# QoS技术白皮书

**关键词：**QoS，服务模型，IntServ，DiffServ，拥塞管理，拥塞避免，队列技术，流量监管，流量整形，链路效率机制

**摘　要：**本文对Internet的三种服务模型（Best-Effort、IntServ和DiffServ），以及服务模型的发展历程进行了简单介绍，较为详细地介绍了H3C系列数据通信产品所支持的QoS技术，内容包括：流量分类和标记、拥塞管理、拥塞避免、流量监管与流量整形、链路效率机制以及MPLS网络相关QoS技术，并且简要描述了在实际应用中的QoS解决方案。网络运营商及行业用户等通过对这些QoS技术的灵活运用，可以在Internet或任何基于IP的网络上为客户提供有保证的区分服务。

**缩略语：**

| 缩略语 | 英文全名 | 中文解释 |
|---|---|---|
| AF | Assured Forwarding | 确保转发 |
| BE | Best Effort | 尽力转发 |
| CAR | Committed Access Rate | 约定访问速率 |
| CBWFQ | Class Based Weighted Fair Queuing | 基于类的加权公平队列 |
| CQ | Custom Queuing | 定制队列 |
| DiffServ | Differentiated Service | 区分服务 |
| DSCP | Differentiated Services Codepoint | 区分服务编码点 |
| EF | Expedited Forwarding | 加速转发 |
| FEC | Forwarding Equivalance Class | 转发等价类 |
| FIFO | First in First out | 先进先出 |
| GTS | Generic Traffic Shaping | 通用流量整形 |
| IntServ | Integrated Service | 综合服务 |
| IPHC | IP Header Compression | IP头压缩 |
| ISP | Inernet Service Provider | Internet服务提供商 |
| LFI | Link Fragmentation & Interleaving | 链路分片与交叉 |
| LLQ | Low Latency Queuing | 低时延队列 |
| LR | Line Rate | 物理接口总速率限制 |

| 缩略语 | 英文全名 | 中文解释 |
| --- | --- | --- |
| LSP | Label Switched Path | 标签交换路径 |
| MPLS | Multiprotocol Label Switching | 多协议标签交换 |
| PHB | Per-hop Behavior | 单中继段行为，指IP转发中每一跳的转发行为 |
| PQ | Priority Queuing | 优先队列 |
| QoS | Quality of Service | 服务质量，指报文传送的吞吐量、时延、时延抖动、丢失率等性能 |
| RED | Random Early Detection | 随机早期检测 |
| RSVP | Resource Reservation Protocol | 资源预留协议 |
| RTP | Real Time Protocol | 实时协议 |
| SLA | Service Level Agreement | 服务水平协议。是服务使用者和服务提供者之间签订的服务水平协议。服务提供者按此协议向服务使用者提供服务 |
| TE | Traffic Engineering | 流量工程 |
| ToS | Type of Service | 服务类型 |
| VoIP | Voice over IP | 通过IP报文传递语音报文 |
| VPN | Virtual Private Network | 虚拟专用网 |
| WFQ | Weighted Fair Queuing | 加权公平队列 |
| WRED | Weighted Random Early Detection | 加权随机早期检测 |

# 目 录

# 1 概述

## 1.1 产生背景

在传统的IP网络中，所有的报文都被无区别的等同对待，每个转发设备对所有的报文均采用先入先出（FIFO）的策略进行处理，它尽最大的努力（Best-Effort）将报文送到目的地，但对报文传送的可靠性、传送延迟等性能不提供任何保证。

网络发展日新月异，随着IP网络上新应用的不断出现，对IP网络的服务质量也提出了新的要求，例如VoIP等实时业务就对报文的传输延迟提出了较高要求，如果报文传送延时太长，用户将不能接受（相对而言，E-Mail和FTP业务对时间延迟并不敏感）。为了支持具有不同服务需求的语音、视频以及数据等业务，要求网络能够区分出不同的通信，进而为之提供相应的服务。传统IP网络的尽力服务不可能识别和区分出网络中的各种通信类别，而具备通信类别的区分能力正是为不同的通信提供不同服务的前提，所以说传统网络的尽力服务模式已不能满足应用的需要。

QoS技术的出现便致力于解决这个问题。

## 1.2 技术优点

QoS旨在针对各种应用的不同需求，为其提供不同的服务质量。如：

- 可以限制骨干网上 FTP 使用的带宽，也可以给数据库访问以较高优先级。
- 对于 ISP，其用户可能传送语音、视频或其他实时业务，QoS 使 ISP 能区分这些不同的报文，并提供不同服务。
- 可以为时间敏感的多媒体业务提供带宽和低时延保证，而其他业务在使用网络时，也不会影响这些时间敏感的业务。

## 1.3 QoS服务模型简介

通常QoS提供以下三种服务模型（服务模型，是指一组端到端的QoS功能）：

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

### 1.3.1 Best-Effort服务模型

Best-Effort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。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，它通过FIFO队列来实现。

### 1.3.2 IntServ服务模型

IntServ是一个综合服务模型，它可以满足多种QoS需求。这种服务模型在发送报文前，需要向网络申请特定的服务。这个请求是通过信令RSVP来完成的。RSVP是在应用程序开始发送报文之前来为该应用申请网络资源的，所以是带外信令。

应用程序首先通知网络它自己的流量参数和需要的特定服务质量请求，包括带宽、时延等。网络在收到应用程序的资源请求后，执行资源分配检查，即基于应用程序的资源申请和网络现有的资源情况，判断是否为应用程序分配资源。一旦网络确认为应用程序分配资源，则网络将为每个流（Flow，由两端的IP地址、端口号、协议号确定）维护一个状态，并基于这个状态执行报文的分类、流量监管、排队及其调度。应用程序在收到网络的确认信息（即确认网络已经为这个应用程序的报文预留了资源）后，才开始发送报文。只要应用程序的报文控制在流量参数描述的范围内，网络将承诺满足应用程序的QoS需求。

IntServ可以提供以下两种服务：

- 保证服务：它提供保证的带宽和时延限制来满足应用程序的要求。如 VoIP 应用可以预留 10M 带宽和要求不超过 1 秒的时延。
- 负载控制服务：它保证即使在网络过载的情况下，能对报文提供近似于网络未过载类似的服务，即在网络拥塞的情况下，保证某些应用程序的报文低时延和优先通过。

### 1.3.3 DiffServ服务模型

DiffServ是一个多服务模型，它可以满足不同的QoS需求。与IntServ不同，它不需要使用RSVP，即应用程序在发出报文前，不需要通知网络为其预留资源。对DiffServ服务模型，网络不需要为每个流维护状态，它根据每个报文的差分服务类（IP报文头中的差分服务标记字段DSCP值），来提供特定的服务。

在实施DiffServ的网络中，每一个转发设备都会根据报文的DSCP字段执行相应的转发行为，主要包括以下三类转发行为：

- 加速转发（EF）：主要用于低延迟、抖动和丢包率的业务，这类业务一般运行一个相对稳定的速率，需要在转发设备中进行快速转发；
- 确保转发（AF）：采用此转发行为的业务在没有超过最大允许带宽时能够确保转发，一旦超出最大允许带宽，则将转发行为分为 4 类，每类又可划分为 3 个不同的丢弃优先级，其中每一个确保转发类都被分配了不同的带宽资源。IETF 建议使用 4 个不同的队列分别传输 AF1x、AF2x、AF3x、AF4x 业务，并且每个队列提供 3 种不同的丢弃优先级，因此可以构成 12 个有保证转发的 PHB；
- 尽力转发（BE）：主要用于对时延、抖动和丢包不敏感的业务。

区分服务只包含有限数量的业务级别，状态信息的数量少，因此实现简单，扩展性较好。它的不足之处是很难提供基于流的端到端的质量保证。目前，区分服务是业界认同的IP骨干网的QoS解决方案，尽管IETF为每个标准的PHB都定义了推荐的DSCP值，但是设备厂家可以重新定义DSCP与PHB之间的映射关系，因此不同运营商的DiffServ网络之间的互通还存在困难，不同DiffServ网络在互通时必须维护一致的DSCP与PHB映射。

### 1.3.4 IntServ与DiffServ之间的互通

一般来讲，在提供IP网络的QoS时，为了适应不同规模的网络，在IP骨干网往往需要采用DiffServ体系结构；在IP边缘网可以有两种选择：采用DiffServ体系结构或采用IntServ体系结构。目前在IP边缘网络采用哪一种QoS体系结构还没有定论，也许这两种会同时并存于IP边缘网中。在IP边缘网采用DiffServ体系结构的情况下，IP骨干网与IP边缘网之间的互通没有问题。在IP边缘网采用IntServ体系结构的情况下，需要解决IntServ与DiffServ之间的互通问题，包括RSVP在DiffServ域的处理方式、IntServ支持的服务与DiffServ支持的PHB之间的映射。

RSVP在DiffServ域的处理可以有多种可选择的方式。例如下面两种：

- 一种方式为 RSVP 对 DiffServ 域透明，RSVP 在 IntServ 域边界转发设备终结，DiffServ 域对 IntServ 域采用静态资源提供方式。本方式实现相对简单，但可能造成 DiffServ 域资源的浪费。
- 一种方式为 DiffServ 域参与 RSVP 协议处理，DiffServ 域对 IntServ 域采用动态资源提供方式。本方式实现相对复杂，但可以优化 DiffServ 域资源的使用。

根据IntServ支持的服务和DiffServ提供的PHB的特点，IntServ支持的服务与DiffServ支持的PHB之间的映射问题可以通过下面方式解决：

- 将 IntServ 中的保证服务映射为 DiffServ 中的 EF。
- 将 IntServ 中的负载控制服务映射为 DiffServ 中的 AF。

# 2 IP QoS技术实现

目前，H3C的IP网络产品已经全面提供对DiffServ服务模型的支持：

- 完全兼容 DiffServ 服务模型的相关标准，包括 RFC2474、RFC2475、RFC2597、RFC2598 等；
- 支持以 IP Precedence 或 DSCP 作为 QoS 带内信令，可灵活配置；
- 支持 DiffServ 相关的功能组件，包括流量调节器（包括分类器、标记器、测量单元、整形器和丢弃器等）和各类 PHB（拥塞管理、拥塞避免等）。

## 2.1 IP QoS功能总述

IP QoS技术提供了下述功能：

- 流量分类和标记：依据一定的匹配规则识别出对象，是有区别地实施服务的前提，通常作用在接口入方向。
- 拥塞管理：是必须采取的解决资源竞争的措施，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，通常作用在接口出方向。
- 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，通常作用在接口出方向。
- 流量监管：对进入设备的特定流量的规格进行监管，通常作用在接口入方向。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，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，通常作用在接口出方向。
- 链路效率机制：可以改善链路的性能，间接提高网络的 QoS，如降低链路发包的时延（针对特定业务）、调整有效带宽。

在这些QoS技术中，流量分类和标记是基础，是有区别地实施服务的前提；而其他QoS技术则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

思想的具体体现。

网络设备对QoS的支持是通过结合各种QoS技术来实现的。图1描述了各种QoS技术在网络设备中的处理顺序。

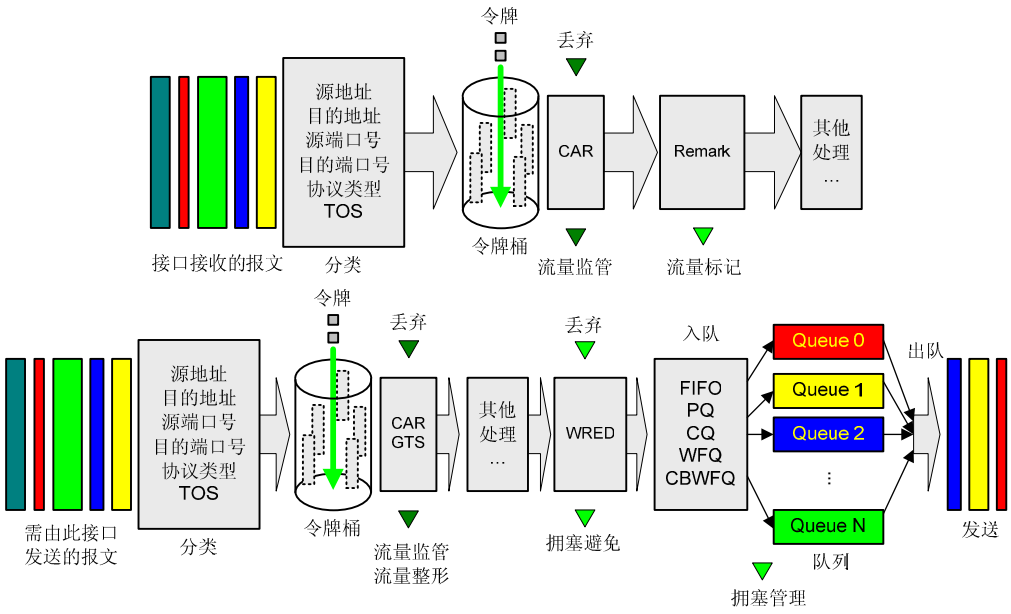


**图1 各QoS技术在同一网络设备中的处理顺序**

首先是流量分类，其后根据报文所属类别再将CAR、GTS、WRED、队列等各种技术运用其上，最终为具有不同网络需求的各种业务提供并保证所承诺的服务。

## 2.1 流量分类和标记

流量分类是将数据报文划分为多个优先级或多个服务类。网络管理者可以设置流量分类的策略，这个策略除可以包括IP报文的IP优先级或DSCP值、802.1p的CoS值等带内信令，还可以包括输入接口、源IP地址、目的IP地址、MAC地址、IP协议或应用程序的端口号等。分类的结果是没有范围限制的，它可以是一个由五元组（源IP地址、源端口号、协议号、目的IP地址、目的端口号）确定的流这样狭小的范围，也可以是到某某网段的所有报文。

下游网络可以选择接受上游网络的分类结果，也可以按照自己的分类标准对数据流量重新进行分类。

下面分别介绍一下在IPv4、IPv6、二层以太网中如何对流量进行分类和标记。

### 2.1.1 IP QoS业务分类

IP优先级和DSCP在报文中的位置如图2所示。

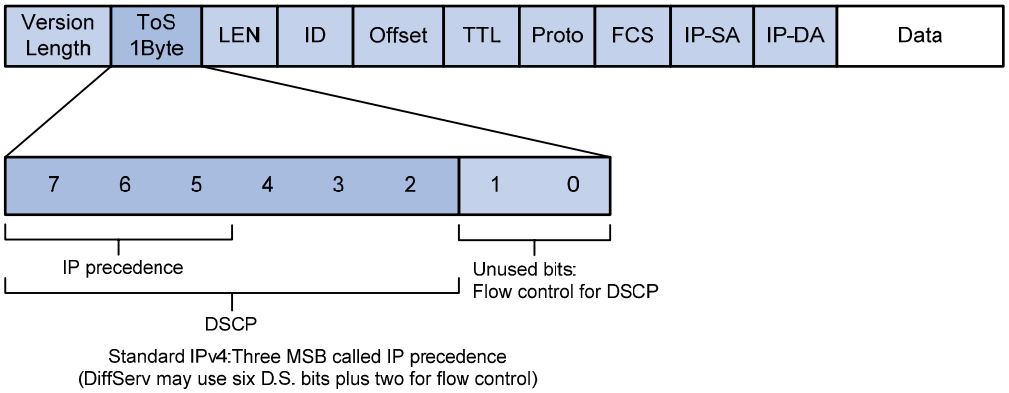


图2 IP/DSCP

(1) 基于 IP 优先级的业务分类

IPv4报文在IP报文头的ToS域中定义了8种IP业务类型，如表1所示。

表1 8种IP业务类型定义

| 业务类型 | IP 优先级 |
|---|---|
| Network Control | 7 |
| Internet work Control | 6 |
| CRITIC/ECP | 5 |
| Flash Override | 4 |
| Flash | 3 |
| Immediate | 2 |
| Priority | 1 |
| Routine | 0 |

(2) 基于 DSCP 的业务分类

DiffServ模型定义了64种业务类型，其中典型的业务类型定义如表2所示。

表2 典型的DSCP PHB定义

| 业务类型 | DSCP PHB | DSCP 值 |
|---|---|---|
| Network Control | CS7(111000) | 56 |
| IP Routing | CS6(110000) | 48 |

| 业务类型 | DSCP PHB | DSCP 值 |
|---|---|---|
| Interactive Voice | EF(101110) | 46 |
| Interactive Video | AF41(100010) | 34 |
| Video control | AF31(011010) | 26 |
| Transactional/Interactive(对应高优先级应用) | AF2x(010xx0) | 18、20、22 |
| Bulk Data(对应中优先级应用) | AF1x(001xx0) | 10、12、14 |
| Streaming Video | CS4(000100) | 4 |
| Telephony Signaling | CS3(000011) | 3 |
| Network Management | CS2(000010) | 2 |
| Scavenger | CS1(000001) | 1 |
| Best Effort | 0 | 0 |

### 2.1.2 IPv6 QoS业务分类

IPv6在报头中保留了类似IPv4的ToS域，称为传输级别域（TC），以继续为IP提供差分QoS服务，可以基于IPv6 ToS域进行业务分类与标记。同时IPv6报头中增加20比特流标签（Flow Label）域，可供后续扩展用。

### 2.1.3 以太网QoS业务分类

以太网在以太网帧头的VLAN TAG中定义了8种业务优先级（CoS，Class of Service），如图3所示。

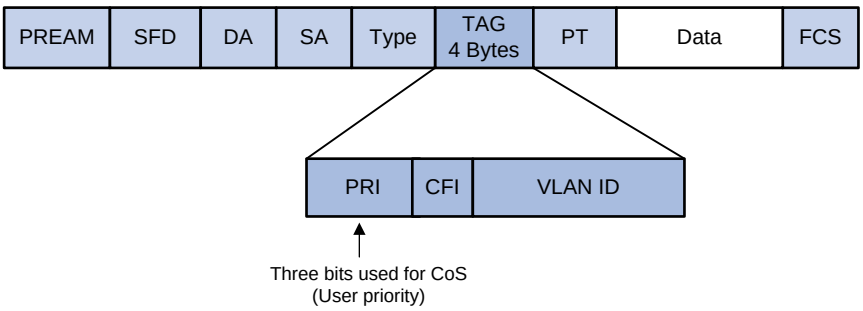


图3 802.1Q CoS

每一种业务映射到出端口的哪个队列转发将关系到该业务的时延、抖动和可获得的带宽。8种以太网业务类型定义如表3所示。

表3 8种以太网CoS业务类型定义

| 业务类型 | 业务特征 | 以太网CoS | 协议举例 |
|---|---|---|---|
| Network Control | 适用于网络维护与管理报文的可靠传输，要求低丢包率 | 7 | BGP, PIM, SNMP |
| Internet work Control | 适用于大型网络中区分于普通流量的网络协议控制报文，要求低丢包率和低时延 | 6 | STP, OSPF, RIP |
| Voice | 适用于语音业务，一般要求时延小于10 ms | 5 | SIP, MGCP |
| Video | 适用于视频业务，一般要求时延小于100 ms | 4 | RTP |
| Critical Applications | 适用于要求确保最小带宽的业务 | 3 | NFS, SMB, RPC |
| Excellent Effort | 也称为“CEO’s best effort”，比best effort的传输优先级稍高一些，用于一般的信息组织向最重要的客户发送信息 | 2 | SQL |
| Best Effort | 缺省使用的业务类型，无优先发送的要求，只要求“尽力而为”的服务质量 | 1 | HTTP, IM, X11 |
| Background | 适用于不影响用户或关键应用的批量传输业务 | 0 | FTP, SMTP |

IEEE 802.1Q推荐的业务类型到队列的映射关系定义如表4所示。

表4 IEEE 802.1Q推荐的CoS类型与队列的映射关系

| Number of queues | Queue ID | 业务类型 |
|---|---|---|
| 1 | 1 | Best Effort, Background, Excellent effort, Critical Applications, Voice, Video, Internet work Control, Network Control |
| 2 | 1 | Best Effort, Background, Excellent effort, Critical Applications |
| | 2 | Voice, Video, Internet work Control, Network Control |
| 3 | 1 | Best Effort, Background, Excellent effort, Critical Applications |
| | 2 | Voice, Video |
| | 3 | Network Control, Internet work Control |
| 4 | 1 | Best Effort, Background |
| | 2 | Critical Applications, Excellent effort |
| | 3 | Voice, Video |
| | 4 | Network Control, Internet work Control |

| Number of queues | Queue ID | 业务类型 |
|---|---|---|
| 5 | 1 | Best Effort, Background |
| | 2 | Critical Applications, Excellent effort |
| | 3 | Voice, Video |
| | 4 | Internet work Control |
| | 5 | Network Control |
| 6 | 1 | Background |
| | 2 | Best Effort |
| | 3 | Critical Applications, Excellent effort |
| | 4 | Voice, Video |
| | 5 | Internet work Control |
| | 6 | Network Control |
| 7 | 1 | Background |
| | 2 | Best Effort |
| | 3 | Excellent effort |
| | 4 | Critical Applications |
| | 5 | Voice, Video |
| | 6 | Internet work Control |
| | 7 | Network Control |
| 8 | 1 | Background |
| | 2 | Best Effort |
| | 3 | Excellent effort |
| | 4 | Critical Applications |
| | 5 | Video |
| | 6 | Voice |
| | 7 | Internet work Control |
| | 8 | Network Control |

## 2.2 拥塞管理

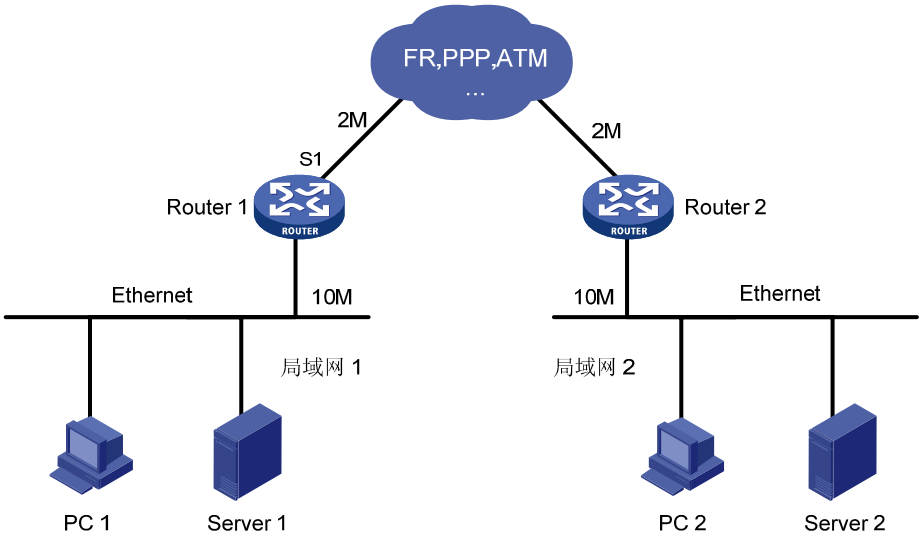


图4 网络拥塞示意图

在计算机数据通信中，通信信道是被多个计算机共享的，并且，广域网的带宽通常要比局域网的带宽小，这样，当一个局域网的计算机向另一个局域网的计算机发送数据时，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转发设备将不能发送某些报文，即网络发生了拥塞。如图4所示，当局域网 1向局域网 2以10M的速度发送数据时，将会使Router 1的串口S1发生拥塞。

拥塞管理是指在网络发生拥塞时，如何进行管理和控制。处理的方法是使用队列技术，具体过程包括队列的创建、报文的分类、将报文送入不同的队列、队列调度等。当接口没有发生拥塞时，报文到达接口后立即被发送出去，当报文到达的速度超过接口发送报文的速度时，接口就发生了拥塞。拥塞管理就会将这些报文进行分类，送入不同的队列；而队列调度将对不同优先级的报文进行分别处理，优先级高的报文会得到优先处理。常用的队列有FIFO、PQ、CQ、WFQ、CBWFQ、RTP优先队列等，下面将对这些队列进行详细的介绍。

### 2.2.1 先进先出队列（FIFO）

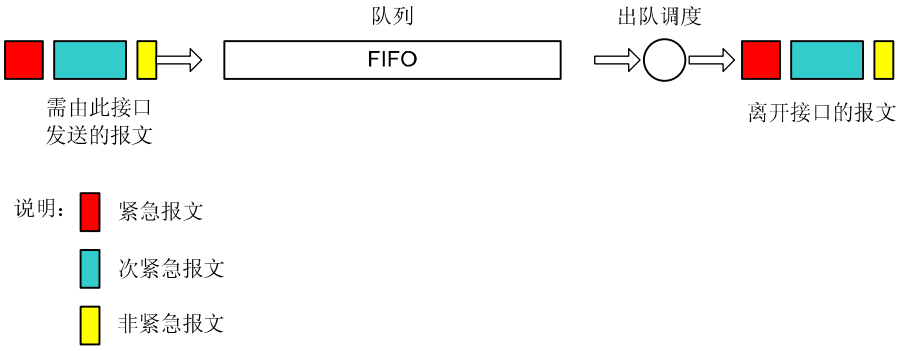


图5 先进先出队列示意图

如图5所示，先进先出队列（以后简称FIFO）不对报文进行分类，当报文进入接口的速度大于接口能发送的速度时，FIFO按报文到达接口的先后顺序让报文进入队列，同时，FIFO在队列的出口让报文按进队的顺序出队，先进的报文将先出队，后进的报文将后出队。

在如图4所示的网络图中，假设局域网 1的Server 1向局域网 2的Server 2发送关键业务的数据，局域网 1的PC 1向局域网 2的PC 2发送非关键业务的数据，则FIFO不会对这两种不同业务的报文做任何区别对待，报文的出队完全依赖于报文到来的先后顺序。

### 2.2.2 优先队列（PQ）

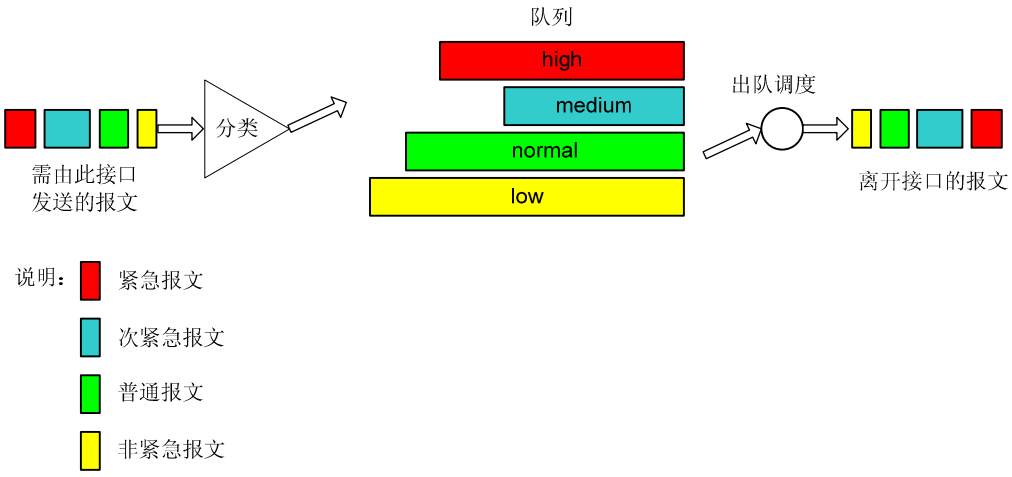


图6 PQ队列示意图

如图6所示，优先队列（以后简称PQ）对报文进行分类，对于IP网络，可以根据IP

报文的优先级/DSCP、五元组等条件进行分类，对于MPLS网络，则根据MPLS报文EXP域值进行分类。最终将所有报文分成最多至4类，分别属于PQ的4个队列中的一个，然后，按报文的类别将报文送入相应的队列。PQ的4个队列分别为：高优先队列、中优先队列、正常优先队列和低优先队列，它们的优先级依次降低。在报文出队的时候，PQ首先让高优先队列中的报文出队并发送，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，同样直到发送完，然后是正常优先队列和低优先队列。这样，分类时属于较高优先级队列的报文将会得到优先发送，而较低优先级的报文将会在发生拥塞时被较高优先级的报文抢先，使得高优先级业务（如VoIP）的报文能够得到优先处理，较低优先级业务（如E-Mail）的报文在网络处理完关键业务后的空闲中得到处理，既保证了高优先级业务的优先，又充分利用了网络资源。

在如图4所示的网络图中，假设局域网 1的Server 1向局域网 2的Server 2发送关键业务的数据，局域网 1的PC 1向局域网 2的PC 2发送非关键业务的数据，如果对Router 1的串口S1配置PQ进行拥塞管理，同时配置Server间的数据流进入较高优先级的队列，PC间的数据流进入较低优先级的队列，则PQ将对这两种不同业务的报文做区别对待，Server间的报文总是被先发送，直到暂时没有Server间的报文，转发设备才发送PC间的报文。

### 2.2.3 定制队列（CQ）

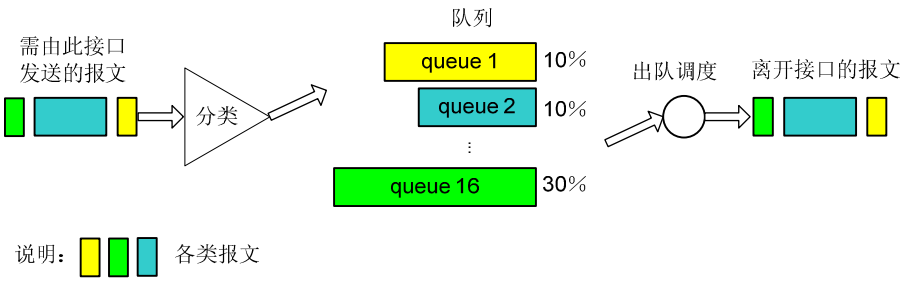


图7 定制队列示意图

如图7所示，定制队列（以后简称CQ）对报文进行分类，对于IP网络，可以根据IP报文的优先级/DSCP、五元组等条件进行分类，对于MPLS网络，则根据MPLS报文EXP域值进行分类。最终将所有报文分成最多至16类，分别属于CQ的16个队列中的一个，然后按报文的类别将报文送入相应的队列。CQ的1到16号队列，可以按用户的定义分配它们能占用接口带宽的比例。在报文出队的时候，CQ按定义的带宽比例分别从1到16号队列中取一定量的报文在接口上发送出去。

现在我们将CQ和PQ做个比较：

- PQ 赋予较高优先级的报文绝对的优先权，这样虽然可以保证关键业务的优先，但在较高优先级的报文的速度总是大于接口的速度时，将会使较低优先级的报文始终得不到发送的机会。采用 CQ，则可以避免这种情况的发生。
- CQ 可以规定每个队列中的报文所占接口带宽的比例，这样，就可以让不同业务的报文获得合理的带宽，从而既保证关键业务能获得较多的带宽，又不至于使非关键业务得不到带宽。但是，由于 CQ 轮询调度各个队列，它对高优先级尤其是实时业务的时延保证不如 PQ。

在如图4所示的网络图中，假设局域网 1的Server 1向局域网 2的Server 2发送关键业务的数据，局域网 1的PC 1向局域网 2的PC 2发送非关键业务的数据，可以在Router 1的串口S1配置如下CQ进行拥塞管理：

- 配置 Server 间的数据流进入队列 1，队列 1 中的报文占有 60%的带宽（例如每次出队 6000 个字节的报文）；
- 配置 PC 间的数据流进入队列 2，队列 2 中的报文占有 20%的带宽（例如每次出队 2000 个字节的报文）；
- 配置其他队列中的报文共占有 20%的带宽（例如每次出队 2000 个字节的报文）。

那么，CQ对这两种不同业务的报文将做区别对待。报文的发送采用轮询调度的方式，首先让队列1中的报文出队并发送，直到此队列中的报文被发送的字节数不少于6000字节，然后才开始发送队列2中的报文，直到此队列中的报文被发送的字节数不少于2000字节，然后是其他队列。各种数据报文占用带宽的情况如下：

- Router 1 的串口 S1 的物理带宽是 2M，则发送关键业务的数据所能占的带宽将至少为 1.2M（2*0.6），发送非关键业务的数据所能占的带宽将至少为 0.4M（2*0.2）。
- 当 S1 中除了上述两个数据流外没有其他数据要发送时，这两种数据流将按比例分享接口的剩余空闲带宽，即发送关键业务的数据所能占的带宽将为 1.5M（2*0.6/（0.2+0.6））， 发送非关键业务的数据所能占的带宽为 0.5M（2*0.2/（0.2+0.6））。
- 当 S1 中只有非关键业务的数据要发送时，发送非关键业务的数据所能占的带宽将可以为 2M。

### 2.2.4 加权公平队列（WFQ）

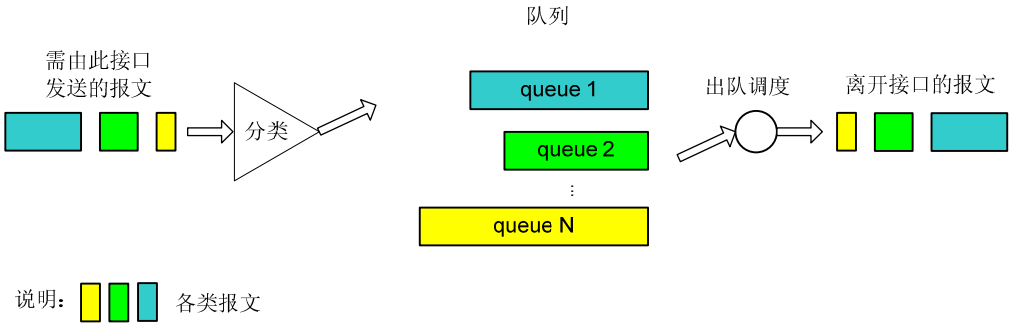


图8 加权公平队列示意图

如图8所示，加权公平队列（以后简称WFQ）对报文按流进行分类：对于IP网络，相同源IP地址、目的IP地址、源端口号、目的端口号、协议号、IP优先级（或DSCP）的报文属于同一个流；而对于MPLS网络，具有相同EXP域值的报文属于同一个流。每一个流被分配到一个队列，尽量将不同的流分入不同的队列。WFQ的队列数目N可以配置。在出队的时候，WFQ按流的IP优先级（或DSCP、EXP域值）来分配每个流占有出口的带宽。优先级的数值越小，所得的带宽越少。优先级的数值越大，所得的带宽越多。这样就保证了同优先级业务之间的公平，体现了不同优先级业务之间的权值。

例如：接口中当前有8个流，它们的优先级分别为0、1、2、3、4、5、6、7。则带宽的总配额将是所有（流的优先级+1）之和，即1+2+3+4+5+6+7+8=36。每个流所占带宽比例为：（自己的优先级数+1）/（所有（流的优先级+1）之和）。即，每个流可得的带宽比例分别为：1/36、2/36、3/36、4/36、5/36、6/36、7/36、8/36。

又如：当前共4个流，3个流的优先级为4，1个流的优先级为5，则带宽的总配额将是：（4+1）*3+（5+1）=21。那么，3个优先级为4的流获得的带宽比例均为5/21，优先级为5的流获得的带宽比例为6/21。

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

### 2.2.5 基于类的加权公平队列（CBWFQ）

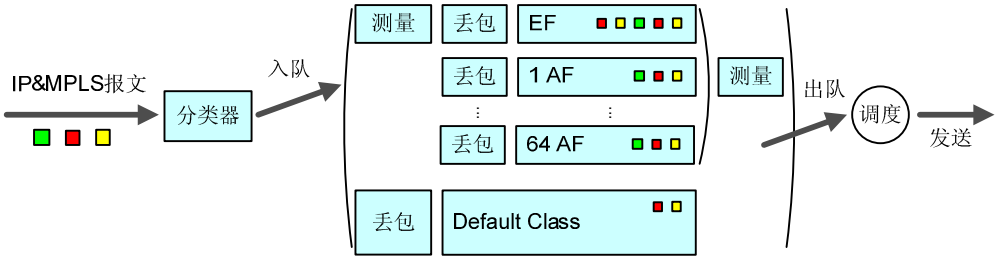


图9 基于类的加权公平队列示意图

如图9所示，基于类的加权公平队列（以后简称CBWFQ）首先根据IP优先级或者DSCP、输入接口、IP报文的五元组等规则来对报文进行分类；对于MPLS网络的LSR，主要是根据EXP域值进行分类，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，报文被送入系统定义的缺省类。

CBWFQ分为三类队列：EF队列、AF队列、BE队列，下面分别进行介绍。

(1) EF 队列

如图9所示，EF队列是一个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（通过此方式模拟出多个EF队列）。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。另外，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。

对于EF队列，由于在接口拥塞的时候流量限制开始起作用，所以用户不必设置队列的长度。由于优先队列中的报文一般是语音报文（VoIP），采用的是UDP报文，所以没有必要采用WRED的丢弃策略，采用尾丢弃策略即可。

(2) AF 队列

图9中，AF队列1到64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。这种队列技术应用了先进的队列调度算法，可以实现各个类的队列的公平调度。当接口中某些类别的报文没有时，AF队列的报文还可以公平地得到空闲的带宽，和时分复用系统相比，大大提高了线路的利用率。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（请参见后面关于WRED的描述）。

(3) 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（请参见后面关于WRED的描述）。

(4) 三种队列的配合

综上所述：

- 低时延队列（EF 队列）用来支撑 EF 类业务，被绝对优先发送，保证时延；
- 带宽保证队列（AF 队列）用来支撑 AF 类业务，可以保证每一个队列的带宽及可控的时延；
- 缺省队列（BE 队列）对应 BE 业务，使用接口剩余带宽进行发送。

对于进入EF队列和AF队列的报文要进行测量，考虑到链路层控制报文的发送链路层封装开销及物理层开销（如ATM信元头），建议EF队列与AF队列占用接口的总带宽不要超过接口带宽的80%。

### 2.2.6 RTP优先队列

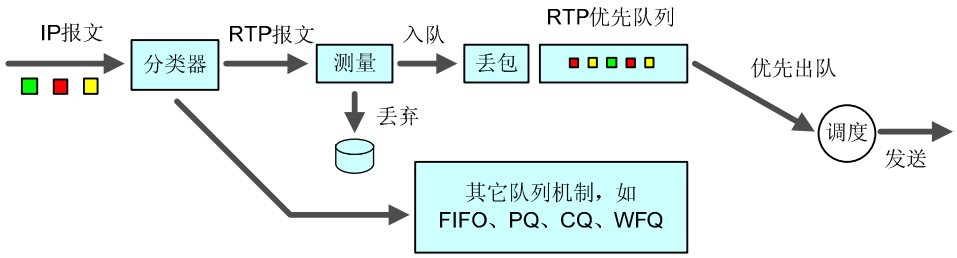


图10 RTP优先队列示意图

RTP优先队列是一种解决实时业务（包括语音与视频业务）服务质量的简单的队列技术。其原理就是将承载语音或视频的RTP报文送入高优先级队列，使其得到优先发送，保证时延和抖动降低为最低限度，从而保证了语音或视频这种对时延敏感业务的服务质量。

如图10所示，RTP报文被送入一个具有较高优先级的RTP优先队列。RTP报文是端口号在一定范围内为偶数的UDP报文，端口号的范围可以配置，一般为16384～32767。RTP优先队列可以同前面所述的各种队列（包括FIFO、PQ、CQ、WFQ）结合使用，它的优先级是最高的。由于CBWFQ中的LLQ完全可以解决实时业务的QoS问题，所以不支持将RTP优先队列与CBWFQ结合应用。

由于对进入RTP优先队列的报文进行了限速，超出规定流量的报文将被丢弃，这样在接口拥塞的情况下，可以保证属于RTP优先队列的报文不会占用超出规定的带宽，保护了其他报文的应得带宽，解决了RTP优先队列的流量可能“饿死”其他数据流量的问题。

### 2.2.7 队列技术对比

上述队列技术，突破了传统IP设备的单一FIFO拥塞管理策略，提供了强大的QoS能力，使得IP设备可以满足不同业务的不同服务质量的要求。为了更好的利用这些队列技术，现对各种队列技术做一比较。

表5 队列技术对比

| 队列名 | 队列数 | 优 点 | 缺 点 |
|---|---|---|---|
| FIFO | 1 | 1 缺省队列机制，队列长度配置简单，易于使用。<br>2 处理简单，处理延迟小。 | 1 所有报文同等对待，报文到来的次序决定了报文可占用的带宽、报文的延迟、报文的丢失。<br>2 对不匹配的数据源（即没有流控机制的流，如UDP报文发送）无约束力，不匹配的数据源会造成匹配的数据源（如TCP报文发送）带宽受损失。<br>3 对时间敏感的实时应用（如VoIP）的延迟得不到保证。 |
| PQ | 4 | 可对不同业务数据提供绝对的优先，对时间敏感的实时应用（如VoIP）的延迟可以得到保证。对优先业务的报文的带宽占用可以绝对优先。 | 1 需配置，处理速度慢。<br>2 如果不对高优先级的报文的带宽加限制，会造成低优先级的报文得不到带宽。 |

| | | | |
|---|---|---|---|
| CQ | 16 | 1 可对不同业务的报文按带宽比例分配带宽。<br>2 当没有某些类别的报文时，能自动增加现存类别的报文可占的带宽。 | 1 需配置，处理速度慢。<br>2 不适于解决对时延敏感的实时业务。 |
| WFQ | 用户决定 | 1 配置简单，流量分类自动完成。<br>2 可以保护匹配（交互）的数据源（如TCP报文发送）的带宽。<br>3 可以使延迟的抖动减小。<br>4 可以减小数据量小的交互式应用的延迟。<br>5 可以为不同优先级的流分配不同的带宽。<br>6 当流的数目减少时，能自动增加现存流可占的带宽。 | 1 处理速度比FIFO要慢，但比PQ、CQ要快（在分类规则较多情况）。<br>2 不适于解决对时延敏感的实时业务。 |
| RTPQ | 1 | 与上述队列配合使用，为语音等低延时流量提供绝对优先调度，保证了实时业务的时延。 | 流分类方式不及CBWFQ丰富。 |
| CBWFQ | 用户决定 | 1 可以对数据根据灵活、多样的分类规则进行划分，分别为EF（加速转发）、AF（确保转发）、BE（尽力转发）业务提供不同的队列调度机制。<br>2 可以为AF业务提供严格、精确的带宽保证，并且保证各类AF业务之间根据权值按一定的比例关系进行队列调度。<br>3 可以为EF业务提供绝对优先的队列调度，确保实时数据的时延；同时通过对高优先级数据流量的限制，克服了PQ的低优先级队列可能“饿死”的弊病。<br>4 对于尽力转发的缺省类数据，提供基于流的公平队列调度机制（WFQ）。 | 当配置的类较多时，系统开销比较大。 |

## 2.3 拥塞避免

过度的拥塞会对网络资源造成极大危害，必须采取某种措施加以解除。拥塞避免是一种流控机制，它可以通过监视网络资源（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。

与端到端的流控相比，这里的流控具有更广泛的意义，它影响到设备中更多的业务

流的负载。设备在丢弃报文时，并不排斥与源端的流控动作（比如TCP流控）的配合，更好地调整网络的流量到一个合理的负载状态。丢包策略和源端流控机制有效的组合，可以使网络的吞吐量和利用效率最大化，并且使报文丢弃和延迟最小化。

### 2.3.1 传统的丢包策略

传统的丢包策略采用尾部丢弃（Tail-Drop）的方法。当队列的长度达到某一最大值后，所有新到来的报文都将被丢弃。

这种丢弃策略会引发TCP全局同步现象——当队列同时丢弃多个TCP连接的报文时，将造成多个TCP连接同时进入拥塞避免和慢启动状态以降低并调整流量，而后又会在某个时间同时出现流量高峰，如此反复，使网络流量不停震荡。

### 2.3.2 RED与WRED

为避免TCP全局同步现象，可使用RED或WRED。

在RED类算法中，为每个队列都设定上限和下限，对队列中的报文进行如下处理：

- 当队列的长度小于下限时，不丢弃报文；
- 当队列的长度超过上限时，丢弃所有到来的报文；
- 当队列的长度在上限和下限之间时，开始随机丢弃到来的报文。队列越长，丢弃概率越高，但有一个最大丢弃概率。

直接采用队列的长度和上限、下限比较并进行丢弃，将会对突发性的数据流造成不公正的待遇，不利于数据流的传输。RED类算法采用平均队列长度和设置的队列上限、下限比较来确定丢弃的概率。计算队列平均长度的公式为：平均队列长度=（以前的平均队列长度×（1-1/（2的n次方））+（当前队列长度×（1/（2的n次方））），其中n可以通过命令配置。队列平均长度既反映了队列的变化趋势，又对队列长度的突发变化不敏感，避免了对突发性数据流的不公正待遇。

WRED算法在RED算法的基础上引入了优先权，它引入IP优先级、DSCP和MPLS EXP区别丢弃策略，考虑了高优先权报文的利益，使其被丢弃的概率相对较小。如果对于所有优先权配置相同的丢弃策略，那么WRED就变成了RED。

RED和WRED通过随机丢弃报文避免了TCP的全局同步现象，使得当某个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，无论什么时候，总有TCP连接在进行较快的发送，提高了线路带宽的利用率。

### 2.3.3 WRED和队列机制的关系

WRED和队列机制的关系如图11所示。

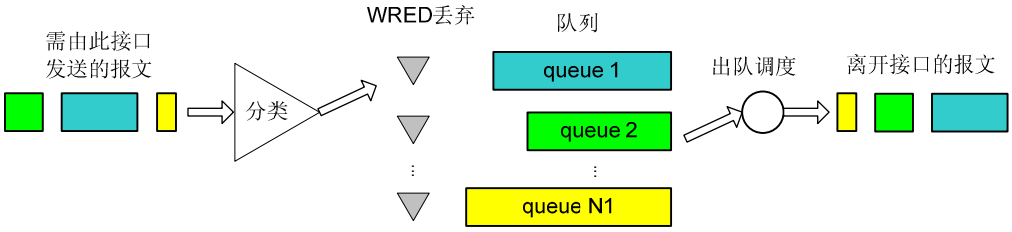


图11 WRED和队列机制关系示意图

当队列机制采用WFQ时：

- WRED 可以为不同优先级的报文设定计算队列平均长度时的指数、上限、下限、丢弃概率，从而对不同优先级的报文提供不同的丢弃特性。
- 可以实现基于流的 WRED。这是因为，在进行分类的时候，不同的流有自己的队列，对于流量小的流，由于其队列长度总是比较小，所以丢弃的概率将比较小。而流量大的流将会有较大的队列长度，从而丢弃较多的报文，保护了流量较小的流的利益。

当队列机制采用FIFO、PQ、CQ时：

- WRED 可以为每个队列设定计算队列平均长度时的指数、上限、下限、丢弃概率，为不同队列的报文提供不同的丢弃特性。
- 对于流量小的流，由于其报文的个数较少，从统计概率来说，被丢弃的概率也会较小，也可以保护流量较小的流的利益。

## 2.4 流量监管与流量整形

流量监管的典型作用是限制进入某一网络的某一连接的流量与突发。在报文满足一定的条件时，如某个连接的报文流量过大，流量监管就可以对该报文采取不同的处理动作，例如丢弃报文，或重新设置报文的优先级等。通常的用法是使用CAR来限制某类报文的流量，例如限制HTTP报文不能占用超过50%的网络带宽。

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

### 2.4.1 约定访问速率（CAR）

对于ISP来说，对用户送入网络中的流量进行控制是十分必要的。对于企业网，对某些应用的流量进行控制也是一个有力的控制网络状况的工具。网络管理者可以使用约定访问速率（以后简称CAR）来对流量（不包括紧急报文、协议报文）进行控制。

CAR利用令牌桶进行流量控制。

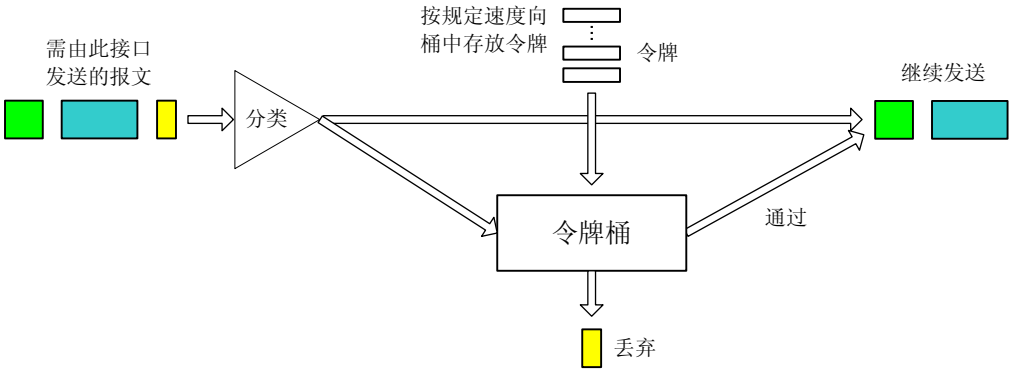


图12 CAR进行流量控制的基本处理过程示意图

图12所示为利用CAR进行流量控制的基本处理过程。首先，根据预先设置的匹配规则来对报文进行分类，对于不符合分类规则的报文，就直接继续发送，并不需要经过令牌桶的处理；对于符合分类规则的报文，则会进入令牌桶中进行处理。令牌桶是一个控制数据流量的很好的工具。令牌桶按用户设定的速度向桶中放置令牌，并且用户可以设置令牌桶的容量，当桶中令牌的量达到桶的容量时，令牌不再增加。桶中的令牌数表示可借贷的数据突发量，这样可以允许数据的突发性传输。如果令牌桶中有足够的令牌可以用来发送报文，则允许报文通过，报文可以被发送出去，同时，令牌桶中的令牌量随报文发送相应减少。如果令牌桶中的令牌少到报文不能再发送时，则报文被丢弃。等到桶中生成了新的令牌，报文才可以被发送出去，这就可以限制报文的流量只能小于等于令牌生成的速度，达到限制流量的目的。

在实际应用中，CAR不仅可以用来进行流量控制，还可以进行报文的标记或重新标记。即，通过CAR可以设置IP报文的优先级或修改IP报文的优先级，达到标记报文的目的。例如，当报文符合流量特性的时候，可以设置报文的优先级为5；当报文不符合流量特性的时候，可以丢弃，也可以设置报文的优先级为1并继续进行发送。这样，后续的处理可以尽量保证不丢弃优先级为5的报文，在网络不拥塞的情

况下，也发送优先级为1的报文；当网络拥塞时，首先丢弃优先级为1的报文，然后才丢弃优先级为5的报文。

此外，CAR还可以对流量进行多次分类处理。例如，可以限制部分报文的流量符合某个流量特性，然后限制所有报文的总流量。

### 2.4.2 通用流量整形（GTS）

通用流量整形（以后简称GTS）可以对不规则或不符合预定流量特性的流量进行整形，以利于网络上下游之间的带宽匹配。

GTS与CAR一样，均采用了令牌桶技术来控制流量。GTS与CAR的主要区别在于：利用CAR进行报文流量控制时，对不符合流量特性的报文进行丢弃；而GTS对于不符合流量特性的报文则是进行缓冲，减少了由突发流量造成的报文的丢弃。

GTS的基本处理过程如图13所示，其中用于缓存报文的队列称为GTS队列。

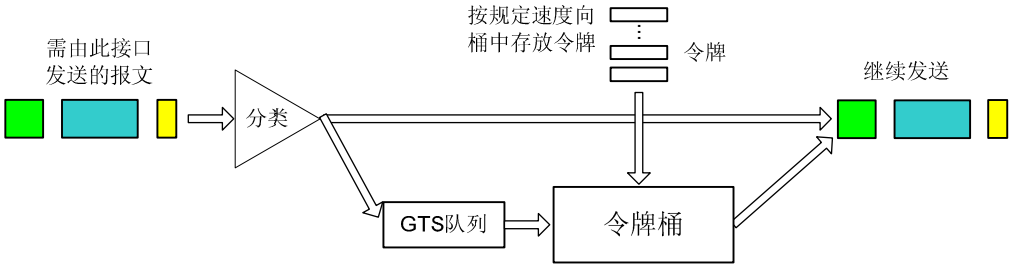


**图13 GTS处理过程示意图**

GTS可以对接口上指定的报文流或所有报文进行整形。当报文到来的时候，首先根据预先设置的匹配规则来对报文进行分类，对于不符合分类规则的报文，就继续发送，不需要经过令牌桶的处理；对于符合分类规则的报文，则会进入令牌桶中进行处理。令牌桶按用户设定的速度向桶中放置令牌，如果令牌桶中有足够的令牌可以用来发送报文，则报文直接被继续发送出去，同时，令牌桶中的令牌量随报文发送相应减少。当令牌桶中的令牌少到报文不能再发送时，报文将被缓存入GTS队列中（后续到达GTS的报文，检测到缓存队列中有报文，直接入队。若队长达到上限，报文被直接丢弃）。当GTS队列中有报文的时候，GTS按一定的周期从队列中取出报文进行发送，每次发送都会与令牌桶中的令牌数作比较，直到令牌桶中的令牌数减少到队列中的报文不能再发送或是队列中的报文全部发送完毕为止。

在如图4所示的网络图中，为了减少由突发流量造成的报文的丢失，可以在Router 1的出口对报文进行GTS处理，对于超出GTS流量特性的报文，将在Router 1中缓

冲；当可以继续发送下一批报文时，GTS再从缓冲队列中取出报文进行发送。这样，发往Router 2的报文将都符合Router 2的流量规定，从而减少报文在Router 2上的丢弃。相反，如果不在Router 1的出口做GTS处理，则所有超出Router 2的CAR流量特性的报文将被Router 2丢弃。

### 2.4.3 物理接口总速率限制（LR）

利用物理接口总速率限制（以后简称LR）可以在一个物理接口上限制接口发送报文（包括紧急报文、协议报文）的总速率。

LR的处理过程仍然采用令牌桶进行流量控制。如果用户在转发设备的某个接口上配置了LR，规定了流量特性，则所有经由该接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可以用来发送报文，则报文可以发送；如果令牌桶中的令牌不满足报文的发送条件，则报文入QoS队列进行拥塞管理。这样，就可以对通过该物理接口的报文流量进行控制。

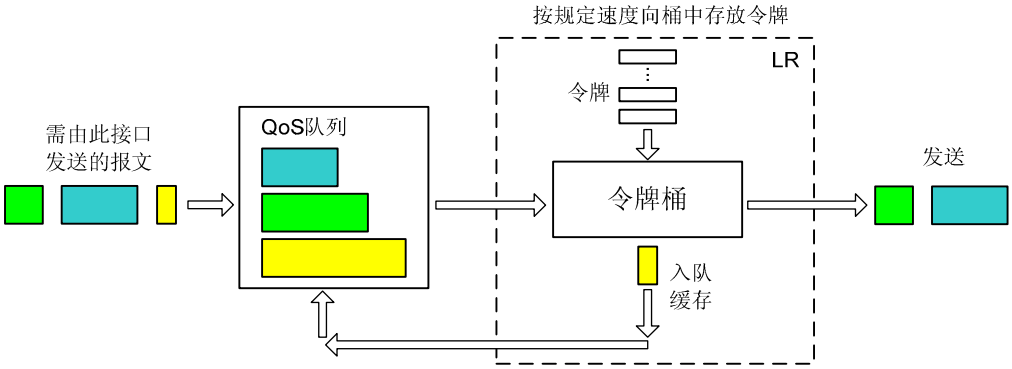


图14 LR处理过程示意图

LR的处理过程如图14所示。同样的，由于采用了令牌桶控制流量，当令牌桶中积存有令牌时，可以允许报文的突发性传输。当令牌桶中没有令牌的时候，报文将不能被发送，只有等到桶中生成了新的令牌，报文才可以发送，这就可以限制报文的流量只能小于等于令牌生成的速度，达到限制流量的同时允许突发流量通过的目的。

由于LR不需对报文进行分类，所以当用户只要求对所有报文进行限速时，可以采用LR，配置比较简单。

相较于GTS，LR不但能够对超过流量限制的报文进行缓存，并且可以利用QoS丰富的队列来缓存报文，而GTS则是将报文缓存在GTS队列中。

## 2.5 链路效率机制

链路效率机制可以改善链路的性能，间接提高网络的QoS，如降低链路发包的时延（针对特定业务）、调整有效带宽。

目前有LFI和IPHC两种链路效率机制，下面将分别进行介绍。

### 2.5.1 链路分片与交叉（LFI）

对于低速链路，即使为语音等实时业务报文配置了高优先级队列（如RTP优先队列或LLQ），也不能够保证其时延与抖动，原因在于，接口在发送其他数据报文的瞬间，语音业务报文只能等待，而对于低速接口发送较大的数据报文要花费相当长的时间。采用LFI以后，数据报文（非RTP优先队列和非LLQ中的报文）在发送前被分片、逐一发送，而此时如果有语音报文到达则被优先发送，从而保证了语音等实时业务的时延与抖动。LFI主要用于低速链路。

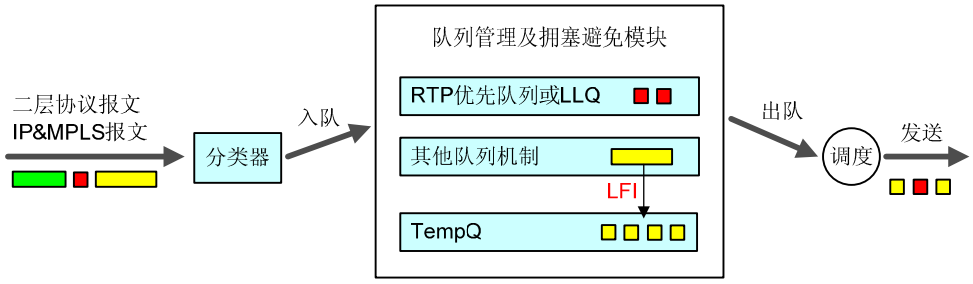


图15 LFI工作原理示意

如图15所示，应用LFI技术，在大报文出队的时候，可以将其分为定制长度的小片报文，这就使RTP优先队列或LLQ中的报文不必等到大报文发完后再得到调度，它等候的时间只是其中小片报文的发送时间，这样就大大降低低速链路因为发送大报文造成的时延。

### 2.5.2 IP报文头压缩（IPHC）

RTP协议用于在IP网络上承载语音、视频等实时多媒体业务。RTP报文包括数据部分和头部分，RTP的数据部分相对小，而RTP的头部分较大。12字节的RTP头，加上20字节的IP头和8字节的UDP头，就是40字节的IP/UDP/RTP头。而RTP典型的负载是20字节到160字节。为了避免不必要的带宽消耗，可以使用IPHC特性对报文头进行压缩。IPHC可以将IP/UDP/RTP头从40字节压缩到2～5字节（不使用校验和可压缩到2字节），对于40字节的负载，头压缩到5字节，压缩比为（40+40）

/（40+5），约为1.78，可见效果是相当可观的。IPHC可以有效的减少链路（尤其是低速链路）带宽的消耗，提高链路的利用率。

# 3 MPLS QoS技术实现

目前，IP网络仅提供对DiffServ服务模型的支持，但MPLS网络可以同时提供对DiffServ服务模型和IntServ服务模型的支持：

- MPLS 网络的 DiffServ 与 IP 网络的 DiffServ 原理相同，不同的是 MPLS 网络是通过 MPLS 报文头中的 EXP 值携带 DiffServ PHB 来实现。
- MPLS 网络的 IntServ 通过 MPLS-TE 技术实现。

本节将对这两种MPLS QoS技术进行简单介绍。

## 3.1 MPLS DiffServ

DiffServ的基本机制是：在网络边缘，根据业务的服务质量要求将该业务映射到一定的业务类别中，利用IP分组中的DS字段（由ToS域而来）唯一的标记该类业务，然后，骨干网络中的各节点根据该字段对各种业务采取预先设定的服务策略，保证相应的服务质量。DiffServ的这种对服务质量的分类和标签机制与MPLS的标签分配十分相似，事实上，基于MPLS的DiffServ就是通过将DS的分配与MPLS的标签分配过程结合来实现的。

MPLS DiffServ通过MPLS报文头中的EXP值携带DiffServ PHB实现，标签交换路由器（LSR）在做出转发决策时要考虑MPLS EXP值。但是DiffServ PHB最多可以支持64个编码值，如何承载在只有3bit的EXP字段中？MPLS DiffServ提供两种解决方案，具体采用哪种方案将取决于具体的应用环境：

- E-LSP 路径，即由 EXP 位决定 PHB 的 LSP。该方法适用于支持少于 8 个 PHB 的网络，特定的 DSCP 直接映射为特定的 EXP，标识到特定的 PHB。在转发过程中，LSP 决定转发路径，但是 EXP 决定在每一跳 LSR 上的调度和丢弃优先级，因此同一条 LSP 可以承载 8 类不同 PHB 的流，通过 MPLS 头部的 EXP 域来进行区分。EXP 可以直接由运营商配置决定，也可以从报文的 DSCP 直接映射得到。这种方法不需要信令协议转的 PHB 信息，而且标签使用率较高，状态易于维护。
- L-LSP 路径，即由标签和 EXP 共同决定 PHB 的 LSP。该方法适用于支持任意数量 PHB 的网络。在转发过程中，标签不仅用于决定转发路径而且决定在 LSR 上的调度行为，而 EXP 位则用于决定数据报文的丢弃优先级。由于通过

标签来区分业务流的类型，因此需要为不同的流建立不同的 LSP。这种方法需要使用更多的标签，保存更多的状态。

H3C网络产品提供的MPLS DiffServ技术采用E-LSP解决方案：

- 以EXP值作为QoS信令，表示MPLS网络中的业务优先级，EXP值将指示报文在传输中所使用的队列号和丢弃优先级（DiffServ PHB与EXP的映射关系建议如 表 6所示）。
- 所有的 DiffServ 功能组件（流量调节器和各类 PHB）均为 EXP 做了扩展。

表6 DiffServ PHB与EXP的映射关系表

| DSCP PHB | EXP 优先级 |
|---|---|
| CS7(111000) | 111 |
| CS6(110000) | 110 |
| EF(101110) | 101 |
| AF4X(100xx0) | 100 |
| AF3X(011xx0) | 011 |
| AF2X(010xx0) | 010 |
| AF1X(001xx0) | 001 |
| Best Effort | 000 |

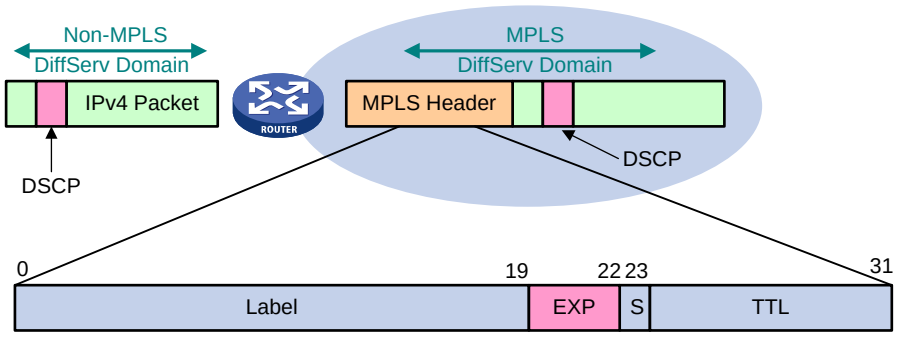


图16 DiffServ PHB与EXP的转换

如图16所示，在MPLS网络的边缘，缺省情况下，将IP报文的IP优先级直接拷贝到MPLS报文的EXP域；但是在下面的情况下，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，则可以根据一定的分类策略，例如IP报文的IP优先级或DSCP、IP报文的五元组、输入接口等，在MPLS网络边缘重新设置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。

## 3.2 MPLS-TE

MPLS-TE是一种间接改善MPLS网络QoS的技术。传统路由协议（如OSPF或IS-IS）主要是保障网络的连通性和可达性，通常选取不是非常灵敏的参数作为SPF计算根据，导致网络负载不均衡、路由动荡等缺陷；MPLS-TE在网络资源有限的前提下，将网络流量合理引导，达到实际网络流量负载与物理网络资源相匹配，间接改善了网络的服务质量。

流量工程的性能指标包括两个方面：

- 面向业务的性能指标：增强业务的 QoS 性能。例如对分组丢失、时延、吞吐量以及服务等级协定 SLA 的影响；
- 面向资源的性能指标：优化资源利用。带宽是一种重要的资源，对带宽资源进行高效管理是流量工程的一项中心任务。MPLS TE 结合了 MPLS 技术与流量工程，通过建立到达指定路径的 LSP 隧道进行资源预留，使网络流量绕开拥塞节点，达到平衡网络流量的目的。

MPLS-TE可以为流建立有带宽保证的路径，其具体方法：每条链路下都可以配置带宽（最大物理带宽、最大可预留带宽），IGP扩展将泛洪这些信息，生成TEDB；在流量入口建立LSP路径时，可以指定LSP所需要的带宽；CSPF进行计算时会计算出满足带宽、时延等要求的路径；最后RSVP-TE根据计算结果建立路径。通过将TE metric设置为链路时延，从而可以按照VoIP等的时延需求计算可用路径。

MPLS-TE的链路优先级可以用于将某些LSP标记为比其他LSP更重要，允许其抢占低优先级LSP的带宽资源，这样确保了：（1）在缺少重要LSP的情况下，可以利用低优先级LSP预留资源；（2）重要LSP始终沿最优路径建立，不受已有预留的影响。MPLS-TE用两个优先级属性来决定是否可以进行抢占：建立优先级和保持优先级，并定义了8个优先级，优先级的范围是0～7，0最重要。建立优先级用于在建立LSP时控制对资源的接入，而保持优先级用于对已建立LSP的资源访问。当新建一条路径Path1时，如果需要与已建立的路径Path2争夺资源，只有当Path1的建立优先级高于Path2的保持优先级时，Path1才能抢占成功。

MPLS-TE的自动带宽调整特性基于隧道进行流量统计，按照指定频率在一定范围

内调节隧道的带宽，可以达到当用户业务增多时，自动调整分配给这些LSP的带宽。

# 4 典型组网方案

## 4.1 企业VPN QoS实施

ISP可以通过IP网络向企业提供VPN业务以降低企业的建网费用/租用线费用，对于企业很有吸引力。VPN可以用于连接出差人员与企业总部、异地分支机构与企业总部、企业合作伙伴与企业总部，提供它们之间的信息传输。但是如果VPN不能保证企业运营数据的及时有效发送（即提供有效的QoS保证），那么VPN将仍然不能有效的为企业服务。例如，往来工作函件数据库访问需要受到优先对待，保证这些应用的带宽要求，而对于与工作无关的E-Mail、WWW访问等则可以按照Best-Effort信息流对待。

H3C提供的丰富QoS机制完全能够满足企业VPN的上述要求：

- 对于不同的业务分别对 IP 优先级/DSCP 进行标记，并且基于 IP 优先级/DSCP 对流量进行分类。
- 通过 CBWFQ 队列调度算法，保证企业运营数据的带宽、时延、时延抖动等 QoS 性能。
- 通过 WRED/尾丢弃机制对于 VPN 信息区别对待，避免网络内部流量振荡。
- 通过流量监管机制限制 VPN 中不同信息流的流量。

在VPN各个站点的CE路由器上对业务流进行分类和着色，例如可以将业务流分为数据库访问、重要工作邮件和WWW访问三类，并且根据需要利用CAR将这三种业务报文的优先级分别标记为高、中、低。同时VPN服务提供商还可在每个CE路由器的接入端口设置CAR和GTS功能，分别用于流量监管和流量整形，以此来控制由各VPN站点进入服务提供商网络的报文流量不会超过承诺的流量上限。LR则可以应用在CE路由器上进行接口总速率限制，裁减和控制CE路由器接入端口的带宽，保护VPN服务提供商的利益。

在VPN服务提供商网络的各PE路由器上，缺省情况下，MPLS EXP会拷贝IP报文的优先级。这样可以在VPN服务提供商网络的各PE和P路由器，通过配置WFQ、CBWFQ等队列来控制报文的调度方式，保证在网络拥塞发生时具有较高优先级的报文能够优先获得服务，以达到低时延、低时延抖动等目的，同时可以设置WRED来避免TCP流量的全局同步现象。

另外，如果ISP希望定义与用户网络不同的服务级别，或者不信任用户网络的IP优先级，也可以采用在PE路由器的入口，根据一定的规则，对MPLS EXP进行重新标记的方式。

## 4.2 VoIP QoS网络设计

VoIP QoS的基本要求：丢包率<1%，端到端单程延时<150ms（不同编码方式要求不同），抖动小于30ms。每路会话的带宽需求为21～106Kbps，主要根据采样速率、压缩算法、二层封装等内容决定。

对于VoIP QoS的网络设计，主要考虑以下几点：

(1) 合理设计网络带宽保证业务传送需求。

合适的带宽是保障业务QoS的重要手段。例如一路比较清晰的IP电话需要占用21～106Kbps范围内的网络带宽，具体带宽占用值依据不同的编解码算法而有所不同。假如一路电话占用21Kbps带宽，则不可能在1条64Kbps的链路上同时承载4路这样的IP电话。在网络建设时，可根据业务模型规划网络，合理配备带宽资源，并根据业务的使用频度来考虑业务对带宽的复用。但如果网络中各种业务都是频繁业务，则只能采用带宽叠加的方法。在实际使用线路时，必须综合考虑运营商网络带宽资源的实际情况加以分析。

(2) 选择合适的语音编解码技术。

语音压缩有多种算法，不同的算法所需传输带宽不同。在网络带宽充裕的情况下，可采用G.711进行压缩，其压缩率很低，基本不影响语音质量，但要求传输带宽很高；在网络带宽紧张的情况下，可采用G.729进行压缩，其对带宽的占用率低，但需要以一定的语音失真为代价（但是在标准规定的范围内）。

(3) 选择合适的 QoS 管理技术。

为了降低语音报文的传输时延，可以将以下技术结合起来：

- 采用LLQ队列调度算法，使得语音报文进入LLQ，保证语音报文在拥塞发生的情况下被优先调度；也可以采用RTP优先队列与WFQ结合的方式，如 图17所示。
- 在低速链路上采用IPHC报文头压缩技术提高链路利用率、降低报文时延，如图 18所示。
- 在低速链路上采用 LFI 技术降低语音报文的时延。

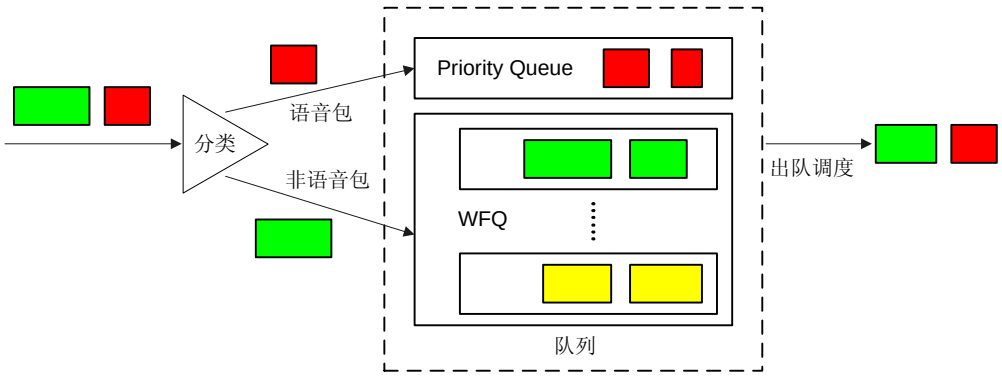


图17 VoIP支持

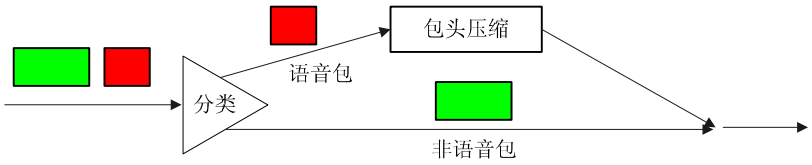


图18 IP报文头压缩

(4) 合理部署接入层、汇聚层和骨干层的 QoS 功能。

- 在接入层网络进行业务规划时，可以考虑将不同的业务划分到不同的 VLAN 中，比如语音业务终端统一在 VLAN 10，视频业务终端统一在 VLAN 20，高速上网业务统一在 VLAN 30 中。这样，接入层设备可以根据不同的 VLAN 号对业务进行排序，优先转发实时性业务，从而保证 QoS。假如接入层设备不支持 VLAN 划分，则在网络规划时，可以考虑将话音、视频与高速上网业务分配在不同的网络中，并在物理层隔离实现，例如可以将不同业务终端的 IP 地址划分在不同的网段内。
- 在汇聚层网络可以通过划分不同的虚拟专网（VPN）保证语音、视频业务的带宽。对于无法划分虚拟专网的网络，则可以在接入层和汇聚层之间对业务进行分类，配置语音、视频业务的优先级高于数据业务，然后在汇聚层按照优先级高低进行数据报文转发，从而避免大量突发数据业务对语音、视频业务的影响。
- 在骨干层网络将语音、视频业务统一归属到一个 VPN 中，和数据业务区分开来，从而保证语音、视频业务的 QoS 和网络安全。在同一个 VPN 内部还可以对业务进行分类，按照优先级高低进行数据报文转发，也可以结合 MPLS 技术，综合全网资源情况，对语音业务提供低时延、有带宽保证的转发路径，保证语音业务的服务质量。

# 5 参考文献


- RFC 1349：Type of Service in the Internet Protocol Suite
- RFC 1633：Integrated Services in the Internet Architecture: an Overview
- RFC 2205：Resource Reservation Protocol (RSVP)-Version1 Functional Specification
- RFC 2210：The use of RSVP with IETF Integrated Services
- RFC 2211：Specification of the Controlled-Load Network Element Service
- RFC 2212：Specification of Guaranteed Quality of Service
- RFC 2215：General Characterization Parameters for Integrated Service Network Elements
- RFC 2474：Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers
- RFC 2475：An Architecture for Differentiated Services
- RFC 2597：Assured Forwarding PHB Group
- RFC 2598：An Expedited Forwarding PHB (Per-Hop Behavior)
- RFC 2697：A single rate three color marker
- RFC 2698：A two rate three color marker
- RFC 3270 ： Multi-Protocol Label Switching (MPLS) Support of Differentiated Services
- IEEE 802.1Q-REV/D5.0 Annex G