提供有界延迟服务的网络结构*

孙利民 窦文华 周兴铭

(国防科技大学计算机系 长沙 410073)

摘 要 将来的计算机网络必须支持具有不同通信量和不同服务质量(QoS)要求的应用,有界延迟服务保证所有应用包的延迟都不超过给定延迟上界。本文首先提出有界延迟实时服务网络的框架,详细说明了其关键部件的功能和工作原理,并阐述了分析它们性能的技术,最后讨论了设计有界延迟服务网络在性能与实现复杂性之间的折衷。

关键词 有界延迟服务,网络结构,确定性通信量模型,包调度原理,连接准入控制,通信量监控
分类号 TP393

The Architecture of Network with a Bounded Delay Service

Sun Limin Dou Wenhua Zhou Xingming

(Department of Computer Science, National University of Defense Technology, Changsha, 410073)

Abstract Future integrated-services networks are expected to support applications with a wide range of service requirements. Bounded delay service supports deterministic guarantees on maximum delay for connections. This paper first presents the architecture of network with a bounded delay service and introduces the functions and work principles of the key components of the network in tail. Various issues and tradeoffs in designing bounded delay network are presented finally.

Key words bounded delay service, network architecture, deterministic tuaffic model, packet scheduling, admission control, traffic policing

随着高速计算机网络的飞速发展,网络支持具有不同通信量和不同服务质量(QoS)要求的应用成为可能。为此,Internet提出了预测(predictive)服务和保证(guaranteed)服务[1,2],预测服务基于对网络性能的测量估计已有网络的通信量,提供应用可伸缩的网络性能;保证服务分为确定性服务和统计性服务,确定性服务基于应用的最大通信量说明,保证所有包在给定延迟内传输到目的,统计性服务在给定延迟内传输一定概率的包到达目的。ATM论坛定义了五种通信量模式和服务策略:固定位速率(CBR)、实时可变位速率(rt-VBR)、非实时可变位速率(nrt-VBR)、可用位速率(ABR)和不定位速率(UBR),其中CBR和rt-VBR用于实时服务类型。

网络服务质量(QoS)参数主要包括网络延迟、延迟抖动、吞吐量和丢失率,网络延迟是实时服务网络最重要的参数。Internet和ATM都包含了实时服务类型,我们称实时服务类型中保证所有包的延迟都不超过给定延迟上界的服务为有界延迟服务。本文主要讨论分组交换网络中提供有界延迟服务的框架,首先给出了网络结构的关键部件,说明它们的功能和工作原理,并阐述了分析它们性能的技术,最后讨论设计有界延迟服务网络在效率和实现复杂度之间的折衷。

1 提供有界延迟服务的网络结构[2,3]

假设在分组交换网络中,链路和交换机为任意拓扑结构,包在链路上的传输延迟是有界的,交换

* 1997年12月2日收稿
国防预研基金资助项目
第一作者:孙利民,男,1966年生,博士生

机非阻塞，即包到达输入链路时，直接路由到相应的输出链路。包在交换机内没有交换冲突发生，到不同输出链路的包互不干扰，在交换机输出端口排队输出。

网络提供有界延迟服务采用如下模式。在通信前，用户向网络说明连接的通信量特性以及关于延迟、延迟抖动等性能要求，如果有充分的资源保证满足已有连接和请求连接的性能要求，网络才接受该连接；在通信中，连接发送包的速度不能超过说明的通信量，否则网络不能保证提供有界延迟服务。有界延迟服务网络必须面向连接（connection- oriented），通过资源预约机制限制连接的个数和通信量，保证连接的 QoS 性能。资源预约机制是网络设计的关键。这是由于若分配太少的资源，连接不能获得希望的 QoS，若分配连接太多的资源，会导致网络利用率的降低。

网络分配资源基于连接的通信量和 QoS 要求。有界延迟服务需要确定性通信量模型(deterministic traffic model) 描述连接的最大通信量。连接建立前，在路由的每个结点上都要进行准入控制(admission control) 测试，判定网络能否在不降低已有连接性能的前提下满足新的性能要求。如果有一个结点不能同时满足新的连接和已有连接的性能要求，则网络拒绝接收新的连接。准入控制测试的基础是通信量模型和包调度（packet scheduling），包调度决定包在输出链路上的服务顺序，不同的包调度原理对于不同的准入控制测试条件。网络接收连接后要执行通信量监管 (traffic policing)，使得连接进入网络的通信量符合建立前的通信量说明，确保所有连接的网络性能。图 1 所示为提供有界延迟服务的网络框架，发送者与接收者之间的箭头指示路由经过的链路和交换机。

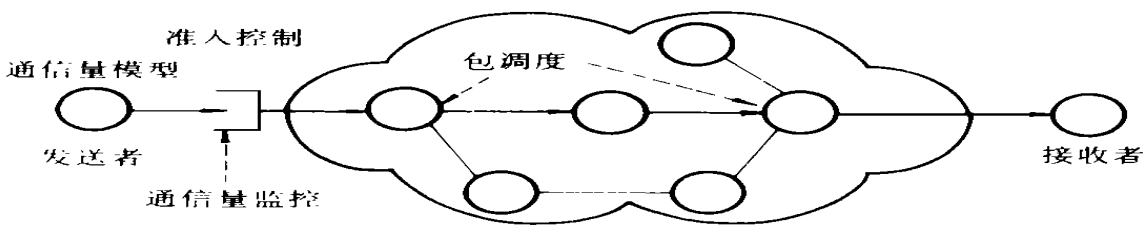

图 1 提供有界延迟服务的网络结构

2 确定性通信量模型

有界延迟服务的通信量模型必须是确定性通信量模型。好的确定性通信量模型应具有以下几个基本特点[4,5]：必须描述信源的最大通信量特性；必须为参数化的模型，使得信源可以向网络高效地说明它的通信量特性；一定要尽量精确地表示信源的通信量特性，使得准入控制不会过分估计需要的网络资源；必须容易监控，以便网络可以实时监视信源的实际通信量。

假设函数 A 表示连接的实际通信量，$A[\tau, \tau+t]$ 表示在时间间隔 $[\tau, \tau+t]$ 内到达的通信量。函数 A 的一个上界 A^* 称为通信量约束函数 (traffic constraint function)[6,7]，如果它满足两个重要特性：(时间无关性和叠加性)。通信量约束函数 A^* 的时间无关性指对于任意的 $\tau \geq 0$ 和 $t \geq 0$，均满足：

$$A[\tau, \tau+t] \leq A^*(t) \tag{1}$$

通信量约束函数的时间无关性使得延迟边界测试独立于连接的开始时间。通信量约束函数 A^* 的叠加性指对于任意的 t_1，$t_2 \geq 0$，均满足：

$$A^*(t_1) + A^*(t_2) \geq A^*(t_1 + t_2) \tag{2}$$

一个给定连接可能有无限多个通信量约束函数，确定性通信量模型只是其中的一个参数化族。由于所有确定性通信量模型都有对应的参数化约束函数，我们通过比较约束函数来比较不同模型的精确性。对于给定连接的到达过程 $A[0, t]$，在任意时间间隔为 t 的到达流中，最精确的时间无关的到达流上界为

$$E^*(t) = \sup_{\tau \geq 0} A[\tau, \tau+t] \tag{3}$$

$E^*(t)$ 称为经验封装 (empirical envelop)，它是到达序列 $A[0, t]$ 最精确的时间无关的确定性通信

量函数。对于任意的通信量约束函数 A^*，都满足 $A^*(t) \geq E^*(t)$.

对于给定的包调度原理，在使用精确的准入控制和经验封装时网络利用率最高。虽然经验封装缺少使用特性，不能高效说明或监控，但它可以作为评价其它确定性通信量模型的标准。通信量模型的约束函数越接近经验封装 $E^*(t)$，网络利用率就越高。

几个主要的通信量模型为: (1) 峰值速率 peak- rate 模型; (2) (r, T) - 模型; (3) (σ, ρ) - 模型; (4) $(\vec{\sigma}, \vec{\rho})$ - 模型; (5) $(X_{\min}, X_{ave}, I, S^{max})$ - 模型; (6) $D-BIND$ 模型。它们的约束函数参见表 1。

表 1 确定性通信量模型的监控机制

通信量模型	监控机制	通信量约束函数
峰值速率	包分隔	$A^*(t) = (\frac{t}{X_{\min}} + 1) s^{max}$
(r, T)	跳动窗口	$A^*(t) = (\frac{t}{T} + 1) rT$
(σ, ρ)	漏 桶	$A^*(t) = \sigma + \rho t$
$(\vec{\sigma}, \vec{\rho})$	多个漏桶	$A^*(t) = \max_{1 \leq i \leq m} \{\sigma_i + \rho_i t\}$
$(X_{\min}, X_{ave}, I, s^{max})$	移动窗口和包分隔	$A^*(t) = \frac{t}{T} \frac{I \cdot s^{max}}{X_{ave}} + \min \{ (\frac{t}{T} - \frac{t}{T}) \frac{I}{X_{\min}}, \frac{I}{X_{ave}} \} s^{max}$
$D-BIND$	多个移动窗口	$A^*(t) = \frac{R_i I_i - R_{i-1} I_{i-1}}{I_i - I_{i-1}} + R_i I_i$

3 包调度原理[4, 8]

在交换机同一输出链路上不同连接的包相互作用，如果不进行适当控制，这些相互作用可严重影响网络性能。包调度决定不同连接包的发送顺序，同一个连接的包按先来先服务的方式服务。包调度是网络提供有界延迟服务的核心部分。目前通常采用 FCFS (First Come First Served) 调度原理，它虽然实现简单，但提供所有连接相同的延迟边界，包的发送顺序完全由它们的到达顺序决定，灵活性差且网络资源利用率低。一个性能良好的包调度原理应满足下列标准: (1) 高效: 网络通过准入控制限制接收的个数和通信量保证延迟边界，高效的调度原理支持更多的延迟有界的连接; (2) 灵活: 调度原理要能满足各种应用的不同通信量的不同延迟要求; (3) 低复杂度: 操作简单以便于高速实现; (4) 防火特性: 保护行为良好的用户 (按通信量说明发送信息的用户) 获得应有的保证服务，不应受到网络负载波动、不良行为用户及 best- effort 通信量的影响; (4) 可分析性强: 调度条件是准入控制的核心，它验证包的最大延迟是否超过连接的延迟边界。如果调度条件不精确，准入控制将没有必要地限制网络可支持的连接数。

一个包调度原理不可能优化上述所有性能，特别是调度的高效与低复杂度相互矛盾，每个调度原理根据特定的条件，在满足上述要求时采用了折衷方案。我们将包调度分为基于延迟的调度和基于速率的调度。基于延迟的调度赋予每个连接延迟上界，根据其延迟约束优先级排序到达的包来保证连接的延迟上界。基于速率的调度提供连接最小吞吐率，给每个连接分配整个带宽的一部份，根据连接的通信量计算延迟上界。

3.1 基于速率的调度原理

所有基于速率的调度原理模拟以下两个系统: (1) 时分多路复用系统 (TDM): 把时间分为固定大小的帧，每个帧分成时间槽，分配时间槽给每个连接; (2) 一般处理机共享系统 (GPS): 分配服务共享因子给每个连接，提供给连接的服务正比于它的共享因子。

停走 (Stop- and- Go)[9] 调度、层次轮询 HRR[10] (Hierarchical- Round- Robin) 调度和虚时钟

VC[11] (Virtual Clock) 调度模拟 TDM 系统, 连接间相互隔离, 每个连接分配固定的带宽。Stop- and - Go 和 HRR 调度采用分帧机制, 在一帧期间到达的包排队等待在下一帧发送, 不同连接的包在一帧内的发送顺序为任意的。分帧机制的缺点类似于 TDM, 若连接没有用完分配的带宽, 剩余的带宽会浪费掉。VC 调度采用统计复用技术, 每个包赋予虚拟完成时间, 虚拟完成时间是包若在 TDM 系统中的完成时间, 包按虚拟完成时间增加的顺序发送。加权公平排队 WFQ (Weighted- Fair- Queueing)[14] 调度如 VC 给每个包赋予虚拟完成时间, 但它模拟 GPS 系统。

(1) VC 调度原理

假设分配连接 i 的服务速率为 r_i, 第 k 个包长度为 $L_{k,j}$, 到达时间为 $t_{k,j}$, 它的虚拟完成时间为 $F_{k,j}$, 那么:

$$F_{1,i} = t_{1,i} + \frac{L_{1,i}}{r_i} \tag{4}$$

$$F_{k,i} = \max\{t_{k,i}, F_{k-1,i}\} + \frac{L_{k,i}}{r_i} \qquad K > 1; \tag{5}$$

包的虚拟完成时间 $F_{k,i}$ 等同于连接在速率为 r_i 的专用参考服务器上的完成时间。到达的包根据它的虚拟完成时间插入排序队列中, 排序复杂性为 O (LogN), N 为队列中的包个数。VC 调度没有分帧调度的带宽浪费, 但存在惩罚利用剩余带宽连接的问题。VC 调度的两个扩展是及时离开 (Liave- in- Time)[12] 调度和突发调度 (Burst Scheduling)[13]。

(2) WFS 调度原理

WFS 调度又称为 PGPS 调度 (Packet-by-packet Generalized Processor Sharing), 它模拟 GPS 系统。假设 GPS 系统有 N 个连接, 输出链路的发送速率为 1, 连接 i 赋予服务共享因子 ϕ_i, 服务器在时间 t 提供连接 i 的服务速率为 $\frac{\phi_i}{\sum_{j\in B(t)}\phi_j}$, 其中 $B(t) \subset N$ 表示在时间 t 有等待包发送的连接集合。因此, GPS 给有包发送的连接 i 分配带宽正比于它们的服务共享因子 ϕ_i。最坏情况下 $B(t) = N$, 每个连接的最小保证速率为 $g_i = \frac{\phi_i}{\sum_{j\in N}\phi_j}$, 但 GPS 实际上不可能实现, 因为它同时服务所有非空连接, 实际的包作为整体发送而非逐位的复用。

WPS 调度用 VC 模拟 TDM 相同的方法模拟 GPS 系统。假设分配连接 i 的服务共享因子 ϕ_i, 连接 i 的第 k 个包在时间 $t_{k,i}$ 到达, 包的发送时间为 s, 它的虚拟完成时间为:

$$F_{1,i} = t_{1,i} \tag{6}$$

$$F_{k,i} = \max\{t_{k,i}, F_{k-1,i}\} + \frac{s\sum_{j\in B(t)}\phi_j}{\phi_i} \qquad i > 1 \tag{7}$$

比较公式 (7) 和 (5), WPS 调度的虚拟完成时间依赖于有等待包的连接, 因此需要跟踪连接的活动性, 实现上比 VC 调度复杂, 但它提供良好的公平性。WPS 调度具有如下性质: 若包在时间 t 离开 GPS, 它离开 WPS 不迟于 $t + s^{max}$, s^{max} 是系统中最大包的发送时间。WPS 调度的两个扩展是 SCFQ 调度 (Self- Clocked Fair Queueing)[15] 和 WF^2Q 调度 (Worst- case Fair Weighted Fair Queueing)[16]

3.2 基于延迟的调度原理

(1) EDF (Earlist- Deadline- First) 调度[17]

EDF 调度给每个到达包赋予调度期限 (deadline), 即到达时间与延迟边界之和, 选择最小调度期限的包发送, 正在发送的包不能被中断。EDF 需要维持按包调度期限递增顺序排列的发送队列, 当新的包到达时, 首先计算包的调度期限, 然后查找在队列中的正确位置并插入队列。EDF 调度是优化的调度: 如果任何其它包调度可调度一组延迟约束的连接, EDF 调度也可调度该组连接。EDF 调度有两个扩展: Delay- EDD 调度[17] 和 Jitter- EDD 调度[18]。

(2) SP (Static- Priority) 调度[19]

EDF 调度因查找排序队列而操作复杂性高, 为降低复杂度提出了 SP 调度。它将连接集 C 分成 P

个子集 $\{C_p\}$ $1 \leq p \leq P$，子集 C_p 中的所有连接具有相同的延迟边界 d_p，$d_q < d_q$，$P < q$. SP 调度维持 P 个具有优先级的 FIFO 队列，FIFO 1，FIFO 2，…FIFO P，FIFO 1 的优先级最高，FIFO P 的优先级最低，优先级高的延迟边界小。子集 C_p 中所有连接的包排到FIFO P 队列中，SP 调度选择优先级最高的非空队列的包发送。SP 调度只有少量的固定的操作，性能介于 FCFS 调度和 EDF 调度之间。

4 准入控制

为了提供有界延迟服务，网络必须根据最坏情况进行预约资源，限制连接的个数和每个连接的通信量。有界延迟服务的准入控制，要根据调度原理决定连接之间在网络内相互作用的最坏情况，结合确定性通信量模型来判定网络是否能够保证已有连接和请求连接的延迟上界。仅当所有路由结点中请求连接和已有连接的性能要求都满足时，网络才接收新的连接。

准入控制测试的内容包括传输能力、CPU 速度、buffer 空间等，有界延迟服务最重要的准入控制测试是延迟边界测试。延迟边界测试验证所有连接包的延迟是否大于给定的延迟上界，其它的测试如 buffer 空间都可从延迟边界测试中直接推导出来。

4.1 延迟分析[20, 21]

准入控制条件依赖于延迟分析技术，图 2 给出延迟分析的几个主要概念。水平轴是时间，垂直轴是信息位，上一条曲线表示在时间 t 时到达输出链路发送队列的总位数，下一条曲线表示到时间 t 时已经发送的总位数，两条曲线之差是队列中正在排队等待发送的位数，称为积压 (backlog) 函数。当积压函数为 0，即两条曲线相遇时，队列中没有信息表示忙周期结束。分析的关键在于如果上一条曲线是确定性有界曲线，最大延迟如何表示为两个曲线的函数。例如，最大忙周期提供任何连续型 (work-conserving) 调度一个延迟上界，最大的积压值除以链路速度提供 FCFS 调度的延迟上界。

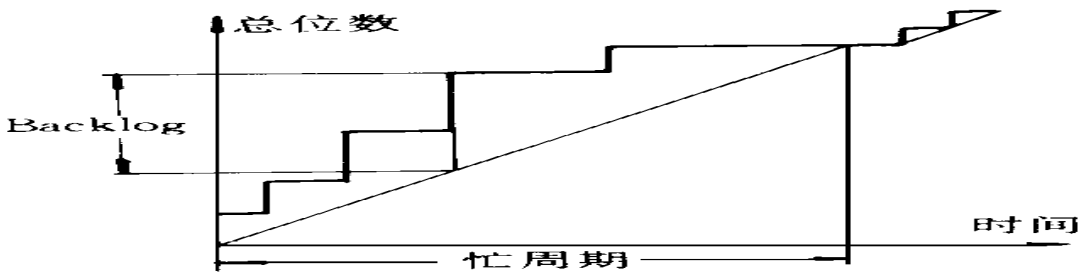

图 2 概念：延迟，忙周期和积压函数

通信量约束函数 $A_i^*(t)$ 给出了时间间隔为 t 的通信量上界，各个信源约束函数的会聚形成图 2 的上一条曲线。对于 FCFS 调度，连接 $i = 1, \cdots, n$ 的约束函数为 $A_i^*(t)$，链路速率为 l，最大包长度为 s，所有连接的延迟上界为：

$$d = \frac{1}{l} \max_{t \geq 0} \{ \sum_{i=1}^{n} A_i^*(t) - lt + s \} \tag{8}$$

4.2 延迟边界的测试条件[5, 6]

基于延迟的调度原理存在延迟边界测试。表 2 给出了 EDF 调度和SP 调度延迟边界测试的充分必要条件。调度的充分条件测试复杂度低，但可能拒绝网络能够保证性能的连接。延迟边界测试的精确性和复杂性之间存在折衷。

5 通信量监控[1, 4]

网络通过准入控制限制连接的个数和通信量，提供连接有界延迟服务。如果连接发送的包超过它说明的通信量，不仅它本身的延迟得不到保证，而且影响其它连接的性能。网络应该保护请求有界延迟服务的用户免受其它用户的影响，需要监控每个信源的通信量，保证它们发送的通信量不超过说明的通信量。这种在网络边沿的访问控制功能称为通信量监控，如图 3 所示。输入到监控器的通信量来

自信源，监控器的输出到网络。监控器确保它输出到网络的通信量满足信源参数模型说明的通信量约束函数。当输入到监控器的通信量超过约束函数规定的限制时，监控器缓存或丢失多余包。几个常用确定性通信量模型相应的通信量监控机制如表 1 所示。

表 2 延迟边界测试的充分必要条件

调度原理	准入控制条件-延迟边界测试
EDF	对所有的 $t \geq d_1$，有 $t \geq \sum_{i \in N} A_i^*(t - d_i) + \max_{k, d_k > t} s_k^{\max}$
SP	对所有的 p，$t \geq 0$，$\exists\ \tau \leq d_p - s_p^{\min}$ 使得： $t + \tau \geq \sum_{j \in C_p} A_j^*(t) + \sum_{q=1}^{p-1} \sum_{j \in C_q} A_j^*(t + \tau - s_p^{\min}) + \max_{r > p} s_r$

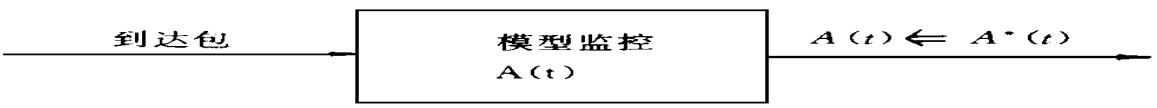

图 3 通信量约束函数 $A^*(t)$ 的监控器

6 结束语

传统电话网、有线电视网传输单一信号。QoS 网络提供集成服务功能，同时传输各种通信量的不同服务质量要求的信息，在保证服务质量的前提下，主要研究问题是提高网络的利用率。有界延迟服务网络的关键部件是确定性通信量模型、包调度原理、连接准入控制及通信量监控，存在着实现的简单性和高的网络利用率之间的折衷问题，如确定性通信量模型的简洁性、易监控与精确性之间的折衷，包调度原理的实现简单性、调度条件的易分析性与高效性之间的折衷。设计有界延迟服务网络时，要结合不同应用特点提出高效的通信量模型或新的调度原理，根据具体条件处理好各种折衷，从整体上提高网络利用率。

参考文献

1 Shenker D. Fundamental design issues for the future intemet. IEEE J. on selected area in communication, 1995, 13(7): 1176–1187

2 Clark D, et al. Supporting real-time application in an integrated services packet network: architecture and mechanism, ACM SIGCOMM´92, Baltimore, USA, Aug. 1992

3 Campbell A, et al. A quality of service architecture. In proc. ACM SIGCOMM´94, Aug. 1994

4 Georgiadis L, et al. Effcient network QoS provisioning based on per node traffic shaping. In proc. IEEE INFOCOM´96, Mar. 1996

5 Liebeherr J, et al. Exact admission control for networks with a bounded delay service. IEEE/ACM transactions on networking, 1996, 4(6) 885-9017

6 Wrege E, et al. Deterministic delay bounds for VBR video in packet-switching networks: fundamental limits and practical tradeoffs. IEEE/ACM transactions on networking, Jun, 1996, 14(3): 352– 362

7 Knightly W. and Zhang H. D-BIND: An accurate traffic model for providing QoS guarantees to VBR traffic. IEEE/ACM transactions on networking, Jun. 1996, 15(2): 219-231

8 Zhang H. Service disciplines for guaranteed perfoumance service in packet-switching networks. In proc. of the IEEE, 1995, 83(10): 1373-1396

9 Golestani S. J. A stop-and-go queueing framework for congestion management. In proc. ACM SIGCOMM´90, Sep. 1990

10 Kalmanek C. R, et al. Rate controlled servers for very high-speed network, In IEEE Global Telecommunications Conference, Dec. 1990

11 Zhang L. VirtualClock: Ynew traffic control algorithm for packet switching networks. In proc. ACM SIGCOMM´90, Sept. 1990

12 Figueira R. and Pasquale J. Leave-in-Time: a new service discipline for real-time communications in a pacdet-switching network. In proc. ACM SIGCOMM'95, 1995

13 Lam S. and Xie G. Burst scheduling networks. In proc. IEEE INFOCOM'95, Apr. 1995

14 Parekh A. K. and Gallager R. G. Ageneralized processor sharing approach to flow control in integrated services networks: The singlen-ode case. IEEE/ACM transactions on networking, 1993, 1 (3) 344-357

15 Golestani S. J. A self-clocked fair queueing scheme for broadband applications. In proc. IEEE INFOCOM'94, Jun. 1994

16 Bennett J. and Zhang H. WF^2Q: Worst-case fair weighted fair queueing. In proc. IEEE INFOCOM'96, Mar. 1996

17 Ferrari D. and Verma D. A scheme for real-time channel establishment in wide-area networks. IEEE J. on selected area in communica-tion, Apr. 1990, 8 (3): 368-379

18 Verma D, Zhang H. and Ferrari D. Guaranteeing delay jitter bounds in packet switching networks. In proc. Tricomm'91, Apr. 1991. 368-379

19 Zhang H. Rate-controlled static-priority queueing. In proc. IEEE INFOCOM'93, Apr. 1993

20 Cruz R. A calculus for network delay, part I: network element in isolation. IEEE Transaction on Information Theory, Jan. 1991, 37 (1) 114-131

21 Cruz R. A calculus for network delay, part II: network analysis. IEEE Transaction on Information Theory, 1991, 37 (1) 132-141