不仅在单个报文中体现，某些协议报文的特征是分布在多个报文中的，需要对多个报文进行采集分析，才能够识别出协议类型。

- 关联识别

  对于某些数据流，控制通道和数据通道是分开的（如 FTP、SIP、H.323 等），会在网络中建立两个会话连接。这类应用需要先识别出控制流，再根据控制流的信息识别出相应的数据流。关联识别可以将同一应用协议的控制流和数据流关联起来。通过对控制流的分析，分析出通讯双方将要在哪个通道上建立何种类型的数据流，并在协议识别时将控制通道流和该控制通道流协商出来的数据通道流关联起来。

  DPI 在对控制通道流进行深度解析时，提取出其中协商的数据通道流的源三元组和目的三元组信息，并加入关联表。在后续识别过程中，可以通过该关联表项对数据通道流快速识别。

- 行为识别

  行为模式识别技术必须先对终端的各种行为进行研究，并在此基础上建立行为识别模型，基于行为识别模型，行为模式识别技术即根据客户已经实施的行为，判断用户正在进行的动作或者即将实施的动作。行为模式识别技术通常用于那些无法由协议本身就能判别的业务，例如：从电子邮件的内容看，垃圾邮件和普通邮件的业务流两者间根本没有区别，只有进一步分析，具体根据发送邮件的大小、频率，目的邮件和源邮件地址、变化的频率和被拒绝的频率等综合分析，建立综合识别模型，才能判断是否为垃圾邮件。

特征库

应用识别对于大多协议都是通过特征码来匹配（如：BT、电驴等 P2P 软件），但应用软件会不断升级和更新，其特征码也会发生变化，导致原有特征码无法正确或精确匹配应用协议，特征码需要及时更新，如果在产品软件包中固化特征码，就要更换新的软件版本，用户使用很不方便。将报文数据特征以特征文件库的方式提供给用户，特征码与软件包分离，实现特征码的动态加载，用户不需要更新产品软件版本，不需要重新启动设备，用户就可以很方便地更新升级应用识别能力。另外，当出现一种新应用时，直接通过特征库的加载就能够动态支持。

应用统计

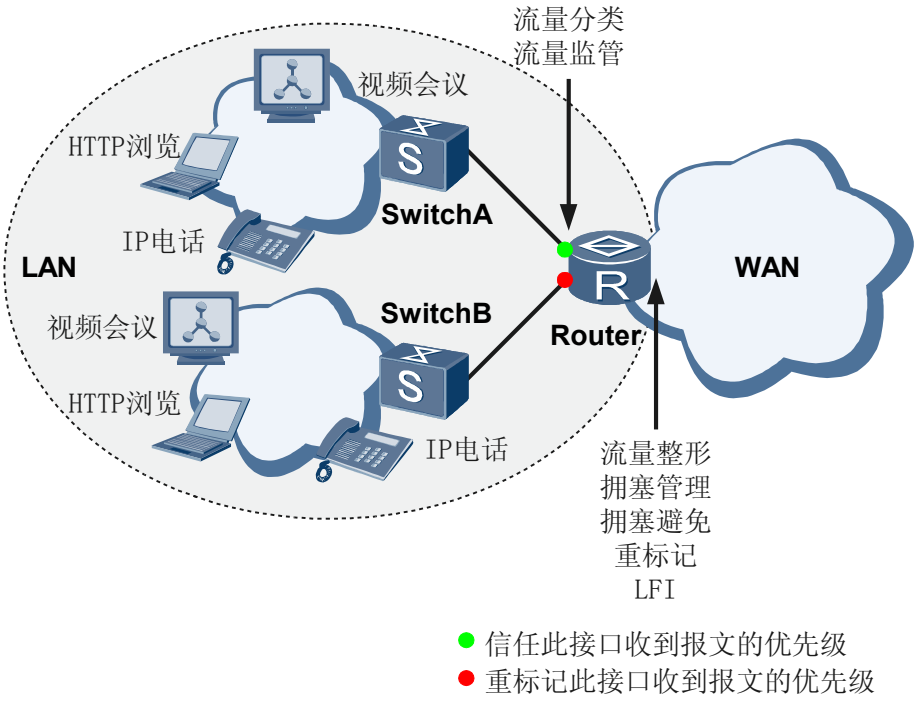
在接口使能应用统计功能时，设备会自动识别经过接口的流量并进行分类，对不同应用的流量进行统计。网络管理员可以及时掌握网络流量特征，从而优化网络部署，合理分配带宽。

1.5 应用

1.5.1 基本 QoS 的应用

如图 1-23 所示，AR1200 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 AR1200 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。

图 1-23 AR1200 在企业网的部署



AR1200 LAN 侧

来自外部网络的各类数据通过 LAN 接口发送到企业网内部不同部门，企业内部的一些数据也需要通过 LAN 接口进行转发。

- 流分类：在接口上根据业务流的 QoS 服务质量要求为业务流进行流分类。在 LAN 接口上可以信任上游的流量分类，直接根据 8021p 或 DSCP 标识业务优先级。也可以由设备重标记报文优先级，标识语音、视频、数据等业务的优先级，以便提供差分服务。
- 流量监管：在接口的入方向上针对业务流进行流量监管，惩罚超出速率限制的突发流量。

AR1200 WAN 侧

企业网内部数据通过 WAN 接口发送到外部网络，WAN 接口的速率较 LAN 接口的速率低。在 WAN 接口的出方向；在发出报文前对接口出方向上进行流量整形和队列调度。

- 流量整形：在发出报文前对接口出方向上的流量进行限速，输出一条比较平整的流量。
- 拥塞管理：通过对队列实施各种调度策略进行拥塞管理。
- 拥塞避免：在每个队列的尾部，通过 WRED 技术进行拥塞避免，提前丢弃报文，解决 TCP 全局同步引起重要业务的抖动。
- 重标记：根据需要重标记报文流的某些字段，以便报文的新标记值能与外部网络的要求相匹配。

- LFI：WAN 链路是低速链路时，为了保证语音报文低时延，可以采用 LFI（Link Fragmentation and Interleaving）技术，实现报文的分片和重组，防止大报文阻塞语音报文。

  说明

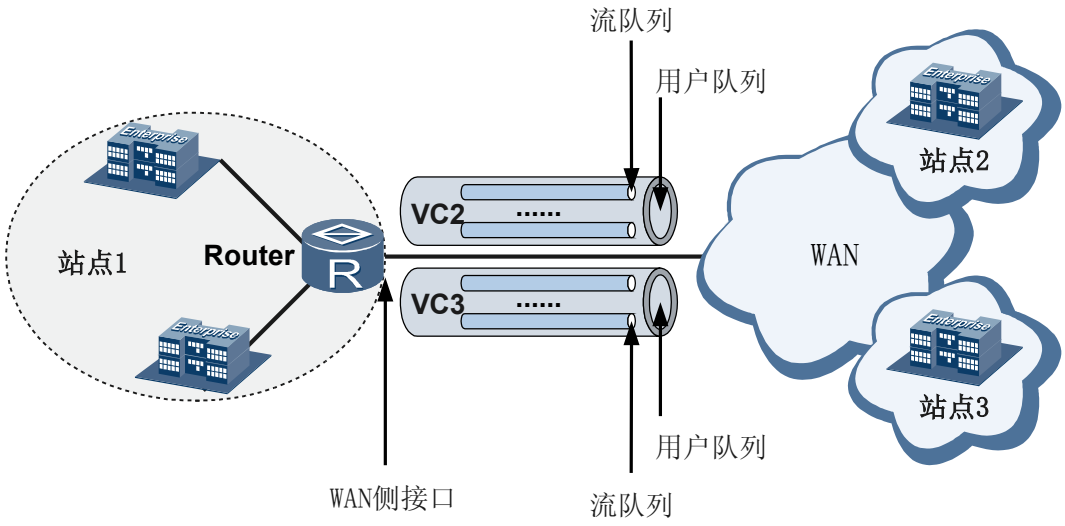
  关于 LFI 的更多内容，请参见《Huawei AR1200 系列企业路由器 特性描述-广域网互联》中“PPP 和 MP”部分。

1.5.2 HQoS 的应用

图 1-24 所示为企业网的典型应用组网,站点 1 为总部，站点 2 和站点 3 为 2 个子部门，子部门与总部通过两条子链路连接。各部门都有语音、视频、数据等业务流。

要求各部门有各自的保证带宽，并且能共享端口的最大带宽；对于不同的业务流，语音报文要保证优先发送，视频、数据报文要保证带宽。

图 1-24 HQoS 在 WAN 侧接口的部署



在 AR1200 WAN 侧接口的出方向上，可以通过部署 HQoS 来实现上述需求。在子接口上应用嵌套流策略，父策略的流分类通过 VLAN、PVC、DLCI 等区分不同的用户，即子接口一级的用户队列；子策略的流分类区分不同的业务类型，即流队列。而 CBQ 提供的 EF 队列可以实现语音报文的优先发送，AF 队列可以实现带宽保证。

1.5.3 SAC 的应用

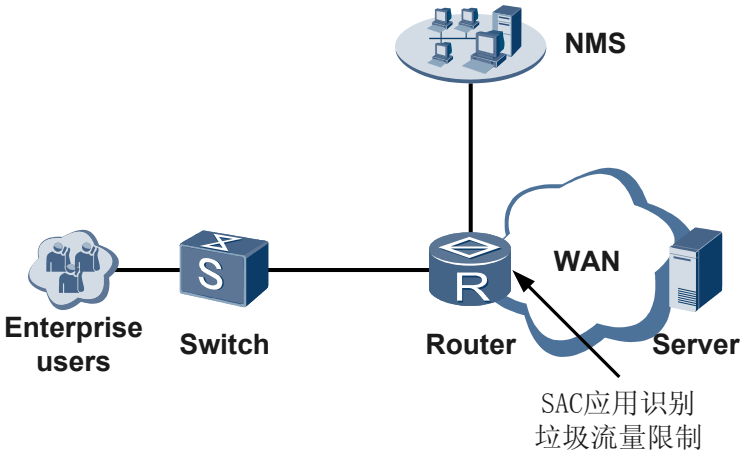
SAC 基于报文第 4 ～ 7 层内容对流量进行智能分类，识别出关键业务和应用。根据识别结果结合流量的重标记、队列调度、流量监管和流量整形等技术，可以保障关键业务的高优先级和带宽。

如图 1-25 所示，AR1200 部署在企业网出口，对经过的流量进行识别、分类和统计，并将统计数据上报给网管系统，网管记录并分析各业务应用及带宽的分布情况，形成统计报表。

企业网中，对于识别出的 P2P、聊天、游戏等与工作无关的流量（称为企业垃圾流量），可以采用如下的策略限制企业员工从事与工作无关的事务，规范员工上网行为。

- 阻断垃圾流量。
- 对垃圾流量进行限速。
- 基于时间段控制流量。

图 1-25 SAC 在 AR1200 的部署



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 包含承诺信息速率（CIR）、承诺突发尺寸（CBS）和超出突发尺寸（EBS）三个指标，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。

术语	解释
公平队列	英文全称是 Fair Queue。一种尽可能使队列公平的分享网络资源，使所有的流的延迟和延迟抖动达到最优的队列调度机制。
IP 优先级	英文全称是 IP-Precedence。IP 报文中 TOS 字段中的前三个比特位用于表示报文的 IP 优先级，是 QoS 中进行流分类的一个依据。
加权公平队列	英文全称是 Weighted Fair Queue。其特点是可以自动进行流分类，并且均衡各个流的延迟和延迟抖动。WFQ 与 FQ 相比，考虑了高优先级报文的利益。
加权随机早期检测	英文全称是 Weighted Random Early Detection。一种用于拥塞避免的丢包算法，可以避免传统的尾部丢包所带来的 TCP 全局同步现象，并在计算报文的丢包概率时，考虑了高优先级报文的利益。
加速转发	英文全称是 Expedited Forwarding。从任何 DS 节点发出的信息流速率在任何情况下必须等于或大于设定的速率。确保报文的低延迟和带宽保证。 A mechanism in which messages from any DS node should be sent at an equal or more rate than what has specified. This can ensure little delay and enough bandwidth.
尽力而为	英文全称是 Best-Effort。传统的 IP 分组投递服务。其特点是依照报文到达时间的先后顺序采用先来先服务的原则处理报文的转发，所有用户的报文共同分享网络和路由器的带宽资源，至于得到资源的多少完全取决于报文到达的时机。Best-Effort 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。
流分类	英文全称是 Traffic Classifier。依据一定的匹配规则识别出感兴趣的报文，是有区别地提供服务的前提和基础。
流量监管	英文全称是 Traffic policin。一种监督特定流量进入路由器的规格的机制，当流量超出规格时，可以采取限制或惩罚措施，以保护运营商的商业利益和网络资源不受损害。
流量整形	英文全称是 Traffic Shaping。是一种主动调整流的输出速率的流控措施，通常是为了使流量适配下游路由器可供给的网络资源，避免不必要的报文丢弃和拥塞。
确保转发	英文全称是 Assured Service。在为用户提供服务时允许业务量大于客户订购的规格，对不超过所订购规格的流量要求确保转发的质量，对超出规格的流量减低服务待遇继续转发，而不是简单的丢弃。
随机早期检测	英文全称是 Random Early Detection。一种用于拥塞避免的丢包算法，通过设定队列丢弃报文的低门限、高门限，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，可以避免传统的尾部丢包所带来的 TCP 全局同步现象。
尾部丢弃	英文全称是 Tail-Drop。一种队列的丢弃机制，通过设定队列长度，当队列长度达到设定值后，丢弃所有后来的报文。
拥塞避免	英文全称是 Congestion Avoidance。一种通过监视网络资源的使用情况，在拥塞已经产生并且有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载的流控机制。

术语	解释
拥塞管理	英文全称是 Congestion Management。一种解决网络资源竞争的流控措施。它在网络发生拥塞时将报文放入队列中缓存，并采取某种调度策略决定报文的转发次序。
优先队列	英文全称是 Priority Queue。根据优先级进行排队的策略。特点是如果同时存在多种优先级的报文，高优先级的报文先被分配资源。
五元组	源 IP 地址、源端口、目的 IP 地址、目的端口和传输层协议类型组成的集合。

缩略语

缩略语	英文全称	中文全称
BA	Behavior Aggregate	行为聚合
CAR	Committed Access Rate	承诺接入速率
CBQ	Class-Based Queueing	基于类的队列
DPI	Deep Packet Inspection	深度报文检测
DRR	Deficit Round Robin	赤字轮循队列调度
FIFO	First In First Out	先进先出队列
FTP	File Transfer Protocol	文件传输协议
GTS	Generic Traffic Shaping	通用流量整形
HQoS	Hierarchical QoS	层次化 QoS
HTTP	Hypertext Transfer Protocol	超文本传输协议
IP	Internet Protocol	互联网协议
ISP	Internet Service Provider	互联网服务提供方
LLQ	Low Latency Queue	低延时队列
MQC	Modular QoS Command	模块化 QoS 命令
OSS	Operations Support System	运营支撑系统（即网管系统）
P2P	Peer to Peer	点对点
PHB	Per Hop Behavior	每一跳行为
QoS	Quality of Service	服务质量
RTP	Real Time Protocol	实时协议
SAC	Smart Application Control	智能应用控制
SIP	Session Initiation Protocol	会话发起协议

缩略语	英文全称	中文全称
SP	Strict Priority	严格优先级
SPI	Shallow Protocol Inspection	普通报文检测
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
WAN	Wide Area Network	广域网
WFQ	Weighted Fair Queue	加权公平队列
WRED	Weighted Random Early Discard	加权随机早期丢弃
WRR	Weighted Round Robin	加权轮循队列调度
VLAN	Virtual Local Area Network	虚拟局域网