

文章编号 :1001 - 2486(2007)02 - 0085 - 06

支持高效组播的高性能路由器 QoS 机制研究与实现^{*}

吴纯青 杨学军

(国防科学技术大学 计算机学院 ,湖南 长沙 410073)

摘 要 研究并提出一种有效支持组播的高性能路由器 QoS 机制 QoS _ RESM ,该机制以多阈值机制为基础 ,考虑远程转发引擎的拥塞状态信息 ,引入全局控制反馈机制 ,有效完成了高性能路由器上支持组播的流量控制。最后描述 QoS _ RESM 在基于网络处理器的核心路由器上的实现方法。

关键词 组播 ;流量控制 ;拥塞控制 ;高性能路由器 ;网络处理器

中图分类号 :TP393 **文献标识码** :A

Research and Implementation of a New QoS Mechanism with Effective Multicast Supporting in High Performance Routers

WU Chun-qing ,YANG Xue-jun

(College of Computer ,National Univ. of Defense Technology ,Changsha 410073 ,China)

Abstract :A new QoS mechanism with effective multicast supporting in high performance routers—QoS _ RESM is presented and studied in this paper. Based on multi-threshold mechanism ,it considers the congestion information of remote forward engines by introducing the mechanism of global control information feedback. It ends with a description of the implementation method of QoS _ RESM in the core router on the basis of network processor.

Key words :multicast ;flow control ;congestion control ;high performance router ;network processor

当今 Internet 的应用 ,如流媒体、在线信息追踪以及软件或代理的 cache 更新等 ,都要求更高的带宽和更大的分布规模。在这类应用中 ,组播技术有效地降低了传输成本。在这样的应用环境下 ,必须研究支持组播的拥塞控制策略。近年来 ,有关支持组播的 QoS 机制已成为研究热点 ,但支持组播 QoS 机制的研究多集中在端系统 ,有关路由器上的支持组播拥塞控制的研究相对较少。

本文提出一种支持高效组播的高性能路由器 QoS 机制 QoS _ RESM (a QoS mechanism on high performance Routers with Effective Supporting of Multicast) ,该机制以多阈值机制为基础 ,考虑远程转发引擎的拥塞状态信息 ,引入全局控制反馈机制 ,并结合现有路由器的拥塞控制机制 ,有效完成了高性能路由器上支持组播的流量控制。

1 相关工作

1.1 RED 及其发展

IP 网络的区分服务主要依靠队列管理机制、多队列的调度机制等实现。流量控制的目的是为了保证金能充分利用网络的带宽 ,保护高优先级的报文优先传输 ,同时又能避免造成网络拥塞。传统上流量控制通常在端系统中实现 ,但人们发现网络内部的流量控制意义更大。为此 ,人们提出了主动队列管理机制 (AQM)。RED^[1] 算法或者变种^[2-3] ,采用如图 1 所示的方法。

当队列长度小于阈值 1 时 ,正常传送 ,当队列长度大于阈值 1 小于阈值 2 时 ,开始按照概率丢弃 ;当

^{*} 收稿日期 2006 - 09 - 05

基金项目 :国家自然科学基金资助项目(90604006) ,国家部委资助项目(2003CB314802)

作者简介 :吴纯青(1964—) ,女 ,研究员 ,博士生。

队列长度大于阈值 2 时,丢弃所有报文。

WRED^[2]和 RIO^[3]可以提供多种丢弃优先级。FRED(Fair RED)^[4]也可以提供多种优先级。由于 FRED 依赖每个流的信息,需要较多的存储器和处理能力。

IETF 研究 DiffServ 体系结构下的各种情况。一种情况是研究如何防止非响应式 UDP 流在获得 AF 类中非预约带宽的超过公平性^[5]。文献 [6-7]研究了需要 2 级或者 3 级丢弃优先级。文献 [8]则研究了三种丢弃优先级如何分配各非响应 UDP 流和响应式 TCP 流,以解决公平性。另一种则是研究如何获得 AF 类内部流间的公平性。文献 [9]提出一种新的公平性标记算法。

针对利用队列长度确定拥塞程度问题,文献 [10]提出一种称为 Blue 的主动队列管理算法,Blue 采用报文丢失率和链路利用率管理拥塞技术。实验表明,对于交互式视频和音频应用,Blue 比 RED 能够提供较低的报文丢失率和较少的排队延迟。

1.2 路由器对组播拥塞控制的支持

就端系统而言,针对 UDP 协议缺少拥塞控制机制的研究主要集中在针对组播拥塞控制的需求,设计替代 UDP 的新型协议。例如,近一两年来由 S. Floyd 等人提出来的 DCCP 协议^[11-13],已经被提议作为 IETF 标准。DCCP 建立在一个类似于 UDP 的基础上,具有支持拥塞控制的基本机制,这些机制与 TCP 相似,考虑到了 TCP 协议设计的各个方面,但没有可靠性方面的要求。

在路由器系统中,设某组播流的流量为 f ,要发送到 k 个端口,则路由器为整个网络贡献的流量是 kf ,而在路由器的内部处理中,只有在转发到某个端口时才进行必要的报文复制。因此,在路由器内部,就拥塞控制而言,通常没有特别区分单播流量和组播流量。

2 QoS _ RESM 机制

QoS _ RESM 控制机制包括了 AQM,同时它还能根据系统资源的使用状况来进行流量控制。

QoS _ RESM 对 QoS 控制的支撑机制包括硬件机制和控制软件两方面,软件和硬件配合实现 QoS 控制算法。QoS 控制算法根据流队列或目标端口队列的拥塞状态、报文到达速率、数据存储区状态、远程转发引擎的当前状态来确定某个报文流的报文丢弃概率。

QoS _ RESM 机制由区分服务支持机制、路由器队列机制、多阈值机制、基于输出端拥塞信息反馈的丢弃机制等组成。这里重点阐述后面两种机制。

为了解决组播环境下的 QoS 支持等问题,在总结相关研究的基础上,根据我们研究的路由器体系结构,提出如图 2 所示的 QoS 解决方案。

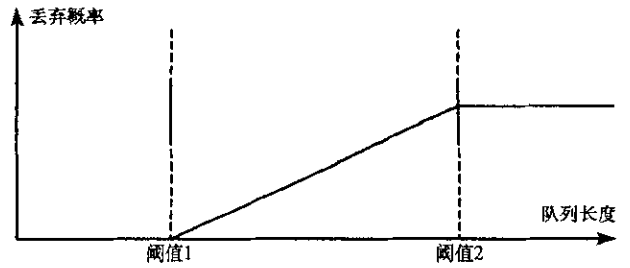


图 1 丢弃概率与平均队列长度
Fig. 1 Drop probability and average length of queue

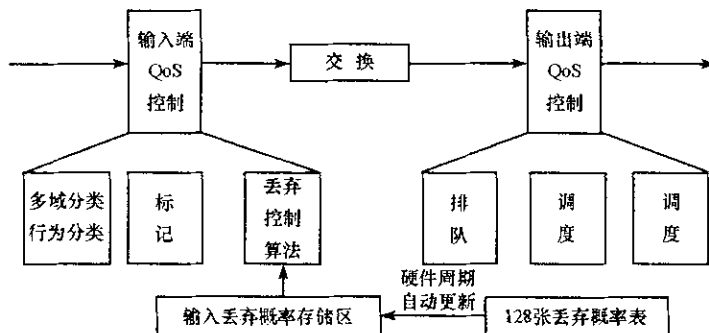


图 2 路由器 QoS 处理模型
Fig. 2 The model of router QoS processing

2.1 队列指数加权移动平均

由于当前队列长度具有不稳定性等特点,而一段时间的指数平均值具有较好的稳定性^[1],通常认为队列长度的指数加权移动平均值能够较平滑地反映队列长度的变化^[2]。因此我们在实现时采用队列的指数加权移动平均值监视系统的拥塞情况。

设系统当前队列的长度为 Q_Len ,其最新指数加权移动平均值为 $EWMA_Q$,第 i 次计算的 $EWMA_Q$ 记为 $EWMA_Q_i$ 则

$$EWMA_Q_{i+1}=(1-K)*EWMA_Q_i+K*Q_Len$$

其中 $K \in \{1, 1/2, 1/4, 1/8\}$; $i \in \{0, 1, 2, \dots\}$; $EWMA_Q_0=0$ 。

2.2 多阈值机制

在研究一般 QoS 算法时,没有考虑报文的自然属性。实际上在路由器内部存在两种自然属性不同的报文:一种是路由器需要转发的报文,另一种是路由器内部报文。为了区分起见,我们称前者为用户报文,后者为控制报文。针对组播环境,我们将上述的用户报文细分为组播用户报文和一般用户报文。为此我们提出多阈值控制算法。多阈值由控制报文阈值、组播报文阈值、一般用户报文阈值、新报文阈值和预警阈值五个值组成。如图 3 所示。

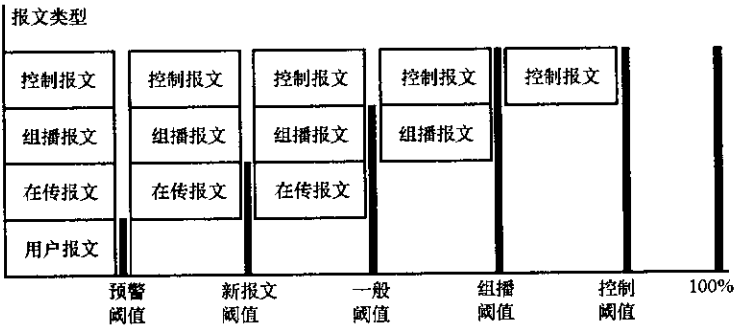


图 3 各类阈值条件下传输的报文类型

Fig. 3 The transmission packet types under various thresholds

(1) 预警阈值(A_Th :Alarm Threshold):当控制门限大于该值时,产生一个提示信号,提醒缓冲区数量进入该阈值;

(2) 新报文阈值(NI_Th :New Incoming Threshold):当控制门限大于该值时,不再为新的用户流量分配数据分配缓冲区,控制信息、组播信息和已经在传输的报文可以获得缓冲区;

(3) 一般用户报文阈值(GUP_Th :General User Packet Threshold):当控制门限大于该值时,不再为新的用户流量分配数据分配缓冲区,控制信息和组播信息可以获得缓冲区;

(4) 组播报文阈值(MUP_Th :Multicast User Packet Threshold):当控制门限大于该值时,不再为组播报文分配缓冲区,但是控制信息可以获得缓冲区;

(5) 控制报文阈值(CP_Th :Control Packet Threshold):当控制门限大于该值时,不再为任何输入数据(含控制报文)分配缓冲区。

通常每个队列至少应配置两个阈值:低阈值 Q_{min} 和一个高阈值 Q_{max} 。队列长度低于 Q_{min} 时,表明系统处于畅通状态;队列长度高于 Q_{min} 低于 Q_{max} 时,表明系统已出现轻度拥塞;而当队列长度高于 Q_{max} 时,表明系统处于严重拥塞状态。不难看出,如果 Q_{min} 和 Q_{max} 相同或几乎相同,则系统呈现尾丢弃流控。尾丢弃的问题是拥塞期间产生的丢弃具有突发性,从而造成系统吞吐率的剧烈震荡。普遍认为,对连续报文的突发性丢弃机制会降低 TCP 效率。所以 Q_{min} 和 Q_{max} 应该有相对的差距,例如 $Q_{max} = 10Q_{min}$ 或 $Q_{max} = 8Q_{min}$ 。

多阈值机制是队列阈值的补充,用以区分不同报文在系统拥塞情况下的处理。一般来说我们有:

$$A_Th \leq NI_Th \leq GUP_Th \leq MUP_Th \leq CP_Th$$

阈值的选择主要依据用户场景进行配置。在实际应用中可以采用相关经验和运营网络需要保障的对象进行合理配置,作者建议根据用户对组播的重视程度进行相关参数的调谐。

在具体实现时,需要选择特定对象作为控制因素。为了高效实现该功能还需要硬件支持。由于可用的空闲输入块数目是一定的,与报文队列的长度正好成为互补关系,而且也表征了队列的剩余长度。

2.3 全局拥塞控制反馈机制

目前实现的拥塞避免主要依据本地的信息进行判断和决策,本文在转发引擎间增加拥塞控制指示,向本路由器的上游转发引擎报告拥塞控制状态,称为 FE-Cogestion,如图 4 所示。该标识由转发引擎定时更新,通过交换网络直接反映到其他转发引擎,为转发引擎的拥塞判断提供支持。

在 3.2 节中我们将看到,有了全局的拥塞信息,可以在丢弃调度中一起利用和考虑该信息,能够更好地支持全局的调度。

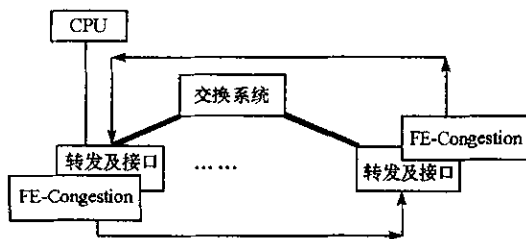


图 4 全局拥塞控制反馈机制

Fig. 4 The feedback mechanism of global congestion control

3 实现与评估

我们在高性能路由器实现验证了上述研究成果。作为一个完整的系统,其流分类、漏桶算法等直接采用其他研究成果,如文献[14]等。本文主要讨论其中的特殊处理,主要包括支持组播 QoS 的调度机制、报文丢弃概率计算机制、DiffServ 的 EF 和 AF 特性的支持和报文转发流程等。

3.1 丢弃机制

原理上报文经过路由器中只进行一次 RED 处理,实际上路由器内部只要存在排队都需要采用引入丢弃概率计算机制,以实现拥塞避免。在本系统中,QoS _ RESM 在输入端和输出端都实现了 QoS 控制机制。输入端的 QoS 控制机制设置在两个时刻:IP 报文到达网络处理器时、IP 报文到目标端口队列排队时。

当 IP 报文到达网络处理器时,QoS 控制机制会检查输入数据存储区的使用状况,根据其拥塞程度和拥塞阈值决定是否接收 IP 报文。当 IP 报文到目标端口队列排队时 QoS 控制机制将检查 IP 报文的优先级,根据阈值或丢弃概率表确定是否接收 IP 报文。按照概率丢弃的算法如图 5 所示。

WRED 算法考虑了报文的加权因子。如何选择加权因子进行计算既方便又节省空间,特别是如何利于高速网络处理中的硬件实现,对于提高算法效率是非常重要的。

与文献[15]类似,我们选择基于概率丢弃办法,其关键是如何设计合适的丢弃概率表,以便确定具体的丢弃概率。为了对报文丢弃实现精细的控制,将系统拥塞程度划分为 128 个等级。每个拥塞等级对应不同的流量有不同的处理,报文要素主要考虑它的 DSCP 值、是否是组播报文、远端的网络处理器是否拥塞等情况,分为 128 种处理。也就是说,每张传输概率表有 128 个表项,每个项对应于某种拥塞程度下不同类型 IP 报文的丢弃概率。系统周期性地根据当前的网络处理器拥塞情况选择一张丢弃概率表,作为当前的丢弃概率表保存到丢弃概率存储区。表结构的定义如图 6 所示。

丢弃概率表有 128 个表项,对应不同的 IP 报文类型,表项的索引值为“QQQTMCC”,其中 QQQ 表示 DSCP 中的 QoS 类型,T 表示 Remote TB Status,如果目标网络处理拥塞,那么该位为 1。M 表示是否为组播报文。CC 表示 DSCP 中的丢弃优先级(Color)。

在输出部分,传输概率表也有 128 个表项,对应不同的 IP 报文类型,表项的索引值为“TTCCMF”,其中 TT 表示报文的类型,CC 表示 DSCP 中的丢弃优先级(Color),M 表示是否为组播报文,F 表示流队列阈值溢出(QCB Threshold),P 表示目标队列阈值溢出(Target port TH _ PQ + FQ)。输出端不会改变该值。输出端传输概率表的结构与输入端的类似,x 轴为 FP,y 轴为 TTCC,z 轴为 M。

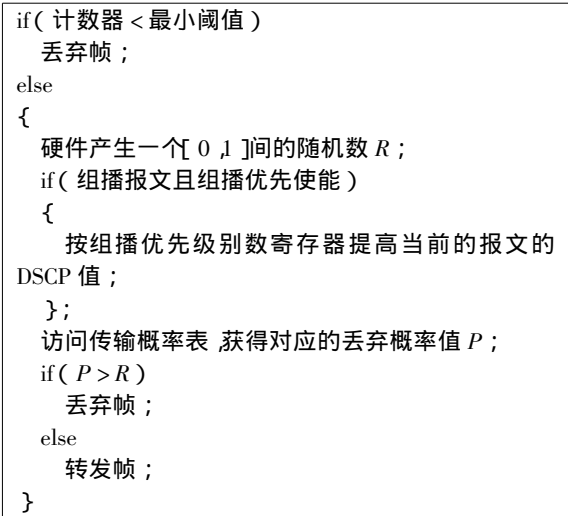


图 5 报文传输与丢弃策略

Fig. 5 Packet transmission and dropping policy

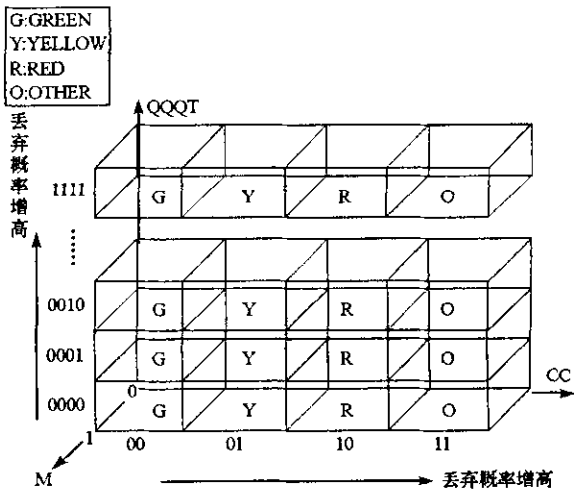


图 6 丢弃概率表结构

Fig. 6 The table structure of dropping probability

3.2 对 DiffServ 的 EF 与 AF 特性的支持

在区分服务体系结构提供的服务类型中 , 奖赏服务和确保服务得到了最多的关注和研究。奖赏服务的 EF PHB 希望在 diffServ 域中的实现低丢失率、低延迟、低抖动 , 事先确定的带宽、端到端服务。确保服务的 AF PHB 值为用户提供至少确保最低带宽保证的低丢失率的传输服务。

为了实现上述目标 , 在路由器队列设计中 , 专门设计了 EF 队列 EF-Queue、AF 队列 AF-Queue 和 PBS 队列 PBS-Queue , 既可以实现 EF PHB 的要求 , 也可以实现峰值速度等业务的需求 , 报文队列结构如图 7 所示。

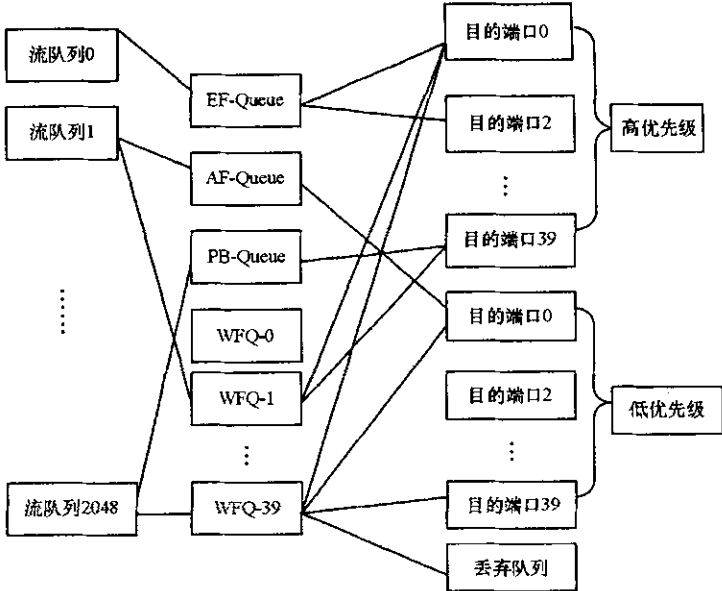


图 7 报文队列结构

Fig. 7 The packet queue structure

队列调度器负责保证实现流队列的预约带宽 , 以及完成剩余带宽在各条流之间的加权平均分配。调度器包括调度方法、流队列、目标端口队列、丢弃队列等。

调度方法分为基于时间片的调度和 WFQ 调度。基于时间片的调度主要应用于确保带宽服务(LLS 和 NLS)和峰值带宽服务(PBS) , WFQ 调度用于按权值分配剩余带宽 , 调度方法的选择顺序是 : 基于时间的调度和 WFQ 调度。

3.3 测试结果

我们对实现了 QoS _ RESM 的高性能 QoS 路由器原型系统进行了测量 ,其中嵌入式控制平台配置 600MHz PowerPC750 CPU ,128MB DRAM ,转发平台由基于 IBM 网络处理器 NP4GS3C 的各类转发接口板组成。测试使用 Spirent^[16]公司的 AX4000 网络测试仪。

表 1 和表 2 分别给出了千兆以太网接口和 ATM622M POS 接口的性能测试结果。

表 1 千兆以太网接口性能测试结果

Tab. 1 The testing results of GE interface

测试帧长 (Byte)	吞吐量 (%)	时延 (μs)	丢包率(%)	
			80% 负载	100% 负载
64	100%	10. 35	0. 0	4. 96
65	100%	10. 77	0. 0	0. 0
128	100%	11. 05	0. 0	0. 0
256	100%	11. 55	0. 0	0. 0
512	100%	12. 11	0. 0	0. 0
1024	100%	14. 78	0. 0	0. 0
1280	100%	16. 85	0. 0	0. 0

表 2 ATM622M POS 接口性能测试结果

Tab. 2 The testing results of ATM622 POS interface

测试帧长 (Byte)	吞吐量 (%)	时延 (μs)	丢包率(%)	
			80% 负载	100% 负载
64	100%	21. 07	0. 0	4. 96
65	100%	19. 24	0. 0	0. 0
128	100%	20. 43	0. 0	0. 0
256	100%	69. 78	0. 0	0. 0
512	100%	49. 63	0. 0	0. 0
1024	100%	64. 28	0. 0	0. 0
1280	100%	76. 80	0. 0	0. 0

4 结 论

本文研究了高性能路由器上支持组播的 QoS 控制机制 QoS _ RESM ,并在基于网络处理器的核心路由器上实现了该机制。QoS _ RESM 的多阈值机制、全局拥塞控制反馈机制、支持组播的丢弃概率表以及调度队列等的设计充分利用了共路控制信令的特点 ,实践表明具有合理性 ,在重负载情况下可提高系统吞吐量。

参 考 文 献 :

[1] Floyd S , Jacobson V. Random Early Detection Gateways for Congestion Avoidance[J]. IEEE/ACM Transactions on Networking , 1997 , 1 (4).

[2] Technical Specification from Cisco , Weighted Random Early Detection on the Cisco 12000 Series Router[Z]. URL : http://www.cisco.com/univercd/cc/td/doc/product/software/ios112/ios112p/gsr/wred_gs.htm.

[3] Clark D D , Fang W J. Explicit Allocation of Best-effort Packet Delivery Service[J]. IEEE/ACM Transactions on Networking , 1998 , 6 (4).

[4] Lin D , Morris R. Dynamics of Random Early Detection[C]//SIGCOMM97 , 1997.

[5] Heinanen J et al. Assured Forwarding PHB Group[S]. IETF RFC 2597 , June 1999.

[6] Jain R , Liu C , Goyal M et al. Performance Analysis of Assured Forwarding[S]. IETF DRAFT , February 2000.

[7] Elloumi O , De Cnodder S , Pauwels K. Usefulness of Three Drop Precedences in Assured Forwarding service[S]. IETF DRAFT , July 1999.

[8] Seddigh N , Nandy B , Piedra P. Study of TCP and UDP Interaction for the AF PHB[S]. IETF DRAFT , September 1999.

[9] Kim H. A Fair Marke[S]. IETF DRAFT , April 1999.

[10] Feng W , Kandlur D , Saha D , et al. A New Class of Active Queue Management Algorithms[R]. Technical Report , U. Michigan CSE-TR-387 - 99 , <http://www.eecs.umich.edu/~wuchang/blue/> , 1999.

[11] Floyd S , Kohler E. Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 2 : TCP-like Congestion Control[S]. RFC 4341 , Internet Engineering Task Force , Mar. 2006.

[12] Floyd S , Kohler K , Padhye J. Profile for Datagram Congestion Control Protocol (DCCP) Congestion Control ID 3 : TCP-Friendly Rate Control (TFRC) [S]. RFC 4342 , Internet Engineering Task Force , Mar. 2006.

[13] Kohler E , Handley M , Floyd S. Datagram Congestion Control Proto[S]. RFC 4340 , Internet Engineering Task Force , Mar. 2006.

[14] Lo L , Ou F. A Generic Traffic Conditioner[S]. IETF Internet Draft < draft-lin-diffserv-gtc-01. txt > , August 1999.

[15] IBM Power NP4GS3 Network Processor[Z]. <http://www.chips.ibm.com/techlib>.

[16] Spirent homepage[Z]. <http://www.spirentcom.com/>

