网络大爬虫
Network Addicts

QoS专题

2011年第1期总第4期

H³Care俱乐部VIP客户专属

H3C

编者寄语

邮递P游记

邮递P乃A国人氏，司邮差一职，外形短小精悍，双腿日行千里，人送外号飞毛腿。一天国王召见邮递P，令其将一封鸡毛信件交与E国国王，速去速回，不得有误，否则以军法处置。

A国与E国相距甚远，途经BCD三国方能到达。邮递P领命后，不敢怠慢，将鸡毛信好生封装，拂晓时分启程。邮递P家门口的ADSL小路有些狭窄，但路上行人稀少，一眨眼的功夫，邮递P已穿越边境进入B国领地，好的开始是成功的一半，邮递P心中不免有些得意，甚至有点后悔出发太早了。

显然邮递P有些乐观过头了，此时的B国正处于兵荒马乱的年代，社会秩序混乱不堪，政府还在尽力而为的维持现状。人流如潮水般汹涌，邮递P夹杂其中，数次被人推倒在地，亏得邮递P身经百战，竟然杀出重围，狼狈不堪的逃离B国来到C国领地，邮递P惊魂未定，连忙查看随身信件，所幸完好无损。

此时邮递P心情一落千丈，如果不能及时抵达E国，恐自己小命难保。C国与A国素有往来，闻知邮递P途径此地，C国国王非常重视，要求为邮递P保驾护航。有关部门马上对邮递P经过路段实行管制，平民百姓不得进入主路，邮递P尾随警车呼啸而过，很快来到CD国交界处，邮递P悬着的心略微放松，不过把自己的快乐建立在他人痛苦之上，邮递P内心还是有些不安。

D国是一个高度发达的国家，实行的是区分式服务的治国方针，政府根据每个人的社会价值分配优先级，优先级越高，社会福利越好。邮递P作为外国使节，得到了EF级的礼遇，尽管马路上车水马龙，但邮递P所到之处，大家都自觉为其让路，丝毫不影响邮递P行进速度，邮递P很快便抵达目的地E国，回想D国百姓安居乐业的情形，邮递P心中也无比快乐。

见到E国国王，邮递P毕恭毕敬的将鸡毛信呈交上去，国王十分高兴，盛情挽留邮递P，希望其小住几日再走，不过邮递P以还有要务在身为由，拒绝了国王的好意，行色匆匆的踏上了归途。

邮递P一路马不停蹄，终于赶在日落之前回到故土，邮递P任务完成出色，国王龙颜大悦，大加犒赏，不在话下。不过邮递P没有居功自傲，连夜奋笔疾书，将沿途见闻记录成册，美其名曰“《网络大爬虫》之QoS专刊”，早朝之时献给国王，国王翻阅后，惊为天人之作，遂提拔邮递P为宰相，辅佐朝政，并钦定《网络大爬虫》为本国治国宝典。

许青邦

CONTENTS 目录

2011年第1期总第4期　QoS专题

序　言

基础知识

深入探讨

扩展应用

QoS缩略语列表

缩略语	英文解释	中文解释
ACL	Access Control List	访问控制列表
AF	Assured Forwarding	确保转发
CAR	Committed Access Rate	承诺访问速率
CBWFQ	Class Based Weighted Fair Queuing	基于类的加权公平队列
CIR	Committed Information Rate	承诺信息速率
CoS	Class of Service	服务等级
CQ	Custom Queuing	定制队列
DiffServ	Differentiated Serve	区分服务
DSCP	Differentiated Services Codepoint	区分服务码点
DS-TE	DiffServ-Aware Traffic Engineering	DiffServ感知的流量工程
DTS	Distributed Traffic Shaping	分布式流量整型
EF	Expedited Forwarding	加速转发
FIFO	First in First out	先入先出
FRTS	Frame Relay Traffic Shaping	帧中继流量整型
GTS	Generic Traffic Shaping	通用流量整形
IntServ	Integrated Service	综合服务

缩略语	英文解释	中文解释
IPHC	IP Header Compression	IP头压缩
LFI	Link Fragmentation & Interleaving	链路分片与交叉
LLQ	Low Latency Queueing	低时延队列
MPLS	Multiprotocol Label Switching	多协议标签交换
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、丢包率等性能
RED	Random Early Detection	随机早期丢弃
RSVP	Resource Reservation Protocol	资源预留协议
RTP	Real Time Protocol	实时协议
SQC	Structual QoS CLI	结构化命令行
TE	Traffic Engineering	流量工程
ToS	Type of Service	服务类型
VoIP	Voice over IP	通过IP网络承载语音服务
VPN	Virtual Private Network	虚拟专用网络
WFQ	Weighted Fair Queuing	加权公平队列
WRED	Weighted Random Early Detection	加权随机早期丢弃

QoS发展史

文/余卉

随着IP技术和网络的发展，IP网络已经从当初的单一数据网络向集成数据、语音、视频、图像的多业务网络转变。为了实现端到端QoS，IP QoS目前的研究主要集中在以下方面：

- 为多业务网络定义合理可行的QoS业务分类标准
- 为端到端QoS建立可实施的整网IP QoS模型

当前国际上各个研究组织都在为自己所关注的业务设计IP QoS模型。本文将逐一分析IETF（互联网工程工作组）、ITU-T（国际电信同盟）、ETSI（欧洲标准化组织）、MSF（多业务交换论坛）、TIPHON/TISPAN（传输平台功能体）等提出的QoS业务分类标准以及几种QoS应用模型发展的概况。

QoS业务分类标准

业务优先级分类的基本模型是区分不同类型的业务，在数据包头的特定域携带该优先级，然后网络节点根据包头携带的优先级实施不同的转发处理。目前，优先级分类根据各种网络所关注的业务类型已经出现多种不同的标准，相关标准可以参考：

- RFC791，Internet Protocol（根据各IP应用的特点，将业务分为Network Control、Internetwork Control、CRITIC/ECP、Flash Override、Flash、Immediate、Priority、Routine共8类优先级。其中，Routine优先级最低，Network Control优先级最高）；
- RFC1349，Type of Service in the Internet Protocol Suite（将业务按照ToS的定义分为16类优先级，ToS使用4个bit位分别表示：minimize delay、maximize throughput、minimize monetary cost、maximize reliability，并建议了各IP应用应该如何取ToS值，例如，FTP CONTROL报文建议其ToS取值为minimize delay）；
- RFC1490（被RFC2427替代），Multiprotocol Interconnect over Frame Relay（将业务按照Frame Relay Discard Eligibility bit的定义分为2类丢弃优先级）；
- RFC1483（被RFC2684替代），Multiprotocol Encapsulation over ATM Adaptation Layer 5（将业务按照ATM Cell Loss Priority bit的定义分为2类丢弃优先级）；
- RFC2474, Definition of the Differentiated Services Field（DS Field） in the IPv4 and IPv6 Header（DiffServ网络定义了

四类PHB（per-hop behavior）：EF（Expedited Forwarding）PHB适用于低时延、低丢失、低抖动、确保带宽的优先业务；AF（Assured Forwarding）PHB分为四类，每个AF类又分为三个丢弃优先级，可以对相应业务进行等级细分，QoS性能参数低于EF类型；CS（class selector）PHB是从IP ToS字段演变而来，共8类；BE PHB是CS中特殊一类，没有任何保证，现有IP网络流量也都默认为此类）；

- IEEE802.5，Token ring access method and Physical Layer specifications（令牌环网的优先级，可以将业务根据Access Priority的定义为8类优先级）；
- IEEE802.1p，Class of Service（以太网优先级，可以将业务根据802.1P Priority的定义分为8类优先级，0类至7类优先级相应递增，0类是BE业务，尽力传输）。

除了IETF，其它从事IP网络QoS标准研究的主要组织，例如ITU-T、ETSI等也都根据其业务定义的QoS业务分类标准。

ITU-T 13组建议Y.1541，主要根据IPTD（传输时延）、IPDV（时延变化）、IPLR（丢包率）、IPER（错误率）四个方面将业务划分为5类，0类至5类优先级相应递减，第5类是BE业务，对性能无保证。其中0类和2类对时延要求很严格，并且0类对抖动还有限制；1类和3类的时延要求比较严格，1类对抖动有限制；4类对时延要求比较宽松，且没有定义抖动限制；除了第5类外都对丢包率和错误率有要求。相关标准可以参考：

- ITU-T Recommendation Y.1541，Network Performance Objectives for IP-Based Services

ITU-T H.323 Annex N定义的业务类别分为两大类：GSC和CSC。前者对时延和抖动敏感，后者则无要求。其中GSC又分为GSC1、2、3、4。GSC1和2适用于CBR类型的流量，区别在于1对错误率有要求，而2没有；GSC3和GSC4适用于VBR类型的流量，区别在于3对于错误率有要求，而4没有。CSC也分为CSC1、2、3、4。CSC1和CSC2适用于nrt-VBR类型的流量，区别在于1对错误率有要求，而2没有；CSC3和CSC4适用于ABR类型的流量，区别在于3对错误率有要求，而4没有。相关标准可以参考：

- ITU-T Recommendation ANNEX N of H.323, End to End Quality of Service（QoS）and Service Priority Control and Sig-nalling in H.323 systems

ETSI 3GPP主要针对移动网络，它将业务类别分为conversational、streaming、interactive、background四大类，分类的主要依据是业务对时延的敏感度。Conversational类对时延非常敏感，依次递减，background对时延最不敏感。Conversational和streaming主要用于实时流量业务，区别只在于对时延的容许程度。Interactive和background主要用于传统的IP应用，两者都定义了一定的误码率要求，区别在于前者更多用于交互式场合，而后者主要用于后台业务。相关标准可以参考：

- 3GPP TS 23.107，QoS Concept and Architecture

TIPHON基于VoIP，将业务分为3大类，wideband、narrowband、BE，分类的依据是端到端时延。三类业务的时延限值依次递增，对应于用户感知的语音质量的满意度则是依次递减。其中narrowband又根据时延大小细分为三类：high、medium、acceptable，对应于narrowband中有等级区别的应用。相关标准可以参考：

- ETSI TS102 024-2，Telecommunications and Internet Protocol Harmonization Over Networks（TIPHON）Release 4；End-to-end Quality of Service in TIPHON Systems；Part 2：Definition of Speech Quality of Service（QoS）Classes

IETF的Inter-Serv模型和Diff-Serv模型

Inter-Serv模型

1994年，IETF出版RFC1633提出Inter-Serv模型，该模型使用资源预留（RSVP）协议，RSVP运行在从源端到目的端的每个路由器上，可以监视每个流，以防止其消耗比其请求、预留和预先购买的要多的资源。这种体系能够明确区分并保证每一个业务流的服务质量，为网络提供最细粒度化的服务质量区分。相关标准可以参考：

- RFC1633，Integrated Services in the Internet Architecture: An Overview
- RFC2205，Resource Reservation Protocol
- RFC2206，RSVP Management Information Base using SMIv2
- RFC2210，RSVP with IETF Integrated Services
- RFC2211，Controlled-Load Network Element Service
- RFC2212，Specification of Guaranteed Quality of Service

- RFC2215，General Characterization Parameters for Integrated Service Network Elements
- RFC2748，The COPS（Common Open Policy Service）Protocol
- RFC2749，COPS Usage for RSVP

Inter-Serv模型能够在IP网上提供端到端的QoS保证。但是，Inter-Serv模型对路由器的要求很高，当网络中的数据流数量很大时，路由器的存储和处理能力会遇到很大的压力。因此，Inter-Serv模型可扩展性很差，难以在Internet核心网络实施，目前业界普遍认为Inter-Serv模型有可能会应用在网络的边缘上。

Diff-Serv模型

区分服务（DiffServ）是IETF工作组为了克服Inter-Serv的可扩展性差在1998年提出的另一个服务模型，目的是制定一个可扩展性相对较强的方法来保证IP的服务质量。相关标准可以参考：

- RFC2474，Definition of the Differentiated Services Field（DS Field）in the IPv4 and IPv6 Headers
- RFC2475，An Architecture for Differentiated Services Framework
- RFC2597，Assured Forwarding PHB
- RFC2598，An Expedited Forwarding PHB
- RFC2983，Differentiated Services and Tunnels
- RFC3086，Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification
- RFC3140，Per Hop Behavior Identification Codes
- RFC3246，An Expedited Forwarding PHB
- RFC3247，Supplemental Information for the New Definition of the EF PHB
- RFC3248，A Delay Bound alternative revision of RFC2598

一同发展的还有许多与DiffServ有关的因特网草案（Internet-Drafts）可以参考：

- An Informal Management Model for Diffserv Routers
- Management Information Base for the Differentiated Services Architecture
- New Terminology and Clarification for Diffserv
- Differentiated Services Quality of Service Policy Information Base
- An Assured Rate Per-Domain Behaviour for Differentiated Services

在Diff-Serv模型中，业务流被划分成不同的差分服务类。一个业务流的差分服务类由其IP包头中的差分服务标记字段（Different Service Code Point，简称DSCP）来表示。在实施DiffServ的网络中，每一个路由器都会根据数据包的DSCP字段执行相应的PHB行为。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，但是由于标准还不够详尽，不同运营商的DiffServ网络之间的互通还存在困难。IETF RSVP和DiffServ两个工作组都正在研究RSVP与DiffServ相结合的问题，以进一步扩大DiffServ与现有系统的可兼容性，此外在业务分类、业务性能的量化描述以及域间业务类型映射等问题上，DiffServ模型也需进一步明确和开展的研究。

QoS路由

现在的Internet路由协议（OSPF、RIP等）都采用单个测度（如跳数、成本）来计算最短路由，没有考虑多个QoS参数的要求。QoS路由根据多种不同的度量参数（如带宽、成本、每一跳开销、时延、可靠性等）来选择路由。QoS路由包括三个主要功能：链路状态信息发布，路由计算和路由表存储。QoS路由能够满足业务的QoS要求，同时提高网络的资源利用率。但是QoS路由的计算十分复杂，增加了网络的开销，目前实用的QoS路由算法还不多见。相关标准可以参考：

- RFC2386，A Framework for QoS-based Routing in the Internet
- RFC2676，QoS Routing Mechanisms and OSPF Extensions

此外，还有一些草案可以参考，例如：

- ETF-draft-boucadair-QoS-bgp-spec-01，QoS-Enhanced Border Gateway Protocol

MPLS/TE

1997年，以Cisco公司为主的几家公司（包括Ipsilon、IBM、Cascade、Toshiba）提出了MPLS（Multiprotocol Label Switch）技术。MPLS技术产生的初衷就是为了综合利用网络核心的交换技术和网络边缘的IP路由技术各自的优点。对于骨干网业务提供者来说，MPLS已成为实现TE（Traffic Engineering）的重要手段，并且与DiffServ结合成为提供QoS的重要手段。流量工程的主要目的是合理分布数据流通过网络的路径，以避免不均匀地使用网络而导致拥

塞。一般的动态路由协议都会导致不均匀的通信分布，因为它们总是选择最短路径转发包，直接导致两个网络节点之间的最短路径上可能发生了拥塞，而较长路径上的路由器和链路却是空闲的。当网络中存在多条并行或可选的路径时，流量工程就显得非常重要了。面对复杂网络时，由于流量分布的变化是实时的，而且变化较大，因此流量工程需要能够自动化和适应业务流量分布的变化，这样才能保证业务的服务质量。可以参考的标准与草案比较多，此处不再敷述。

子网带宽管理SBM

由于数据包的发送能要经过中间某个网络的子网，为了实现端到端QoS，在子网内也要保证高优先级的数据帧获得高级别的服务。某些链路层的技术已经可以支持QoS了，例如异步传输模式ATM。而其它更多的LAN技术（如以太网技术）最初并非为支持QoS设计的。为此，IETF的ISSLL小组定义了上层QoS协议和服务与以太网之类的数据链路层技术之间的映射关系，并且提出了子网带宽管理（Subnet Bandwidth Management，SBM）的方案，它适用于802.1LAN，如以太网、令牌环和FDDI等，相关标准可以参考：

- RFC2814，SBM（Subnet Bandwidth Manager）：A Protocol for RSVP-based Admission Control over IEEE802-style networks

SBM的主要构件有以下三个部分，在其体系结构中提供了RM-to-BA以及BA-to-BA的信令机制来请求资源、改变或删除分配资源。

- 请求模块（Request Module，RM）：请求模块驻留在每个端系统中而不驻留在任何交换机中。请求模块根据管理员所定义的策略，将高层的QoS协议参数映射到第二层的优先级别；
- 带宽分配器（Bandwidth Allocator，BA）：带宽分配器保存子网内资源的分配状态，并且根据可用资源的情况以及管理员所定义的策略来执行接入控制；
- 通信协议（Communication Protocols，CP）：通信协议用于在SBM系统中，各个不同的组件之间进行通信。

上述这些QoS基本模型可以互为补充，在不同的网络层次上组合使用，例如IntServ和Diffserv结合，在核心网采用Diffserv，在接入网采用IntServ。又如MPLS和Diffserv结合，或MPLS和QoS路由结合。目前MPLS+DiffServ技术最有可能成为IP网络运营商首选的QoS方案。

但是这些成果没有解决全网的QoS问题，仍然缺乏一个可以整网实施的QoS机制与模型。与此同时，Internet2、MSF、ETSI TIPHON、ITU等组织目前也都在大力研究和制订可整网实施的IP QoS模型，以及无线领域的QoS模型，如IEEE 802.11e为无线局域网制订的QoS模型，美国有线电视工业的标准组织CableLab为Cable运营商研究制订的IP QoS模型，3GPP为下一代无线核心网络研究制订的QoS模型。

Internet2的BB（Bandwidth Broker）

Internet2研究的一个主要目标是创建一个可伸缩、可交互操作和可管理的QoS体系结构。以实现一些在现有的互联网上不能实现的应用，如远程医疗、数字化图书馆及虚拟实验室等。为此，在1997年，参加Internet2工程的大学和研究机构共同成立了QoS工作组，建立了QBone计划，以进行下一代互联网的QoS测试、开发和部署工作。Internet2开发的BB模型（Bandwidth Broker）是在DiffServ架构上运用带宽资源管理，其主要机制是在IP骨干网上使用DiffServ，每个DS域引入BB收集网络的拓扑和节点及链路状态信息，管理网络资源并结合策略服务器规定的策略进行接入控制。BB负责处理所有带宽申请请求，并根据当前网络的资源预留状况和配置的策略以及与用户签订的业务SLA，确定是否允许用户的带宽申请。DS域之间通过BB进行SLA协商，使DiffServ能够实现端到端的接入控制和QoS保障。相关的协议草案可以参考：

- QBone Architecture（V1.0），Ben Teitelbaum et al。Internet 2 QoS Working Group Draft，August 1999，Work-in-progress。

TIPHON/MSF的QoS标准化研究

欧洲标准化组织ETSI的TIPHON工作组，致力于发展电信与IP融合的下一代电信网络架构。ETSI TIPHON的QoS架构中，在IP网络的核心层引入了TRM（Transmission Resource Manager）来动态管理核心网的资源调度，实现实时的流量工程能力。而TRM接受接入层的业务资源申请，为业务分配和管理核心骨干网的资源和转发路径，并控制边缘或网关路由器，识别用户

的业务流，让用户的业务流按照TRM分配的路由和资源转发。

美国多业务交换论坛MSF研究NGN架构，提出类似的电信级IP网络。MSF的QoS架构与ETSI类似，在核心骨干层引入了Bandwidth Manager来动态管理核心网的资源调度，实现实时的流量工程能力。

ETSI/MSF等组织定义的模型目前遇到的问题是如何在目前的网络条件下实现承载业务按照所分配的路径去转发，还有各类接口的标准化过程以及运营中的网络管理体系建立等。由于这是一个全新的网络运营环境，还有很多细致的工作要做。

ITU-T的研究

ITU-T SG13组负责对NGN的总体研究，如果要在NGN上支持各种业务（实时业务/流媒体业务/非实时业务/多媒体业务……），那么NGN就必须为各种QoS级别的业务提供可预见的、一致的、端到端的QoS保证。ITU-T SG13组目前正在研究制订的Y.QoSar和Y.123.QoS两篇标准草案。草案Y.QoSar明确定义了基本QoS构建模块（接入控制、拥塞反馈、计量和测量、策略及策略配置、队列和调度、资源预留、服务等级管理, 费率表征和流量标识等），通过不同的方式把这些块组织起来，就可以控制网络提供业务所要求的性能。同时也考虑了实现QoS对安全的影响及相应机制。从建议草案的完整性角度看，Y.QoSar仍处于编辑及完善阶段，在2003年7月21日至8月1日ITU-T SG13 2001-2004研究期的第五次会议上取得了较大进展。该次会议对建议草案Y.QoSar做了以下完善：

- 完善了QoS路由、资源预留、QoS信令方面的内容；
- 在案例方法中补充了集中资源管理下的预配置叠加MPLS网络（Pre-provisioned overlay MPLS networks with centralized resource management）。（注：这一案例方法是由中国电信和华为公司联合在本次会议提交的）中国电信和华为联合提交的两篇文稿成为本次会议中重点讨论的内容，这两篇文稿分别是：
- A Carrier-class QoS Solution Framework for IP-based Backbone Networks（该文稿提出了用于基于IP骨干网络的运营级QoS解决框架）；
- A Carrier-class QoS Solution Framework for IP-based Access Network（该文稿提出了用于基于以太网IP接入网络的运营级QoS解决框架）。

其中第二篇文稿的基本内容作为了新建议草案Y.123.QoS的基本参考模型。Y.123.QoS是纳入到Y.QoSar统一框架中的一个基于以太的IP接入网的QoS架构。它强调了呼叫控制和承载分离的思想，是在接入段保证用户服务质量的重要方法。

我国行业标准的研究状况

我国电信标准协会网络与交换标准技术委员会已经研究制订了《IP网络技术要求——网络性能参数与指标》标准（编号为YD/T 1171-2001）。该标准规定了支持IPv4的IP网络性能参数和临时指标，其中有些指标与用户所选择的服务质量（QoS）类型相关，还规定了满足推荐的、端到端国际IP网通信的、性能指标的每个网络段，应该提供的性能指标要求。适用于具有一个或多个网络段的端到端路径，所定义的QoS类型适用于终端用户与网络服务提供商之间以及网络服务提供商之间的IP网通信，可作为IP网网络规划、工程设计、运行维护以及相应设备的引进、开发的技术依据。该标准定义的QoS类型主要基于对下列应用的支持：点到点电话、多媒体会议和数据传输。

今后的研究方向

从这些标准组织的研究方向和思路来看，一方面是统一实施整网的IP QoS，但是目前的研究成果主要是一些比较笼统的框架性文件，在具体的实施技术规范上还没有显著的成果，这将是国际标准组织未来几年内的研究和制订的重点。国内外的运营商/研究机构和设备厂商也在积极研究制订IP网络的QoS机制和技术。另一方面， IP网与电信网络的运维、规划的统一融合也成为新的研究方向。国际电信联盟2008年世界电信标准化全会（WTSA-08）确定了2009年～2012年研究期的工作计划及重点。伴随着国内外运营商对于NGN网络的部署和各类多媒体业务的发展，对IP网络的QoS机制的研究和部署变得日益迫切起来，而电信级IP QoS是在IP网络上开展可运营、可管理、高质量的电信业务的关键技术，是实现电信和IP网络融合的基础。电信级IP QoS要求能够在IP网络上承载电信级多媒体业务，如IP电话、视讯会议、视频点播等，并为3G、软交换等提供承载层的服务质量保证。目前，各个标准化组织都在积极研究和制订电信级IP QoS的相关标准，各个厂家也纷纷提出相关的技术实现方案。

QoS的基本原理

文/胡国华

QoS（Quality of Service）是服务质量的简称。对于网络业务来说，服务质量包括哪些方面呢？从传统意义上来讲，无非就是传输的带宽、传送的时延、数据的丢包率等，而提高服务质量无非也就是保证传输的带宽，降低传送的时延，降低数据的丢包率以及时延抖动等。广义上讲，服务质量涉及网络应用的方方面面，只要是对网络应用有利的措施，其实都是在提高服务质量。因此，从这个意义上来说，防火墙、策略路由、快速转发等也都是提高网络业务服务质量的措施之一。

服务质量相对网络业务而言，在保证某类业务服务质量的同时，可能就是在损害其它业务的服务质量。因为网络资源总是有限的，只要存在抢夺网络资源的情况，就会出现服务质量的要求。比如，网络总带宽为100Mbps，而BT下载占用了90Mbps，其他业务就只能占用剩下的10Mbps。而如果限制BT下载占用的最大带宽为50Mbps，也就提高了其他业务的服务质量，使其他业务能够占用最少50Mbps的带宽，但这是在损害BT业务的服务质量为前提的。

QoS模型

网络中的通信都是由各种应用流组成的，这些应用对网络服务和性能的要求各不相同，比如FTP下载业务希望能获取尽量多的带宽，而VoIP语音业务则希望能保证尽量少的延迟和抖动等。但是所有这些应用的特殊要求又取决于网络所能提供的QoS能力，根据网络对应用的控制能力的不同，可以把网络的QoS能力分为三种模型。

Best Effort模型

Best Effort（尽力而为）模型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，网络尽最大的可能性来发送报文，对带宽、时延、抖动和可靠性等不提供任何保证。

Best Effort是Internet的缺省服务模型，通过FIFO（First In First Out，先进先出）队列来实现。

尽力而为的服务实质上并不属于QoS的范畴，因为在转发尽力而为的通信时，并没有提供任何服务或转发保证。

DiffServ模型

DiffServ（Differentiated Service，区分服务）模型由RFC2475定义，在区分服务中，根据服务要求对不同业务的数据进行分类，对报文按类进行优先级标记，然后有差别地提供服务。

区分服务一般用来为一些重要的应用提供端到端的QoS，它通过下列技术来实现：

- 流量标记与控制技术：它根据报文的CoS（Class of Service，服务等级）域、ToS域（对于IP报文是指IP优先级或者DSCP）、IP报文的五元组（协议、源地址、目的地址、源端口号、目的端口号）等信息进行报文分类，完成报文的标记和流量监管。目前实现流量监管技术多采用令牌桶机制；
- 拥塞管理与拥塞避免技术：WRED、PQ、CQ、WFQ、CBQ等队列技术对拥塞的报文进行缓存和调度，实现拥塞管理与拥塞避免。

上述这些技术的主要实现原理将在下文的QoS基本原理中进行重点介绍。

IntServ模型

IntServ（Integrated Service，综合服务）模型由RFC1633定义，在这种模型中，节点在发送报文前，需要向网络申请资源预留，确保网络能够满足数据流的特定服务要求。

IntServ可以提供保证服务和负载控制服务两种服务，保证服务提供保证的延迟和带宽来满足应用程序的要求；负载控制服务保证即使在网络过载的情况下，也能对报文提供与网络未过载时类似的服务。

在IntServ模型中，网络资源的申请是通过信令来完成的，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，应用程序一般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。同时应用程序发出的报文应该控制在流量参数描述的范围以内。负责完成保证服务的信令为RSVP（Resource Reservation Protocol，资源预留协议），它通知网络设备应用程序的QoS需求。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。

保证服务要求为单个流预先保留所有连接路径上的网络资源，而当前在Internet主干网络上有着成千上万条应用流，保证服务如果要为每一条流提供QoS服务就变得不可想象了。因此，IntServ模型很难独立应用于大规模的网络，目前主要与MPLS TE（Traffic Engineering，流量工程）结合使用。

QoS基本原理

流量分类与标记

流量分类，就是将流量划分为多个优先级或多个服务类，如使用以太网帧中802.1Q头保留的User Priority（用户优先级）字段标记服务级别，可以将以太网帧最多分成2^3=8类；使用IP报文头的ToS（Type of service，服务类型）字段的前三位（即IP优先级）来标记报文，可以将报文最多分成2^3=8类；使用DSCP（Differentiated Services Codepoint，区分服务编码点，ToS域的前6位），则最多可分成2^6=64类。在报文分类后，就可以将其它的QoS特性应用到

不同的分类，实现基于类的拥塞管理、流量整形等。

对于MPLS网络报文，则一般是根据MPLS报文中的EXP域进行处理。EXP域包括3位，虽然RFC3032把它叫做实验域，但它通常作为MPLS报文的CoS域，与IP网络的ToS或DSCP域等效。

对于流量的分类，上面提到的关于以太网帧的CoS域、IP报文的ToS域等与MPLS报文的EXP域等仅是分类的一种情况，其实几乎可以对报文的任何信息段进行分类，比如也可以根据源IP地址、目的IP地址、源端口号、目的端口号、协议ID等进行流量的分类。

虽然流量分类几乎可以根据报文的任何字段进行，但是流量分类标记则一般只对802.1Q以太网帧的CoS域、IP报文的ToS域、MPLS报文的EXP域进行标记。流量的标记主要的目的就是让其他处理此报文的应用系统或设备知道该报文的类别，并根据这种类别对报文进行一些事先约定了的处理。

例如，在网络的边界做如下分类和标记：

- 所有VoIP数据报文聚合为EF业务类，将报文的IP优先级标记为5，或者将DSCP值标记为EF；
- 所有VoIP控制报文聚合AF业务类，将报文的IP优先级标记为4，或者将DSCP值标记为AF31。

当报文在网络边界被标记分类之后，在网络的中间节点，就可以根据标记，对不同类别的流量给予差别服务了。例如：对上述例子中的EF类业务保证时延和减少抖动，同时进行流量监管；对AF业务类在网络拥塞时仍然保证一定的带宽，等等。

拥塞管理技术原理

拥塞管理基本概念

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路由器将不能发送一些报文，即网络发生了拥塞。

如下图1所示，当公司分支1向公司总部以100M的速度发送数据时，将会使Router 2的串口S0/1发生拥塞。

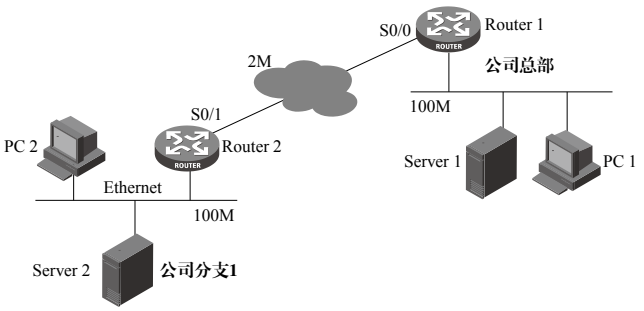


图1 实际应用中的拥塞实例

拥塞管理是指网络在发生拥塞时，如何进行管理和控制。处理的方法是使用队列技术。将所有要从一个接口发出的报文进入多个队列，按照各个队列的优先级进行处理。不同的队列算法用来解决不同的问题，并产生不同的效果。常用的队列技术有FIFO、PQ、CQ、WFQ、CBWFQ等，下文逐一介绍这些常用队列技术的基本原理。

FIFO队列原理简述

FIFO（First In First Out，先进先出）队列示意图如图2所示：

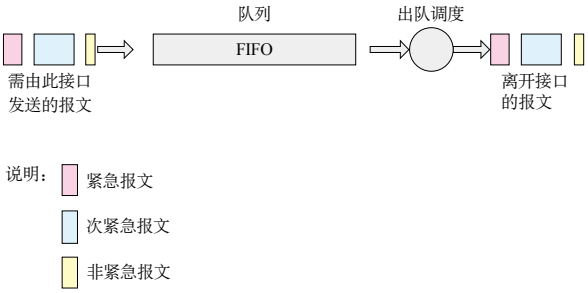


图2 FIFO队列示意图

FIFO队列不对报文进行分类，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。

FIFO队列具有处理简单，开销小的优点。但FIFO不区分报文类型，采用尽力而为的转发模式，使对时间敏感的实时应用（如VoIP）的延迟得不到保证，关键业务的带宽也不能得到保证。

PQ原理简述

PQ（Priority Queuing，优先队列）示意图如图3所示：

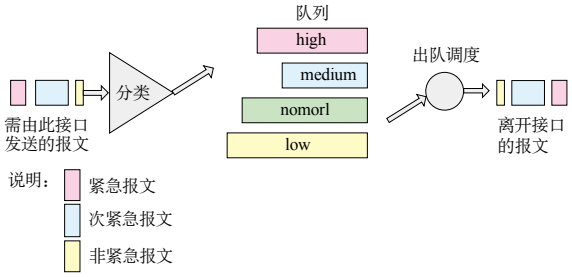


图3 PQ队列示意图

PQ队列是针对关键业务应用设计的。关键业务有一个重要特点，需要在拥塞发生时要求优先获得服务以减少响应的延迟。PQ可以根据网络协议（如IP、IPX）、数据流入接口、报文长短、IP报文的ToS、五元组（协议ID、源IP地址、目的IP地址、源端口号、目的端口号）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4类，分别属于PQ的4个队列中的一个，然后，按报文所属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占。这样会使得实时业务（如VoIP）的报文能够得到优先处理，非实时业务（如E-Mail）的报文在网络处理完关键业务后的空闲间隙得到处理，既保证了实时业务的优先，又充分利用了网络资源。

PQ的缺点是，当较高优先级队列中总有报文存在时，则低优先级队列中的报文将一直得不到服务，出现队列“饿死”现象。

CQ原理简述

CQ（Custom Queuing，定制队列）示意图如图4所示：

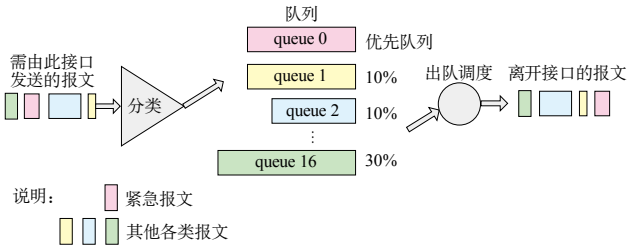


图4 CQ队列示意图

CQ的分类方法和PQ基本相同，不同的是它最终将所有报文分成最多至17类，每类报文对应CQ中的一个队列，接口拥塞时，报文按匹配规则被送入对应的队列；如果报文不匹配任何规则，则被送入缺省队列（缺省队列默认为1，可配置修改缺省队列）。

CQ的17个队列中，0号队列是优先队列，路由器总是先把0号队列中的报文发送完，然后才处理1到16号队列中的报文，所以0号队列一般作为系统队列，把实时性要求高的交互式协议报文放到0号队列。1到16号队列调度采用轮询方式，按照用户预先配置的额度依次从1到16号用户队列中取出一定数量的报文发送。如果轮询到某队列时该队列恰好为空，则立即转而轮询下一个队列。

CQ把报文分类，然后按类别将报文分配到CQ的一个队列中去，而对每个队列，又可以规定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。但由于采用轮询调度各个队列，CQ无法保证任何数据流的延迟。

WFQ原理简述

WFQ（Weighted Fair Queuing，加权公平队列）示意图如图5所示：

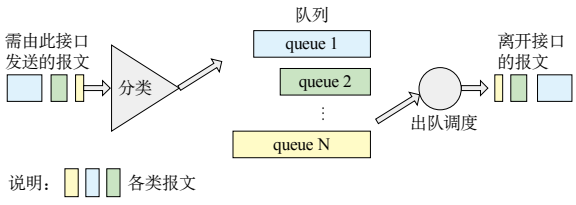


图5 WFQ队列示意图

WFQ对报文按流特征进行分类，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、ToS的报文属于同一个流，

而对于MPLS网络，具有相同的标签和EXP域值的报文属于同一个流。每一个流被分配到一个队列，该过程称为散列，采用HASH算法来自动完成，这种方式会尽量将不同特征的流分入不同的队列中。每个队列类别可以看作是一类流，其报文进入WFQ中的同一个队列。WFQ允许的队列数目是有限的，用户可以根据需要配置该值。

在出队的时候，WFQ按流的优先级（precedence）来分配每个流应占有出口的带宽。优先级的数值越小，所得的带宽越少。优先级的数值越大，所得的带宽越多。这样就保证了相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。

WFQ优点在于配置简单，有利于小包的转发，每条流都可以获得公平调度，同时照顾高优先级报文的利益。但由于流是自动分类，无法手工干预，故缺乏一定的灵活性，且受资源限制，当多个流进入同一个队列时无法提供精确服务，无法保证每个流获得的实际资源量。WFQ均衡各个流的延迟与抖动，同样也不适合延迟敏感的业务应用。

CBQ原理简述

CBQ（Class Based Queuing，基于类的队列）示意图如图6所示：

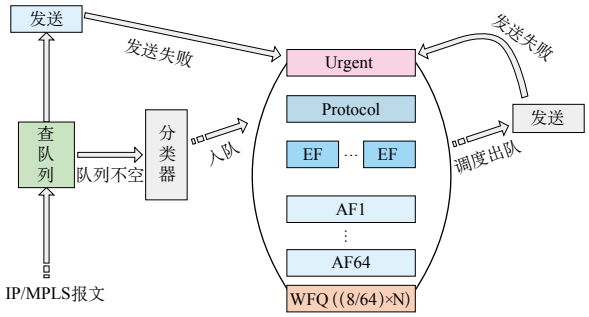


图6 CBQ队列示意图

CBQ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。

CBQ包括一个低时延队列LLQ（Low Latency Queuing，低时延队列），用来支撑EF（Expedited Forwarding，快速转发）类业务，被绝对优先发送，保证时延。进入EF的报文在接口没有发生拥塞的时候（此时所有队列中都没有报文），所有属于EF的报文都可以被发送。在接口发生拥塞的时候（队列中有报文时），进入EF的报文被限速，超出规定流量的报文将被丢弃。

另外有64个BQ队列（Bandwidth Queuing，带宽保证队列），用来支撑AF（Assured Forwarding，确保转发）类业务，可以保证每一个队列的带宽及可控的时延。系统调度报文出队列的时候，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队列没有报文时，BQ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。

最后还有一个WFQ队列，对应BE（Best Effort，尽力传送）业务，使用接口剩余带宽进行发送。

CBQ可根据报文的输入接口、满足ACL情况、IP Precedence、DSCP、EXP、Label等规则对报文进行分类、进入相应队列。对于进入EF和AF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与AF占用接口的总带宽不要超过接口带宽的75%。

CBQ可为不同的业务定义不同的调度策略（如带宽、时延等），由于涉及到复杂的流分类，对于高速接口（GE以上）启用CBQ特性系统资源存在一定的开销。

RTP原理简述

RTP优先队列（Real Time Protocol Priority Queuing）示意图如图7所示：

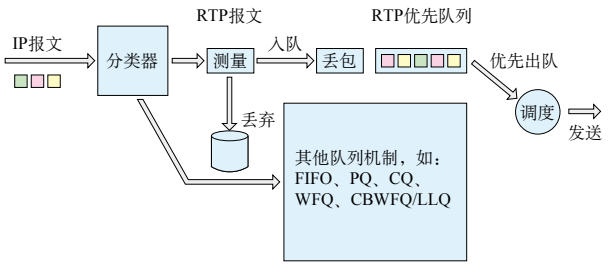


图7 RTP队列示意图

RTP优先队列是一种保证实时业务（包括语音与视频业务）服务质量的简单队列技术。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。

RTP优先队列将RTP报文送入一个具有较高优先级的队列，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任何一种队列（包括FIFO、PQ、CQ、WFQ与CBQ）结合使用，它的优先级是最高的。由于CBQ中的EF完全可以解决实时业务，所以不推荐将RTP优先队列与CBQ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PQ的高优先级队列的流量可能“饿死”低优先级流量的问题。

拥塞避免原理

受限于设备的内存资源，按照传统的处理方法，当队列的长度达到规定的最大长度时，所有到来的报文都被丢弃。对于TCP报文，如果大量的报文被丢弃，将造成TCP超时，从而引发TCP的慢启动和拥塞避免机制，使TCP减少报文的发送。当队列同时丢弃多个TCP连接的报文时，将造成多个TCP连接同时进入慢启动和拥塞避免，称之为：TCP全局同步。这样多个TCP连接发向队列的报文将同时减少，使得发向队列的报文的量不及线路发送的速度，减少了线路带宽的利用。并且，发向队列的报文的流量总是忽大忽小，使线路的上的流量总在极少和饱满之间波动。

为了避免这种情况的发生，队列可以采用加权随机早期检测WRED（Weighted Random Early Detection）的报文丢弃策略（WRED与RED的区别在于前者引入IP优先权，DSCP值，和MPLS EXP来区别丢弃策略）。采用WRED时，用户可以设定队列的阈值（threshold）。当队列的长度小于低阈值时，不丢弃报文；当队列的长度在低阈值和高阈值之间时，WRED开始随机丢弃报文（队列的长度越长，丢弃的概率越高）；当队列的长度大于高阈值时，丢弃所有的报文。

WRED和队列机制的关系如图8所示：

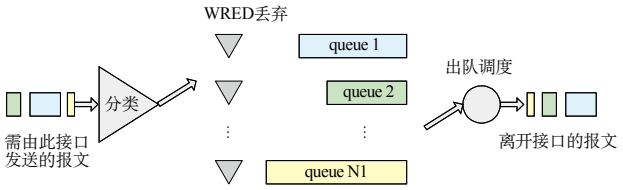


图8 WRED处理方式示意图

流量监管原理

流量监管（Commit Access Rate，简称CAR）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。

CAR利用令牌桶（Token Bucket，简称TB）进行流量控制。如图9所示为利用CAR进行流量控制的基本处理过程：

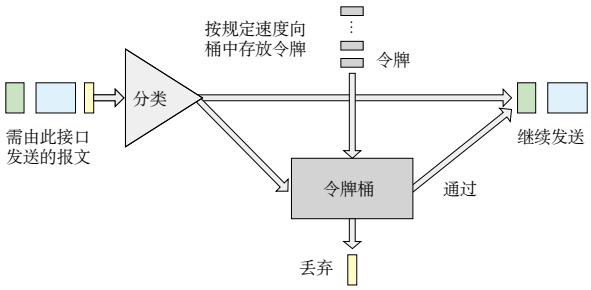


图9 CAR处理方式示意图

首先，根据预先设置的匹配规则来对报文进行分类，如果是没有规定流量特性的报文，就直接继续发送，并不需要经过令牌桶的处理；如果是需要进行流量控制的报文，则会进入令牌桶中进行处理。如果令牌桶中有足够的令牌可以用来发送报文，则允许报文通过，报文可以被继续发送下去。如果令牌桶中的令牌不满足报文的发送条件，则报文被丢弃。这样，就可以对某类报文的流量进行控制。

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（mark）或重新标记（re-mark）。具体来讲就是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。

流量整型原理

通用流量整形（Generic Traffic Shaping，简称GTS）可以对不规则或不符合预定流量特性的流量进行整形，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在接口的出、入方向进行报文的流量控制，对不符合流量特性的报文进行丢弃；而GTS只在接口的出方向对于不符合流量特性的报文进行缓冲，减少了报文的丢弃，同时满足报文的流量特性，但增加了报文的延迟。

GTS的基本处理过程如图10所示，其中用于缓存报文的队列称为GTS队列。

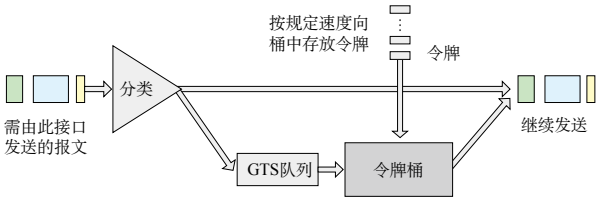


图10 GTS处理过程示意图

物理接口总速率限制原理

利用物理接口总速率限制（Line Rate，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）的总速率。

LR的处理过程仍然采用令牌桶进行流量控制。如果用户在路由器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。

LR的基本处理过程如图11所示：

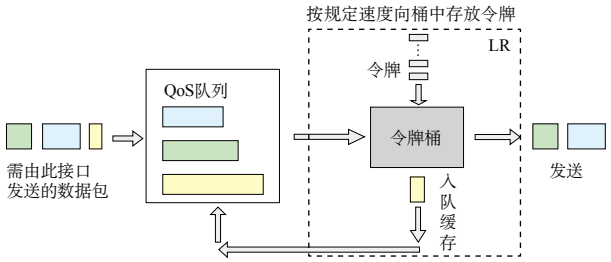


图11 LR处理过程示意图

其他提高QoS的技术

链路效率机制

链路效率机制，用于改善链路的性能，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。链路效率机制有很多种，下面介绍两种比较典型的链路效率机制及其基本原理。

链路分片与交叉（Link Fragment & Interleave，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实时队列和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。

链路效率机制的工作原理图如图12所示：

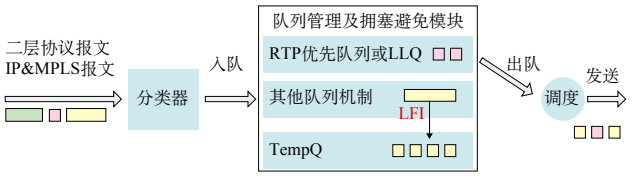


图12 LFI处理过程示意图

如图12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

到大片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，这样就很大程度的降低了低速链路因为发送大片报文造成的时延。

RTP报文头压缩（RTP Header Compression，CRTP）

CRTP用于RTP（Real-time Transport Protocol）协议，对IP头、UDP头和RTP头进行压缩，通常在低速链路上使用。可将40字节的IP/UDP/RTP头压缩到2～4个字节（不使用UDP校验和可到2字节），提高链路带宽的利用率。CRTP主要得益于同一会话的语音分组头和语音分组头之间的差别往往是不变的，因此只需传递增量。

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报文包括头部分和数据部分，RTP的头部分包括：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；RTP数据部分典型载荷是20字节到160字节。为了避免不必要的带宽消耗，可以使用CRTP特性对报文头进行压缩。CRTP可以将IP/UDP/RTP头从40字节压缩到2～4字节，对于40字节的载荷，头压缩到4字节，压缩比为（40+40）/（40+4），约为1.82，可见效果是相当可观的，可以有效的减少链路带宽的消耗，尤其是低速链路。

RTP报文头压缩的处理过程如下图所示：

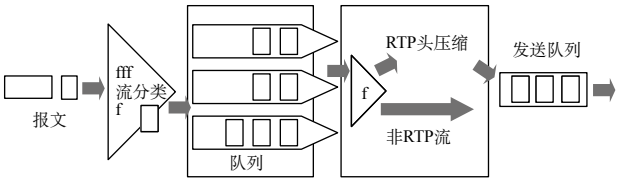


图13 CRTP处理过程示意图

链路层QoS技术

链路层QoS技术主要针对ATM（Asynchronous Transfer Mode，异步传输模式）、帧中继等链路层协议支持QoS。ATM作为一种面向连接的技术，提供对QoS最强有力的支持，而且可以基于每个连接提供特定的QoS保证；帧中继网络确保连接的CIR（Committed Information Rate，承诺信息速率）最小，即在网络拥塞时，传输速度不能小于这个值。

ATM QoS

ATM是异步传输模式（Asynchronous Transfer Mode）的简称，以信元为基本单位进行信息传输、复接和交换。ATM信元具有53字节的固定长度，其中5个字节构成信元头部，主要用来标识虚连接，另外也完成了一些功能有限的流量控制，拥塞控制，差错控制等功能，其余48个字节是有效载荷。ATM是面向连接的交换，其连接是逻辑连接，即虚电路。每条虚电路（Virtual Circuit，VC）用虚路径标识符（Virtual Path Identifier，VPI）和虚通道标识符（Virtual Channel Identifier，VCI）来标识。一个VPI/VCI值对只具有本地意义，不具有全局有效性。它在ATM节点上被翻译。当一个连接被释放时，与此相关的VPI/VCI值对也被释放，它被放回资源表，供其它连接使用。

ATM中每一条VC都有一定的QoS保障，这是由ATM的连接管理来实现的。当用户与网络或网络与网络建立一个连接的时候，双方就确定了一份通信契约，契约中包括流量参数和QoS参数两部分。此通信契约为双方所共识，双方必须遵守。流量参数包括峰值信元速率（PCR，Peak Cell Rate）、持续信元速率（SCR，Sustained Cell Rate）、最小信元速率（MCR，Minimum Cell Rate）以及最大突发量（MBS，Maximum Burst Size），它们描述业务本身的流量特性，又称为源流量参数。QoS参数主要包括最大信元传递时延（MCTD，MeanCell Transfer Delay）、信元抖动容限（CDVT，CellDelayVariationTolerance）和信元丢失率（CLR，Cell Loss Ratio），MCTD是信元从一个端点到另一个端点所需要的时间，CDVT是信元间隔的上限，CLR是可以接受的因网络拥塞而导致信元丢失比例。

ATM端系统负责确保传输的流量符合QoS合同。ATM端系统通过缓冲数据来对流量进行整形，并按约定的QoS参数传输通信。ATM交换机控制每个用户的通信指标，并将其与QoS合同进行比较。对于超过了QoS合同的通信，ATM节点可以设置信元的CLP（Cell Loss Priority，信元丢弃优先级）位。在网络拥塞时，CLP置位的信元被丢弃的可能性更大。

ATM网络拥塞管理的基本思想在于：引入预防性控制措施，不再是出现拥塞之后再采取措施来消除拥塞，而是通过精心管理网络资源来避免拥塞的出现。

FR QoS

FR（Frame Relay，帧中继）是一种统计复用的协议，它能够在单一物理传输线路上提供多条虚电路。每条虚电路用DLCI（Data Link Connection Identifier，数据链路连接标识）来标识。每条虚电路通过LMI（Local Management Interface，本地管理接口）协议检测和维护虚电路的状态。

帧中继采用VC（Virtual Circuit）虚电路技术，即帧中继传送数据使用的传输链路是逻辑连接，而不是物理连接。虚电路是面向连接的，可以保证用户帧按顺序传送至目的地。根据虚电路建立方式的不同，将帧中继虚电路分为两种类型：永久虚电路（PVC，Permanent Virtual Circuit）和交换虚电路（SVC，Switched Virtual Circuit）。PVC是手工设置产生的虚电路，而SVC是通过协议协商自动创建和删除的虚电路。

帧中继报头中的3个位提供了帧中继网络中的拥塞控制机制，这3个位分别叫做向前显式拥塞通知（FECN，Forward Explicit Congestion Notification）位、向后显式拥塞通知（BECN，Backward Explicit Congestion Notification）位和丢弃合格（DE，Discard Eligible）位。可以通过帧中继交换机将FECN位置1来告知诸如路由器等目标数据终端设备（DTE，Data Terminal Equipment），在帧从源传送到目的地的方向发生了拥塞。帧中继交换机将BECN位置1则告知目标路由器，在帧从源传送到目的地的反方向上发生了拥塞。DE位由路由器或其他DTE设备设置，指出被标记的帧没有传输的其他帧那么重要，它在帧中继网络中提供了一种基本的优先级机制，如果发生拥塞时，DE位置位的帧会被优先丢弃。

帧中继流量整形（FRTS，Frame Relay Traffic Shaping）对从帧中继VC输出的通信进行整形，使之与配置速率一致，它将超出平均速率的分组放到缓冲区来使突发通信变得平滑。根据配置的排队机制，当有足够的可用资源时，这些缓冲的分组出队并等候被传输。排队算法是基于单个VC配置的，它只能针对接口的出站通信进行设置。FRTS可对每个VC的流量进行整形，将其峰值速率整形为承诺信息速率（CIR，Committed Information Rate）或其他定义的值，如超额信息速率（EIR，Excess Information Rate）。自适应模式的FRTS还能够根据收到的网络BECN拥塞指示符降低帧中继VC的输出量，将PVC的输出流量整形为与网络的可用带宽一致。

QoS技术展望

文/余卉

伴随网络技术、多媒体技术的飞速发展，IP网在承载现有WWW、FTP、E-mail等服务的基础上，开始越来越多地承载VoIP以及交互式多媒体通信业务，而每种业务要求的传输时延和抖动等服务参数不尽相同。因此，为用户提供服务质量（QoS）成为Internet发展的重要挑战。网络QoS技术作为公认的新一代网络的核心技术之一，是当前网络研究和开发的热点。IP QoS的研究涉及许多内容，IETF已经提出了许多服务模型与机制来满足QoS需求，典型的有：集成服务（IntServ）/RSVP模型、区分服务（DiffServ）模型、多协议标签交换（MPLS）与流量工程（Traffic Engineering）等，而ITU-T则更偏重于QoS整体框架的制定和IP性能指标体系的构建。此外，MSF（城域网交换论坛）、Internet II和IPCablecom也提供了多种的多种端到端IP QoS解决方案。

上述这些技术和解决方案的基本情况已经在《QoS发展史》一文中有了初步的介绍，本文将重点介绍QoS框架发展的几大关键趋势，作为对QoS标准发展概况的进一步深入介绍。

技术展望

通过QoS支持多业务IP网络

实现统一的支持QoS能力的电信级多业务IP网络是发展的趋势。

电信运营商一直致力于采用基于Internet理念的IP网技术来作为下一代网络技术的统一平台，但是随着时间的推移，人们设想中的基于IP和MPLS的多业务网并未真正得以实施。QoS、流量管理、组播和安全性问题基本上没有解决，这些问题日益影响运营商在此单一而庞大的IP网络上能够提供越来越多的新型增值服务。对电信运营商而言，现在面临的挑战是如何以一种非常有效而现实的方式来为不同的业务提供满意的端到端QoS保证，同时还要充分考虑到整个网络的性能，并且要充分考虑承载级QoS架构所需的可扩展性、可靠性、可操作性和可管理性。如何在统一的IP QoS架构中提供一组通用的网络运行机制来控制网络对某一业务需求进行正确的响应成为研究的关键问题。目前研究较多的IP QoS电信网络架构，一般将QoS关键构件归属于三个平面，即控制平面、数据平面和管理平面。IP QoS电信网络通用架构如图所示。

电信网络架构

其中，控制平面包含了与业务流路径相关的QoS关键构件，主要包括了允许控制、QoS路由及资源预留；数据平面包含了与用户数据流处理相关的QoS关键构件，这些机制包括了缓存器管理、拥塞避免、包标记队列和调度、业务流分类、业务流策略和流量整形；管理平面包含了与运营、管理相关的QoS构件，主要包括了服务等级合同（SLA）、业务流恢复，计量和记录。

如何有效地组织这些QoS基本构件模块，并充分利用IETF和ITU-T地现有研究成果成为关键问题。在ITU-T建议草案Y.QoSar明确定义了基本QoS构建模块（接入控制、拥塞反馈、计量和测量、策略及策略配置、队列和调度、资源预留、服务等级管理，费率表征和流量标识等），通过不同的方式把这些块组织起来，就可以控制网络提供业务所要求的性能。同时也考虑了实现QoS对安全的影响及相应机制。

IPv6 QoS

IPv6在报头中保留了类似IPv4的ToS域，称为传输级别域，以继续为IP提供区分QoS服务。同时IPv6报头中增加20比特流标签（Flow Label）域。流标签更好支持综合QoS服务，可以直接标识流，并配合RSVP实现资源预留。IPv4的流分类器是根据源地址、目的地址、源端口号、目的端口号和传输协议类型的五元组确定。由于分组的拆分或加密，有些域往往难以获得，对高层报头的访问，也可能会阻碍新协议的引入。IPv6中一个流可以由源IPv6地址和非空的流标签唯一地标识。源可以通过逐跳扩展头或控制协议RSVP等向转发路径的中间节点建立流状态。IPv6节点接收到一个有标记的IPv6分组时，可以用流标记、源地址将分组分类到某个流。根据在一系列IPv6节点上建立的流状态，可以对分组提供一些流特殊处理。IPv6 QoS具体实施还在草案讨论制定中，还有一些具体应用问题需要考虑。

基于QoS的路由

整网或局部网络的QoS控制通常通过对路由与信令的控制达到对业务流传输的直接控制，因此路由直接关系到网络性能，所

以QoS路由成为解决QoS问题的一项关键技术。QoS路由研究中需要遇到的问题包括以下几个方面：

■ 实时应用往往会对延时，延时抖动，带宽，丢失率，业务代价等多个参数同时提出性能要求，例如，实时多媒体业务会对延时和延时抖动同时提出要求，这些参数相互独立时，选择满足多个参数限制的路由直接关系到路由算法的可实现性；

■ 同时承载多种QoS要求不同的业务时，网络性能优化困难，扩展困难，尤其是QoS和尽力而为（Best Effort）的业务独立共存时，很难确定最优的操作点；

■ 每个路由节点状态信息的存储量大。QoS路由中，每个路由节点需记录的状态参量将增多，如状态信息的存储量随网络节点个数的增加而指数性增加，将限制网络的扩展；

■ 传输负载的抖动等动态信息都可能导致网络状态变化，这些变化因素直接影响全网状态信息的准确性，同时也直接影响算法的性能。

除了上述主要研究之外，其他QoS标准化的研究已经在《QoS发展史》中介绍过。此外，QoS的研究领域还广泛涉及到以下问题：

■ 如何为应用层协议定制QoS服务。在QoS体系结构中一个最基本的问题：QoS是否应该基于每个应用层服务为传输提供服务保证；是否应该将QoS作为应用层协议的一个传输选项；每一个应用层服务是否应该能通过不同的方式扩展QoS体系结构，以便于应用层服务能够根据网络的QoS响应调整应用层的服务行为；

■ 如何综合协调TCP拥塞控制与QoS服务级别。由于TCP协议自身的拥塞控制机制可以使用反向确认的ACK报文即时调整数据传输的速率，因此最终得到的业务服务质量就成为正向传输数据和反向传输ACK报文的综合结果。如果反向的ACK流在网络中处于一个不同的服务级别，那么极有可能导致后续数据流量的高速突发；

■ 在IntServ和DiffServ体系结构中使用基于流的QoS识别，还是使用基于报文的QoS识别；

■ 如何测量QoS服务参数，在一条特定的转发路径上如何测量其带宽等可用资源，以便更好的进行网络接入控制；

■ 如何对区分QoS服务实施不同的计费，至今仍然没有一个详细定义的计费模型可以用于获取资源使用状况相关的数据。

试题说明

本试题用于对IP QoS知识体系掌握情况的考查或自测。题目范围以三层QoS为主，涵盖了部分二层QoS的基本概念。出于对试题篇幅和难度的考虑，本测试题未涉及MPLS QoS、帧中继QoS以及链路效率机制等方面的内容。

试题整体难度为中级，适用于对IP QoS体系已经有所掌握的人。

IP QoS测试题

一、选择题

1、数据网络的服务质量通常因传送的业务不同而有不同的要求，与数据业务相比，语音和Video业务更注重于：（　）

A. 时延　　B. 时延抖动　　C. 吞吐量　　D. 可靠性

2、当前IP QoS的主要技术方案都包括：（　）

A. Int-Serv　　B. 带宽代理（BB）　　C. MPLS-TE&QoS　　D. Diff-Serv

3、局域网上的QoS主要依靠在以太网帧头上加入优先级字段来实现，这定义在以下哪个标准中：（　）

A. RFC 2474　　B. RFC 2475　　C. IEEE802.1p/q　　D. IEEE802.1X

4、Diff-Serv模型中定义了哪几种不同的服务类型：（　）

A. 确保服务（AF）　　B. 奖赏服务（EF）　　C. 尽力而为服务（BE）　　D. 先进先出服务（FIFO）

5、用令牌桶进行流量的复杂评估时（采用双桶算法），第二个桶的大小应等于：（　）

A. CBS　　B. CIR　　C. EBS　　D. CBS＋EBS

6、与流量监管相比，流量整形会引入额外的：（　）

A. 丢包　　B. 时延　　C. 负载　　D. 时延抖动

7、在流量监管中对匹配的流量实施的监管动作包括：（　）

A. 转发　　B. 丢弃　　C. 着色　　D. 下一级监管

8、接口中当前共有4个流，它们的优先级分别为1、2、3、4。如果启用WFQ队列，那么优先级为3的流所占带宽比例为：（　）

A. 1/5　　B. 2/7　　C. 3/14　　D. 3/10

9、管理员要对接口E/1/0/1所有出去的流量进行限速，他最好的办法是在该接口上实施：（ ）

A. CAR　　B. GTS　　C. ACL　　D. LR

10、某运营商的接入路由器上既要承载语音和视频等实时业务，同时还连接了一个数据中心，此外对于付费不同的企业用户还要提供不同的QoS保证。这时运营商的QoS策略可以基于以下哪种方式：（ ）

A. FIFO + CQ　　B. WFQ　　C. CBQ　　D. RTP + CQ

二、填空题

1、实施流量管理的基础是对所有进入网络的流量进行正确的（ ）。

2、IP报文头中的ToS字段共（ ）bits，提供了（ ）个优先级和（ ）个DSCP值。

3、在定制队列CQ中根据一定的规则，报文最多可以进入（ ）个队列。

4、在拥塞避免中传统的尾丢弃方式会带来（ ）的问题。

5、WRED只能和（ ）队列共同使用，不能单独使用或和其他队列共同使用。

三、判断题

1、Int-Serv模型为用户提供了端到端的绝对的QoS保障，而Diff-Serv模型只能承诺相对的服务质量。（ ）

2、用令牌桶进行流量评估时其突发尺寸可以大于、小于或等于最大报文长度。（ ）

3、流量监管和流量整形是最常用的QoS手段之一，他们既可以在流量的入接口也可以在出接口上实现限速等多种流量控制功能。（ ）

4、CQ的一个缺点是对各个用户队列分配固定的时间片，当某一队列为空时也必须经过相应时间后才能切换到下一个队列。（ ）

5、在CBQ中配置的最大接口可用带宽，其值最大不能超过物理接口的实际带宽。（ ）

四、问答题

1、试通过分析Int-Serv和Diff-Serv模型的优缺点说明Int-Serv模型没有得到成功商用的原因。

2、除了流分类外，Comware中还包括哪几种主要的流量管理技术？它们各自的功能是什么？

3、请简述采用WRED方式进行拥塞避免的原理。

参考答案

一、选择题

1、AB　2、ABCD　3、C　4、ABC　5、D　6、BD　7、ABCD　8、B　9、D　10、CD

二、填空题

1、流分类　2、8、8、64　3、17　4、TCP全局同步　5、WFQ

三、判断题

1、√　2、×　3、×　4、×　5、×

四、问答题：

1、答：Int-Serv模型的优点在于采用资源预留的方式，可以为互联网上的每条流提供端到端的绝对的QoS保证。其缺点在于每个结点都要保留所有流的状态信息，这就导致核心路由器负担太重，可扩展性很差。而且资源预留与路由协议在选路上出现的矛盾也是一个很难解决的问题。这就导致了至今为止也没有一个真正意义上的Int-Serv模型的网络成功商用。而Diff-Serv模型的优点表现在业务流状态信息的保存和流控机制的实现都在网络边界结点进行，内部结点与状态无关，只保存简单的DSCP和PHB的对应机制，因此实现简单，扩展性好。其缺点在于只能提供相对的服务质量。另外，当拥塞发生时只能采取丢弃报文的方式，而不能进行旁路等处理。

2、答：还包括以下几种流量管理技术：

- 流量监管：一种通过对用户流量进行监督来限制流量及其资源使用的流控策略。典型的应用是对超出监管的流量进行“惩罚”；
- 流量整形：是一种主动调整流量输出速率的措施。典型应用是基于下游网络结点的TP指标来控制本地流量的输出；
- 拥塞管理：用于当拥塞发生时制定一个资源的调度策略，决定报文转发的处理次序。其核心思想是队列调度技术；
- 拥塞避免：指通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞有加剧的趋势时，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。

3、答：在WRED中为不同的优先级设定一个队列的低限值和高限值，并规定：

- 当队列的平均长度小于低限时，不丢弃报文；
- 当队列的平均长度超过高限时，丢弃所有到来的报文；
- 当队列的平均长度在低限和高限之间时，开始随机丢弃到来的报文。方法是为每个到来的报文赋予一随机数，并用该随机数与当前队列的丢弃概率比较，如果小于丢弃概率则被丢弃。

平均队列长度的计算公式为：$average=old\_average\times(1-2^{-n})+current\_queue\_size\times 2^{-n}$，其中n为用户配置值，即Weighting-constant。

当前队列的丢弃概率：$\frac{(average-MinTreshold)\times 4\times 10^{6}}{(MaxThreshold-MinThreshold)\times Discardprobability}\times X$，其中X是从1开始的计数器，如报文未丢弃则X加1；否则，X被归1。

队列调度机制简介

文/史计达

队列调度机制在QoS技术体系中属于拥塞管理的范畴。虽然企业和运营商想尽一切办法去扩展自己的链路带宽，但是现实网络上各种应用对带宽的消耗速度远远超出企业和运营商带宽扩充能力，也就是说网络的拥塞是无法避免的，这也决定了拥塞管理这一技术的重要性。拥塞管理是指网络发生拥塞时，如何进行管理和控制，处理的方法是使用合适的队列技术来确保关键业务的优先保障。在出接口发生拥塞时，通过适当的队列调度机制，可以优先保证某种类型的报文的QoS参数，例如带宽、时延、抖动等。我们这里所说的队列是指出队列，其实就是指向指定缓存的一系列指针，其作用是在接口有能力发送报文之前先将报文在缓存中保留下来，直到接口可以继续发送报文，所以队列调度机制都是在出端口发生拥塞情况下产生作用，另外一个主要作用就是将报文重新排序（FIFO除外）。

队列是一个比较容易理解的概念，我们在日常生活中也用到类似技术。例如我们去电影院买票的时候，大家排成几队顺序买票，排在前面的先拿到票（FIFO）；有时突然冲出一个人跑到队伍的最前面拿出VIP证件马上就拿到了票（PQ），这类人属于特权阶级需要优先处理，后面的人只能等这类人买完票才能继续排队买票。

FIFO

FIFO是队列机制中最简单的，每个接口上只有一个FIFO队列，表面上看FIFO队列并没有提供什么QoS保证，甚至很多人认为FIFO严格意义上不算做一种队列技术，实则不然，FIFO是其它队列的基础，FIFO也会影响到衡量QoS的关键指标：报文的丢弃、延时、抖动。既然只有一个队列，自然不需要考虑如何对报文进行复杂的流量分类，也不用考虑下一个报文怎么拿、拿多少的问题，即FIFO无需流分类、调度机制，而且因为按顺序取报文，FIFO无需对报文重新排序。简化了这些实现其实也就提高了对报文时延的保证。

FIFO关心的就是队列长度问题，队列长度会影响到时延、抖动、丢包率。因为队列长度是有限的，有可能被填满，这就涉及到该机制的丢弃原则，FIFO使用Tail Drop机制。如果定义了较长的队列长度，那么队列不容易填满，被丢弃的报文也就少了，但是队列长度太长了会出现时延的问题，一般情况下时延的增加会导致抖动也增加；如果定义了较短的队列，时延的问题可以得到解决，但是发生Tail Drop的报文就变多了。类似的问题其它排队方法也存在。

Tail Drop机制简单的说就是如果该队列如果已经满了，那么后续进入的报文被丢弃，而没有什么机制来保证后续的报文可以挤掉已经在队列内的报文。

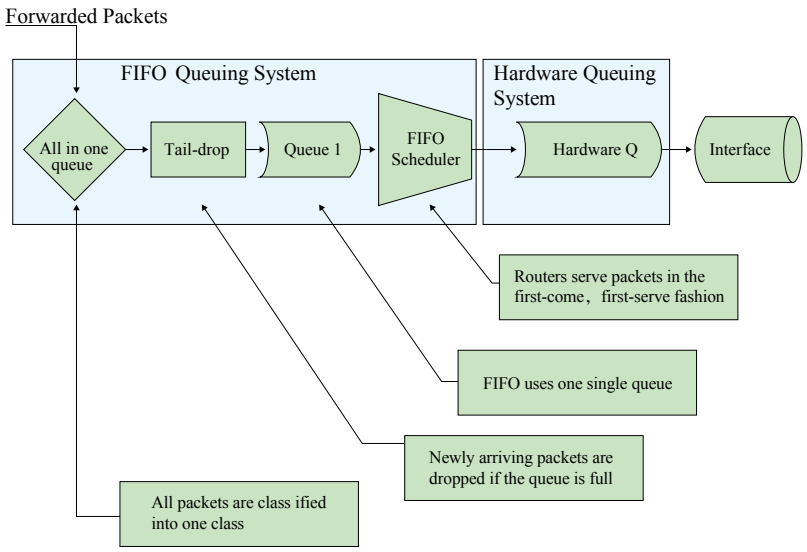


图1 FIFO流程图

Priority Queuing

阶级社会将不同的人划入到不同的阶层，不同的阶层享受的关照度不同，如果把FIFO比作原始社会的话，那么PQ就进入了封建社会了。

在报文到达接口后，首先对报文进行分类，然后按照报文所属类别让报文进入所属队列尾部，在报文发送时，按照优先级，总是在所有优先级较高队列中报文发送完毕后，再发送低优先级队列中报文，这样，在每次发送报文时，总是将优先级高的报文先发出去，保证了属于较高优先级队列报文有较低时延，所以PQ的优缺点是很明显的：优点是可以保证高优先级队列的报文可以得到较大带宽、较低的时延、较小的抖动；缺点是低优先级队列的报文不能得到及时的调度，甚至得不到调度，即会出现“饿死”现象。

PQ具有如下特征：

- 报文丢弃策略采用Tail Drop机制；
- 每个队列内部使用FIFO逻辑；
- 当从队列调度报文时，先从高优先级的队列调度报文。

从上面可以看出，PQ一般的应用场合是保证某类流量尽可能得到最好的服务，而不管其它流量的“死活”。

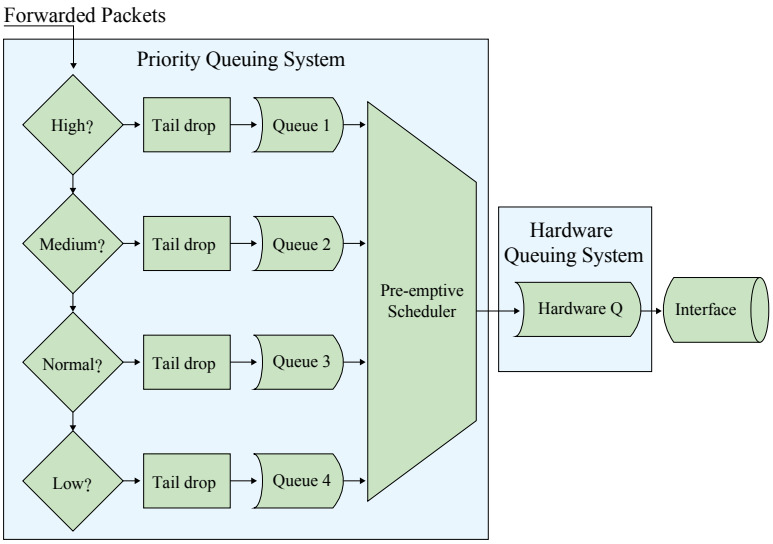


图2 PQ流程图

Custom Queuing

CQ可以说是对PQ的一种改进，解决了PQ“饿死”的重大缺点，能够确保使所有的队列都得到服务。CQ可以把报文分类，然后按照类别将报文分配到CQ的一个队列中去，而对于每个队列，也可以规定队列中报文所占接口的带宽比例，这样，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。

CQ有0～16个队列，其中0队列是优先级队列，只有0队列的报文处理完才会去处理1～16队列，所以0队列一般用做系统队列。

CQ采用Round Robin调度方式，从队列1开始，从每个队列取出指定数目的报文，直到报文数目满足或者超出设置的范围，当从该队列取出了足够的报文或者队列中没有报文的话，开始对下一个队列进行类似的操作。CQ不会配置确切的链路带宽比例，而是配置字节数目，可以根据配置的每个队列应取得的字节数目计算出每个队列占用的链路带宽，公式为：该队列应取得的字节数目/所有队列应取得的字节数目＝该队列占用的链路带宽。如果一段时间内某个队列为空，剩余的队列按照比例分配该空队列所占用的带宽。举个例子来说：现在配置了5个队列，每次取的字节数分别为5000、5000、10000、10000、20000，如果5个队列都有充足的报文需要发送，那么每个队列分配的带宽比例为10%、10%、20%、20%、40%；假设队列4有一段时间内没有报文发送，即队列为空，那么剩余的4个队列把这20%的带宽按照1:1:2:4的比例进行分配，所以在这段时间内这四个队列实际分配的带宽为12.5%、12.5%、25%、50%。

CQ不能将报文进行分片，例如要从队列1拿出1500字节的报文，可能会出现如下情况：

前面拿了1499bytes，还需要拿1byte，但是紧接着的一个报文大小是1500bytes，那么实际上从该队列拿出了1499＋1500＝2999bytes了，所以从局部来看的话，调度的比例和预期设置的结果有出入。

CQ有如下特点：

- Tail Drop是唯一的丢弃机制；
- 最大16个队列（因为0队列用做系统队列，这里不计算在内）；
- 队列内部使用FIFO逻辑；

■ 在对各队列进行调度时，使用Round-Robin对各队列按字节数调度。

CQ可以应用在对抖动要求不严格同时能够根据需要对不同分类的流量保留指定链路带宽的情况。CQ没有像PQ一样对某类流量提供低时延的服务，但是它可以保证在发生拥塞时1-16队列能够分配到指定额度的带宽。

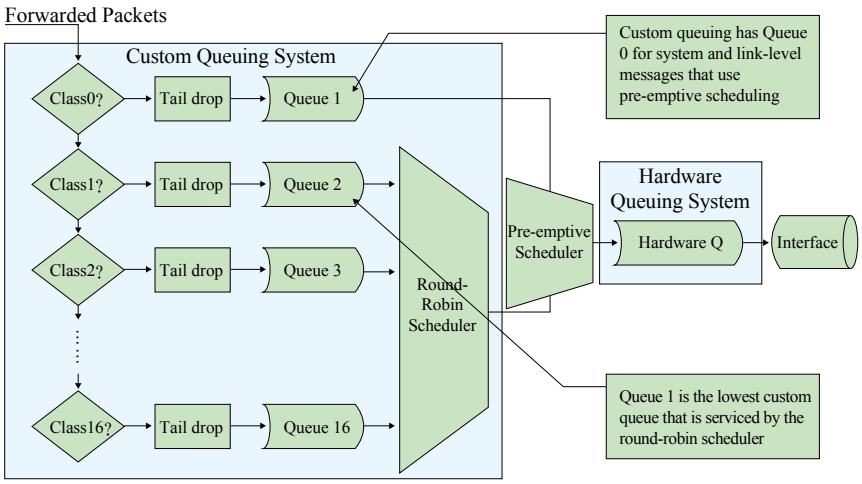
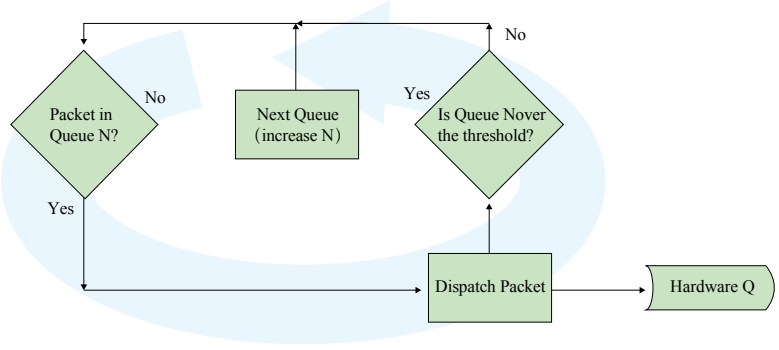


图3 CQ流程图



－Custom queuing uses a round-robin service policy
－Each queue is allowed to forward a configurable amount of bytes（threshold）in one round

图4 CQ的RR调度流程图

Weight Fair Queuing

WFQ是根据流对报文进行动态分类，对于IP网络，五元组（源IP地址、目的IP地址、源端口号、目的端口号、协议号）和IP优先级或者DSCP相同的报文属于同一个流。在接入层的网络中，通常使用IP优先级和五元组配合进行流分类；在汇聚层网络中通常使用DSCP值和五元组配合进行流分类，具有相同特性的报文属于同一个流，使用Hash算法映射到不同的队列中；另外的一个区别就是如果使用WFQ，那么low-volume（字节数小的报文）、higher-precedence（优先级高的报文）的流会比large-volume、lower-precedence的流更先处理。因为WFQ是基于流的，每个流使用不同的队列，这就要求WFQ能够支持很大数目的队列——WFQ最大可以在每个接口支持到4096个队列。

WFQ与CQ主要区别如下：

■ CQ可以自定义ACL规则来对报文进行分类，而WFQ只能根据元组对报文进行动态分类；

■ WFQ和CQ的队列调度方式不一样，CQ的调度方式是RR，而WFQ的调度机制是WFQ调度机制；

■ WFQ和CQ的报文丢弃机制不一样：CQ使用Tail Drop机制，WFQ使用WFQ丢弃机制，该机制是对Tail Drop的一种改进。

要想理解WFQ，必须了解这个机制出现的目的是什么，即使用WFQ是为了达到什么目的？WFQ调度主要是为了两个主要的目的，一个是在各个流之间提供公平的调度即WFQ名字中的F（fairness）的含义，另外一个就是保证高IP precedence的流能够得到更多带宽即WFQ名字中的W（weighted）的含义。

为了提供各个流之间的公平调度，WFQ给每个流分配的带宽是相同的。例如一个接口有10条流，该接口带宽为128Kbps，那么每个流得到的带宽为128/10=12.8Kbps。从某种意义上讲，有些类似于时分复用机制（TDM）。WFQ允许其它流使用某条流的剩余带宽，例如接口带宽为128kbps，共10条流，则每条流分配的带宽为12.8kbps，可能实际上某条流例如流1只有5kbps，而流2有20kbps，那么其它的流就可以分配流1所剩余下的12.8-5=7.8kbps的带宽。

WFQ的加权是根据流中的IP precedence进行的，保证高IP precedence的流分配到更多的带宽。算法为（IP precedence＋1）/Sum（IP precedence＋1），例如有四个流，其IP precedence分别为1、2、3、4，那么每个流占用的带宽分别为2/14、3/14、4/14、5/14。

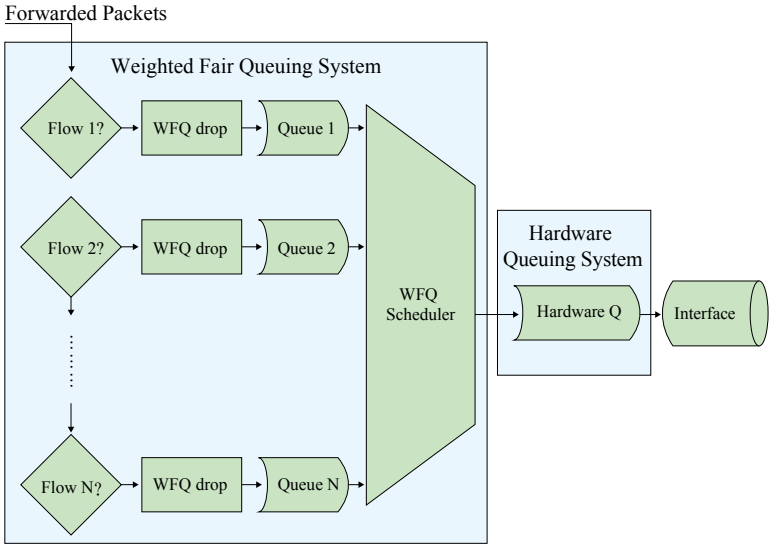


图5 WFQ流程图

要想理解WFQ的报文丢弃机制和队列调度机制，需要理解WFQ中的一个重要概念：序列号SN（不同的文档可能采用不同的参数，不管使用什么参数都应该达到小字节、高IP优先级的流被优先调度），报文在经过流分类后，在决定该报文是入队列还是丢弃之前，都要赋予一个SN。SN的计算公式为SN＝Previous_SN＋weight×new_packet_length，WFQ进行报文调度时都是先调度SN小的报文，为了保证IP Precedence大的能够获得更多的带宽，从SN的计算公式就可以看出Weight应与Precedence成反比。

其中Previous_SN分为两种情况：

- 如果报文进入的队列为非空，使用该队列中最近进入队列报文的SN作为Previous_SN；
- 如果报文进入的队列为空，使用发送队列最近发送的报文的SN作为Previous_SN。

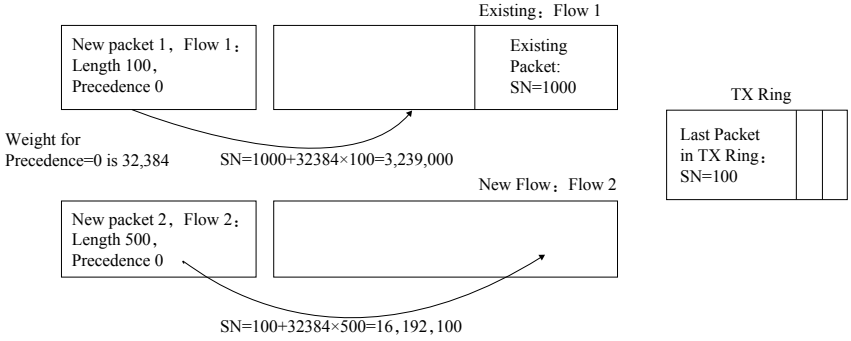


图6 Previous_SN的选择

WFQ在进行报文丢弃和入队列时也是根据SN的大小来进行的：

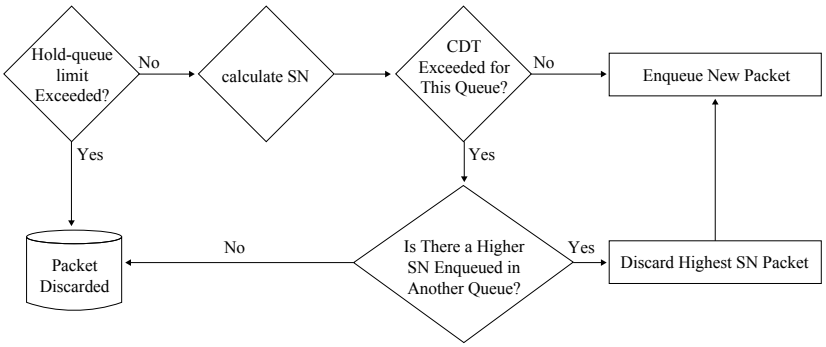


图7 报文丢弃原理

HQL（Hold-Queue-Limit）：限制了所有队列中能够存放的报文总数目；

CDT（Congestive Discard threshold）：限制了每个队列中能够存放的报文数目。

				Flow 1
1500 byte, Precedence 0	Packet4 SN=194,304,000	Packet3 SN=145,728,000	Packet2 SN=97,152,000	Packet1 SN=48,576,000
				Flow 2
1000 byte, Precedence 0	Packet8 SN=129,536,000	Packet7 SN=97,152,000	Packet6 SN=65,536,000	Packet5 SN=32,384,000
				Flow 3
500 byte, Precedence 0	Packet12 SN=65,536,000	Packet11 SN=48,576,000	Packet10 SN=32,384,000	Packet9 SN=16,192,000
				Flow 4
100 byte, Precedence 0	Packet16 SN=12,954,600	Packet15 SN=9,715,200	Packet14 SN=6,553,600	Packet13 SN=3,238,400

图8 SN的选择

上图中：有四条流，每条流的precedence相同都为0，只是报文的大小不同，Flow 1到Flow 4的报文长度从大到小，按照SN的计算公式，报文长度小的SN小，所以Flow 4中的报文应该被优先调度出去，当然最终的决定因素还是SN的大小，对于SN相同的报文实行顺序调度，如本例所示：Packet 5和Packet 10的SN相同、Packet 1和Packet 11的SN相同，按照顺序调度规则，应该是Packet 5在Packet 10前，Packet 1在Packet 11前。最终的调度的结果是：13，14，15，16，9，5，10，1，11，6，12，2，7，8，3，4。

WFQ使用WFQ丢弃机制，该机制是对Tail Drop的一种改进，其中的一个决定因素也是SN，另外WFQ还使用HQL和CDT来决定如何对报文进行丢弃。如果一个新的报文达到时HQL已经到达最大值，该报文直接被丢弃；如果此时HQL没有到达最大值，WFQ将该报文根据WFQ的分类原则进行分类决定进入到哪个队列并计算出SN，剩下的丢弃机制还会由CDT决定，CDT是每个队列自己的丢弃阀值，如果此时CDT没有到达最大值报文直接进入该队列，如果CDT已经达到阀值，则判断其它队列是否有SN比新进入的报文SN大，如果没有直接丢弃新进入的报文，如果其他队列有SN大于当前入队列的报文，WFQ会选择丢弃SN大最大的报文。简单的说就是当某个队列的报文数目已经超过该队列CDT，WFQ可以选择丢弃其它队列中SN最大的报文！其流程图如图8所示。

将WFQ的特点可以总结为如下特点：

- 基于元组对报文进行流分类，不支持用户自定义的分类；
- WFQ丢弃机制，是对Tail Drop的改进；
- 队列内部使用FIFO；
- WFQ调度：优先服务低SN的报文。

Class-Based WFQ

CBWFQ从名字来看像是CQ和WFQ的混合体。和CQ类似的，CBWFQ可以为每个队列保留最小带宽，使用和CQ类似的报文分类，但是与CQ不同的是，用户可以配置CBWFQ实际占有的流量百分比，而不是字节数；和WFQ相比，CBWFQ可以在一个特定的队列里面使用WFQ机制：CBWFQ有一个特殊的队列，即缺省队列，只有该特殊队列可以采用WFQ机制。

CBWFQ支持两种丢弃机制：Tail Drop和WRED。可以配置任何一个队列的丢弃机制为WRED，但是并不是所有的队列配置WRED丢弃机制都是有益的，WRED可以用在对丢包不是很敏感的数据队列；如果该队列是存放语音、视频报文，这类业务报文对丢包比较敏感就不适合采用WRED了。

CBWFQ有0～64队列，0队列是缺省队列，该队列是自动配置、不可人工干预。可以使

用流分类将不同类型的报文映射到1～64队列，可以设置每个队列所占用的带宽比例；如果进入的报文不能匹配任何流分类，进入缺省队列，缺省队列可以使用FIFO或者WFQ机制，而1～64只能使用FIFO机制。为什么只有缺省队列0可以采用WFQ机制？这样有什么好处呢？前面已经提到CBWFQ可以根据分类将报文放入到指定队列，通过配置该队列的带宽比例获取相应的服务，如果在一段时间某个队列不需要该带宽，由其它队列共享；对于那些无法进行分类的报文统统放入到队列0，通过在0队列使用WFQ机制可以使该队列中的所有报文能够得到公平的调度。

CBWFQ的有一个严重的缺点就是没有一个队列可以满足那些对时延有特殊要求的报文，例如语音、视频流，也就是缺乏类似于PQ之类的严格优先级队列。

Low Latency Queuing

从名字我们就可以大致知道这个队列的作用，就是为了保证某类报文的低时延，目前的实现方式都是通过严格优先级队列来保证该类报文被优先处理，从而对时延加以保证。实际上LLQ严格意义上并不是一个独立的队列机制，可以认为它是CBWFQ队列机制的一个变种。通过在在CBWFQ队列中加入了一个或者几个优先级队列来实现，以保证这些队列的优先处理，从而保证进入该分类的报文较低的时延；而通过设置带宽阀值，又能防止出现“饿死”现象。

LLQ在调度时一直首先检查低时延队列，从该队列取报文，如果该队列没有报文时才转向其它非低时延队列。LLQ可以设置多个低延时队列，多个低延时队列之间的关系是平等的，采用的FIFO逻辑，用户可以根据需要将不同的业务放入到不同的低延时队列，例如语音流放到一个低延时队列，视频流放到另外一个低延时队列，可以更好的保证两种业务能够同时得到相应的服务。低延时队列和非低延时队列之间的关系类似PQ，既然类似PQ就不可避免的出现“饿死”现象。LLQ通过设置低时延队列的带宽值来防止“饿死”现象出现。

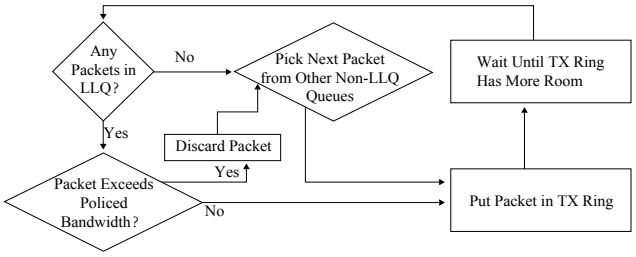


图9 LLQ原理图

从上图中可以看出，为了保护其它非LLQ队列得到调度，LLQ采用丢弃LLQ队列超过指定带宽的报文的方式来实现该目的，这样的负面影响也是很明显的：语音、视频流量需要进入LLQ队列来保证低时延、低抖动，它们同样对报文丢弃很敏感，这样反而失去了LLQ的本来意义，有点矛盾，唯一的办法就是合理安排好LLQ队列所占用的带宽比例，尽可能的保证该队列的报文不出现丢包。

IP RTP Prioritization

IP RTP和LLQ类似，但又有一些不同点。它通过在WFQ或者CBWFQ队列中加入严格优先级队列来实现的，它通过区分UDP报文的目的端口号来对VoIP报文进行分类，从中选择出UDP目的端口号在一定范围之内且为偶数的流量。因为IP RTP是严格优先级队列，所以会被最先处理，并且通过一定的策略防止这个严格优先级队列占用太多的带宽，也就是说该严格优先级队列占用的带宽是有额度的，超过限制的流量被丢弃。

通过RTP的实现可以看出，RTP具有如下特点：

- 在CBWFQ中增加了一个低时延队列，保证VoIP报文的及时处理；
- 限制了优先级队列带宽大小，防止出现“饿死”现象；
- 流分类手段贫乏。

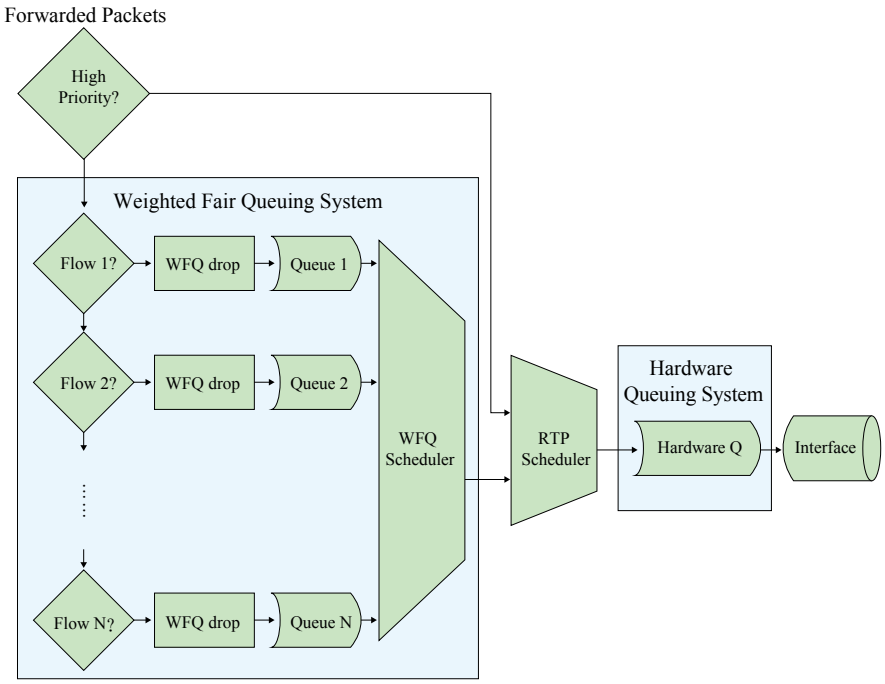


图10 RTP流程图

总结

对于队列机制而言，它的最重要的两个特性就是调度方式和丢弃机制，在学习队列机制时要从这两个特性去考察、对比不同队列机制之间的共同点和不同点，明白不同队列机制不同的应用场合。

队列调度机制	调度方式	丢弃机制
FIFO	顺序调度	尾丢弃
PQ	首先调度高优先级队列	尾丢弃
CQ	从每个队列取指定字节的报文，队列之间采用RR机制	尾丢弃
WFQ	先调度SN小的报文	改进的尾丢弃
CBWFQ	保证每个队列的设定带宽	尾丢弃或WRED
LLQ	首先服务低时延队列，但是低时延队列有阀值	尾丢弃或WRED
IP RTP priority	首先服务低时延队列，但是低时延队列有阀值	尾丢弃

图11各队列的对比图

流量监管和流量整形

文/吴秀

本文主要阐述QoS技术中流量监管和流量整形的实现机制。本文描述现今IETF对流量规格度量的两种算法来了解令牌桶的工作原理；主要讲述H3C路由器流量监管和流量整形的实现机制，由于流量限速也用到相同的令牌桶处理机制，所以也讲述了流量限速的实现机制。

流量监管和流量整形简介

在提供QoS服务时，网络边界路由器与内部路由器功能有所侧重，并像一个整体一样相互协作。Diff-Serv将复杂的流分类和流量控制都推至边界路由器来完成。边界路由器主要完成复杂流分类、为分组打DSCP标记、流量的接入速率监管、访问控制等动作。区域内部路由器只需进行简单流分类，对同一类流实施流量控制。这样做避免了Int-Serv模型中的基于每个流（Per-Flow）的复杂流分类及流控，从而使得区分网络内部的转发操作可以得到高效的实现。也就是说流量监管和流量整形主要是在Diff-Serv中的边缘设备上进行。

从高速链路向低速链路传输数据时，带宽会在低速链路接口处出现瓶颈，导致数据丢失严重，特别是会影响到低延时要求的数据

如语音等。流量监管（traffic policing）的典型作用是限制进入或流出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（traffic shaping）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，使这类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。

IETF的两种令牌桶算法

IETF建议采用srTCM（A Single Rate Three Color Marker，RFC2697）算法或trTCM（A Two Rate Three Color Marker，RFC2698）算法对流量进行测评，根据评估结果为报文打颜色标记，即绿色、黄色和红色。对于AF业务的报文，可根据评估结果按照报文的颜色，将报文重新标记为不同的丢弃优先级。

srTCM与trTCM算法均采用两个令牌桶对到达的报文进行评估，他们允许流量在某种级别上突发——但关注的侧重点不同，srTCM更关注报文尺寸的突发，trTCM则关注速率上的突发。srTCM与trTCM算法有两种工作模式，色盲模式（Color-Blind）与感色模式（Color-Aware），其中色盲模式是较常用的。下面我们简单介绍一下这两个算法。

srTCM算法（RFC 2697）

单速率三色标记器（srTCM）能够度量IP分组流，并把分组标记为绿色、黄色或红色。如果到达的分组未超过承诺突发尺寸，则把它标记为绿色；如果超过了承诺突发尺寸而未超过超额突发尺寸，则把它标记为黄色；否则，标记为红色。单速率三色标记器可以用在网络入口处来管制服务。

单速率三色标记器有两种工作模式:色盲模式和感色模式。在色盲模式下，假定所有的分组都是未经标记的。在感色模式下，假定所有输入的分组已经被标记为绿色、黄色或红色。

配置单速率三色标记器时要指定3个参数：承诺信息速率CIR、承诺突发尺寸CBS和超额突发尺寸EBS。其中，CBS和EBS要大于0，并且至少应该大于等于最大的分组长度。

为方便，将两个令牌桶称为C桶和E桶，用Tc和Te表示桶中的令牌数量，Tc和Te初始化等于CBS和EBS。CBS比EBS要小。

Tc和Te在每秒钟内更新CIR次，更新时遵循以下规则：

- 如果Tc<CBS，则Tc增加1，否则；
- 如果Te<EBS，则Te增加1，否则；
- Tc和Te都不增加。

色盲模式下，在对到达报文（假设报文大小为B）进行评估时，遵循以下规则：

- 如果Tc-B>=0，则报文被标记为绿色，且Tc降低B，否则；
- 如果Te-B>=0，则报文被标记为黄色，且Te降低B，否则；
- 报文被标记为红色且Tc和Te都不降低。

非色盲模式下，在对到达报文（假设报文大小为B）进行评估时，遵循以下规则：

- 如果报文已被标记为绿色且Tc-B>=0，则报文被标记为绿色，且Tc降低B，否则；
- 如果报文已被标记为绿色或黄色且Te-B>=0，则报文被标记为黄色，且Te降低B，否则报文被标记为红色且Tc和Te都不降低。

trTCM算法（RFC 2698）

配置单速率三色标记器时要指定4个参数：承诺信息速率CIR、峰值信息速率PIR、承诺突发尺寸CBS和超额突发尺寸EBS。trTCM算法中两个令牌桶的填充令牌的速率不同，分别为承诺的平均速率CIR（Committed Information Rate）和峰值速率PIR（Peak Information Rate）。为方便将这两个令牌桶称为C桶和P桶，这两个桶的尺寸分别为承诺突发尺寸CBS（Committed Burst Size）和峰值突发尺寸PBS（Peak Burst Size）。用Tc和Tp表示桶中的令牌数量，Tc和Tp初始化等于CBS和PBS。Tc和Tp在每秒钟内分别更

新CIR和PIR次，每次更新增加一个令牌（除非桶满）。

在色盲模式下，在对到达报文（假设报文大小为B）进行评估时，遵循以下规则：

- 如果Tp-B<0，则报文被标记为红色，否则；
- 如果Tc-B<0，则报文被标记为黄色，且Tp降低B，否则；
- 报文被标记为绿色且Tc和Tp都降低B。

在非色盲模式下，在对到达报文（假设报文大小为B）进行评估时，遵循以下规则：

- 如果报文已被标记为红色或者Tp-B<0，则报文被标记为红色，否则；
- 如果报文已被标记为黄色或者Tc-B<0，则报文被标记为黄色，且Tp降低B，否则；
- 报文被标记为绿色且Tc和Tp都降低B。

流量监管和流量整形的实现机制

流量监管（Committed Access Rate，CAR）

H3C路由器实现的令牌桶算法是色盲模式srTCM，但是算法上作了一定的改进,所谓的单桶双色算法。CAR和GTS中EBS可以配置，但是没有起应有的作用，而是把这个桶放到CBS中，起加深CBS桶的作用。CBS和EBS是合起来算的，也就是说实际的CBS=配置的CBS+EBS，初始时令牌是满的。那么当有报文来时，如果报文长度B大于实际的CBS就是红，小于实际的CBS就是绿，没有黄色报文。值得注意的是：CAR和GTS中报文长度B包含二层信息的长度，如果是PPPOE链路，在VT上配置CAR和GTS报文长度应包含PPP头部。令牌桶如图1所示。

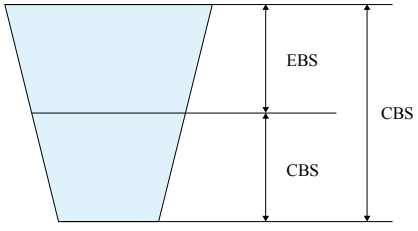


图1 令牌桶的构成图图表

CAR进行流量控制的基本处理过程。

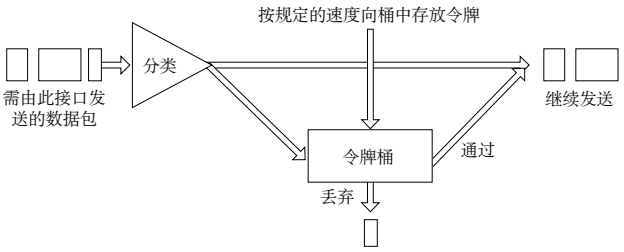


图2 CAR示意图

令牌桶初始时是满的，TB=CBS+EBS。首先，根据预先设置的匹配规则来对报文进行分类，如果是没有规定流量特性的报文，就直接继续发送，并不需要经过令牌桶的处理。

其次，如果是需要进行流量控制的报文，则会进入令牌桶中进行处理。具体来说，进入令牌桶处理的包长度B–TB<0，则报文是绿色的，反之报文是红色的。具体对不同颜色报文的处理行为可以通过针对定义不同颜色的处理行为来决定，被判断为可以通过的报文被送出，在桶中取走与报文长度（bit为单位）等量的令牌，所剩令牌为TB=TB–B报文长度。

最后，如果令牌桶中的令牌不满足报文的发送条件，则报文被丢弃。

针对上面的处理流程，例如，如果对进入令牌桶处理的报文长度为800bits。

TB=CBS+EBS=30000bits,这时30000-800>0,所以报文是绿色的，否则报文是红色的。

H3C向令牌桶中补充令牌不是周期性加的，当桶中令牌不够发送报文时给令牌桶加（当前的时间 —— 上次加令牌的时间）×CIR个令牌，溢出的令牌丢弃。这样，就可以对某类报文的流量进行控制。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。

令牌桶是一个控制数据流量的很好的工具。当令牌桶中充满令牌的时候，桶中所有的令牌代表的报文都可以被发送，这样可以允许数据的突发性传输。突发就是任何一段时间T1、T2（T1<T2）内通过的流量都不可能大于（T2-T1）×cir+cbs。

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是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，VRP的CAR不仅可以用来进行流量控制，还可以进行报文的标记（mark）或重新标记（re-mark）。具体来讲就是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。

CAR可以为不同类别的报文设置不同的流量特性和标记特性。即，首先对报文进行分类，然后不同类别的报文有不同的流量特性和标记特性。

CAR能在出接口和入接口上生效。

通用流量整形（Generic Traffic Shaping, 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，减少了报文的丢弃。

GTS的基本处理过程如图3所示，其中用于缓存报文的队列称为GTS队列。

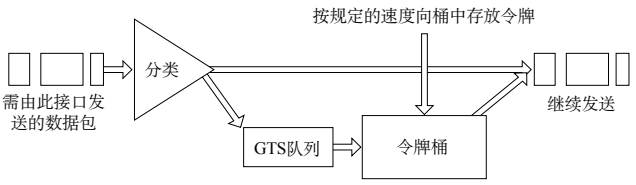


图3 GTS处理过程图

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，首先对报文进行分类，如果报文不需要进行GTS处理，就继续发送，不需要经过令牌桶的处理；流量整形的令牌桶的构成同CAR，如果报文需要进行GTS处理，则与令牌桶中的令牌进行比较，进入令牌桶处理的包长度B–TB<0则报文被发送，否则报文被缓存，等到令牌桶中有足够的令牌时继续发送报文。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下去，同时，令牌桶中的令牌量按报文的长度做相应的减少。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，这里的队列是FIFO队列，与接口上的FIFO不同，当然队列有一定的长度（以包为单位），当需要缓存的报文个数大于队列长度时报文因无法缓存而丢弃。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。另外，GTS也允许有突发。GTS只能在出接口上生效。

GTS可用来进行网络上下游之间的带宽匹配。例如在图4所示的应用中，假设路由器1向路由器2发送报文，路由器2对路由器1发送来的报文进行了CAR流量限制。

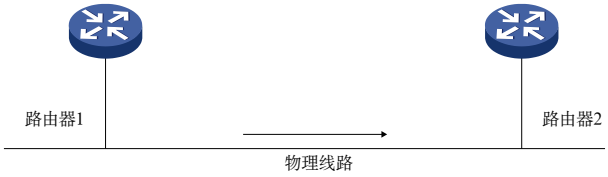


图4 GTS的应用示例图

为了减少报文的丢失，可以在路由器1的出口对报文进行GTS处理，对于超出GTS流量特性的报文，将在路由器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，发往路由器2的报文将都符合路由器2的流量规定，从而减少报文在路由器2上的丢弃。相反，如果不在路由器1的出口做GTS处理，则所有超出路由器2的CAR流量特性的报文将被路由器2丢弃。

物理接口总速率限制（Line rate, 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）的总速率。

LR的处理过程仍然是采用令牌桶进行流量控制。如果用户在路由器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，则报文入紧急队列等待下一次发送机会。这样，就可以对通过该物理接口的报文流量进行控制。LR的处理过程如图5所示。

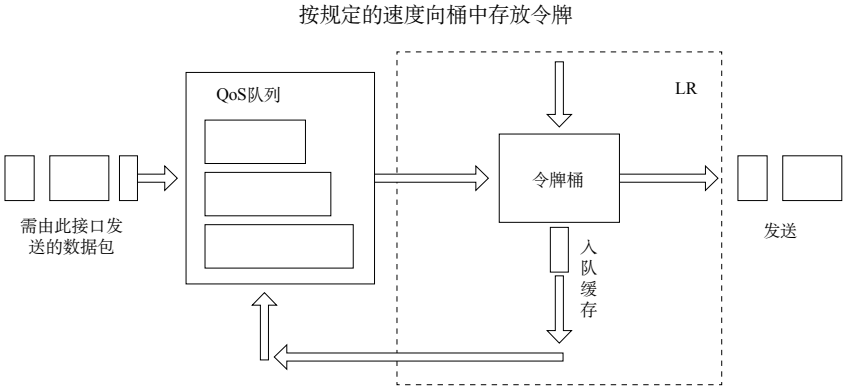


图5 LR处理过程图

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是小于等于令牌生成的速度，具有限制流量，同时允许突发流量通过的目的。

注意：图5中只描述得不到令牌的报文需要入队列，也就是说能够得到令牌的报文不用入队列，直接就发送出去。

LR能够限制在物理接口上通过的所有报文。LR相比较于GTS，不但能够对超过流量限制的报文进行缓存，而且还因为进入了QoS队列机制进行处理，所以队列调度机制更灵活。

在用户只要求对所有报文限速时，使用LR所需的配置操作简单。对于网络建设投资者，可以对客户隐藏实际带宽，客户只能严格按所购买的带宽来使用。

QoS队列调度算法概述

文／常慧锋

队列调度算法是实现网络QoS（Quality of Service，服务质量）控制的核心机制之一，是网络资源管理的重要内容，通过控制不同类型的分组对链路带宽的使用，使不同的数据流得到不同等级的服务。

通常调度算法的工作模式可以分为两种：工作保留模式（work-conserving）和非工作保留模式（non-work-conserving）。如果队列中有数据包等待发送服务器就工作的调度算法称为工作保留调度算法；如果队列中有数据包等待发送但服务器仍然处于空闲状态的调度算法称为非工作保留调度算法，例如，即使服务器处于空闲状态同时队列中有数据包等待发送，但是为了等待下一个高优先级的数据包服务器也会推迟当前数据包的传输，这种调度算法就属于非工作保留调度算法。当数据包的传输时间很短时，非工作保留调度算法几乎是不公平的。

调度算法的另一种分类方法是根据调度算法的内部结构来划分的，主要有两种：基于优先级分类的调度算法和基于帧结构的调度算法。在基于优先级的调度算法中有一个称为虚拟时间（virtual time）的全局变量。调度算法根据该变量为每个数据包计算一个时间戳，然后根据时间戳对数据包排序和调度。虚拟时钟，加权公平队列都属于这种结构。基于优先级的调度算法的实现复杂度取决于两个因素：更新优先级列表算法和选择最高优先级数据包算法的复杂度（至少是O（log$V$），其中 是共享输出链路的队列数）和计算时间戳算法的复杂度（这主要取决于所采用的调度算法，加权公平队列（WFQ）的时间戳的计算复杂度为O（V），虚拟时钟的计算复杂度只为O（1））。

在基于帧结构的调度算法中，时间被分为固定长度或可变长度的帧。每个数据流所能使用的带宽资源就是每一帧中所允许传输业务量的最大值。存储转发队列是帧长度固定的基于帧结构的调度算法，在这种结构中，如果一帧中数据流的业务量小于预留带宽，服务器就会空闲。加权循环队列，差额循环队列允许帧长度可变，同时，如果一个数据流的业务量小于预留带宽时，下一个数据流就可以提前被调度。基于帧结构的调度算法最大的优点是实现简单，成本低，最大的缺点是缺乏灵活性和扩展性。

典型的调度算法简介

先进先出队列（FIFO）

FIFO队列是最简单的基于优先级的调度算法。在FIFO队列中数据包的时间戳就是数据包的到达时间。FIFO队列提供了基本的存储转发功能，也是目前因特网中使用最广泛的一种方式，它采用默认的排队方法，不需要配置。其优点是实现简单，成本低，缺点是不能提供QoS功能和隔离技术，缺乏公平性，易于受到非法用户的攻击。

严格优先级调度算法（PQ）

严格优先级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（如图1所示）。假设队列1比队列2具有更高的优先权，队列2比队列3具有更高的优先权等等。只要链路能够传输分组，队列1尽可能快地被服务。只有当队列1为空，调度器才考虑队列2。当队列2有分组等待传输且队列1为空时，队列2以链路速率接受类似地服务。当队列1和队列2为空时，队列3以链路速率接收服务等等。

然而该调度机制会使低优先级队列处于饥饿状态。例如，如果影射到队列1的数据流在一段时间内以100%的输出链路速率到达，调度器将从不为队列2、3、4服务。避免队列饥饿需要上游路由器精心规定数据流的业务特性以确保映射到队列1的业务类不超出输出链路容量的一定比例，这样可以使队列1常常为空，允许调度器为低优先级队列服务。

严格优先级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，如果调度器是空闲的，则分组被立即服务。

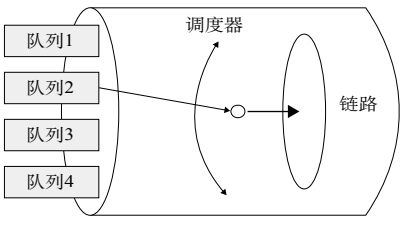


图1 严格优先权调度器

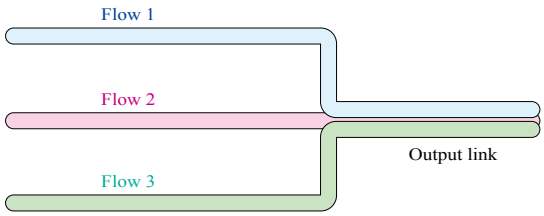


图2 通用处理器共享

通用处理器共享算法（GPS）和加权公平队列算法（WFQ）

通用处理器共享（Generalized Processor Sharing）

GPS算法是一种理想的调度算法（如图2所示），是根据流模型定义的，也就是假设数据包是可以被无限细分的。在GPS算法中，假设服务器的处理速率恒为r。在t时刻，如果数据流i的数据包在队列中等待处理，就认为数据流i在t时刻处于激活状态。假设有正整数$\varphi_1, \varphi_2, \varphi_3\cdots\cdots$，$\varphi_N$代表各数据流的权重，$S_i(\tau, t)$是时间间隔$(\tau, t]$中服务器为数据流$i$提供的服务，那么对于时间间隔$(\tau, t]$内任何暂存在服务器中的数据流$i$获得的服务$S_i(\tau, t)$，GPS算法定义为 $\frac{S_i(\tau,t)}{S_j(\tau,t)} \geq \frac{\phi_i}{\phi_j}, j = 1,2,.....N$ 。

假设r为服务器的处理速率，将数据流j累加可得 $S_i(\tau,t)\sum_j \phi_j \geq (t-\tau)r\phi_i$ 。根据轮换对称性，任意数据流i的保证速率（最小服务速率）为 $g_i = \frac{\phi_i}{\sum_j \phi_j} r$ 。

GPS算法具有如下特性：

- 假设r_i是数据流i的平均速率，只要$r_i \leq g_i$，就可以保证数据流i独立于其它数据流的吞吐率ρ_i。还可以保证数据流i的瞬时数据以大于等于g_i的速率被处理；
- 由于数据流i在任意时刻获得的服务独立于其他数据流，数据流i的时延抖动是自己队列长度和达到时间的函数，独立于其它连接的队列长度和到达时间。其它调度算法如FIFO和PQ没有这种性质；
- 通过改变φ_i我们可以以不同的方式处理不同的数据流。例如，当所有的φ_i都相同时，GPS就退化为均衡处理器共享（Uniform Processor Sharing）。另外只要数据流的平均速率之和小于r，不论怎么分配φ_i，系统总是稳定的；
- 通过增加φ_i就可以减少数据包经历的时延，虽然这种方式是以牺牲其它数据包的时延作为代价的，但是当激活数据流稳定时，

这种代价并不是很大。所以GPS算法和速率整形器联合使用时可以得到性能优良的调度器，为数据流提供最坏情况下的时延和时延抖动保证。

加权公平队列算法（WFQ）

GPS算法最大的缺点是不能处理长度可变的数据包。WFQ算法是GPS算法的近似。假设F_p是数据包p在GPS算法中的离开时间，那么WFQ算法就是一个模拟GPS算法并按F_p升序调度数据包的工作保留算法，也就是说WFQ算法总是选择F_p值最小的数据包进行调度。下面介绍WFQ算法的虚拟时间实现。

假设时间t_j是事件j^{th}的发生时间（同时发生的事件可以任意排序），服务器的处理速率为r。在服务器中，第一个发生事件的时间记为$t_1=0$。可以看出在时间间隔（t_{j-1}，t_j）内处于激活状态的数据流是固定的，我们将它记为集合B_j。当服务器空闲时，虚拟时间$V(t)$记为0，那么WFQ的$V(t)$计算如下：

$$\begin{cases} V(0)=0 \\ V(t_{j-1}+\tau)=V(t_{j-1})+\dfrac{\tau}{\sum_{i\in B_j}\phi_i}r, \tau \le t_j-t_{j-1}, j=2,3,... \end{cases}$$

$V(t)$的变化率$\dfrac{\partial V(t_j+\tau)}{\partial \tau}$为$\dfrac{1}{\sum_{i\in B_j}\phi_i}r$，每个暂存的数据流接收到的处理速度为$\phi_i\dfrac{\partial V(t_j+\tau)}{\partial \tau}$。假设数据流$i$的第$k$个数据包的到达时间为$a_i^k$，长度为$L_i^k$，$S_i^k$和$F_i^k$分别表示这个数据包的开始虚拟时间和结束虚拟时间，$F_i^o=0$，那么，我们可以得到：$\begin{cases} S_i^k=\max\{F_i^{k-1},V(a_i^k)\} \\ F_i^k=S_i^k+\dfrac{L_i^k}{\phi_i} \end{cases}$

从实现的角度来看，WFQ算法的虚拟时间实现有两个重要优点：

- 数据包的完成时间决定于数据包的到达时间和上一个数据包的完成时间；
- 数据包根据完成时间升序被处理。

WFQ使用虚拟时间存在的缺点：跟踪集合B_j需要花费很大的开销。

虚拟时钟算法（Virtual Clock）

虚拟时钟算法根据数据包的到达时间和用户定义的保留速率计算数据包的时间戳。假设TS_i^k是数据流i的第k个数据包的时间戳，ρ_i是数据流i的保留速率，AT是长度为L_i^k的数据包的实际到达时间，那么数据包的时间戳定义如下：

$$TS_i^K \leftarrow Max(AT, TS_i^{K-1})+\frac{L_i^K}{\rho_i}$$

如果数据包比预期的到达时间晚，那么经过最大延迟L_i^k/ρ_i后数据包被传输；如果数据包比预期的到达时间早，在最坏情况下，数据包被传输的时间为$TS_i^{k-1}-1+L_i^k/\rho_i$。最坏情况下的服务质量不受其它连接行为的影响。

存储转发队列（Stop-and-Go）

存储转发队列将输入和输出链路的时间轴分为固定长度的时隙“帧”。在两帧之间到达的数据包只能在下一帧中被传输，在同一帧中的数据包可以以任何次序传输，因此每个数据包都被引入一个固定的时延θ。如果数据包的最大到达速率小于帧中保留的时间片，那么这种算法能确保有限的时延和时延抖动。存储转发队列有两个问题：因为算法是非工作保留的，所以没有静态统计复用收益；时延θ与帧分配的颗粒度有关，选择小的帧长度可以获得小的θ，但是，为了得到好的带宽利用率应该选择大一些的帧长度。

循环队列（round-robin）

循环队列通过循环服务避免局部队列饥饿。调度器总是顺序地移到下一个有分组要发送的队列（空队列被跳过）。如果每个队列都有分组等待发送，调度顺序和队列顺序匹配；如果一些队列为空，则其它队列被频繁地服务。在极端情况下，如果其它队列都为空，单个队列就可以使用全部链路带宽。当分组进入一个空队列时，该队列在下一个循环中被服务，这样就可以避免队列“饥饿”。

循环调度的缺点是分组时延难于改进，它不可能为低时延业务分配专用队列。每个队列的服务间隔完全依赖于那段时间内其他队列中有多少分组等待发送以及这些分组的长度，这些变量难以准确预测，所以RR调度容易产生时延抖动。调度器可以通过改变服务顺序（例如采用顺序1，2，3，2，4，2，1，2，…）更频繁地调度某些队列以给这些队列更频繁的传送机会，然而分组大小的随机分布仍然会造成时延抖动问题。

差额循环队列（DRR）和加权循环队列（WRR）

DRR算法是RR算法的扩展。DRR算法为每个队列分配一个常量Q_N（以权重为比例的时间片）和一个变量D_N（差额）。Q_N反应了该队列可以发送的长期平均字节数。D_N的初始值为零且当队列为空时复位为0。当DRR算法服务一个新队列时，调度器复位计数器B_{sent}（表示该循环已经从队列中发送的字节数）。当下面两个条件满足时，DRR算法从队列中发送分组：

- 队列中有分组等待发送；
- （Q_N+D_N）大于等于（B_{sent}+队列中下一个分组的长度）。

否则，该队列的差额D_{N+1}被置为$Q_N+D_N-B_{sent}$，调度器按顺序移到下一个队列。Q_N+D_N表示在服务时间间隔内队列能够发送的最大字节数，在一定程度上DN可以平滑数据流的突发。队列通过Q_N可以获得长期的相对带宽分配。如果激活队列的数目小于N，则激活队列可以根据Q_N值共享未用的输出链路带宽。

WRR算法非常类似于DRR算法。WRR算法采用类似的时间片和差额的概念，但是算法稍有不同。在WRR中，当队列发送（B_{sent}）的字节数超过队列允许的限制时（仍为Q_N+D_N），才对下一个队列进行服务。因此，差额是一个负数值（超出Q_N+D_N的数量）且被当作下一个循环该队列发送的字节数的减少量。

WFQ算法与其他调度算法的比较

在DRR算法中，每个队列都有一个权值W_i。服务器按照预先规定的顺序以速率$w_i/\sum_j wj$轮询每个队列。如果遇到一个空队列，服务器立即移到下一个队列。如果队列错过了它的传输时序就只能等到下一个属于它的时序才能传输。如果每个队列都在使用，那么该队列的数据包要等到所有的队列都处理完之后才能被处理。WFQ不受这种影响，而且比DRR更适合于处理变长数据包。

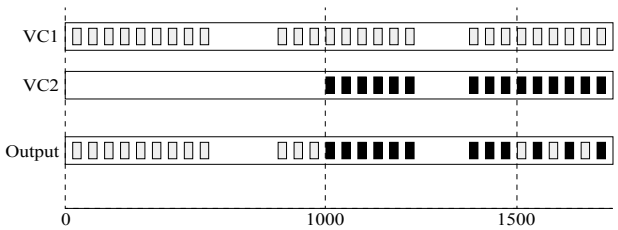


图3 虚拟时钟算法（VC）的局部不公平性

虚拟时钟算法（VC）可以为数据流提供带宽和平均时延保证，但是，如果当系统处于轻负荷状态时，某个数据流发送了大量的突发分组，那么当其他数据流被激活以后，这个数据流会受到惩罚（称为局部不公平性，如图3所示），WFQ算法不会产生局部不公平性，这样数据流就可以充分利用网络轻载时的系统资源。

Stop-and-Go队列采用基于全网范围的时隙划分结构，有两个优点：可以提供好的时延抖动控制和易于实现。但是Stop-and-Go队列没有WFQ调度算法灵活，而且它属于非工作保留调度算法，无法充分利用服务器的资源，而WFQ算法属于工作保留调度算法，所以WFQ算法能够比Stop-and-Go队列提供好的多的平均时延控制。

总结

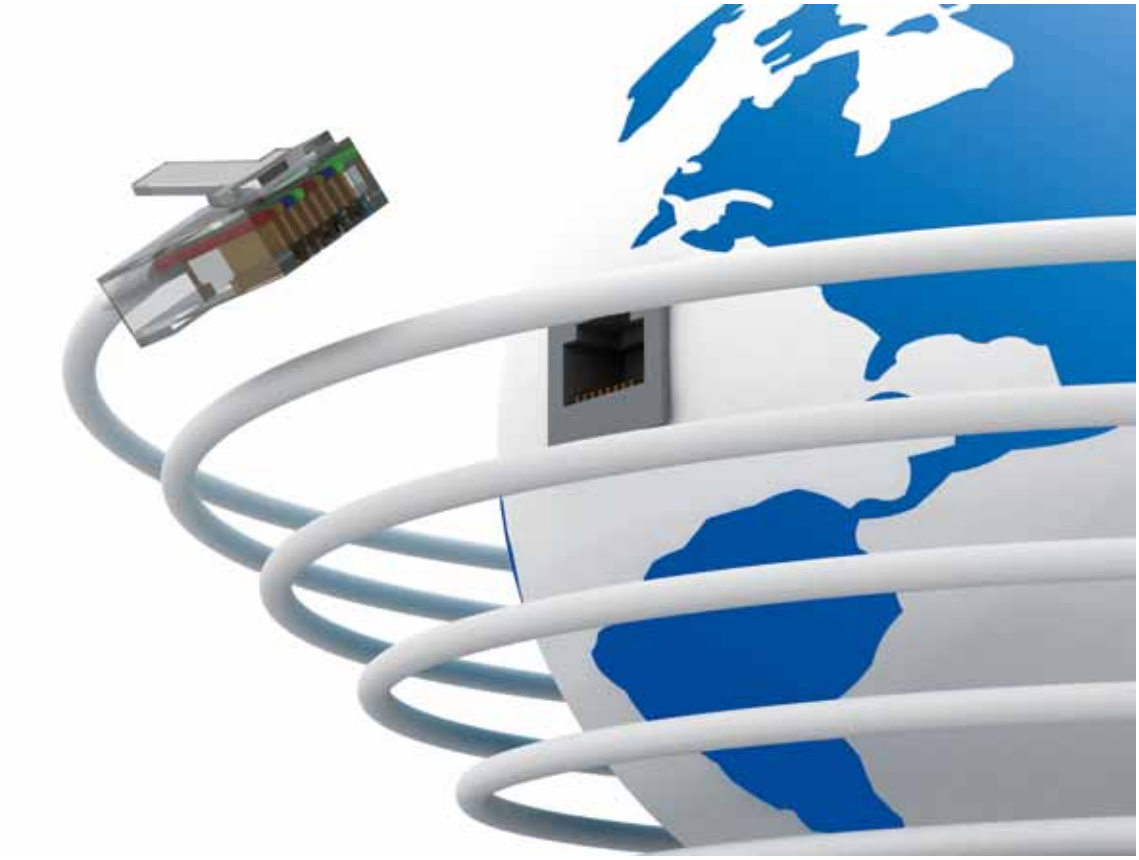
为了平衡影响调度算法设计的各种因素，获得比较好的性价比，调度算法的设计必须根据每类业务的特点选择不同的算法，但是任何调度算法都必须具有以下基本特征：

- 调度算法必须能够隔离不同数据流之间的相互影响，也就是说，即使是在有恶意数据流存在的情况下，调度算法也能够提供最基本的QoS保证；
- 调度算法必须在不过分影响网络资源利用率的前提下，为单个数据流提供端到端时延保证。同时，调度算法还应该具有能够通过只控制保留带宽资源来控制数据流的时延上下限的能力；
- 调度算法必须能够有效地利用共享链路带宽，这意味着调度算法必须充分利用数据流的统计复用特性，有效地处理突发信源；
- 共享链路带宽必须在队列中加权公平分配，不公平的调度算法在短时间间隔内可能会为两个具有相同保留带宽的数据流提供差别很大的服务速率；
- 实现的简单性；
- 可扩展性：调度算法必须适用于巨大的数据流和变化范围很大的链路速率。

网络的发展趋势是业务多样化，在实际网络中，彼此有分层关系的业务流常常共享链路，这时单一的调度算法无法满足链路带宽共享的需求，需要考虑使用综合结构的调度算法。也就是说，在追求简单性和易实现性的同时，考虑算法的综合性能。

MPLS QoS实现介绍

文/陶豆



MPLS，即多协议标签交换（Multiprotocol Label Switching），它使用标签转发替代了传统的路由转发，路由功能强大、灵活，可以满足各种新应用对网络的要求，而且其核心技术可扩展到多种网络协议（IPv6、IPX等）。目前这种技术被广泛地应用于大规模网络的组建，在MPLS网络中实现服务质量（QoS）也就成为必须考虑的问题。

MPLS QoS实现介绍

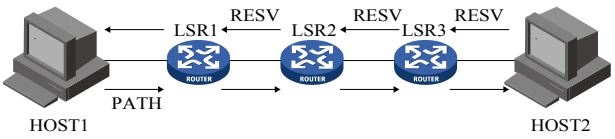
对于网络业务来说，服务质量（QoS）包括传输的带宽、传送的时延、数据的丢包率等，根据网络对应用的控制能力的不同，可以把网络QoS能力分为以下三种等级：尽力而为的服务、区分服务、保证服务。

保证服务（IntServ）

保证服务是通过信令（signal）来完成的，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，应用程序一般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，同时应用程

序发出的报文应该控制在流量参数描述的范围以内。负责完成保证服务的信令为RSVP（Resource Reservation Protocol，资源预留协议），它通知路由器应用程序的QoS需求。

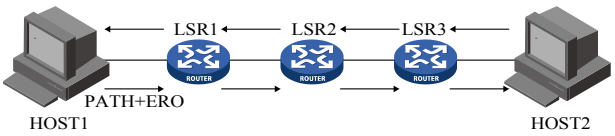
在MPLS中，InteServ的实现过程也是类似的，下面详细说明：



在上图的环境中，LSR1、LSR2、LSR3之间为MPLS网络。如果LSR1想建立一条到LSR3的预留路径，它就会经过LSR2向LSR3发送一个RSVP的PATH消息。LSR3收到这个RSVP PATH消息后，就从它的标签池中分配一个标签（7），并向LSR2发出一条RESV消息，消息携带分配的标签（7）。同时LSR3在自己的LFIB中指定该标签（7）为输入标签。LSR2收到这个RESV消息后，在LFIB中将该标签（7）作为输出标签，同时它还会分配一个新标签（3）作为输入标签，并将该标签（3）随RESV消息发送给LSR1。如此这样，随着各节点对携带标签的RESV消息的处理，沿着RESV路由建立起一条LSP，每台LSR可以把QoS资源和LSP建立关联。

当LSR2从LSR1收到一个具有标签3的数据包时，它可以在LFIB中查询标签，并应用与这个LSP关联的所有QoS机制，而不需要检查IP或传送层包头。

另外，还有一种方式实现保证服务，这种方式需要在MPLS QoS网络入口处进行约束计算，然后发起携带显示路由的资源预留申请。如下图所示：



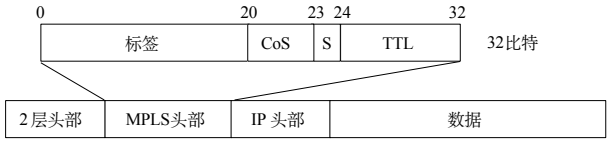
如果希望沿着从LSR1到LSR3的路径建立起一个预留，LSR1就查询它的链路状态数据库，并在向节点LSR3发送一条PATH消息之前，选择一条到达LSR3的路径。这条路径将需要满足穿越所有链路的带宽需求限制，以支持这个预留，而且需要在中间节点上具有足够的缓冲空间，以适应预留数据流的突发。在获得这条路径后，LSR1就把一个显示路由对象插入到这条PATH消息中，确保LSP将会沿着选择的路径建立。这样的LSP被成为“保证带宽的LSP”。

区分服务（DifferServ）

IP QoS实现区分服务的方法是:在网络的边缘进行流量分类，将流量划分为多个优先级或多个服务类，如使用IP报文头的ToS（Type of service，服务类型）字段的前三位（即IP优先级）来标记报文，可以将报文最多分成2^3=8类；若使用DSCP（Differentiated Services Codepoint，区分服务编码点，ToS域的前6位），则最多可分成2^6=64类。在报文发送途径中的每一跳（HOP）通过检查报文的DSCP或ToS字段，确定该数据包要求的QoS，这种行为称为“每跳行为（Per-Hop Behavior，PHB）”。

而在MPLS网络中，依靠标签转发的路由器可能不会检查IP报文头的内容，也就无法通过ToS或DSCP字段来进行流量分类。

在MPLS报文头中，有一个3bit的扩展（exp）字段，这一字段被用来承载DiffServ信息。



因此在MPLS网络中，流量分类主要通过如下两种方案实行：

第一种解决方案适用于少于8个PHB的网络，映射直接进行：特定的DSCP对应特定的EXP组合，映射到特定的PHB中。转发时报文根据标签转发，而由EXP位决定PHB。EXP可以从承载在LSP中的IP数据包的DSCP或ToS拷贝，也可以由MPLS网络运营商设置。我们称这种从EXP位推导得出PHB的LSP为E-LSP。

第二种解决方案适用于支持8个以上PHB的网络。在此EXP位不能单独承载所有的必要信息区分不同的PHB。MPLS报头中可用于该目的的唯一一个字段只有标签。在转发期间，标签决定数据包转发路径，并为其分配调度行为；而EXP位则携带分配给数据包的丢弃优先级信息，因此PHB由标签和EXP位来决定。由于标签与逐跳行为之间存在联系，因此当建立LSP信令时需要传输此类信息，这

种使用标签来传输所需PHB信息的LSP被称为L-LSP。L-LSP可传输来自单一PHB的数据包，也可以传输采用相同调度政策．但丢弃优先级不同的多个PHB的数据包。

按照这样的方法实现了流量分类后，流量整形、流量监管、拥塞避免等等QoS实现方法就和IP网络中的完全相同了。

区分服务的隧道化模式

MPLS网络实质上是为其承载的业务提供了一种隧道化服务，在RFC3270中，定义了三种MPLS区分服务的隧道化模式。第一种为管道模型（Pipe Model）。在该模式中，MPLS隧道的区分服务信息（即LSP区分服务信息，比如MPLS标签的EXP字段）与业务数据的区分服务信息（称为隧道化的区分服务信息，比如IP头的优先级字段）的操作是完全独立的。因此，业务数据的信息可以透明地从MPLS网络的一个站点传送到另一个站点。

具体说来，管道模式以如下方式决定QoS行为并处理区分服务信息。

在入站LSR上，LSP区分服务信息可以手工设置，也可以从隧道化的区分服务信息中获得。

在转发路径中的LSR上，出站标签的LSP区分服务信息来自入站标签的LSP区分信息。

在出站LSR上，依据LSP区分服务信息决定报文转发的QoS行为，但LSP区分服务信息不会复制到隧道化的区分服务信息中去。

第二种模型是管道模型的一个变种，称为短管道模型（Short Pipe Model）。它与管道模型的区别在于：在出站LSR上，决定报文转发的QoS行为的依据是隧道化区分服务信息，而不是LSP区分服务信息。当然，与管道模型一样，LSP区分服务信息也不会复制到隧道化的区分服务信息中去。

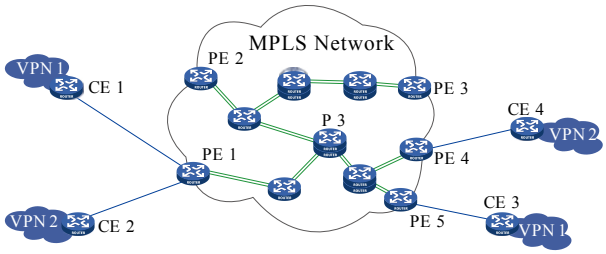
RFC3270描述的最后一种模型叫做统一模型（Uniform Model）。所谓统一是指LSP区分服务信息与隧道化区分服务信息一致，即最外层标签也包含了业务数据的QoS信息。这种模型的具体操作规则为：

在入站LSR上，隧道化区分服务信息必须要复制到LSP区分服务信息中去。

在转发路径中的LSR上，出站标签的LSP区分服务信息来自入站标签的LSP区分信息。

在出站LSR上，LSP区分服务信息必须要复制到隧道化区分服务信息中去。

在RFC3270中规定，管道模型是必须支持的模型。事实上，目前，主流厂商都通过QoS命令行的配置方式来支持三种模型的操作，可以看出，关键是通过控制LSP区分服务信息和隧道化区分服务的信息之间的映射，以在MPLS网络环境中实施与纯IP环境相类似的QoS手段。4MPLS QoS实现举例：



在上面的MPLS VPN网络中：
用户从一个CE 2站点向另外一个CE 4站点发送IP数据流。数据流被发送到了网络边缘的PE路由器后，根据IP报文头中的ToS或DSCP字段，数据流被分类并实现了IP网络中的QoS，如CAR等。然后报文被打上MPLS标签，标签中的exp字段可以从IP报文头中的ToS字段拷贝，也可以按照MPLS网络策略的配置，根据数据包的类型，输入接口等其他因素而被重新赋值。然后打标签的数据报文将被发送到MPLS网络中，网络中的P路由器可以根据MPLS报文头的exp字段对报文提供WFQ和WRED等QoS服务。而报文到达对端的PE 4后，MPLS标签弹出，该报文在IP网络中再次根据IP报文头中的ToS（DSCP）分类实现IP QoS。

MPLS TE

文/邹旭东

MPLS和流量工程简介

MPLS（Multiprotocol Label Switching）是多协议标签交换的简称，它用短而定长的标签来封装网络层分组。MPLS从各种链路层（如PPP、ATM、帧中继、以太网等）得到链路层服务，又为网络层提供面向连接的服务。MPLS能从IP路由协议和控制协议中得到支持，同时，还支持基于策略的约束路由，它路由功能强大、灵活，可以满足各种新应用对网络的要求。这种技术起源于IPv4，但其核心技术可扩展到多种网络协议（IPv6、IPX等）。MPLS最初是为提高路由器的转发速度而提出的一个协议，但是，它的用途已不仅仅局限于此，而是广泛地应用于流量工程（TE－Traffic Engineering）、VPN、QoS等方面，从而日益成为大规模IP网络的重要标准。

基于MPLS的流程工程即MPLS TE，正在成为一种重要的QoS工具，能够提供网络流量管理、减少拥塞等功能。同时，MPLS快速重路由（Fast Re-Route）技术在提高MPLS网络可靠性中扮演了重要角色。这种技术借助MPLS流量工程（Traffic Engineer）的能力，为LSP提供快速保护倒换。MPLS快速重路由事先建立本地备份路径，保护LSP不会受链路/节点故障的影响。当故障发生时，检测到链路/节点故障的设备可以快速将业务从故障链路切换到备份路径上，从而减少数据丢失。快速响应、及时切换是MPLS快速重路由的特点，它可以保证业务数据的平滑过渡，不会导致业务中断；同时，LSP的头节点会尝试寻找新的路径来重新建立LSP，并将数据切换到新路径上，在新的LSP建立成功之前，业务数据会一直通过保护路径转发。

MPLS基本概念

转发等价类（FEC）

FEC（Forwarding Equivalence Class）是MPLS中的一个重要概念。MPLS实际上是一种分类转发技术，它将具有相同转发处理方式（目的地相同、使用转发路径相同、具有相同的服务等级等）的分组归为一类，称为转发等价类，属于相同转发等价类的分组在MPLS网络中将获得完全相同的处理。

标签

标签是一个长度固定、具有本地意义的短标识符，用于标识一个FEC。当分组到达MPLS网络边界时，入口路由器按一定规则划分分组所属的FEC，将对应的标签嵌入到分组头部。这样，MPLS在网络中，按标签进行分组转发即可。标签的结构如图1所示。

Label	Exp	S	TTL

图1 标签的结构

标签位于链路层包头和网络层分组之间，长度为4个字节。标签共有4个域：

- Label：标签值字段，长度为20bits，用于转发的指针；
- Exp：3bits，保留，协议中没有明确规定，通常用于CoS；
- S：1bit，MPLS支持标签的分层结构，即多重标签。值为1时表明为最底层标签；
- TTL：8bits，和IP分组中的TTL意义相同。

MPLS TE及其四个构件

传统的路由器选择最短的路径进行路由，不考虑带宽等因素，这样，即使某条路径发生拥塞，也不会将流量切换到其他的路径上。在网络流量比较小的情况下，问题不是很严重，但是随着Internet的应用越来越广泛，传统的最短路径优先路由的问题暴露无遗。

MPLS TE是一种将流量工程技术与MPLS模型相叠加结合的技术。通过MPLS TE，可以建立指定路径的LSP隧道，进行资源预留；并且可以进行定时优化，在资源紧张的情况下，依据优先级和抢占参数，抢占低优先级LSP隧道的带宽资源等。同时，MPLS TE还可以通过备份路径和快速重路由技术，在链路或节点失败的情况下，提供路由保护。

MPLS TE的实现需要四个部分：

- 网络信息搜集，通过OSPF/ISIS的TE扩展实现；
- 路径计算，通过CSPF来实现；
- 建立LSP的信令，采用RSVP TE或CR-LDP协议；
- MPLS转发。

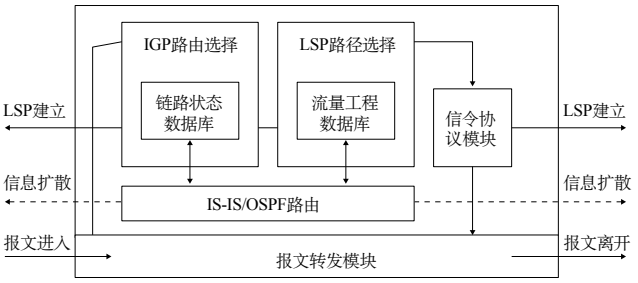


图2 MPLS TE的四个组件

报文转发组件

MPLS TE报文转发组件是基于标签的，通过标签交换沿着预先建立好的LSP进行报文转发。由于LSP隧道的路径可以指定，因而可以避免IGP的弊端。

信息发布组件

除了网络的拓扑信息外，流量工程还需要知道网络的动态负载信息。为此，引入信息发布组件，通过对现有的IGP进行扩展来发布链路状态信息，比如在IS-IS协议中引入新的TLV，或者在OSPF中引入新的LSA。具体来说这些链路状态信息主要包括：

- 本端IP地址
- 对端IP地址
- 链路的带宽
- 链路的最大可预留带宽
- 链路的着色
- 链路的8个优先级的未被预留带宽

其中最重要的是链路的最大可预留带宽和每个优先级的链路未被预留带宽，它们携带了链路的主要带宽信息。每个路由器都在本区域内发布自己的链路流量工程信息，同时学习其他路由器的信息，形成流量工程数据库TED。

路径选择组件

路径选择组件通过CSPF算法进行约束路由计算。约束路由计算的输入有两个：一个就是通过ISIS或OSPF流量工程扩展产生的TED。另一个是通过约束路由指定的数据转发路径，即在每个入口路由器上指定LSP隧道经过的路径。这种约束路由可以是严格的，也可以是松散的；可以指定必须经过某个路由器，或者不经过某个路由器；可以逐跳指定，也可以指定部分跳。此外，约束路由还可指定带宽、着色、抢占/保持优先级等约束条件。

有了以上两个输入，约束路由计算从逻辑上来说就是针对LSP的要求，对流量工程数据库中的链路进行裁剪：把不满足带宽要求的链路剪掉，把不满足颜色要求的链路也剪掉。然后，采用OSPF或ISIS的最短路径算法（SPF算法）在裁剪以后的拓扑中进行计算，得到一条满足LSP的约束条件的最短路径。这里值得注意的是传统ISIS或OSPF的SPF计算得出某个路由的下一跳就是直接的下一跳，每一个路由器都需要运行SPF算法。CSPF计算的结果是一条满足约束条件的完全约束路径，计算过程通常只在需要建立LSP的入口点进行。这条约束路径要起作用必须通过MPLS信令建立起LSP，MPLS信令把CSPF计算出来的严格约束路径通过信令中的约束路径扩展信息传到下游节点。LSP建立成功以后，入口路由器把需要进入这个LSP的IP包打上相应的MPLS标签，沿着LSP向下转发，直到到达LSP的出口。

信令组件

信令组件的作用是预留资源，建立LSP。LSP隧道的建立可以通过CR-LDP或RSVP-TE协议完成。这两种信令都可以支持LSP的建立、约束路由、资源信息携带等功能。

以RSVP-TE为例，为了能够建立LSP隧道，对RSVP协议进行扩展，在RSVP PATH消息中引入Label Request对象，支持发起标签请求；在RSVP RESV消息中引入Label对象支持标签分配，这样就可以建立LSP隧道了。为了支持约束路由，在RSVP RESV消息中又引入Explicit Route对象。

MPLS TE与QoS

人们希望IP网络能够提供高带宽、低延时的服务，这就要求在IP网上实现一定的QoS功能。由于IP网天生是一种面向无连接的网络，IP网中不可能实现和ATM网一样强大的QoS功能。在IP网中实现QoS的一个基本要求就是对现有网络结构改变尽量小。IP网中影响QoS的最主要原因是网络存在拥塞，IP包的网络延时主要来自于路由器调度时的排队延时。在一个没有拥塞的网络中QoS是可以得到很好的保证的，而流量工程的作用就是调配网络流量使得LSP的流量能够避开网络拥塞点，从而达到均衡网络流量，减少网络的拥塞的目的。从这个意义上讲，流量工程正是一种QoS机制！

另外采用MPLS流量工程可以支持快速重路由。在发现链路或路由器故障以后，通过硬件直接切换链路，可以做到从链路故障到快速重路由切换成功的延时小于50ms。

RSVP协议简介

文/张志飞

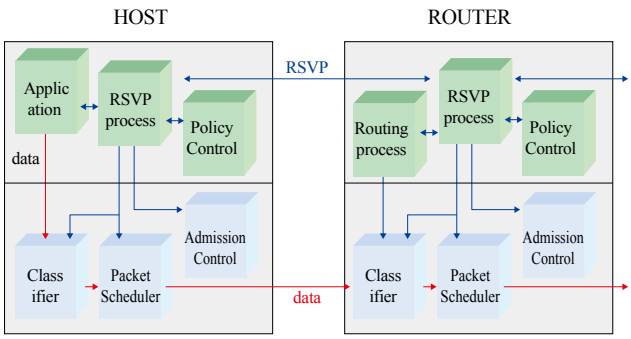
互联网是当今应用最广泛、发展最迅速的IP数据分组交换通信网络。基于IP的数据、音频和视频等业务以其低廉的费用、随处可接入性等优点越来越来引人注目，在电信业务中所占的比重也越来越大。网络的发展推动业务的发展，基于IP的多媒体通信异军突起，发展势头极为迅猛，随着多媒体技术的成熟以及计算能力的提高，已经能够在互联网上提供WWW浏览、IP电话、视频点播、视频会议、远程教学等多媒体业务。

80年代中后期以来，在国际计算机网络研究领域广泛地开展了以支持实时多媒体通信传输为目标的新型网络体系结构。互联网工程任务组（IETF）在服务区分方面提出的第一个体系结构是集成服务体系结构，集成服务体系结构对传统互联网进行扩展以支持多媒体实时应用。它不仅可以提供无服务性能要求的传统尽力传输服务模式，还可以提供支持完全服务性能保证的服务模式。在服务层次上，它提供端到端的质量保证型服务或可控负载型服务。典型应用如远程教学、视频点播等交互式音频和视频应用。在实现层次上，它需要所有路由器在控制路径上处理每个流的信令消息并维护每个流的路径状态和资源预留状态，在数据路径上执行流的分类、调度和缓冲区管理。具体而言，集成服务依靠资源预留协议（RSVP）逐节点地建立或拆除每个流的资源预留软状态；依靠接纳控制决定链路或网络节点是否有足够的资源满足QoS要求；依靠传输控制将IP包分成传输流，并根据每个流的状态对分组的传输实施QoS路由、传输调度等控制。

最早构思资源预留协议（RSVP）的是南加利弗尼亚大学（USC）信息科学院和施乐Palo Alto研究中心的研究人员。它提供了一种有效的资源预留方式，可以有效的描述应用程序对资源的需求。资源预留协议（RSVP）建立在IP协议之上，可以利用IP数据包传输RSVP消息；RSVP是一个单工协议，只在一个方向上预留资源；RSVP是一个面向客户端协议，由信宿负责资源预留；RSVP可以满足点到多点群通信中客户端异构的需求，每个客户端可以预订不同数量的资源，接收不同的数据流；RSVP还提供了动态适应成员关系变化、路由变化的能力。为了建立并维护分组数据传输通道中各个交换机的状态，RSVP建立了异构信宿树。简而言之，

RSVP协议就是通过在中间结点传输预留信息以创建和维护预留状态，从而实现资源预留和释放。

RSVP协议基本架构包含决策控制（Policy）、接纳控制（Admission）、分类控制器（Classifier）、分组调度器（Scheduler）与RSVP处理模块等几个主要成分。决策控制用来判断用户是否拥有资源预留的许可权；接纳控制则用来判断可用资源是否满足应用的需求，主要用来减少网络负荷；分类控制器用来决定数据分组的通信服务等级，主要用来实现分组过滤；分组调度器则根据服务等级进行优先级排序，主要用来实现资源配置以满足特定的QoS。当决策控制或接纳控制未能获得许可时，RSVP处理模块将产生预留错误消息并传送给收发端点；否则将由RSVP处理模块设定分类与调度控制器所需的通信服务质量参数。



RSVP in Hosts and Routers

RSVP协议基本架构图

流（Flow）是以单播或多播方式在信源和信宿间传输的数据码流，它为不同服务提供类似连接的逻辑通道。在RSVP协议中，发送端点简单地以多播方式传送数据；接收端点如欲接收数据，将由网络路由协议系统（IGMP协议等）负责形成在源宿间转发数据的路由，也就是由路由协议配合形成数据码流。流在RSVP协议中占有至关重要的位置，RSVP协议的所有操作几乎都是围绕流而进行的。

RSVP支持四种基本的消息：资源预留请求消息、路径消息、错误和确认消息、拆链消息。

资源预留请求消息（Reservation-Request Messages）：一个资源预留请求消息由接收方主机向发送方主机发送。资源预留请求消息使用同数据报路由方向相反的方向传送，直至到达发送方主机。一个资源预留请求消息必须到达发送方主机，只有这样，发送方才能为传输的第一跳设置合适的控制参数。

路径消息（Path Message）：一个路径消息由发送方通过单播或组播路由向外发送。路径消息用于存储每个结点的路径状态（PS）。资源预留请求正是通过这些路径状态才能从相反方向回到发送方的。

错误和确认消息（Error and Confirmation Messages）：错误消息有两种类型：PathErr和ResvErr。PathErr由路径消息引起，并传送到发送者。ResvErr消息由预留消息引起，并传送到相关的接收者。

拆链消息（Teardown Messages）：RSVP拆链用于超时之前删除路径和预留状态。拆链消息有两种类型：PathTear和ResvTear。PathTear删除从消息发出的节点到所有的接收者路径上的预约状态，PathTear的路由和路径消息的路由严格一致。ResvTear删除从消息发出的节点到所有发送者路径上的预约状态，ResvTear的路由和预留消息的路由严格一致。ResvTear消息可以由一个接收者，或一个状态超时或预约被剥夺的节点产生。节点上状态的删除可能会引起本节点相关预约状态的更新。

RSVP协议的基本工作原理如下：数据流的源主机为将要发送的数据流做出一个规范的描述Tspec，包括传输数据流所需要占用带宽的上限和下限，时间延迟和延迟抖动。主机中的RSVP信令模块则向目的主机发送Path消息，其中包括Tspec信息。在源主机到目的主机的下行线路上的每一个支持RSVP的路由器在收到Path消息时都在内部建立起链路状态标识。为使下游节点了解流的来源，上游节点将Path消息中Lasthop（上级节点）域改写为该节点的IP地址，Resv消息正是利用Path消息中Lasthop的信息实现逐级向上游节点预留资源。

为建立起资源预留，目的端主机在上行线路上发送Resv消息，包括预留服务的种类及数据流描述符。当上行线路上的路由器收到Resv消息时，路由器上的许可控制器来验证是否有足够的资源来满足该请求，然后被送到策略控制器来判断用户是否有权预约资源。如果两个验证都成功，则分配给该请求资源，并把请求送给下一个节点，否则返回错误给提出请求的应用程序即发送错误信息至目的端主机；反之则向上行线路的下一跳路由器发送Resv消息。

当上行线路上的最后一个路由器资源预留成功时，则向目的端主机发送确认信息。结束RSVP控制的基本工作原理与建立RSVP控制的基本工作原理类似。

语音质量与语音质量的测量

文/刘先楠

声音是日常生活中最常见的信息传递方式，人们通过声音彼此交流联系，人的声音的频率范围是20Hz到20KHz，我们通常说话时的频率大部分都集中在300～3300Hz的范围内。早期的电话采用炭精麦克风、电池和磁性耳机组成，人的声波通过空气压迫炭精电阻器的膜片，可以使炭精的电阻发生变化，使得流过炭精的电流发生变化，这样就把声音信号转换为强弱不同的电流信号，强弱不同的电流通过电话线到达对端，耳机接收到大小不同的电流，引起耳机膜片进行不同频率的震动，这种震动通过空气反馈到人耳。这也是我们平常使用的普通电话的工作原理。随着IP无处不在，越来越多的话音以IP为承载，VoIP不再是仅仅为了通过广域IP链路承载来节省长途通话费，它现已发展成可以提供更多更丰富的话音业务，成为运营商的获利方式，企业降低通讯成本与增强协同工作的工具。但是包交换网络与生俱来的特点决定了VoIP的话音质量面临比PSTN更多的问题，本文简要介绍了在VoIP网络中话音质量的定义和测量。下文中所提及的语音质量均是指在IP网络中的语音质量。

语音质量

语音质量对于大多数情况来说是一个主观因素，对于清晰、低时延、低抖动的话音我们认为其质量是可以接受的。比如我们通常的对外测试，请客户听一听话音的质量，客户的评价就是语音质量的主观评价。当然ITU也有可以量化的语音质量评判标准，这在后文中会有说明。无论是主观的评判还是客观的量化的指标，清晰、低时延、低抖动都是高质量话音的保证。那么为了得到清晰、低时延、低抖动的话音，我们如何从设备侧和网络侧来保证上述的要素呢？那我们先看看影响语音质量的因素有哪些？

影响语音质量的因素

时延

什么是时延（Delay）？时延很好理解，就是我说话你听见的这个时间段。对于VoIP网络可以认为是端到端包传递的时间。如果大家以前用卫星线路打过电话或前些年的免费PC-TO-PHONE的VoIP国际长途电话，大家会切身感觉到时延对我们通话的影响。ITU G.114规范建议，在传输语音流量时，单向语音包端到端延迟要低于150ms（对于国际长途呼叫，特别是卫星传输时，可接受的单向延迟为300ms。如果超过300ms则通话的质量会变的让人不能忍受。过多的包延迟可以引起通话声音不清晰、不连贯或破碎。例如，当通话的一方不能及时接收到期望的回复时，说话者可能会重复所说的话，这样会与远端延迟的回复碰撞，导致重复。大的时延也往往说明承载网络的某个地方发生了拥塞，队列中的报文等待时间过长，拥塞不仅仅是增加了包的时延，而很可能导致部分包被丢弃，这时听者会感觉到声音会发生异变、破碎。大多数用户察觉不到小于100毫秒的延迟，当延迟在100毫秒和300毫秒之间时，说话者可以察觉到对方回复的轻微停顿。这种停顿可以影响到通话双方的交流。超过300毫秒，延迟就会很明显，用户开始互相等待对方的回复，通话过程变成类似对讲机式的模式。而且较长的时延也会将回声问题的影响放大。

时延的产生有多种因素，下面列出了主要的时延源：

编码的处理：模拟形式的声音信号在CODEC被采样和量化为PCM信号，DSP对PCM信号进行压缩处理所产生的时延为编码处理时延。这种时延产生在设备侧，如果设备的编码器固定，则编码时延也固定。

各编码方式时延：

编码技术	比特率（K bps）	MOS值	编码时延（ms）
G.711	64	4.4	0.75
G.723.1（5.3K）	5.3	3.6	30
G.723.1（6.3K）	6.3	3.4	30
G.729	8	4.0	10

包化（Packetization）：包化就是将编码器输出的语音净荷放置到RTP/UDP/IP包中的过程，相对于编码的时延，包化的时延很小，因为包化的过程没有复杂的运算，仅仅是增加包头和计算校验和，而编码则有大量的数学运算。

队列（Queuing）：语音的净荷放置到IP包中后，要被设备转发到目的地，这些包会在设备的出接口队列中，等待被调度。转发设备不同的队列机制对IP包的处理有很大不同。可以通过合理的配置来减少语音包在队列中等待的时间，进而减少队列时延。

串行化（Serialization）：接口队列中的语音IP包，被送离设备前会放置到接口的物理队列当中，如果物理队列中有一个较大分组，还在发送状态，则语音分组必须等待这个较大的分组发送完毕后才能发送，这个等待的时间就是串行化时延。比如一个时钟速率为64kbps的链路要发送一个1600Bytes大小的FTP分组，则串行化产生的时延会达到200ms（1600×8/64000×1000）。这对于后面等待的语音包来说已经是很大的时延了。

广域网时延：对于ISP提供的广域网链路，对于用户来说只是一个黑盒子，除了上述的编码时延外，构成广域网链路的路由器交换机都会产生包化、队列、串行化的时延。而且到达同一目的的路径不同，其每个包的时延也不同，而这些时延对于用户来说是不可控的，当然我们在租用ISP的线路时，可以要求ISP提供符合时延要求的线路。

抖动

变化的时延被称作抖动（Jitter），抖动大多起源于网络中的队列或缓冲，尤其是在低速链路时。而且抖动的产生是随机的，比如

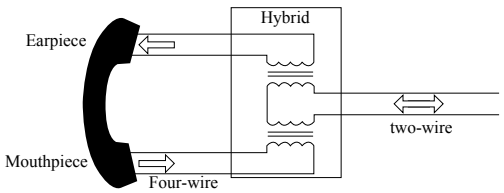
你无法预测在语音包前的数据包的大小，既便你使用LLQ（低延迟队列），如果大数据包正在传输过程中，当语音分组到达时，它还是要等待数据分组被发送完。而在低速的链路中，语音数据混传时，抖动是不可避免的。通常使用LFI（链路分段和交叉）将大包拆小，来减少大包对时延的影响。

丢包率

丢包（Packet Loss）是影响语音质量的重要的因素，当丢包超过一定的比率时，语音质量会变的不可接收，听者会听到，含混、爆破似的声音，在VoIP通话时，接收方的编解码器能够接受一定程度的丢包律，一旦检测到有分组丢失，接收方的编码器就会对丢失时间内的波形进行推算。绝大多数的编解码器都能接受随机5%的丢包率，而不会明显影响通话的质量。这里说随机是因为如果这5%的包是连续丢失的，也会对语音质量造成很大问题。

回声

回声（Echo）一般分为说话者的回声和倾听者的回声。说话者的回声就是在通话过程中，说话者听到了自己的声音。倾听者的回声就是倾听者重复听到说话者的声音。回声的产生的原因一般分为两种，一种是电信号回声，一种是声学回声。电信号回声产生于电话网络的模拟部分。hybrid是4线变2线的混合器，大部分的电信号回声是由于hybrid的阻抗不匹配引起。声学回声是指，声音从扩音器（如耳机、听筒）扩散到扬声器（话筒、麦克）。下图说明了hybrid的作用。



4-Wire to 2-Wire Converter-Hybrid

在大多数的PSTN环境中，回声是存在的，但是回声的产生时间是如此接近正在讲话的人的声音，以至于给人的感觉,这只是简单适度的侧音，就像我们平常说话时，我们通过颚骨听到自己的声音，这样的侧音对我们来说是习惯的。但当回声通过网络以过大的延迟返回时，回声不利的影响就会被说话者察觉，语音质量就变得有问题了。注意回声虽然只能在网络的模拟部分产生，但是IP网络累积的延迟能够导致回声从合适的侧音变成令人困扰的语音质量问题。为了消除回声干扰，可以在尽可能靠近回声源的地方部署被称为回声抑制器的装置。回音抑制（Echo Cancellation Algorithm）其定义由ITU-T G.168给出。回音抑制的功能是用相位补偿的方法抵消串入远端发送信号中的远端接收信号。其目标是消除时延超过25毫秒的回声，因为当回音超过25毫秒时，说话方就能够听到反射回来、滞后的自己的声音。

沿切割

丢失说话人的第一个或最后一个音节或单词的现象叫沿切割（Clipping），沿切割由延迟和IP网络中的静音抑制机制的使用而引起。人们为了更有效地使用网络带宽，会经常使用静音检测机制，例如静音检测（VAD）或者舒适噪音生成器（CNG）。静音检测器（VAD=Voice Activation Detector）该器件在信号电平低于某一特定的门限值时，将限制数据包的传输，此时其提供空闲（Idle）或者“舒适噪音”(Comfort Noise) 以避免电话用户感到“断线”（Dead Air）。网关中相应部件对从静音～讲话和从讲话～静音之间的过渡状态的响应速度如果过慢，就会丢失开始或最后的音节。VAD是产生“前沿切割”（“Leading Edge Clipping”）和“后沿切割”（“Trailing Edge Clipping”）的主要因素。有时这对于商务通话来说是不可接收的，开始和结束时礼貌的问候如果都没有让对方听到，会给对方不好的印象。

许多影响语音质量的因素不是单一产生的，而时常常伴随出现，比如说时延，抖动，和沿切割。

编解码

不同编解码（Codec）对语音质量的影响，下表为不同编解码的MOS值：

ITU Standard	Coding Method Name	MOS
G.711	PCM	4.4
G.723.1	H.323	3.98 to 3.5
G.726	ADCPM	4.2
G.728	CELP	4.2
G.729	ACELP	4.2

MOS值越高，代表语音质量越好，当带宽允许时，选用MOS较高的编解码可以获得较好的通话质量。

使用QoS机制来保障语音质量

QoS的实现需要两个前提，一是流量所经过的所有设备支持QoS机制并进行了合理的配置，二是只有在线路发生拥塞时QoS机制才起作用。线路不拥塞时，QoS是不生效的。提高语音数据的优先级，保证语音数据的带宽是保证语音质量的前提。

在VoIP网络中，为了保证语音包优先我们经常使用的QoS机制有CBQ、RTPQ，CBQ保证语音信令流的带宽和优先级，CBQ也可以保证实时的RTP流的带宽和优先级。RTPQ保证实时语音流的带宽和优先级。

在CBQ的配置中，用ACL精确的定义了数据报文的源端口号目的端口号，屏蔽掉了那些源端口号匹配但是目的端口不匹配的TCP报文，弥补了PQ的不足。同时，对于有些应用如3Com NBX的语音流使用UDP端口号从2093～2096的报文，CBQ都能够匹配上从而进入设置好的队列，不会出现某些端口不能匹配从而报文入不了队列中的情况，这也弥补了RTPQ的一些缺陷，因此也保证了语音质量。

RTPQ进行匹配时，匹配的目的端口号都是偶数，因此配置的RTPQ只对偶数的目的端口号的UDP报文生效。在有的应用中，如3Com NBX的语音流使用UDP端口号从2093～2096的报文，因此配置RTPQ的时候，会导致那些端口为2093，2095的报文不进RTP队列，得不到优先发送。既然RTPQ只对偶数的端口号匹配，奇数端口号的报文不进队列，为何大多数情况都用RTPQ保证语音业务呢？因为通常VoIP使用的呼叫信令是H.323或SIP协议，语音流使用RTP承载，其UDP端口号范围为16384～32768。然而在实际中，用到的UDP端口号是从16384开始以每4位增加，也就是说所用到的UDP端口号都是偶数（尤其16388被用到的非常多），不会出现奇数。这就是为什么做VoIP的QoS时用RTPQ保证语音质量的原因了。

当然，接口启用了qmtoken也是减小语音延迟的原因之一，qmtoken提供了一种底层队列的流量控制机制，它可以根据令牌的数量控制向底层接口队列发送的报文数量。对于语音应用，我们一般在端口上设置qmtoken为1。

语音质量的测量

语音质量测量的目的是通过主观或客观的测量方法，即通过人为的测量项或基于软硬件的测量工具，对一种或多种以上的呼叫质量类别给出一个可信的估计。

主观质量测量

随着IP电话技术的发展，人们不断寻求语音质量的测试方法，以便能规范IP电话设备的技术标准。ITU-T建议P.830描述了一种对语音的主观评定方法-MOS（Mean Opinion Score）方法。根据P.830建议的要求，特定的发话者与听话者在特定的环境下，通过收集测试者在各种不同情景下的主观感受，根据P.830的分析法则得出该语音的品质。P.830对测试的要求非常严格，所有的操作都要严格地服从操作流程，对录音系统、语音采样、语音输入级别、听者级别、不同发话者（8男、8女、8儿童，至少在16人以上）、多发话者（多人同时讲话）、差错处理、不同语音编码方式的兼容性、过失、环境噪音、音乐等等，都做出了详细严格的规定。测试者的主观感受结果也被分为很多不同的范畴，如听者感受的网络质量结果、质量降级结果、比较结果等。做出语音质量的判别标准：5：最佳；4：好（4.5～4.0＝可收费电信级）；3：中级（4.0～3.5＝可通话通信级）；2：较差（3.5～2.5＝可建立连接级）；1：差。很显然，MOS方法是一种模糊的评估方法，但是它能体现出使用者对IP语音质量的最直观的判断。

客观质量测量

MOS方法是一种模糊的评估方法，其测试结果很难对VoIP系统的改进和不同VoIP设备之间性能的比较，做出有实际意义的判别。因此，有人提出借用ITU-T在P.861中建议的PSQM（Perceptual Speech Quality Measurement，ITU-T P.861）方法，用来作为客观质量度量的评估。

ITU开发了P.861（PSQM）和更新的P.862（PESQ），用成本更低的客观测量法来作为主观收听质量测量的补充。采用这些测量技术，可以通过比较送入系统中的一个原始参考文件与输出的受损文件之间的差异，得到由传输系统或CODEC引入的失真。这些技术的初衷是为了CODEC的实验室测量，但在VoIP网络测量中也得到了广泛使用。

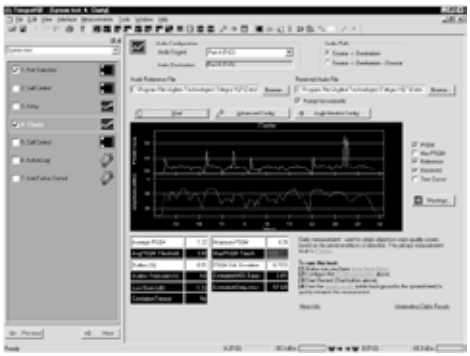
P.861和P.862算法将参考信号和受损信号都分成较短的交迭样本块，计算每一块的傅立叶变换系数，并比较他们的系数。P.862算法最后给出一个PESQ得分，该得分与MOS的范围相近，但它并不是MOS的准确映射。新的PESQ-LQ得分更接近收听质量MOS。这些算法都同时要求访问源文件和输出文件才能测量后者相对于前者的失真。

PESQ（Perceptual Evaluation of Speech Quality，ITU-T P.862）是评价各类端对端网络条件和语音编码与解码的较新的标准。PESQ可以根据一些感知标准来客观地评价语音信号的质量，从而提供可以完全量化的语音质量衡量方法，而这些衡量标准又是与人类对语音质量的感受完全吻合的。PESQ由荷兰的KPN公司和英国电信公司协作开发的，比其前身PSQM有了较大的进展。

2004年，ITU制定了P.563标准，这是一个单端客观测量算法，能够只对接收到的音频流进行操作。P.563测量得到的MOS得分比P.862更广，要使结果更稳定，必须多次测量并对结果进行平均。这一方法并不适合测量个别呼叫，但在测量多个呼叫的服务质量时，能够得到可信的测量结果。

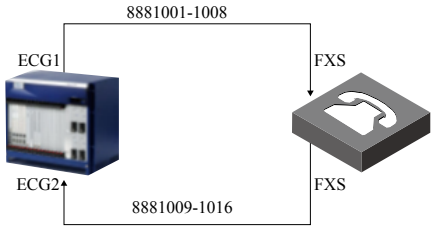
语音测试仪器

常用的语音质量测试仪器有Agilent公司的Telgra VQT，VQT提供FXO和E&M接口可以很方便的接入到被测设备，VQT可以完成End-to-End delay、Voice Signal Clarity、Voice Activity Detector measurement的测试，可以提供标准的PSQM测量。下图是VQT的Application的应用界面，测试过程中的数据可以实时的显示在图形化的界面。

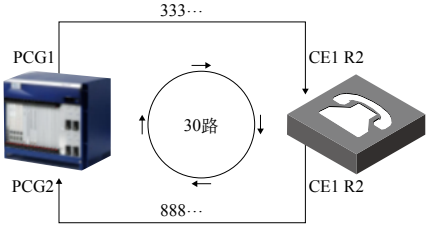


Abacus 5000是Spirent公司的语音测试产品，其接口类型丰富、容量大。它可以模拟用户侧、交换侧以及模拟话音交换。主要是用于语音产品的性能测试，同时可以提供多种标准的语音质量测试，比如PSQM、PSQM+、PESQ等。下面的两个典型的场景可以用来测试FXS口和EIVI接口设备的语音质量。具体的测试方法可以参考文献《Abacus5000配置初步》（张宇翔）。

下图是对于单设备环境下使用Abacus5000进行测试的环境。Abacus5000使用ECG来模拟普通电话，向被测试的语音网关发起呼叫，和被测试语音网关上pots实体匹配后，连通和Abacus5000连接的另外8路模拟语音用户线，完成一个通话的过程。

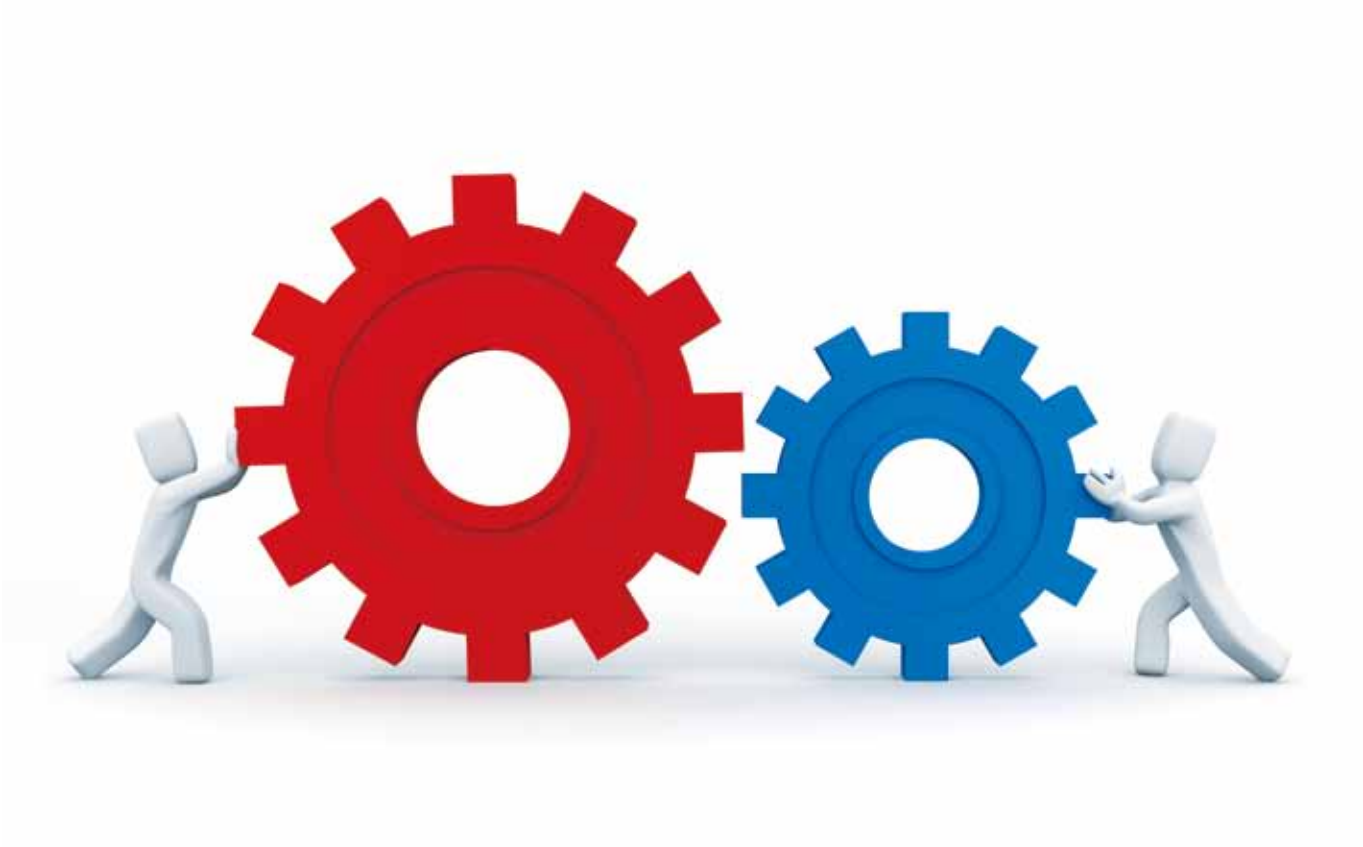


下图是使用PCG卡来模拟中继线动作，用Abacus的一块PCG3卡上的两个CE1接口，分别模拟30路R2信令中继呼出和30路R2信令中继呼入。呼叫由入中继进入被测试设备，从另外一条中继线呼出到Abacus，从而形成一个中继呼叫的环路。



QoS技术应用实例

文/蔡金龙



QoS技术概述

QoS（Quality Of Service）技术顾名思义就是对各种服务提供传输质量保证的技术。任何能够对传输质量进行保证的技术我们都可以称之为QoS技术。当然，任何应用能正常传输的基础都是有足够的网络带宽，如果传输带宽不能满足，再优秀的QoS技术也不能对所有的应用做出保障。所以最为实用的QoS技术是扩充现有线路的传输带宽，当然这个目标的实现有较大的困难，因此，目前我们的网络通信的服务质量还是需要有一定的QoS机制来对重要的应用作出保障。我们通常所指的QoS技术主要有如下几种模型。

尽力而为的服务（Best-effort Service）模型

路由器及交换机等数据通信设备都是分组交换设备，转发过程中每个分组独立选择传输路径，采用的是统计复用的方式，不像时分复用（TDM）那样具有专用连接的概念。传统的IP只提供单一的服务类型——尽力而为的服务，在这种服务方式下，所有经过网络传输的分组具有相同的优先级。尽力而为意味着IP网络会尽一切可能将分组正确完整的送到目的地，但不能保证分组在传输过程中不发生丢弃、损坏、重复、失序及错送等现象。另外也不对分组传输质量相关的传输特性（如传输时延、时延抖动等）作出任何承诺。

严格意义上说尽力而为的服务并不能归类为QoS技术，但这是目前整个Internet所使用的主要服务模型，所以我们需要对它有一些了解。如果网络只通过尽力而为的方式为客户提供服务，网络通讯势必会出现很多问题，比如分组丢弃，还好TCP这种智能的传输层协议通过滑动窗口、报文重传等机制有效解决了报文丢弃的问题，才使得尽力而为服务能够基本上满足大部分业务的通讯需求。尽力而为并不是一个贬义词，正因为这种服务模型才使得Internet有今天的发展，当然随着Internet的发展，这种尽力而为的服务模型已经不能完全满足越来越广泛的应用需求，因此服务提供商们有必要在现有的尽力而为服务基础上提供多种服务类型，使得每一种服务类型能够满足特定的数据通讯要求。这些新的服务类型就是我们下面要介绍的保证服务模型和区分服务模型。

保证服务（IntServ又称之为综合服务）模型

保证服务模型是IETF于1993年开发的一种在IP网络上支持多种服务的机制，它的目标是在IP网络中同时支持实时服务和传统的尽力而为服务，它是一种基于为每个信息流预留资源的模型。

保证服务定义了一个参考模型，在该参考模型中指定了若干不同的构件及这些构件之间的相互作用。

- 资源预留协议（RSVP）允许单个应用向网络设备请求资源，并沿着分组传输路径，为每个单独的数据流设置状态；
- 定义了两个新的服务模型——保证服务和控制负载服务。保证服务通过严格的许可控制、带宽分配和公平队列服务，为有严格传输时延和带宽要求的应用提供所要求的服务；控制负载服务并不能保证传输的时延和带宽的范围，但提供类似轻负载下尽力而为网络所提供的服务；
- 提供用于描述流说明的语法，允许应用程序通过流说明指定特定的资源要求；
- 分组分类通过检测输入分组，确定适用于每一个分组的服务类型；
- 许可控制基于本地和网络的可用资源，确定是否支持所请求的资源预留；
- 管制和整形进程监测每一个信息流，强迫这些信息流符合其信息流描述；
- 分组调度进程将网络资源分配给不同的信息流；
- 保证服务模型要求源和目的主机通过交换RSVP信令消息，在源和目的主机之间传输路径上的每一个节点中建立分组分类和转发状态。

保证服务模型需要为每一个流维持一个转发状态，因此可扩展性较差，而且Internet上有上百万的流量，为每个流维护状态对设备消耗巨大，因此保证服务模型一直没有真正的投入使用。近来对RSVP进行了修改，使其支持资源预留合并，并可以和区分服务配合使用，特别是MPLS VPN技术的发展，使得RSVP又有了新的发展。但在QoS技术上，保证服务模型在实际应用中还是没有被广泛的应用。而区分服务模型恰恰解决了保证服务模型的弊端，成为目前使用最广泛的QoS技术。

区分服务（DiffServ）模型

由于保证服务模型存在着很大的扩展性弊端，1995年前后，服务提供商和各种研究机构开始提出新的支持多种服务的机制，而且这种机制的前提条件是具有良好的可扩展性。到了1997年，IETF已经认识到当前的网络服务模型已经不能适合当前的网络运行，应该有一种更便于分类信息流及更容易扩展的办法为特定的用户和应用提供区分服务，这样IETF组织成立了专门的工作组，开发了区分服务模型，这种模型可以相对简单更方便分类的方法为Internet信息流提供区分服务，并支持多种类型的应用和商业模式。区分服务模型的主要技术有：

- 流量分类与标记技术
- 拥塞管理技术（主要是各种队列技术）
- 拥塞避免技术
- 流量监管技术
- 流量整型技术
- 链路层QoS技术
- 链路效率机制

通过这些技术，可以灵活的对各种应用进行传输服务上的保证及区别对待，基本满足了目前各种类型应用的要求。

QoS技术应用背景

目前，我们的网络数据传输主要是使用基于分组交换的统计复用方式，默认情况下，使用的是尽力而为的服务模型，这种模型前面介绍过，它不能对分组的传输的相关质量做出任何承诺。随着语音、视讯等新兴应用的不断推出，用户对网络能够提供的传输服务质量提出了更高的要求。尽力而为的服务模型已远不能满足用户对网络传输质量的要求，因此，QoS技术被广泛的使用，用来在不能提供绝对充足的带宽环境中对重要的应用做出通信质量上的保证和承诺。能够影响通信质量的因素主要有以下几点：

吞吐量

吞吐量用于描述系统传输数据的能力。目前，各种网络设备及传输介质的吞吐量是一定的，除非更换新的能够提供更高传输带宽的接口及介质，否则任何的技术都不能够增加吞吐量，所以涉及到的一个问题是，如何在现有的吞吐量一定的链路上尽可能的满足应用的要求，并对重要应用做出通信质量的保证。

时延

时延是指分组从网络中某个点传输到网络中另一个点所需要的时间，有多个因素影响分组的传输时延，分别是：转发时延、排队时延、传播时延和串行发送时延。在这四个因素中，唯一可以人工控制的是排队时延，其他因素主要是设备和传输介质自身决定的，除非更换否则无法控制。QoS技术对时延的保证也仅是对排队时延的优化控制。

时延抖动

时延抖动是指属于同一类型信息流的不同分组其相同两个端点之间的传输时延发生变化的过程。像语音和视频的应用对时延抖动是比较敏感的，但时延抖动是由端到端链路决定的，通过QoS技术在端到端链路上实施可以尽量减少时延抖动。

分组丢失率

分组丢失往往是由三种原因导致：

- 物理链路中断导致无法传输分组；
- 分组在传输过程中损坏，下游节点通过校验码获知分组已经损坏，并丢弃该分组；
- 网络拥塞导致缓冲器溢出。

QoS技术可以对第三种情况作一定的控制，主要是通过各种队列技术，控制不同应用的传输质量。

综上所述，目前我们使用QoS技术，主要是为了对像语音、视讯等这类对时延、分组丢失十分敏感的应用提供传输质量的保证，同时保证其他各种网络应用按照其重要性获得相应的服务质量。

QoS技术应用实例分析

下面我们通过分析几个比较典型的组网应用，来看一下在实际应用中如何使用QoS保障各种业务的通信质量。

中小企业组网中QoS的典型应用

中小企业组网的主要特点是网络结构相对简单，各个分支节点通过运营商专线或者Internet VPN方式与中心节点互联。随着网络应用的发展，企业组网往往需要实现三网合一（语音、视频、数据），以最大程度的对现有IP网络进行利用并节约成本。由于所有的流量都从单一出口流出，所以在网络出口处必须做QoS策略对带宽、时延敏感的应用做出保障，下面我们来分析一下中小企业组网中QoS的典型部署情况。

专线用户的QoS部署

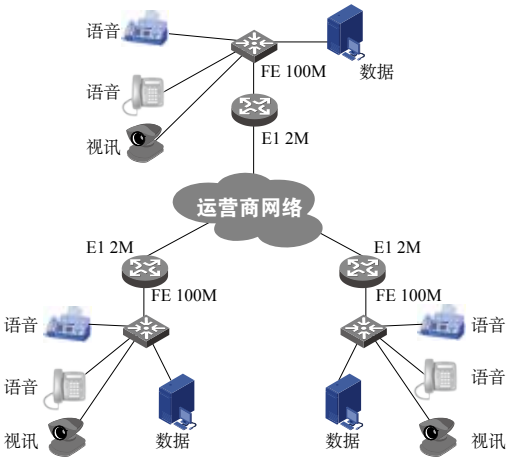


图1 专线用户企业组网图

组网如图1所示，中心节点通过专线连接各个分支节点，，全网运行语音、视频及传统的数据应用实现三网合一。网络的局域网带宽为100M远远大于广域网的出口带宽（2M），所以在广域网出口处会造成数据转发的瓶颈，因此网络各个场点都需要部署QoS来保证语音、视频应用的带宽和时延要求，并保证重要数据的传输带宽要求。在介绍如何部署QoS之前，我们还需要再强调一点，组网之前一定要考虑好网络所有应用所需要带宽的情况，QoS技术本身不能创造新的带宽，它只能在现有带宽的基础上保证特定业务的通信质量，所以如果申请的带宽本身就不能满足所有重要应用在正常网络状况下的要求，那么部署QoS也无济于事。我们假设申请的专线带宽可以满足所有网络应用的要求，那我们如何在现有的网络中通过QoS技术对关键应用做出保障，需要我们从端到端的角度进行部署。现有网络会对服务质量造成影响的因素主要有以下几个方面：

流量的分类

在部署QoS时，首先要考虑的问题是网络中到底在跑哪些流量，这些流量的重要性如何，他们对网络都有什么要求，这是部署任何QoS策略的基础。就像只跑FTP和HTTP的网络，与除了跑FTP这种带宽时延都不敏感的业务同时还跑语音、视频及数据库等流量对带宽及时延都有要求的应用的网络，对QoS的部署的要求是完全不同的。同时，任何QoS策略的部署都会对转发效率造成影响（特别是软件转发的路由器），因此对现有网络的流量进行分析并做好分类及标记，可以最优化的部署QoS，减少QoS对网络造成的转发效率上的影响。确定好需要保障的业务流量，打好标记，作为下一步部署QoS策略的基础。其他所有的流量只需要设备尽力而为的转发即可。对流量进行标记和分类的办法有很多，目前通常是使用五元组和优先级对流量进行分类。

运营商

运营商的网络情况一般对中小企业用户都是透明的，而且一般来说运营商也不会参与到中小企业网络的策略部署中，所以，在中小企业网络中只能假设运营商的网络都能够提供承诺的带宽值，企业网络数据的转发不会在运营商处出现瓶颈。

局域网交换机与路由器之间处理性能上的差异

交换机设备使用硬件芯片对数据进行线速转发，而路由器设备通常使用软件进行转发，往往达不到线速，所以在交换机和路由器之间会造成转发的瓶颈，交换机的线速流量造会成路由器CPU繁忙，导致路由器工作异常，所以在部署全网QoS策略时需要考虑这个问题。通常的解决办法是在交换机与路由器上连的接口上配置限速。具体速率限制的大小分两种场景，第一种是当路由器出口带宽和入口带宽相等时，如企业向运营商申请了100M的以太网专线，路由器下联交换机也是通过100M以太网链路，此时需要对路由器的转发性能进行实测。如果测试出路由器的以太网接口间的转发性能为80M，则在交换机的上连路由器的接口处最好限制速率为70-75M，以保证交换机的数据流量不会对路由器造成冲击。同时在配置了限速的交换机端口上必须再配合以PQ一类的QoS队列调度技术，来保证重要业务的优先传输。第二种情况是当企业向运营商申请的链路带宽小于路由器局域网入口带宽时，此时交换机可以配置上连路由器接口的速率限制为路由器出口带宽（本例中为2M），也可以按照第一种情况配置为路由器接口的转发能力范围内的速率。如果交换机上连口配置为2M，然后配合PQ一类的QoS队列调度技术保证重要业务的优先传输可以大大减轻路由器对QoS策略的处理负担，但是交换机的QoS支持能力往往不够完善，如使用PQ，可能会因为重要业务长时间占用2M带宽导致其他业务流量得不到带宽而被饿死。所以实际应用中还是推荐使用将交换机的上连口速率限制为路由器接口的转发能力范围内的速率，然后在路由器出接口处做QoS策略来保证重要业务的传输。

采用的QoS模型

目前可以使用的QoS模型主要是区分服务模型。对于语音视频一类的应用可以使用PQ和CBQ技术来保障通信的带宽和时延。PQ由于不能对应用的带宽进行一定的限制经常导致非关键流量无法得到传输，所以目前在三网合一的网络中最常使用的QoS技术还是CBQ队列调度技术，通过CBQ队列技术提供的LLQ队列来保证语音视频通信的带宽和时延，同时可以限制这类应用的带宽，以避免该应用占用太多的带宽导致其他应用的流量被饿死。

端到端的考虑

由于每一种应用都是端到端的应用，所以在本例的网络模型中每个节点都需要对QoS进行相同的规划，以保证重要数据在全网都可以正常传输。

VPN用户的QoS部署

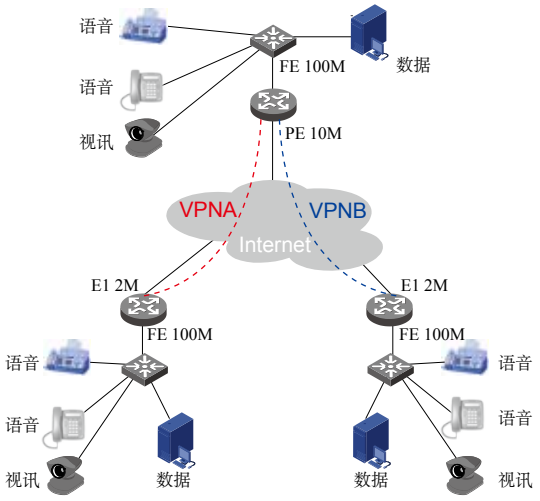


图2 VPN用户企业组网图

组网如图2所示，中心节点通过VPN技术（如IPsec）连接各个分支节点，全网运行语音、视频及传统的数据应用实现三网合一。网络的局域网带宽为100M远远大于广域网的出口带宽，所以在广域网出口处会造成数据转发的瓶颈，因此网络各个场点需要部署QoS来保证语音、视频应用的带宽和时延要求，并保证重要数据的传输带宽要求。和上例一样，组网之前一定要考虑好网络所有应用所需要带宽的情况。我们假设申请的专线带宽可以满足所有网络应用的要求，那我们如何在现有的网络中通过QoS技术对关键应用做出保障，需要我们从端到端的角度进行部署。现有VPN网络会对服务质量造成影响的主要因素和上例相同，本例中只对与上例不同之处进行介绍，其他因素和上例的部署要求一致，在此不再赘述。

流量的分类

与上例要求相同，请参考上例的介绍。

运营商

与上例要求相同，请参考上例的介绍。

局域网交换机与路由器之间处理性能上的差异

与上例要求相同，请参考上例的介绍。

采用的QoS模型

与上例要求相同，请参考上例的介绍。

端到端的考虑

除上例中需要注意的地方外，在VPN组网中往往会出现场点间广域网带宽不一致的情况，如图所示，中心使用10M链路，分支都采用2M链路，此时如果按照前面的例子进行部署则会出现问题。比方说中心向分支发送了超过2M的流量，但并没有超过10M，此时在运营商的网络中就会将超过2M的流量进行丢弃，而且在中心的10M链路上配置的队列并没有生效（因为没有拥塞发生），此时任何的QoS策略都无法对重要业务进行保证（除非运营商可以配合在其设备上也做相关QoS策略，但这种场景下运营商一般不会做，因为客户申请的是Internet线路），企业网所可以使用的技术上对这个问题没有什么好的解决办法。所以组网时一定要注意这种情况，在组网之前充分考虑网络要运行的业务及申请带宽的大小，进行合理的组网以避免出现这种情况。

总结

总之，QoS技术的部署是十分灵活的，不同的应用场景往往使用不同技术的组合来对关键业务进行通信质量的保证，并没有一个固定的模式。目前随着各种应用的出现，QoS已成为组网中必须要考虑的一个重要因素，同时QoS技术也必须继续发展以适应不断变化的应用通讯要求。除了上面讲到的企业网络，运营商的NGN网络能否得以普遍的实施很大程度上也取决于QoS技术的发展程度，目前端到端的QoS技术一直是NGN网络实现的瓶颈，为能顺利部署NGN网络，QoS技术必然会有日新月异的发展和变化。当然，部署了合理的QoS并不表示通信就一定不会出现异常，网络攻击的存在，特别是内部的攻击对网络的破坏是QoS无法保证和避免的，比方说，内部有用户模拟大量的具有语音流特征的报文发往出口路由器，即使做了QoS策略保证语音通信，同样无法避免真正的语音流量受到影响。所以，要想真正使关键业务得到通信质量的保证，也必须加强网络应用的管理力度，但QoS作为通信质量保证的技术基础，地位还是无可替代的！

分层CAR技术简介

文/尹建华

流量监管是差分服务QoS体系的五种技术之一，主要用于流量限速，不比业务识别和队列调度技术，被业界普遍关注和研究，而流量整形和拥塞避免则相对关注得比较少。如今，随着H3C对广域网链路资源通道化思想的提出，综合权衡链路资源与业务质量，实现“预先避免业务拥塞，提升带宽效率和服务质量”的新一代智能流量调度设计，传统QoS技术已经不能满足要求，而通过对流量监管技术创新实现的分层CAR，则让这一思想变为现实。

分层CAR原理

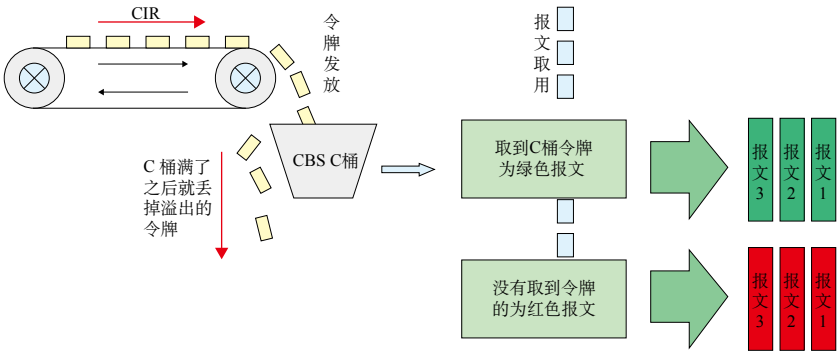


图1 普通CAR原理示意图（单速双色）

CAR作为流量监管的技术，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（如图1中的红色报文）进行丢弃“惩罚”，使进入的流量被限制在一个合理的范围之内，以保护网络资源和用户的利益。

CAR技术采用令牌桶控制流量，当令牌桶中存有令牌时，可以允许报文取令牌进行传输；当令牌桶中没有令牌时，必须等到桶中生成新的令牌后才可以继续发送报

文。即报文的流量不能大于令牌生成的速度，以此达到限制流量的目的。例如，可以限制HTTP报文不能占用超过50%的网络带宽。如果发现某个连接的流量超标，流量监管可以选择丢弃报文，或重新配置报文的优先级。

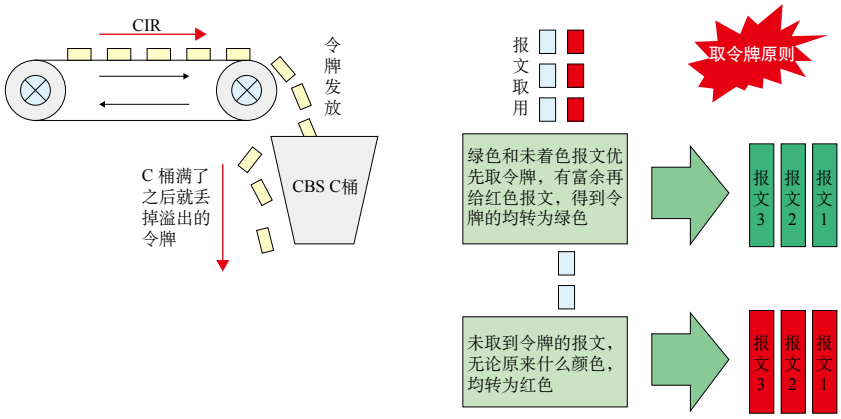


图2 分层CAR原理示意图（单速双色）

相比普通CAR技术，分层CAR是一种更精细的流量监管技术（如图2所示）。它对取C桶令牌的报文进行了细分，根据报文颜色（红色、绿色或未着色的报文）和命令行配置先后决定取令牌的优先顺序，这是与普通CAR的根本区别，因为普通CAR仅按照报文到的时间先后取令牌，是不区分颜色和配置顺序的，因此也没有优先获取令牌的概念。

下面以一个入接口流量调度的例子来说明分层CAR的处理过程和效果。

客户需求：有三种业务A、B、C，分别为视频业务、生产业务和办公业务。业务重要程度是A大于B，B大于C，但是A业务带宽需要限制在10M以内，避免过多视频流量对其他业务产生冲击；B业务为第二优先级业务，要保证20M带宽；C业务优先级最低，保证30M带宽。如果某个业务瞬时实际流量小于其保证带宽，空余的带宽可以被其他业务超出保证带宽的流量占用，实现带宽最大化利用。比如视频业务瞬时流量低于10M时，B业务超出20M的那部分流量优先于C业务超出30M的流量获取令牌，优先进行转发。

配置示意：

```
qos car acl3000 cir 10240Kbps green continue red discard
// acl3000视频业务限定在10M以内，超出部分的流量丢弃。
qos car acl3001 cir 20480Kbps green continue red continue
 //acl3001生产业务保证20M，超出部分的流量不丢弃，选择继续进行二次令牌获取。
qos car acl3002 cir 30720Kbps green continue red continue
//acl3002办公业务保证30M，超出部分的流量不丢弃，选择继续进行二次令牌获取。
qos car acl3003 cir 61440Kbps green pass red discard
```

//acl3003同时包含三种业务，总共保证带宽60M，按照配置顺序和报文颜色，依次是视频业务绿色报文、生产业务绿色报文、办公业务绿色报文获取令牌。如果三种业务的绿色报文瞬间流量之和小于60M，则B和C的红色报文按顺序分别获取令牌，重新成为绿色报文进行转发，未获取到令牌的B或C红色报文依然是红色报文，被丢弃掉。

说明：分层CAR是一种内部令牌优先选取机制，命令字还是CAR。

调度过程：最后一条CAR命令对报文的令牌发放顺序：业务A、B、C绿色报文，业务B、C红色报文。其中由于A业务的红色报文在第一条CAR命令中被discard直接丢弃处理了，因此在最后一条CAR中就没有二次令牌获取的机会了。

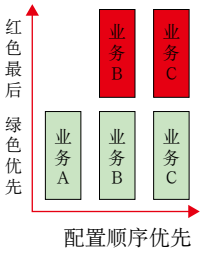


图3 分层CAR令牌发放顺序

利用分层CAR这一针对获取令牌环节的优先顺序的改进，用户就可以在为每个流单独配置CAR动作的基础上，再通过分层CAR对多个业务的流量总和进行限制，实现带宽的二次分配了。

调度效果：

业务瞬间进入流量			实际转发的流量		
业务A	业务B	业务C	业务A	业务B	业务C
10M	20M	30M	10M	20M	30M
5M	25M	30M	5M	25M	30M
0M	25M	35M	0M	25M	35M
15M	15M	50M	10M	15M	35M
10M	50M	30M	10M	20M	30M

从上面表格中可以看出，分层CAR对令牌的发放顺序对各个业务之间的带宽分配（通过令牌分发）起到了关键作用。正因为分层CAR这种令牌发放原则，使得其成为了QoS队列的一种替代设计。在本例中，ABC业务的普通CAR实际上是一种按比例分配带宽的CQ机制，而ABC的分层CAR则体现了在CQ机制上的，超出流量的优先抢占的PQ关系。

注意事项：由于分层CAR优先选择绿色报文和未着色报文进行令牌分配（当分层CAR的处理对象全部为未着色报文时，其效果与普通CAR处理是一样的），绿色报文和未着色报文之间是先到先得令牌的关系，因此不建议把未着色的报文和着色报文进行统一分层CAR操作，避免流量调度效果不清晰。

分层CAR令牌分发和处理原则：

首先，分层CAR的处理对象是：普通CAR处理后的，且动作选择为continue的报文。因为只有经过了普通CAR处理后，报文才有红绿颜色之分；并且只有选择了continue，才能进入分层CAR的二次令牌发放过程。（动作选择为pass或者discard的报文直接被转发或丢弃，没有机会进入分层CAR的二次令牌选取过程，continue是因分层CAR技术新产生的报文动作类型。）

其次，分层CAR的令牌发放遵循两个原则：第一、先给绿色和未着色报文发放，后给红色报文发放；第二，在第一原则基础上，对红绿报文，均按照各业务普通CAR的配置顺序进行发放，直到发完为止。

分层CAR优势

从技术上讲，分层CAR在接口出入双方向都可以部署。当在入接口进行部署时，因为能够对业务进行实时流量监测和立即决定是否转发，当瞬时流量超出目标出端口带宽能力时，可以提前丢弃优先级较低的报文，从而避免出端口带宽拥塞，提高调度效率。这是因为出端口拥塞会产生较大的时延和抖动（报文需要被排队后再丢弃或转发，产生了较大的时延和抖动）。因此，分层CAR的流量调度机制，具有实时性好，报文不用排队等待就可以立即判断是否丢弃的能力，有效避免了排队时延和抖动。另外，根据实际测试，在单物理端口上可支持1000条以上CAR操作，相比QoS队列来说，极具性能优势。与传统QoS队列的优势比较如表1所示：

由于分层CAR具备了如表1中的诸多优势，使其逐渐成为了广域网智能流量调度的核心技术。智能流量调度技术，旨在通过流量调度的创新，实现精确的流量监管、处理和调度，从而提升业务质量和链路资源利用效率。其设计理念就是：流量通道化处理，预

	QoS队列	分层CAR
调度机制	排队调度方式（入队+调度+令牌发放）	令牌发放方式（两次令牌发放）
调度方式	支持PQ、CQ、WFQ、CBQ、H-QoS等队列调度机制	支持与PQ、CQ、CBQ（除WFQ队列）、H-QoS相同的调度机制。具体机制由分层CAR的流量分组和配置顺序决定（就像搭积木），不需要其他技术支撑
分组数量	PQ：4个 CQ：16个 CBQ：64个	没有设计限制，可以上千，取决于设备性能
部署位置	出端口	入端口、出端口、全局。可以根据分流情况与其他技术结合形成丰富方案，如跨端口流量调度
策略时效	端口拥塞后生效	实时有效，可实现非拥塞下的流量控制，如对BT限流，带宽预留
业务抖动	根据拥塞和队列长度情况，会产生较严重抖动，比如几百毫米或秒级	令牌实时发放，立即决策通过或丢弃，不额外产生抖动，对媒体业务极为有效。如需防止突发流量丢弃，可用队列技术配合完善
设备性能	主要影响性能的环节：WRED丢包、报文排队、调度、令牌发放	主要是令牌二次分发。如果在入口做分层car，可以提前丢包，减少后续环节的处理，如路由查表
易用性	设计、部署和流量分析非常复杂	设计、部署和流量分析都比较简单

表1 分层CAR调度和QoS队列调度的比较

先避免目标端口拥塞、实现精确业务调度，使网络QoS部署简单。其应用价值主要体现在以下方面：

高效的业务调度：对令牌机制提出创新的分层CAR技术，不仅实现了PQ、CQ、CBQ等传统的调度方式，在简化了QoS调度设计的同时，能够几倍、十几倍的降低了业务的时延和抖动，显著提升了业务传输质量水平。

虚拟的带宽资源：入接口分层CAR部署，配合策略路由动作，创新实现了多端口或链路流量的统一调度设计，从而避免了单链路的带宽拥塞，降低了拥塞时的丢包和时延，是对单链路QoS设计的一个飞跃。

兼有SDH带宽独占与IP带宽共享的特质：入接口分层CAR和出接口共享带宽限制设计，为企业提供了在多部门间带宽预留和共享的方案。该方案不仅继承了IP承载的带宽共享特征，保证了带宽利用效率，而且体现了SDH传输的带宽独占的可靠承载要求。

动态批量分层CAR，实现用户带宽均分：基于在线IP用户报文检测的动态分层CAR技术，支持成百上千在线IP终端用户的带宽公平分配和保证。这是QoS队列所无法做到的。

关于分层CAR在这些方面的应用案例，可以参考《广域网智能流量调度》相关介绍，这里不再详细阐述。

CAR应用潜力探讨

通过对CAR令牌分发方式的改进，分层CAR实现了类似QoS队列调度的效果，这种创新的流量调度技术与其天然的精确流量评估能力结合，推动产生了一系列的智能流量调度设计方案，并已经在广域网的诸多应用场景和项目中得到了良好应用。

通过技术分析和验证发现，流量监管技术CAR拥有一些天然的优势：比如对业务流量精确的实时监管，对网络设备性能要求低，而且部署位置灵活，还有分层CAR的流量调度能力。所以，利用这些优势，如果再结合和流量处理相关的技术，应该还能做更多的创新设计。比如，把业务的流量监管与路由设计进行结合，就可以实现基于业务流量情况的性能路由，跳出传统QoS仅能在单一链路上调度的局限，实现多出口、多链路的带宽资源动态统一调配，并同时保证服务质量。因此，顺着路由和流量结合这个思路，分层CAR还可以有更多发挥作用的地方，从而设计出更为完善的流量调度方案。

总结

通过本文对分层CAR技术的介绍，我们看到，看似简单的CAR技术，通过令牌分发环节的一点改进，就可以实现优秀的流量调度能力，而且在广域网的诸多方案中已经得到成功应用。沿着流量调度、服务质量和业务路由这条网络设计主线，我相信，通过对技术的点滴改进还可以创新出更多更优秀的方案，帮助用户提升广域网链路的利用效率、业务质量，并优化路由设计。

缩略语：

CAR：Committed Access Rate，承诺访问速率

CIR：Committed Information Rate，承诺信息速率

CBS：Committed Burst Size，承诺突发尺寸

H3C广域网QoS方案设计技术简介

文/尹建华

广域网QoS部署流程

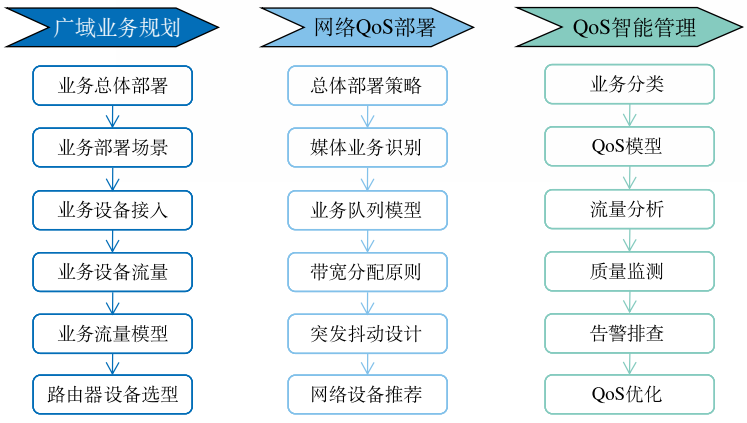


图1 QoS设计基本流程

全面的QoS部署设计应该包括以上三个部分。针对一个具体项目，根据客户实际要求，可能不会涉及到以上所有设计过程，而会对某些步骤省略或者简化：

广域业务规划：理解客户网络业务规划是设计QoS方案的基本前提。业务规划主要涉及两大方面，一是客户的行政划分和相关路由、VPN、拓扑设计；二是客户的重要业务，比如视频会议、语音业务、生产业务、办公业务等等。这两方面的因素是QoS方案设计的基本前提。在项目中，我们需要与客户进行充分的沟通，了解这两方面的真实要求是做好广域网QoS方案设计的前提。

QoS策略制定：在充分了解客户对广域网业务规划的前提下，确定网络中的带宽瓶颈节点，制定适合的QoS方案是达成客户对业务流量和质量保证目标的关键。在某些时候为了满足客户的整体QoS要求，也需要对网络设计作出适当的改进。

QoS部署和监管：广域网络规模大、业务种类多，QoS方案无论是部署，还是实施效果监管都是比较复杂的。通过H3C的iMC网络管理平台的QoS管理组件可以帮助客户简化QoS部署过程，提供QoS方案的准实时流量和质量监管。

广域网QoS部署准备

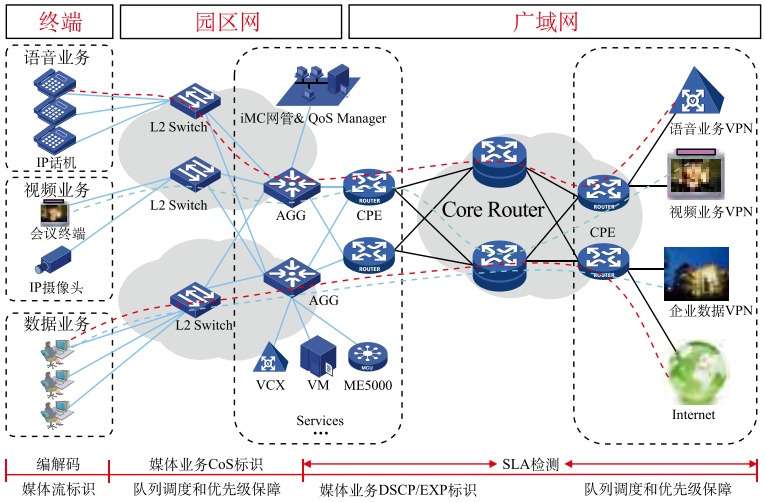


图2 QoS部署典型架构

客户网络的行政需求

客户网络的行政需求会因行业、规模、地域、部门构成等因素有很大差异，其广域组网技术也包括纯IP技术和MPLS VPN，组网形式也多种多样，因此作为广域网带宽分配和业务质量保障的QoS方案要因势而成。针对不同网络架构和技术的广域网，均有相应的QoS设计特点，在本文的后续部分将会详细讲述。

网络业务的自然需求

度量参数	高清	实时视频	监控点播	监控存储	语音	园区网分配	广域网分配
时延	150ms	150ms	1000ms	500ms	150ms	20%	80%
抖动	10ms	10ms	100ms	30ms	30ms	50%	50%
丢包	0.05%	0.10%	0.10%	3%	1%	50%	50%
净荷带宽	3～16M	128K～4M	128K～8M	128K～8M	8K、64K	——	——
实占带宽	4.2～22M	250K～10M	250K～11M	250K～11M	24K、80K	——	——

表1 媒体业务特征和质量要求

业务和网络评估

- 媒体业务：确定客户网络中服务质量要求最高的媒体业务和相应带宽、时延、抖动和丢包要求，整体初步评估网络的可满足性；
- 网络瓶颈：根据媒体业务和容量要求，确定网络关键拥塞点；
- 准入控制：对网络拥塞点的媒体业务量进行呼叫准入控制，防止媒体业务自身拥塞。

链路和设备评估

- 链路评估：对于低速链路如DSL/E1/Serial，不建议视频和数据业务混合转发，或者要求视频业务等优先级业务实占带宽不能超过总带宽的1/3。对于低速接口捆绑链路，不建议进行视频业务转发，尤其对高码率视频；
- 设备评估：需要重点评估QoS性能和接口Buffer，在后面会有详细介绍。

QoS设计和计算

- 业务识别：业务识别的原则是越早越好，以减轻网络设备业务识别负担。最好媒体终端自行标记业务，一般建议在园区网进行业务识别，广域网只需进行优先级映射并进行队列调度即可；
- 流量监管：对网络瓶颈点上游入端口进行流量限速，防止非优先级业务冲击导致的网络设备性能下降；
- 拥塞避免：根据队列内业务分类和权重进行拥塞时的丢包处理；
- 队列调度：根据业务种类和各自带宽需求进行出口带宽评估，确定业务带宽比例和优先级差别；
- 流量整形：针对视频类业务一般不推荐。对视频业务，权衡自身时延、突发流量情况和下游设备缓冲能力后，可以进行整形。

广域网业务分类模型

QoS的核心任务有两点：一是带宽分配，二是确定调度优先顺序。在广域网中，带宽分配大多数情况是基于客户的地域分布、网络结构和部门构成等行政因素进行划分的，而对调度顺序的决定往往是根据业务的服务质量需求进行划分的。所以，广域网QoS业务分类模型的建立必须结合这两个方面（行政属性和业务属性）。

按照广域网流量的行政属性划分，一般基于设备、端口、源和目的IP进行区分，进而实施不同的QoS带宽策略。基于行政属性划分的业务流量一般不会有优先级保证要求，而主要关注各行政单位的带宽分配。

按照广域网流量的业务属性划分，根据RFC4594标准分类，主要分为以下几类：

媒体分类	PHB（DSCP）	CoS	EXP	业务说明	媒体业务举例
Network Control	CS6（48）	6	6	网络控制平面的业务	EIGRP、OSPF、BGP、VRRP
VoIP Telephony	EF（46）	5	5	IP电话业务，包括G.711、G.729等语音流	VCX IP语音业务
Broadcast Video	CS5（40）	5	5	广播电视、视频监控业务，特点是丢包敏感，不具备重新发送和流控能力	监控实况流和广播电视
Real-Time Interactive	CS4（32）	4	5	室内部署的交互视频应用，具有语音和视频能力。而其中的数据业务要划归到“Low-Latency Data”媒体类	视频会议、高清视频
Multimedia Conference	AF41（34）	4	4	桌面多媒体协同应用软件，包含语音和视频应用。其数据业务归属于“Low-Latency Data”媒体类	桌面多媒体会议
Multimedia Streaming	AF31（26）	3	3	VOD流媒体业务。这类业务允许一定的时延，丢包能够重传，比广播电视和实时媒体业务更具“弹性”	视频点播
Call-Signaling	CS3（24）	3	3	IP语音和视频业务的信令流	SIP、H323、MGCP、VMP
OAM	CS2（16）	2	2	网络运营、维护和管理类的业务	SNMP、SSH、Syslog
Low-Latency Data	AF21（18）	2	2	指交互性的重要数据业务，业务往往响应时间短，否则会影响到工作和生产效率	VCX IP消息业务、ERP、CRM、DB
High-Throughput Data	AF11（10）	1	1	指非交互性的“背景”业务，其特点是用户不需要等待业务的响应，业务的响应时间不会直接影响工作和生产效率	E-Mail、FTP、应用备份、监控存储流、回放流
Low-Priority Data	CS1（8）	1	1	与公司业务无关，多是娱乐性的业务。如果网络发生拥塞，这类业务将首先被丢弃	BT、eMule、YouTube等
Best Effort	DF（0）	0	0	大多数业务不进行优先级标记，采用默认值（DSCP 0）	其他应用

表2 RFC4594业务分类标准

■ RC4594不仅根据业务性质和对服务质量要求的差异，对业务进行了分类，并且为了提高业务识别效率，对不同类别的业务进行了标准的分类标记。QoS设计原则上要求在最接近数据源端的设备上识别数据流并根据统一的业务模型进行标记，之后的各个节点信任数据流的标记并根据标记进行QoS处理，保证QoS服务质量。但是对于实际组网情况，一般是采取就近（网络拥塞接口前最近的设备或端口）和能够（准确区分业务）结合的原则决定在哪里进行首次业务分类和标记，这样就避免了该拥塞节点前那些设备不必要的分类和标记操作，同时保证了对于拥塞接口的流量分类的完整性和针对性。也有时候会考虑核心设备的业务分类带来的性能压力，而把业务分类放在前面一跳的设备上进行；

■ 根据承载链路是采用二层以太、三层IP还是MPLS链路，可以分别采用CoS、DSCP、EXP进行业务标记。由于端到端链路上有可能承载方式放生变更，比如从IP承载改为了MPLS转发，那么此时就需要在承载层发生变更的设备上使能优先级信任和映射功能，也就是说标记体系之间必须有一个完整的映射关系。这样才能避免业务多次识别，保证端到端业务服务等级的一致性。

业务分类和标记			
媒体类别	PHB	CoS	EXP
Network Control	CS6	6	6
Voip Telephony	EF	5	5
Broadcast Video	CS5	5	5
Real-Time Interactive	CS	4	5
Multimedia Conference	AF41	4	4
Multimedia Streaming	AF31	3	3
Call-Signaling	CS3	3	3
OAM	CS2	2	2
Low-Latency Data	AF21	2	2
High-Throughput Data	AF11	1	1
Low-Priority Data	CS1	1	1
Best Effort	Default	0	0

表3 业务优先级映射关系

媒体类别	典型业务和协议		识别方式		业务流说明
			TCP/UDP	Port	
VoIP Telephony	语音流	RTP、RTCP	UDP	6002、6004、6020 8000～8360 10000～11024 16384～17105 22520、22521	监控语音对讲 VCX、IP Phome语音 DVR Agent双向语音 MSR网关语音 MG9060语音
Broadcast Video	监控实况流	RTP、RTCP	UDP	15000~20000	DVR Agent实况媒体流
Real-Time Interactive	视频会议	RTP、RTCP	UDP	22620、22621 22820、22821	主流媒体通道 辅流媒体通道
Multimedia Conference	桌面会议	RTP、RTCP	UDP	6020、6021、6040	VC客户端语音
Multimedia Streaming	监控回放	RTP、RTCP	UDP	10660~12000	DVR Agent点播媒体流
Call-Signaling	SIP	SIP	TCP、UDP	5060	SIP协议
	H323	RAS、H.225	TCP、UDP	1719、1720	H323协议
		H.225、H.245	TCP	30000~31999 40000~44999	MG9060使用的信令端口 ME8000使用的信令端口
	MGCP	——	UDP	2427、2428、2728	MGCP协议
	RTSP	——	TCP	554	MS/VC/ECR/ISC流媒体协议
	VMP	——	UDP	6060、6063	VMP协议
	GMP	——	TCP	12000	GMP协议

表4 多媒体业务识别

媒体类别	典型业务和协议		识别方式			业务流说明
			IP	TCP/UDP	Port	
Network Control	RIP	——	——	UDP	520	——
	OSPF	——	IP	——	89	——
	BGP	——	——	TCP、UDP	179	——
	EGP	——	IP	——	8	——
	EIGRP	——	IP	——	88	——
	RSVP	——	——	UDP	1698、1699	——
	DHCP	——	——	UDP	67、68	——
	LDAP	——	——	TCP、UDP	389	——
	VRRP	——	IP	——	112	——
	DNS	——	——	TCP、UDP	53	——
	NTP	——	——	TCP、UDP	123	——
OAM	ICMP	——	IP	——	1	——
	Telnet	——	——	TCP	23	——
	rLogin	——	——	TCP	513	——
	SNMP	——	——	UDP	161、162	——
	Web	一般Web配置	IP	TCP	80、8080	——
		多媒体配置	——	TCP	33012~33051 48000~48499 11576、11582	视频会议MG9060配置端口 视频会议ME8000配置端口 监控IP SAN管理
	SSH	——	——	TCP	22	Secure Shell
	Syslog	视频会议	——	UDP	514 19000	标准Syslog监听端口 视频会议Syslog发送端口

表5 控制和管理业务识别

广域网接入业务识别

多媒体业务的识别

广域网中语音、视频等多媒体业务的发展非常迅速，同时语音、视频等业务对于服务质量要求很高，所以对于语音、视频等多媒体业务的识别就是目前广域网中QoS部署的重点。

常见多媒体业务的识别方式可根据端口范围确定，也可根据广域网事先分配的IP地址区分（一般推荐在做IP地址规划时为语音、视频业务专门划分，从而简化后续做QoS策略时对业务的识别）。表4是H3C公司常用媒体业务端口特征：

控制管理业务的识别

业界常见路由协议和网管报文特征（见表5）：

常见数据业务的识别

业界常见数据业务的报文特征（见表6）：

ACL业务识别管理

根据以上业务分类方式和常见业务端口特征，可通过广域网管理平台iMC ACL Manager进行业务流定义，并可以将ACL配置保存为模板资源，供所有设备使用。

媒体类别	典型业务和流量		识别方式			业务流说明
			TCP/UDP	Port	DAR	
Low-Latency Data	即时消息	MSN	TCP	1863	——	——
	视频会议	数据媒体通道	UDP	22720、22721	——	多媒体产品视频会议数据
	企业应用	SAP	TCP	3300~3315 3200~3215 3600~3615	——	SAP公司的ERP软件
		Exchange	TCP	135	——	微软Exchange邮箱服务
	数据库查询	Oracle	TCP、UDP	1521	——	Oracle公司数据库软件
		MS-SQL Server	TCP	1433	——	微软SQL数据库软件
		CIFS	TCP	139、445	——	Common Internet File System
		Postgre	TCP	5432	——	Postgre SQL访问侦听端口
High-Throughput Data	FTP	——	TCP	21	DAR	File Transfer Protocol
	SFTP	——	TCP	990	——	Secure FTP
	TFTP	——	UDP	69	——	Trivial File Transfer Protocol
	E-MAiL	POP2/3	TCP、UDP	110	——	Post Office Protocol
		SMTP	TCP	25	——	Simple Mail Transfer Protocol
		IMAP	TCP、UDP	143，220	——	Internet Message Access Protocol
	Notes	——	TCP	1352	——	Lotus Notes
	监控存储回放流	NFS	TCP、UDP	2049、111、905	——	监控DM/EC/DC/ECR/ISC存储回放
		iSCSI	TCP	3260	——	iSCSI读写目的端口
	文件共享	——	TCP、UDP	135~139、445	——	Windows文件共享服务
Low-Priority Data	BT	——	TCP	6881~6889	DAR	Bit Torrent
	eMule、eDonkey	——	TCP	4662	——	电驴、电骡下载

表6 常见数据业务识别

业务分类和标记		队列模型和带宽分配参考		
媒体类别	PHB	1P7Q（L3 DSCP 为例）	1P5Q（MPLS EXP 为例）	1P3Q（L2 CoS 为例）
Network Control	CS6	Q8（EF、CS5、CS4）20% Priority Queue	Q6（EXP5）30% Priority Queue	Q4（CoS5、CoS4）30% Priority Queue
VoIP Telephony	EF			
Broadcast Video	CS5			
Real-Time Interactive	CS4	Q7（AF41）10%	Q5（EXP4）10%	
Multimedia Conference	AF41	Q6（AF31）10%	Q4（EXP6、EXP3）15%	Q3（CoS6、CoS3、CoS2）35%
Multimedia Streaming	AF31	Q5（CS6、CS3、CS2）10%		
Call-Signaling	CS3			
OAM	CS2		Q3（EXP2）15%	
Low-Latency Data	AF21	Q4（AF21）10%		
High-Throughput Data	AF11	Q3（AF11）10%	Q2（EXP1）5%	Q2（CoS1）10%
Low-Priority Data	CS1	Q2（CS1）5%		
Best Effort	Default	Q1（DF）25%	Q1（EXP0）25%	Q1（CoS0）25%

表7 业务分类和队列调度模型

广域网QoS队列策略

QoS队列模型的选择

RFC4594从标准理论上对业务进行了完整划分，但在项目中，客户往往不会针对每种业务分类指定调度队列，而是根据实际业务构成、端口速率和业务保证的精细度需要，将这些业务流引入到比业务分类更少的队列中做调度策略就够了。常见的业务队列构成有4类（1个优先级队列，2个带宽保证队列，1个尽力转发队列）、6类和8类队列模型（表7）。

- 4类队列模型较为简单，主要应用在广域网低速链路或是企业网接入层；
- 6类队列模型较为复杂，主要应用于广域网中高速链路或是企业网汇聚层/核心层；
- 8类队列模型最全面，主要应用于广域网高速链路或是企业网核心层。

在实际项目中，要根据客户对业务的分类精细化要求，同时参考以上标准的队列模型进行设计，对于大型广域网，一般建议至少采用4类队列模型，以保证语音视频类优先级业务、路由管理类协议、重要数据业务和尽力转发业务的整体QoS效果。

QoS队列参数的设计

在广域网的QoS队列设计中，除了保证语音视频等实时性较强业务的优先转发外，最为关注的是各个队列的带宽分配。带宽分配主要根据客户对业

务的实际带宽需求制定，但为了达到最佳的QoS部署效果，还要遵循两个带宽分配原则和一个视频突发流量参数设计经验。这些是根据业界以及华三公司大量的项目和实验室最佳实践数据得来的经验。

QoS带宽分配原则：基于队列带宽需求进行分配，为达到整体业务最佳QoS效果，推荐遵循以下原则：

PQ<1/3总带宽原则：语音、视频等采用PQ/LLQ队列保证其优先转发的同时，却增加了数据业务的排队时延，甚至丢包。PQ所占带宽比例越大，PQ业务瞬间突发对数据业务的冲击就越大。根据H3C大量项目建设经验和实验室数据，PQ带宽设计少于总带宽的1/3是保证业务总体质量达到最佳的基本要求。

DF>1/4总带宽原则：Default Forwarding默认转发队列，就像一个公共通道，为优先级标记为0的尽力转发业务和优先级业务超出指定带宽的额外流量转发使用，预留充足的带宽能够为其他业务和应对优先级业务流量突发提供很好的保障。所以，一般推荐DF带宽不低于总带宽的1/4。

QoS队列部署和监管

在完成了对各类业务的队列模型选择和参数设计后，可以通过管理平台iMC QoS Manager进行QoS策略的部署。

图3 iMC QoS管理界面

QoS策略部署：结合ACL流分类或者业务优先级，进行向导式部署，包括业务识别、流量监管、拥塞避免、队列调度和流量整形；

在iMC QoS Manager完成了队列策略部署后，还可以分别选择广域网管理平台iMC NTA和SLA组件对不同QoS队列的流量和服务质量进行监测，提供QoS策略优化参考；

- NTA监管与广域网络设备SR88、SR66和MSR路由器系列的NetStream流统计技术配合，可实现广域准实时流量监管；
- NTA可自动生成流量/应用/节点/会话四大类数十种预制报表，也可自定义报表。

借助NTA管理组件的流量监管功能，对广域网设备的流量基于业务优先级标记，或者ACL所设定的IP特征进行统计，可以实现QoS策略部署效果的监管。

- SLA质量监管：SLA监管与网络设备NQA质量分析技术配合，可实现广域网业务质量监管；
- SLA质量报表：SLA提供设备管理、服务等级定义、服务类型定义、实例管理、审计管理和报表管理等功能。

通过SLA管理组件设置检测流量的业务优先级标记，或者ACL所设定的IP特征进行统计，可以实现QoS策略质量效果的监管。

设备指标		广域网		
		边缘路由器	汇聚路由器	核心路由器
推荐设备		MSR	SR66	SR88
推荐板卡		——	——	NP业务板
转发性能		600Kpps	4.5Mpps	586Mpps
ACL资源		3000	5000	64K/NP单板
QoS队列 缓存/内存	类别	设备内存	设备内存	芯片缓存
	容量	256MB/1GB	2GB	512MB/1GB
	分配方式	全局共享	全局共享	板级+端口
	1帧并发视频数	不限制	不限制	不限制
		不限制	不限制	不限制
业务识别	L2~4层	支持	支持	支持
	L4~7层	DAR	DAR	——
队列调度	二层队列	——	——	8/Port，16K/NP
	调度机制	PQ/LLQ/WFQ/CBWFQ	PQ/LLQ/WFQ/CBWFQ	PQ/LLQ/WFQ/CBWFQ
	分层QoS	——	支持	支持
流量监管	端口限速	64K粒度	64K粒度	1K粒度
	ACL流限速	1K粒度	1K粒度	1K粒度
	CAR队列限速	1K粒度	1K粒度	1K粒度
拥塞避免	WRED	支持	支持	支持
流量整形	GTS	支持	支持	支持
QoS监管	流量统计	NetStream	NetStream	NetStream
	SLA监测	NQA	NQA	NQA

表8 广域网QoS方案推荐设备

总结

本文从QoS设计的基本流程，针对广域网进行了介绍。但是广域网千差万别，客户需求多种多样，仅仅了解QoS的基本技术还是不够的，只有通过一些实际案例才能逐步对QoS的设计有深刻体会。后续通过H3C的一些常见QoS的最佳实践，让您进一步理解QoS设计的技巧，将会发现它给客户带来的价值。

H3C广域网智能流控设计与应用

文/尹建华

智能流控概述

广域流量调度的挑战

流量调度是网络基础设计之一，其重要性就如同开车上路，除了关注怎样到目的地之外，还要关注路况，比如道路是否畅通、路面是否湿滑等等。流量调度就是信息高速公路上的路况设计和调度管理，它与链路性质和带宽、路由设计、QoS策略实施、流量和质量监管等这些因素密切相关。

随着企业信息业务的不断丰富，数据大集中已经成为IT建设的趋势，在广域网带宽紧张、成本昂贵的情况下，如何解决带宽与业务之间的矛盾，打造快速、稳定、高质量的广域网络，进行良好的流量调度设计就显得尤为重要。

传统的流量调度设计，主要存在这三方面的问题和挑战：

■ 业务质量：采用一般的QoS队列调度，链路拥塞时，由于报文排队的原因，业务会出现十几、甚至几十倍的延迟和抖动。这就是为什么采用了QoS设计却仍然出现质量明显下降的问题所在。

如何让业务流量始终平滑是所有QoS设计的关键所在；

■ 链路效率：随着企业广域网络的发展，双链路承载已经成为高可靠设计和带宽扩容的不二选择。而传统基于单接口、单链路的QoS设计，往往造成主链路业务繁忙，而备链路利用不足的严重带宽浪费现象。

■ 带宽公平：多数企业都有分支机构或远程接入点，在这些分支机构内，由于某个员工收Email附件或者BT下载导致的其他人员上网慢，甚至业务中断的现象屡见不鲜。如何保证重要业务的质量和公平的给每个员工分配上网带宽，而且能在员工不上网时释放带宽给其他人用，这也是传统广域流量调度存在的一大难题。

本文广域网智能流量调度将针对以上三大挑战进行优化设计。

智能流控的价值理念

智能流控，旨在通过流量调度和管理的技术创新，实现动态、精确的流量控制，从而提升业务质量和链路资源利用效率。

智能流控的价值理念：更高效、更精确、更方便。

智能流控，智在哪里，有何创新?

■ 高效的调度技术：对令牌机制创新的分层CAR技术，实现了PQ、CQ、CBQ等传统的调度策略，不仅简化了QoS调度设计，还几倍、十几倍的降低了业务流的时延和抖动，显著提升了业务传输质量水平。

■ 虚拟的带宽资源：入接口分层CAR部署，配合策略路由动作，创新实现了多端口或链路的流量统一调度设计，虚拟化了链路资源，是对单链路QoS的一个飞跃；

■ SDH与IP的特质：入接口分层CAR和出接口共享带宽限制设计，为企业提供了在多部门间带宽预留和共享的方案。该方案不仅继承了IP承载的带宽共享特征，保证了带宽利用效率，又体现了SDH传输的带宽独占的可靠承载要求；

■ 动态令牌桶技术：基于在线流量检测的动态令牌桶，结合分层CAR，支持500个以上在线IP终端的带宽公平分配和保证。这是QoS队列所无法做到的。

如何在广域网中应用

广域网应用场景	智能流控关键技术		
	分层CAR	动态CAR	负载分担
低时延抖动调度设计	√		
双链路统一调度设计	√		
园区出口公平带宽设计	√	√	
多部门带宽独占和共享	√		
专线双出口负载分担	可选		√

智能流控设计与关键技术

智能流量调度设计和技术，可以应用于广域网基本QoS设计、双链路流量调度设计、广域网园区或分支出口流量调度设计、广域多部分带宽独占和共享需求设计、精细化QoS设计等常见广域网络和业务应用场景中。

智能流控技术

分层CAR

普通CAR技术：CAR作为流量监管的技术，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，以保护网络资源和用户的利益。

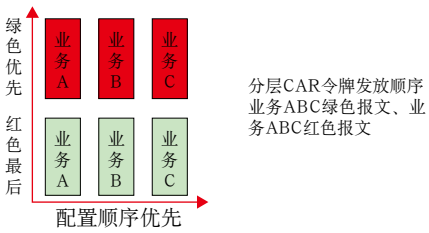
CAR技术原理：采用令牌桶控制流量，当令牌桶中存有令牌时，可以允许报文取令牌进行传输；当令牌桶中没有令牌时，报文必须等到桶中生成了新的令牌后才可以继续发送。这就限制了报文的流量不能大于令牌生成的速度，达到了限制流量，同时允许突发流量通过的目的。 例如可以限制HTTP报文不能占用超过50%的网络带宽。如果发现某个连接的流量超标，流量监管可以选择丢弃报文，或重新配置报文的优先级。

分层CAR技术：相比单层CAR，分层CAR是一种更灵活的流量监管策略，用户可以在为每个流单独配置单层CAR动作的基础上，再通过分层CAR（第二次CAR）对多个业务的流量总和进行限制，实现带宽的二次分配。那么，分层CAR对已经做过普通CAR的各业务流是如何处理的呢?

首先，分层CAR的处理对象是：普通CAR处理后的，且动作选择为continue的报文。因为只有经过了普通CAR处理后，报文才有红绿颜色之分；并且只有选择了continue，而不是pass或者discard，因为这两个都做就直接对报文转发走了，或者丢弃了，没有机会进入分层CAR的二次令牌发放过程。

其次，分层CAR的令牌发放遵循两个原则：第一、先给绿色报文发放，后给红色报文发放；第二，在第一原则基础上，对红绿报文，均按照各业务普通CAR的配置顺序进行发放，直到发完为止。

注意事项：分层CAR是对已经着色的流量进行处理，如果在CAR的处理流量中只是未着色流量，那么该CAR就按照普通CAR进行处理。但如果CAR处理的流量对象既有着色后的流量，也有未经过CAR处理后的未着色流量，那么CAR会同时对绿色流量和未着色流量进行优先处理，然后对红色流量进行令牌分配（如果令牌富余的话），会导致分层CAR的处理混乱，无法按照预先的设计进行绿色流量的准确调度控制，因为中间混合了未着色流量。所以分层CAR的处理对象，务必确保都是已经完全着色，也就是经过至少一次CAR处理后的流量。



举例

配置示意

qos car acl3000 cir 1024Kbps green continue red discard

qos car acl3001 cir 2048Kbps green continue red continue

qos car acl3002 cir 3072Kbps green continue red continue

qos car acl3003 cir 6144Kbps green pass red discard

业务说明

acl3000为业务A，acl3001为业务B，acl3002为业务C，acl3003为业务A+B+C。因为acl3003包含了ABC三种业务，且前面三种业务都有流量选择了continue动作，因此实际上最后一条car对前面continue的流量操作就是分层car的处理。分层car是一种内部处理机制，命令字与普通car是相同的。

转发情况

业务瞬间进入流量			实际转发的流量		
业务A	业务B	业务C	业务A	业务B	业务C
1M	2M	3M	1M	2M	3M
0.5M	2.5M	3M	0.5M	2.5M	3M
0M	2.5M	3.5M	0M	2.5M	3.5M
1.5M	1.5M	5M	1M	1.5M	3.5M
1M	5M	3M	1M	2M	3M

从上面表格中可以看出，分层CAR对令牌的发放顺序对各个业务之间的带宽分配（通过令牌分发）起到了关键作用。正因为分层CAR这种令牌发放原则，使得其成为了QoS队列的一种替代设计。在本例中，ABC业务的普通CAR实际上是一种按比例分配带宽的CQ机制，而ABC的分层CAR则体现了在CQ机制上的，超出流量的优先抢占的PQ关系。

动态CAR

动态CAR是对预先指定的IP地址范围，进行流量的动态检测，如果有该IP范围内新的IP地址的流量产生，系统会动态产生一个基于该IP的令牌桶（CAR），并对其限速；当该IP地址不再产生流量后，该令牌桶自动消失，从而减少对系统令牌桶资源的占用。

另外，动态CAR可以和分层CAR进行结合，实现同一IP地址范围内的在线IP流量之间的公平带宽分配和抢占。

下面举例说明：

举例1：IP地址带宽限制

配置示意

[Sysname] qos carl 1 source-ip-address subnet 1.1.1.0 24 per-address

[Sysname] interface ethernet1/1

[Sysname-Ethernet1/1] qos car outbound carl 1 cir 100 cbs 6250 ebs 0

业务说明

在接口Ethernet1/1的出方向上应用CARL规则1。CARL规则1是对

源地址属于子网1.1.1.0/24内每台PC限速100kbps，网段内各IP地址的流量不共享剩余带宽。

举例2：IP地址带宽公平分配

配置示意

[Sysname] qos carl 2 source-ip-address range 1.1.2.100 to 1.1.2.199 per-address shared-bandwidth

[Sysname] interface ethernet1/1

[Sysname-Ethernet1/1] qos car outbound carl 2 cir 5000 cbs 3125 ebs 31250

业务说明

在接口Ethernet1/1的出方向上应用CARL规则2。CARL规则2是对源地址属于IP地址段1.1.2.100～1.1.2.199内所有PC限速5Mbps，网段内各IP地址的流量共享剩余带宽。比如，初始只有1.1.2.100这台PC上线，网速可以达到5Mbps；接着，1.1.2.199这台PC也上线了，如果后上线这台PC有2.5M以上的流速需求，那么，通过系统内部动态CAR+分层CAR机制，这两台PC流速都会变成2.5Mbps；同样，第三台上线后，每台PC的流量为5Mbps/3=1.67Mbps。如果有一台PC下线了，那么其他的PC流速就会提升上来，达到5Mbps/n，n为实际在线PC数量，而不是配置的IP地址范围的数量。

动态CAR机制是通过CARL的per-address参数体现出来的。而动态CAR结合分层CAR实现带宽公平分配，则还要通过设置shared-bandwidth参数来实现。

负载分担

负载分担的实现方式有以下几种：

- 基于流的负载分担：使能了快速转发功能后只能进行基于流的负载分担。例如，当前设备上存在两条等价路由，如果有一条数据流经过，那么将从其中一条路由上转发；如果有两条数据流经过，那么这两条等价路由分别转发这两条数据流；
- 基于报文的负载分担：关闭了快速转发功能后进行基于报文负载分担，即将待发送报文均匀分配到两条等价路由上；
- 基于带宽的负载分担：关闭了快速转发功能后，报文按接口物理带宽进行负载分担（即基于报文的负载分担）；当用户为接口配置了指定的负载带宽后，设备将按用户指定的接口带宽进行负载分担，即根据各接口物理带宽比例关系进行分配；
- 基于用户的负载分担：设备上存在多条等价路由时，可以根据报文中的用户信息（源IP地址）对流量进行负载分担。

对于广域网双链路，采用基于流的负载分担，或采用基于带宽的非平衡负载分担（多用于双专线带宽不相等的情况）对提升广域链路利用率有很大意义。但需要注意的是，基于带宽的负载分担不适合Internet出口做NAT情况下的分担，因为基于带宽的负载分担只能基于报文，会出现同一个流采用不同的公网源地址转发的情况。

应用实践设计

低时延低抖动QoS设计

设计描述

本设计为采用分层CAR技术实现的QoS队列调度设计，同时与传统CBQ队列调度技术进行实测性能比较，旨在为用户提供一种全新、简单、具有性能优势的QoS设计方案。

设计组网

单台SR66的单一物理接口（5*E1线路，10M）进行验证。

关键配置

基本业务流定义

```
acl number 3511 //video
rule 10 permit ip source 172.16.2.11 0.0.0.7

acl number 3521 //produce
rule 10 permit ip source 172.16.2.21 0.0.0.7

acl number 3531 //others
rule 10 permit ip source 172.16.2.31 0.0.0.7

acl number 3550 //all, include video/produce/others
rule 10 permit ip source 172.16.2.11 0.0.0.7
rule 20 permit ip source 172.16.2.21 0.0.0.7
```

rule 30 permit ip source 172.16.2.31 0.0.0.7

对比方案一：入接口分层CAR设计

interface Pos4/1/0

link-protocol ppp

ip address 100.1.7.2 255.255.255.0

qos car inbound acl 3511 cir 4096 green continue red discard

qos car inbound acl 3521 cir 4096 green continue red continue

qos car inbound acl 3531 cir 512 green continue red continue

qos car inbound acl 3550 cir 8704 green pass red discard

对比方案二：出接口CBQ设计

traffic classifier video

if-match acl 3511

traffic classifier produce

if-match acl 3521

traffic behavior video

queue ef bandwidth 4096 cbs 102400

traffic behavior produce

queue af bandwidth 4096

qos policy flow-assign

classifier produce behavior produce

classifier video behavior video

interface Mp-group4/0/0

qos reserved-bandwidth pct 100

qos apply policy flow-assign outbound

设计效果

视频流丢包率对比结果：

Video（Mbps）	Produce (Mbps)	Others (Mbps)	Video Packet Loss Ratio	
			分层CAR结果	CBQ结果
1.67	6	4	0%	0%
2.50	6	4	0%	0%
3.33	6	4	0%	0%
4.17	6	4	1.75%	4.59%
5.00	6	4	19.36%	20.59%

生产流丢包率对比结果：

Produce (Mbps)	Video（Mbps）	Others (Mbps)	Produce Packet Loss Ratio	
			分层CAR结果	CBQ结果
1.67	6	4	0%	0%
2.50	6	4	0%	0%
3.33	6	4	0%	0%
4.17	6	4	3.05%	2.57%
5.00	6	4	20.46%	18.82%

丢包率分析：从以上结果看，分层CAR和CBQ设计，对报文的丢包率都比较接近。

视频流时延对比结果：

Video（Mbps）	Produce (Mbps)	Others (Mbps)	Video Packet Latency（ms）	
			分层CAR结果	CBQ结果
1.67	6	4	1.74	6.43
2.50	6	4	1.83	6.57
3.33	6	4	2.08	6.74
4.17	6	4	4.86	6.93
5.00	6	4	6.38	6.94

生产流时延对比结果：

Produce (Mbps)	Video（Mbps）	Others (Mbps)	Produce Packet Latency（ms）	
			分层CAR结果	CBQ结果
1.67	6	4	9.05	7.81
2.50	6	4	6.42	8.00
3.33	6	4	4.02	8.58
4.17	6	4	3.99	18.98
5.00	6	4	5.77	33.50

时延和抖动分析：视频业务，分层CAR比CBQ的EF加速转发的调度时延偏小，但EF队列的平均时延比较稳定。生产业务，分层CAR相比CBQ AF确保转发的调度，时延和抖动都要明显小得多，尤其当生产业务流量超出AF确保带宽4M后，其时延明显放大，而分层CAR设计时延没有明显变化。

从对比结果看：分层CAR作为新型的流量调度技术，拥有与传统队列调度技术相似的调度效果，而且时延和抖动控制方面，尤其相对AF确保转发调度，具有明显的优势。在配置方面，分层CAR设计则更为简单、灵活。

带宽预留和共享下的QoS设计

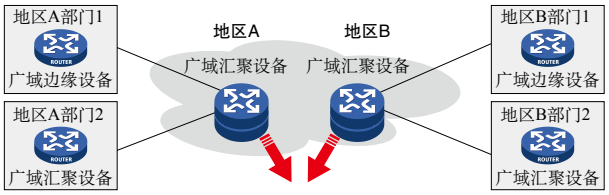
SDH和IP技术在带宽分配方面，前者适合对不同业务分配通道和带宽进行隔离，但各业务之间无法同时做到带宽共享；而IP技术则正相反，业务之间可以共享带宽，但很难实现带宽预留，而不让其他业务占用。对于某些企业的广域网建设，往往既需要各部门都拥有一定的预留带宽，用以保证本部门重要业务，任何时刻不能被其他部门占用；也需要部门之间能够对大部分带宽进行共享，保证总体的带宽利用效率；在此带宽预留和共享的设计之上，如果还要对部门内的不同业务进行良好的QoS设计，一般的IP QoS设计是难以做到的。我们通过下面这个分层CAR的设计来实现上述需求。

设计描述

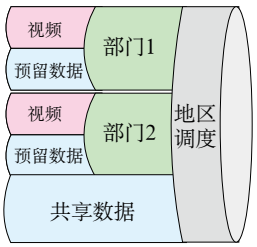
在地区A、B的核心路由器之间的155M链路上，流量调度设计要求如下：

- 为部门1、部门2重要业务预留一定带宽各为40M，超出40M的数据走75M共享数据队列进行共有带宽抢占；
- 各部门视频预留10M并优先转发，超出10M的部分进行丢弃；
- 为保证视频质量的稳定，要求10M视频流占用部门预留带宽，而绝对不能走共享带宽；
- 在部门预留带宽内，视频流量实时变化时，预留带宽内的数据流量可以自动随之调整带宽，使得实时视频流量+预留带宽内的数据流之和保持在40M（除非数据+视频总流量小于40M），而超出总带宽40M的数据流则进入两个部门的共享数据带宽，参与共享带宽的部门间抢占。

设计组网



多部门间共用广域物理线路



单接口部门间带宽独占和共享设计

关键配置

地区汇聚路由器		
入接口（2×GE，2个部门）	出接口（155M POS，2部门）	
普通CAR	分层CAR	队列调度
CAR（EXP5）10M Green pass，Red discard	CAR（EXP5、EXP2、EXP0）40M Green pass remark EXP2 Red pass remark EXP0	Q3（EXP2）CIR保证40M
CAR（EXP2）30M Green pass remark EXP2 Red pass remark EXP0		
CAR（EXP4）10M Green pass， Red discard	CAR（EXP4、EXP1、EXP0）40M Green pass remark EXP1 Red pass remark EXP0	Q2（EXP1）CIR保证40M
CAR（EXP1）30M Green pass remark EXP1 Red pass remark EXP0		
——	——	Q1（EXP 0）GTS上限75M

业务标记说明：通过入接口MPLS报文的优先级标签EXP5代表部门1的视频业务，EXP2代表部门1的数据业务；EXP4代表部门2的视频业务，EXP1代表部门2的数据业务。

流量调度说明：在汇聚路由器的针对部门1和部门2的两个入接口分别进行部门内流量分层CAR调度。实现视频业务优先占用部门预留带宽，富余的预留带宽给本部门数据业务，预留带宽内的视频业务和数据业务均标记为EXP2，而超出预留带宽的数据业务标记为EXP0。在汇聚路由器的155M POS出接口上，对两个部门超出预留带宽的数据业务进行总带宽限制为75M（155M-40M×2），这样就实现了部门的预留带宽，同时保证了预留带宽优先分配给视频业务，其余预留带宽给数据业务的设计。也就是说，带宽预留和共享设计，是通过入口CAR和出口GTS限流配合实现，而预留带宽优先分配给视频，其余带宽给数据的设计是通过分层CAR实现的。

设计效果

在实现企业部门之间带宽预留和共享，实现带宽独立和带宽利用率

双重需求的同时，又保证了部门内视频业务的带宽和服务质量，避免了视频业务走共享数据队列所带来的带宽抢占和丢包风险。

广域双链路流量调度设计

传统QoS的设计，都是基于单接口或链路的，当总业务流量超出该接口带宽后，就要进行业务间QoS调度，并对低优先级和超出承诺带宽的业务进行丢包处理。这种传统QoS技术针对双链路和多链路情况，却没有更好的方案进行跨链路的流量统一QoS调度，而只能做基本的业务分流承载，各自做QoS保障。其结果就是，一条链路可能严重拥塞，而另外一条链路却空闲的很，带宽资源严重浪费。

下面这个设计，就是针对这一现象进行双链路统一QoS流量调度设计，采用的分层CAR技术结合策略路由设计实现。

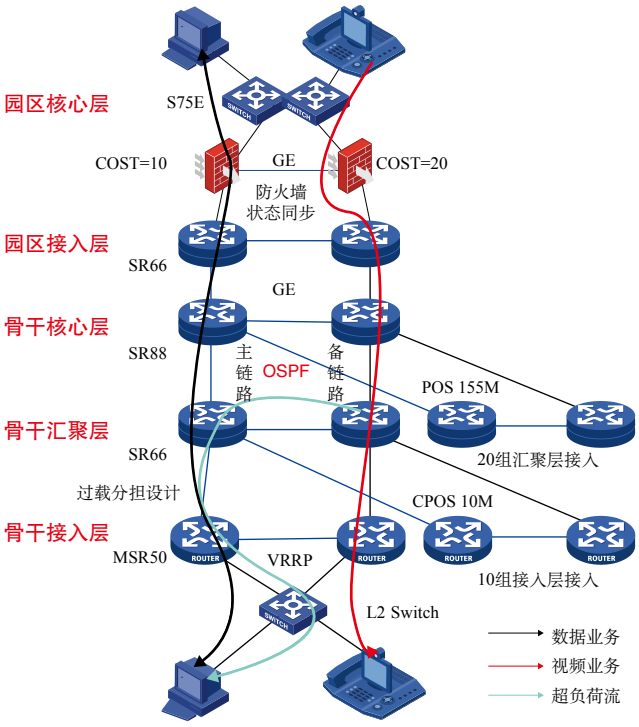
设计描述

业务说明：包括生产、办公和视频业务。一般情况下，通过路由设计，生产业务走主链路，办公和视频业务走备用链路。生产业务虽然平时流量不大，但最为重要，业务不能中断，因此需要尽力保证生产业务不丢包；办公业务重要程度在生产之后，流量较大，需要充分利用双链路空闲带宽；而视频业务偶尔会有，但需要保证带宽并优先转发，链路故障时不对视频业务进行链路切换。

业务流向：全部生产和主要办公业务为骨干各级到园区核心的大集中业务，园区存放服务器。视频业务任何跨层之间都有通信要求。同级各组之间不进行横向业务交流。因此，整体流量主要为纵向流量访问模型。

超负荷流量调度和优先级保证：主链路上，当生产业务下行方向超出10M后，多出10M的流量横向引流到同级备链路路由器，并参与该设备下行方向的QoS调度，进行超负荷流量调度；同样，备链路办公和视频业务总流量超出10M后，仅对办公业务超负荷流量向主链路路由器进行分担，并参与主链路路由器下行出口的QoS调度，而视频业务始终限制在2M之内，为保证视频质量的稳定性，不参与备链路的过载分担。

设计组网



双链路流量过载调度设计

测试组网说明：

- 该组网方式模拟行业用户中常见的广域网建设方式，分为核心园区网、分支园区网络和骨干网络；
- 核心园区网络内是三层网络，网关是SR88，三层设备包括SR66和S75E的三层交换机，在SR66和S75E之间部署防火墙；
- 在分支园区网络中是二层网络，里面是二层交换机；
- 骨干网络由各个层级园区网络的网关设备组成，骨干链路采用CPOS；
- 园区网络和骨干网络采用双链路上行，既保证了链路的可靠性，同时也达到了扩充链路资源的目的。

本设计以双链路分流承载和骨干汇聚层SR66下行方向过载分担设计为主要进行介绍。

关键配置

分流设计配置—骨干汇聚层主链路SR66

配置两个OSPF实例，上行和横向接口属于OSPF 100，下行接

口属于OSPF 200，就是说核心网络属于OSPF 100，而分支网络属于OSPF 200。

```
ospf 100
area 0.0.0.0
network 100.1.5.0 0.0.0.255
network 100.1.6.0 0.0.0.255

ospf 200
area 0.0.0.0
network 100.1.3.0 0.0.0.255

# 配置前缀列表，分别匹配分支的生产业务、办公业务和视频业务网段
ip ip-prefix branch-produce index 10 permit 172.20.0.0 16 less-equal 32
ip ip-prefix branch-office index 10 permit 172.25.0.0 16 less-equal 32
ip ip-prefix branch-video index 10 permit 172.30.0.0 16 less-equal 32

# 配置路由策略，匹配生产业务的路由，cost设定为10，匹配办公和视频业务的路由，cost设定为10000，当OSPF 100引入OSPF 200路由的时候，应用该路由策略。
route-policy branch-to-core permit node 1
if-match ip-prefix branch-produce
apply cost 10
apply cost-type type-1

route-policy branch-to-core permit node 5
if-match ip-prefix branch-office
apply cost 10000
apply cost-type type-1

route-policy branch-to-core permit node 10
if-match ip-prefix branch-video
apply cost 10000
apply cost-type type-1

ospf 100
import-route ospf 200 route-policy branch-to-core

# 配置前缀列表，分别匹配分支的生产业务、办公业务和视频业务网段
ip ip-prefix core-produce index 10 permit 172.16.20.0 24
ip ip-prefix core-office index 10 permit 172.16.25.0 24
ip ip-prefix core-video index 10 permit 172.16.30.0 24

# 配置路由策略，匹配生产业务的路由，cost设定为10，匹配办公和视频业务的路由，cost设定为10000，当OSPF 100引入OSPF 200路由的时候，应用该路由策略。
route-policy core-to-branch permit node 1
if-match ip-prefix core-produce
apply cost 10
apply cost-type type-1

route-policy core-to-branch permit node 5
if-match ip-prefix core-office

apply cost 10000
apply cost-type type-1

route-policy core-to-branch permit node 10
if-match ip-prefix core-video
apply cost 10000
apply cost-type type-1

ospf 200
import-route ospf 100 route-policy core-to-branch
```

分流设计配置—骨干汇聚层备链路SR66

配置基本同主链路SR66，只是在两个OSPF相互引入路由的时候，匹配生产业务路由的cost值为10000，而匹配视频和办公业务的cost的值为10。这样保证了生产业务走左边主链路，而视频和办公业务走右边备份链路。

过载分 担设计配置—骨干汇聚层主链路SR66

```
# 匹配生产业务的ACL
acl number 3100
rule 10 permit ip source 172.16.20.0 0.0.0.255 dest 172.20.1.0
```

0.0.0.255

匹配af11（过载流量标记为af11）的ACL
acl number 3200
rule 0 permit ip dscp af11

配置策略路由，如果是过载流量，将走到R4横向链路
policy-based-route exceed permit node 1
if-match acl 3200
apply ip-address next-hop 100.1.5.2

接口下配置QoS car，承诺速率为9M，超过9M的流量标记为af11，通过策略路由，走横向链路。这里设置为9M而不是10M，主要考虑链路其他流量的存在对下行10M接口的影响，比如路由协议和控制流量，因此当生产业务超过9M时，就需要进行过载分担了。
interface pos 2/1/0
qos car inbound acl 3100 cir 9000 green pass red remark-dscp-continue af11
ip policy-based-route exceed

过载分担设计配置—骨干汇聚层备链路SR66
匹配办公业务的ACL
acl number 3025
rule 10 permit ip source 172.16.25.0 0.0.0.255 dest 172.25.1.0 0.0.0.255

匹配视频业务的ACL
acl number 3030
rule 10 permit ip source 172.16.30.0 0.0.0.255 dest 172.30.1.0 0.0.0.255

办公和视频业务业务都匹配的ACL
acl number 3035
rule 10 permit ip source 172.16.25.0 0.0.0.255 dest 172.25.1.0 0.0.0.255
rule 20 permit ip source 172.16.30.0 0.0.0.255 dest 172.30.1.0 0.0.0.255

匹配af12（过载流量标记为af12）的ACL
acl number 3200
rule 0 permit ip dscp af12

过载流量走到主链路SR66路由器的横向链路
policy-based-route exceed permit node 1
if-match acl 3200
apply ip-address next-hop 100.1.5.1

下行方向入接口配置QoS，如果视频流量超过设定值，直接丢弃；如果办公业务超过设定值，如果视频业务有富余带宽，可以占用。如果视频和生产业务总流量超过设定，通过策略路由，走横向链路到主链路SR66进行转发。
interface Pos4/1/0
qos car inbound acl 3030 cir 4096 green continue red discard
qos car inbound acl 3025 cir 4096 green continue red continue
qos car inbound acl 3035 cir 8192 green pass red remark-dscp-continue af12
ip policy-based-route exceed

注: 以上配置针对一个分支，多个分支的情况下，需要为每个分支配置匹配生产业务的ACL，接口下为每个分支配置QoS car。

下行出口QoS队列调度配置—骨干汇聚层主链路SR66
配置QoS策略，使得生产业务占用AF队列，60%的带宽，办公业务也是AF队列，占用20%的带宽。
traffic classifier produce
if-match acl 3100

traffic behavior produce
queue af bandwidth pct 60

traffic classifier office
if-match acl 3025

traffic behavior office
queue af bandwidth pct 20

```
qos policy flow-assign
classifier produce behavior produce
classifier office behavior office

# 将QoS策略应用在下行出口
interface Mp-group2/0/0
qos reserved-bandwidth pct 100
qos apply policy flow-assign outbound
```

下行出口QoS队列调度配置—骨干汇聚层备链路SR66

```
# 配置QoS策略，视频业务进入ef队列，占用带宽的20%，生产和办公业务占用af队列，分别占用20%和40%的带宽。
traffic classifier produce
 if-match acl 3020

traffic behavior produce
queue af bandwidth pct 20

traffic classifier office
if-match acl 3025

traffic behavior office
queue af bandwidth pct 40

traffic classifier video
if-match acl 3030

traffic behavior video
queue ef bandwidth pct 20

qos policy flow-assign
classifier produce behavior produce
classifier office behavior office
classifier video  behavior video

# 将QoS策略应用在下行出口
interface Mp-group2/0/0
qos reserved-bandwidth pct 100
qos apply policy flow-assign outbound
```

设计效果

双链路统一QoS流量调度，需要统一考虑主备链路过载分担时，对哪些业务进行横向调度，同时也要考虑过载分担过来的流量与本链路原有流量的统一QoS调度设计，避免过载流量对本链路流量进行较大的冲击。

过载分流设计，相比于传统单链路QoS和基本业务分流的组合方案，在带宽利用率方面和QoS带宽保证和优先级设计方面都有比较好的提升。而相对于传统的负载分担设计，最大的优势是尽量保持按设计链路转发，减少转发环节，而且可以做精细的QoS保证设计，比如本设计中的视频业务优先，而不做过载分担这些细节。

广域出口终端带宽控制

企业园区或分支的广域出口带宽相对有限，而园区或企业分支内的终端用户一般都在几十、成百上千的规模，传统的QoS队列设计根本无法做到针对每一个终端进行精确的带宽控制。这样就会出现由于某个人下载大附件导致其他人上网速度慢的情况出现，也有可能影响到其他的重要业务，比如视频会议。通过基于在线IP的动态令牌桶技术，与分层CAR结合，则可以实现上述理想：做到基于终端IP的带宽公平分配和共享，同时对重要业务进行带宽保证和优先转发。

设计描述

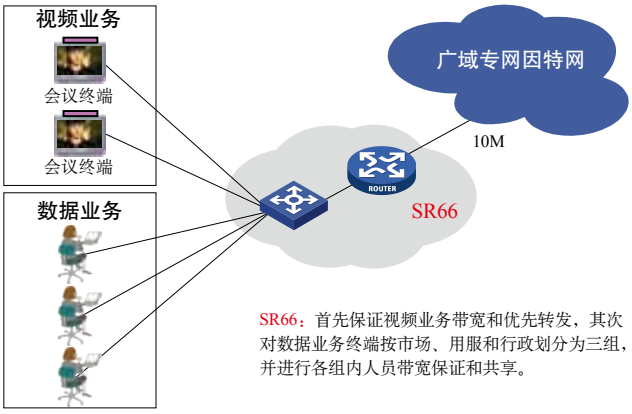
小型园区广域接入，需要考虑园区内的用户和业务分组，及组内/组间的带宽分配。

园区内对业务进行分组：视频业务和数据业务类。其中数据类业务又分为市场/用服/行政三组人员。

在园区出口的入方向（下行流量）进行带宽保证设计，各组内按单用户IP地址进行带宽保证。

当视频不用带宽时，可以留给数据业务组使用。即使数据业务存在富余带宽的时候，视频业务也不能占用。数据业务中，当某个组有富余带宽的时候，可以留给其他部门使用。

设计组网



广域出口终端带宽控制

一个小型园区内，按业务分，包括视频业务和数据业务，其中数据业务的终端用户有200人左右，按照部分划分，包括市场、用服和行政三组人员，各组人员各64人。

关键配置

SR66广域接口的下行方向进行带宽控制配置

配置car list，指定视频地址范围

qos carl 2 destination-ip-address range 192.170.1.1 to 192.170.1.63

配置car list，指定市场组的地址范围，每个在线IP平均分配带宽，并进行富余带宽共享

qos carl 4 destination-ip-address range 192.170.1.64 to 192.170.1.127 per-address shared-bandwidth

配置car list，指定用服组的地址范围，每个在线IP平均分配带宽，并进行富余带宽共享

qos carl 6 destination-ip-address range 192.170.1.128 to 192.170.1.191 per-address shared-bandwidth

配置car list，指定行政组的地址范围，每个在线IP平均分配带宽，并进行富余带宽共享

qos carl 8 destination-ip-address range 192.170.1.192 to 192.170.1.254 per-address shared-bandwidth

配置car list，包括所有的地址范围

qos carl 10 destination-ip-address range 192.170.1.1 to 192.170.1.254

在接口下配置分层CAR，在配置第一层CAR的时候，因为视频业务中超过cir的流量直接discard，因此即使企业业务由富余带宽，也不会抢占。而市场、用服、行政超过cir的流量动作是continue，通过第二层CAR，会进行二次处理，这样当其它业务有富余带宽的时候，会占用。

interface Mp-group4/0/0

qos car inbound carl 2 cir 3000 green continue red discard // 视频

qos car inbound carl 4 cir 2000 green continue red continue // 市场

qos car inbound carl 6 cir 2000 green continue red continue // 用服

qos car inbound carl 8 cir 2000 green continue red continue // 行政

qos car inbound carl 10 cir 9000 green pass red discard // 所有：视频和人员数据

本配置中是通过CARL命令字以及关键参数per-address和shared-bandwidth进行体现的，采用在线动态令牌桶生成技术与分层CAR结合的技术方案。

设计效果

上述设计中，当视频流量在3M以内，视频业务会得到优先带宽保证；而其余带宽则按照每组人员共2M，对内在在线用户进行带宽公平分配，比如市场人员某一时刻在线人员为10人，那么每一个市场人员都可以得到200Kbps的带宽。而另一时刻，市场人员有20人在线，则每个人可以得到的带宽则为100Kbps。其他组人员带宽分配方案与市场人员相同。

如果视频业务流量低于3M，比如为0时，那么视频业务的3M带宽就会被市场、用服和行政人员进行公平分配，而不会有带宽浪费现象出现。

需要注意的是，由于广域网带宽往往由电信运营商进行限定，比如这里限定为10M，流量的丢包控制往往在运营商侧。因此在园区入口进行该方案设计时，总体带宽建议少于运营商给定带宽的

5%～10%左右，这样就为内部的带宽分配提供了控制空间，通过对超额定带宽的业务流进行丢包，达到TCP自动降速或UDP承载的应用层自动降速的目的，实现带宽的公平分配；否则按10M设计，由于进来的流量肯定在10M以内，业务流的丢包控制完全放在了运营商侧，以上设计就会失效。当然，园区上行流量设计则不需要注意该问题。

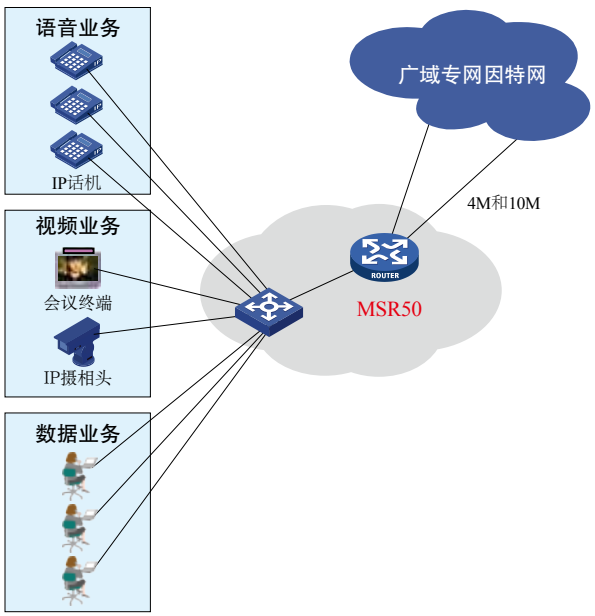
广域专线出口负载分担

对于拥有广域双出口的分支或园区，为达到广域链路资源的充分利用，往往采用负载分担技术。当由于广域网链路带宽可能存在的非对称性，即两条链路带宽不一样时，传统的负载分担技术则存在困难。下面介绍一种基于双出口带宽比例的非对称负载分担设计，则可以帮助企业自由的实现各种比例的负载分担效果。

设计描述

下图MSR50同时拥有4M和10M的双上行链路。设计上行流量基于链路带宽按包进行负载分担。

设计组网



广域出口方向负载分担

关键配置

```
# 配置默认路由，两个出接口是等价路由
ip route-static 0.0.0.0 0.0.0.0 100.1.1.1
ip route-static 0.0.0.0 0.0.0.0 100.1.2.1

# 在两个出接口下关闭快转
interface Mp-group0
undo ip fast-forwarding outbound

interface Serial5/0
undo ip fast-forwarding outbound

# 配置基于端口带宽的负载分担
bandwidth-based-sharing
```

设计效果

上行流量按双出口带宽比例进行逐包负载分担发送。但这种逐包负载分担设计，不适合双Internent经过NAT处理的出口，因为双出口意味着不同的Internet公网地址，而逐包则会导致同一个会话用不同的公网地址转发出去，会话不能维持。

总结

企业广域网链路资源的相对稀缺，以及业务集中承载的趋势，对广域网的业务流量调度和QoS保证提出了更为严格的挑战。本文基于H3C在广域网流量调度领域的创新性设计，对企业广域网的一些典型场景和环节进行了详细阐述。这些场景已经在政府、金融、电力和大企业等各个行业落地，且都是一些行业特色需求，比如带宽预留、超负载分担、数据业务为视频会议让路、统一带宽管理、办事处员工外网带宽保证等要求。

《网络大爬虫》

Network Addicts

《网络大爬虫》是H3C面向H³Care俱乐部VIP会员技术爱好者的专业性技术刊物。本刊主要深入探讨IP相关技术，内容涉及交换、OSPF路由协议、QoS、MPLS VPN等多个技术领域。本刊的文章由H3C相关领域技术专家倾力撰文。如有建议或需求，请您反馈至电子信箱：H³Care_club@h3c.com，感谢您的阅读！

华三服务以满足客户业务需求为导向，依托广博精深的技术实力、完备高效的交付体系，贴近客户，持续进行服务创新，为客户提供温暖感、专业化的服务体验，帮助客户实现 IT 价值的可持续性再生。