高速网络中的两级分布式存储器调度

伊 鹏[1] 王 鹏[2] 郭云飞[1] 李云涛[1]

[1](信息工程大学国家数字交换系统工程技术研究中心 郑州 450002)

[2](清华大学电子工程系微波与数字通信国家重点实验室 北京 100084)

摘 要 在高速网络中,商用存储器的存取速率一直是路由器调度性能提高的制约因素.为此该文提出了两级分布式存储器(TSDM)结构,该结构可以大大降低对商用存储器的存取速率的要求.通过分析给出了该结构模拟输出排队调度所需存储器个数的下界,并从理论上证明了该结构的交换单元无需加速即可模拟输出排队调度.最后文章从工程实现的角度给出了 TSDM 结构的一种工程简化设计方案,并通过仿真对该方案的性能进行了验证.

关键词 交换;调度;输出排队;分布式共享存储器

中图法分类号 TP393

A Two-Stage Distributed Memory Scheduling in High Speed Network

YI Peng[1] WANG Peng[2] GUO Yun-Fei[1] LI Yun-Tao[1]

[1](*National Digital Switching System Engineering & Technology Research Center, Information Engineering University, Zhengzhou* 450002)

[2](*State Key Laboratory on Microwave and Digital Communition, Department of Electronic Engineering, Tsinghua University, Beijing* 100084)

Abstract The commercially available memory rate can hardly keep up with the requirement of building high speed routers since the transmission capacity of the network is greatly improved. While several analytical studies of such a problem are presented, the findings published can not be considered as final. In this paper, we propose a two-stage distributed memory (TSDM) architecture which could decrease the requirement of the rate of commercially available memories without accelerating its switching units. We firstly analyze the lower bound for TSDM to mimic output queued scheduling based on the theory of combinatorics. And then we theoretically prove that the TSDM can emulate output queued scheduling without accelerating its switching units. Finally, we afford an engineering simplified design scheme for TSDM. At the first stage in our engineering simplified scheme, we use a round robin scheduling algorithm to implement port switch based on the output ports of the packets without considering their priority. At the second stage we use a motivated weighted deficit round robin (MWDRR) scheduling algorithm to implement the bandwidth assignments of the priority queues of the same output port. And the performance of the engineering simplified scheme is verified through simulations.

Keywords switch; scheduling; output queued; distributed shared memory

---

收稿日期:2002-11-12;修改稿收到日期:2003-11-17.本课题得到国家"八六三"高技术研究发展计划(2001-AA-12-4-011)资助.**伊 鹏**,男,1977 年生,博士研究生,主要研究方向为路由器交换调度技术.E-mail:yp@mail.ndsc.com.cn.**王 鹏**,男,1976 年生,博士研究生,主要研究方向为高速路由交换、片上系统.**郭云飞**,男,1963 年生,教授,博士生导师,主要研究方向为高速网络中的关键技术.**李云涛**,男,1976 年生,博士研究生,主要研究方向为高速路由器关键技术.

1 引 言

商用存储器的随机访问速率和交换结构交换速率的限制严重影响了路由器性能的进一步提高. Iyer 和 McKeown 等人在文献[1,2]中提出了并行分组交换(Parallel Packet Switch,PPS)结构,该结构可以降低对存储器访问速率的要求. 但是如果子交换结构的层数过多,会使得该系统的实现复杂度及费用都有较大的增加,同时交换机的整机性能维护也较为复杂. 在 2002 年的研究中,Iyer 等又提出了一种分布式共享存储器(Distributed Shared Memory,DSM)结构[3],该交换结构的本质是在交换机的各输入端口中,根据业务分配理论[4~6],将到达数据包的业务负荷分担到一个共享的存储器组中来模拟输出排队[7]. 该交换结构在一定的算法控制机制下可以通过低速存储器实现模拟输出排队调度. 然而 DSM 结构需要其中的 Crossbar 以 4～8 倍的加速比[3]工作于加速状态,依然限制了其在高速环境下的应用. 文献[8]在讨论输出排队调度算法时提出了一种分布式存储器交换结构,然而该结构仅能用于端口速率与存储器的存储速率相同时的情况,不具备普遍性.

为此我们提出了两级分布式存储器(Two Stage Distributed Memory,TSDM)结构,与 DSM 结构相比,TSDM 结构不仅可以降低其对存储器存取速率的要求,而且该结构中的交换单元可以不工作于加速状态,同时还能够支持变长包的处理.

2 TSDM 结构模型

TSDM 结构模型如图 1 所示,数据包首先通过 $N\times M$ 的交换单元写入 M 个读写速率均为 r 的存储器进行排队,再通过 $M\times N$ 的交换单元写入到 N 个输出端缓存,输入及输出端口的线路速率均为 R. 为方便理论分析,我们假定到达包为定长,标记为 *cell*. 在线速率为 R 的条件下,将发送或接收一个 *cell* 的时间称为系统时隙,存储器读入或写出一个 *cell* 的时间称为存储器时隙,而读写速率均为 r 的存储器读入或写出一个 *cell* 所需的系统时隙数称之为存储周期,在数值上,存储周期等于$\lceil R/r\rceil$.

对 TSDM 结构实现模拟输出排队调度的一般情况,为了能够有效降低对存储器访问速率的要求,输入端口的到达 *cell* 会被负荷分担到一个速率较低

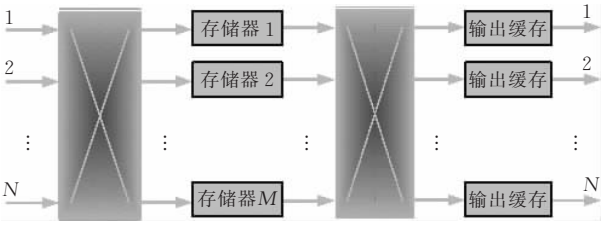

图 1 TSDM 结构模型

的分布式存储器组,当 *cell* 的离去时间到达时,*cell* 会从存储器组读出并送往输出端口. 因此在 TSDM 结构上模拟一般的输出排队调度时,每一个到达 *cell* 将面临以下四种冲突:(1)在 *cell* 写入的一个存储周期内存储器不能被再次写入;(2)同一系统时隙到达的 *cell* 不能被送往相同的存储器;(3)在读取 *cell* 的一个存储周期内存储器不能被再次读取 *cell*;(4)同一系统时隙不能从相同的存储器读取 *cell*.

为了使 TSDM 结构能够模拟输出排队调度,一方面 TSDM 结构中的存储器在数量上必须满足一定要求,另一方面需要引入一定的算法控制机制,才能够使得 TSDM 结构避免发生以上所分析的四种冲突中的任何一种. 通过对 TSDM 结构分析,对于 TSDM 结构中的存储器数量,在定理 1 中给出了所需满足条件的下界.

定理 1. 要使 TSDM 结构能模拟输出排队调度,其结构中存储器数量至少应不小于 $2\lceil R/r\rceil N-1$.

证明. 考虑 TSDM 结构中任意一个到达 *cell* 的情况. 在任意一个系统时隙,冲突 1 最多将导致分布式存储器组中的$(\lceil R/r\rceil-1)N$ 个存储器无法写入 *cell*;冲突 2 最多致使$(N-1)$个存储器无法写入 *cell*. 对于 *cell* 的离去过程有类似的考虑:在任意一个系统时隙,冲突 3 最多将导致分布式存储器组中的$(\lceil R/r\rceil-1)N$ 个存储器无法读取 *cell*;冲突 4 最多致使$(N-1)$个存储器无法读取 *cell*. 最坏情况对应于冲突 1、冲突 2、冲突 3 及冲突 4 中无法读写的存储器数目均取最大值且相互之间的交集均为空集,此时只有当存储器的数量不小于 $2\lceil R/r\rceil N-1$ 时才可能存在可避免冲突的控制算法. 因此要使 TSDM 结构能模拟输出排队调度,其存储器的数量至少应不小于 $2\lceil R/r\rceil N-1$. 证毕.

定理 1 得出了 TSDM 结构模拟输出排队调度时,分布式存储器组所需存储器个数的下界. 第 3 节我们将基于 TSDM 结构,从模拟先入先出(FCFS)输出排队调度与模拟任意入先出(PIFO)输出排队调度的角度,证明满足定理 1 的 TSDM 结构模拟输

出排队调度的可行性.

3　TSDM 结构模拟输出排队调度

为了方便分析,我们引入一个 $N\times N$ 的参考交换结构,其输入输出端口速率均为 R,采取输出排队方式,其中每一输出队列可缓存 L 个 *cell*. 我们首先考虑模拟采取 FCFS 机制的输出排队调度. 对于采用 FCFS 机制的参考交换结构,假定到达 *cell* 按如下规则进行服务:对于所有输出队列中到达时间不同的 *cell*,根据 *cell* 到达时间的先后进行服务;对于任一输出队列中到达时间相同的 *cell*,以到达 *cell* 的输入端口序号为序进行服务.

对于 TSDM 结构模拟采取 FCFS 机制的输出排队调度,在假定存储器数量满足定理 1 的要求的前提下,为避免冲突的发生,我们引入控制机制 1:

(1) 根据参考交换结构中 FCFS 的服务规则确定到达 *cell* 的离去时间,并由 *cell* 的离去时间确定冲突 3 和冲突 4 的冲突域.

(2) 根据冲突 1、冲突 3 和冲突 4 的冲突域确定到达 *cell* 的可用存储器组,即在一个存储周期内没有发生 *cell* 读入和写出事件,且不含有与到达 *cell* 相同离去时间的一组存储器.

(3) 在到达 *cell* 与可用存储器组之间确定其匹配子图,并依据此匹配关系将到达 *cell* 写入存储器组.

(4) 存储器组中的 *cell* 的离去时间到达时即通过第二级 crossbar 送往输出端口.

参照文献[7],在 TSDM 结构中,如果对于在任一系统时隙到达的任一 *cell*,都可以按与采取 FCFS 机制的参考交换结构相同的 *cell* 离去顺序发送出去,则可认为该结构可以模拟采取 FCFS 机制的输出排队调度. 对控制机制 1 进行分析,得到如下定理.

定理 2.　满足定理 1 的 TSDM 结构可以模拟采取 FCFS 机制的输出排队调度.

证明.　根据我们所采用的控制机制可知:对于满足定理 1 的 TSDM 结构,如果任一到达 *cell* 不会被阻塞于系统中,则可以按与采取 FCFS 机制的参考交换结构相同的 *cell* 离去顺序发送出去. 控制机制 1 第 2 步所得的可用存储器组排除了发生冲突 1、冲突 3 和冲突 4 的可能性,第 3 步根据可用存储器组计算匹配子图则进一步排除了发生冲突 2 的可能性. 因此到达 *cell* 不可能被阻塞于系统中. 所以满足定理 1 的 TSDM 结构可以模拟采取 FCFS 机制的输出排队调度.　　证毕.

FCFS 机制只是基于输出排队结构调度的一种特殊情况,还有许多不同的基于输出排队结构的调度算法,如 WFQ[9],VC[10] 等,他们都属于采取 PIFO 机制的调度算法. PIFO 机制指的是新到达 *cell* 可以从队列的任一位置插入队列中,但只能从队列首部出队. TSDM 结构模拟输出排队调度的关键是要能够通过控制,模拟采取 PIFO 机制的输出排队调度.

TSDM 结构模拟采取 PIFO 机制的输出排队调度时,由于模拟的输出排队结构所采取的调度策略不同,使得避免冲突 4 的方法变得复杂. 虽然理论上一个系统时隙中参考交换结构最多有 N 个存储器在读取 *cell*,但由于它采用的是 PIFO 机制,*cell* 到达时刻及其离去时间无法预知,使得 TSDM 结构中潜在的冲突无法避免. 考虑在往存储器写入某一 *cell* 记作 a 的情况,写入时我们只能根据当前系统时隙的情形,挑选不会导致冲突 4 发生的存储器,但由于 *cell* 离去时间的可变性,存储器中原有缓存 *cell* 的离去时间可能变得与 a 的离去时间相同,因而冲突 4 仍可能发生. 用反证法考虑极端的情况可以证明,这种无法预知的冲突是无法通过这种方式完全避免的. 因此我们考虑引人一定的乱序来避免冲突 4 的发生,由于引入的乱序是有限的,在输出端缓存采用简单的算法即可恢复原来顺序. 基于这一考虑,我们引入控制机制 2:

(1) 根据冲突 1 和冲突 2 确定可写存储器组.

(2) 按照连续 N 个相同目的端口的 *cell* 被写入不同存储器的原则,将到达 *cell* 写入到可写存储器组.

(3) 根据冲突 3 确定可读存储器组.

(4) 以 N 个系统时隙作为一个处理单元,在其中每一系统时隙确定可读存储器组和目的端口之间的匹配子图,*cell* 按匹配关系被读出到输出端口缓存.

(5) 在输出端口缓存采取简单的排序算法进行乱序调整.

对于队列长度为 L 的参考交换结构,控制机制 2 中导致的乱序最大为 $2N-1$,因此可通过乱序调整来实现 *cell* 的重新排序,使其按照与参考交换结构中相同的 *cell* 离去顺序离开. 需要说明的是:控制机制 2 没有直接考虑冲突 4,而是采取引入乱序的方式避免冲突 4 的发生,但是这一方式需要在写入过程中连续 N 个相同目的端口的 *cell* 不能被写入相同存储器,这样最多有 $N-1$ 个存储器无法写入 *cell*,在数量上刚好与直接避免冲突 4 一致,因此所

需的存储器总数不变.

同前所述,在 TSDM 结构中,如果对于在任一系统时隙到达的任一 *cell*,都可以按与采取 PIFO 的参考交换结构相同的 *cell* 离去顺序发送出去,则我们认为该结构可以模拟 PIFO 的输出排队[7]. 通过分析控制机制 2,我们得到如下定理.

定理 3. 满足定理 1 的 TSDM 结构可以模拟采取 PIFO 机制的输出排队调度.

证明. 根据我们所采用的控制机制可知:对于满足定理 1 的 TSDM 结构,如果任一到达 *cell* 不会被阻塞于系统中,则可以通过乱序重排,按与采取 PIFO 机制的参考交换结构中相同的 *cell* 离去顺序发送出去. 因为控制机制 2 引入的乱序都是有限的,因此 *cell* 不会被阻塞于乱序重排的缓存中. 控制机制 2 的第 1 步所得的可写存储器组排除了发生冲突 1 和冲突 2 的可能性,第 3 步所得的可读存储器组排除了发生冲突 3 的可能性,第 2 步采用引入乱序的方式避免了冲突 4 的发生,因此到达的 *cell* 也不会被阻塞于系统中. 所以满足定理 1 的 TSDM 结构可以模拟采取 PIFO 机制的输出排队调度. 证毕.

根据定理 2 和定理 3 可知,满足定理 1 的 TSDM 结构可以模拟输出排队.

4 TSDM 工程简化设计

前述分析虽然证明了 TSDM 结构可以完全模拟输出排队结构,实现高性能的调度,然而其控制机制仅是从理论的角度考虑,而且必须采用集中控制式的算法,不容易硬件实现,因此在本节我们按照 TSDM 结构的指导思想,以 8×8 的交换结构为例给出一种易于硬件实现的解决方案,该方案的逻辑结构如图 2 所示. 该方案将存储器组分配到每个输入端口,即在输入端口采取区分目的端口的排队. 虽然这样增加了存储器的总个数,但这样避免了各端口之间的交互,可以采用分离的调度算法,便于硬件设计实现. 对于 8×8 的交换网络,我们在每个输入端口根据其目的端口设置 8 个分离的缓冲队列,分别缓存到达不同目的端口的包. 对于目的端口相同的缓冲队列,我们采用轮循调度策略对到达包进行调度,输出包被送往输出端口的优先级队列进行区分优先级的排队. 在优先级队列的出口通过允许低延迟队列的加权差额循环调度算法(MWDRR),实现对不同优先级业务的带宽分配. 其中针对目的端口的调度是经过二倍加速的,这样可以导致优先级队列的拥塞,从而实现带宽分配.

MWDRR 算法将最高优先级队列设置为低延迟队列,对其它的 7 个 COS 队列采取交替优先级的逐包轮询的机制,轮询机在每一个轮循周期为每一队列分配 N_i 字节输出带宽,为每一队列设置一储蓄计数器 C_i($i=1,2,\cdots,K$;K 为队列数),其中 N_i 根据队列优先级不同设定为不同的值. (1)起始状态所有 $C_i=0$;(2)当轮询到一个队列 i 时,为计数器存入该轮应该输出的字节数 N_i,$C_i=N_i+C_i$,如果队列中有整包,且排在队头的那个包的包长 $L_i^1 \leq C_i$(包长以字节为单位计算),将该包调度输出,并令 $C_i=C_i-L_i^1$,再观察该队列中是否有整包,若有,且排在

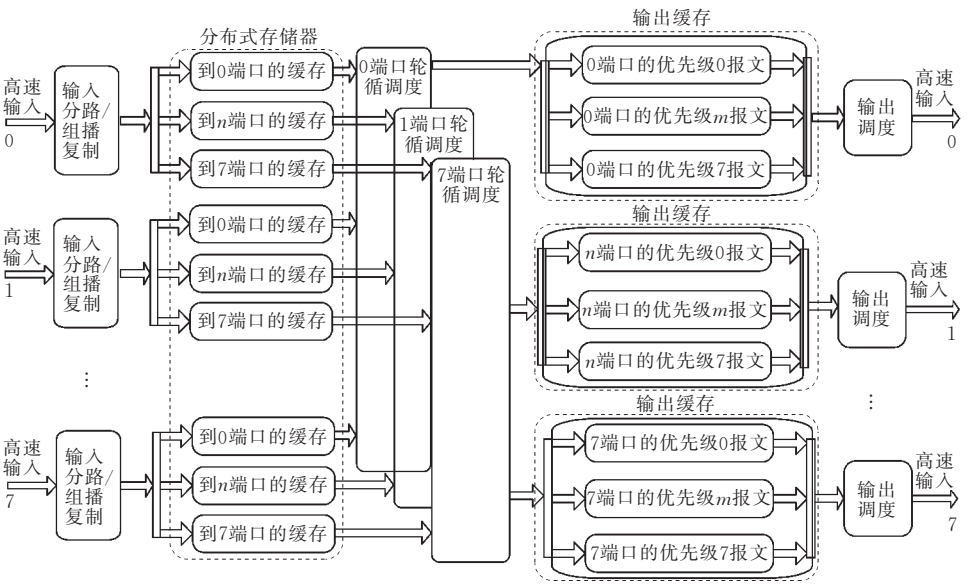

图 2 工程简化实现方案逻辑图

队头的包的长度 $L_i^2 \leq C_i$,则将第二个包调度输出,直至该队列中无整包或计数器的值小于队列中第一个包的长度,转到下一队列;(3)如果轮询到一个队列 i,且为计数器存入该轮应该输出的字节数 N_i($C_i = N_i + C_i$)之后,发现队列中无整包或第一个包的包长 $L_i^1 > C_i$,则将该轮应发的字节数 N_i 储蓄起来,留到下一轮使用,转到下一队列;(4)如果轮询到队列 i 时,队列是空的,则将计数器 C_i 清零,这一做法是为了不让 C_i 无限增大,虽然在一定程度上损害了公平性,但是却能够防止突发的形成和对其它队列的影响.

由于 TSDM 结构工程简化实现方案采用的是尽职工作型(work-conserving)调度算法,而对所有尽职工作型调度算法,其平均队列时延是一样的,因此该方案的平均时延仅为输出排队结构调度的平均时延加上一个常数值的处理时延.所以我们仅对该调度方案的丢包率特性进行仿真分析.

假定每一队列的存储器容量为 16×8K 比特,优先级 i 越小对应队列优先级越高.用户流发送优先级为 i 的包的数量服从参数为 λ_i 的泊松分布,包长服从指数分布,最长为 $Max = 1500$ 字节.在无违反约定流的情况下,λ_i 与对应第 i 优先级所分配的输出带宽成比例.违反约定的流以其分配带宽的 β 倍速率发送包,并假设第 4 个优先级队列中的流为违反约定流,输出带宽为 $C = 2.488$Gbps.优先级为 i 的队列获得的带宽比例为 f_i,每个队列占据 Cf_i 的输出带宽.我们设定 $\beta = 5$,$f_i[7] = [0.26, 0.2, 0.16, 0.13, 0.1, 0.07, 0.05, 0.03]$,在 120s 的时间内仿真了采用 FCFS 输出排队调度的参考交换结构和采用 MWDRR 调度的 TSDM 结构中各个队列的带宽分配情况和包丢失情况.图 3 和图 4 分别给出了轻负荷与重负荷时每一队列的包丢失率.图 5 与图 6 给出了在不同负荷情况下不同队列得到的带宽.

由图 3 和图 4 可见,对于使用 MWDRR 调度算法的 TSDM 结构,无论负载情况如何,遵守约定的流所在队列的包丢失率不会受到违约流的影响,而违约流所在队列(Q_9)的包丢失率总是远高于其它流.与采用 FCFS 调度算法的参考交换结构相比,TSDM 结构使用 MWDRR 调度算法能降低遵守约定的队列中流的包丢失率.图 5 和图 6 显示出,采用 FCFS 算法的参考交换结构中第 4 队列抢占了其它优先级队列的带宽,而采用 MWDRR 算法的 TSDM 结构中其它队列的应得带宽不会受到违约流的影响.图 5 中违约流所在的队列获得了超过其约定的带宽,这是因为在轻负荷时,调度机在保证正常流的包丢失率条件下,还有多余的输出带宽可供使用.

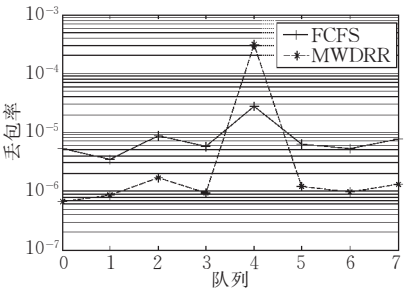

图 3　λ=0.9C 时的丢包率

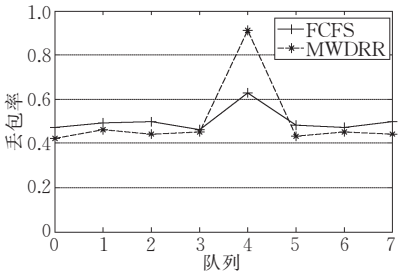

图 4　λ=2C 时的丢包率

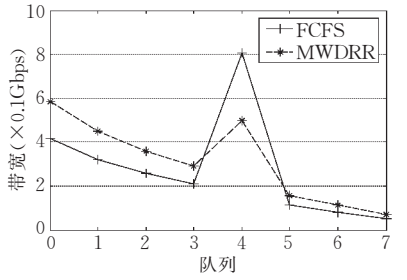

图 5　λ=0.9C 时的队列带宽

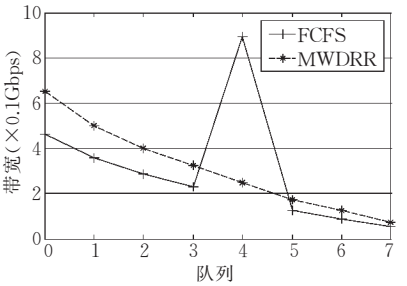

图 6　λ=2C 时的队列带宽

5　结束语

商用存储器的随机访问速率和交换结构的交换速率一直是高速路由器实现的瓶颈,为此需要改善交换结构来降低对存储器存取速率的要求.虽然 DSM 结构能够较好地解决该问题,但其交换结构仍需以 4 倍乃至 8 倍的加速比工作.为此本文对交换结构进行了进一步研究,提出了 TSDM 结构.为了

能使该结构模拟输出排队调度,我们给出了其中分布式存储器组所需的存储器个数的下界,并从理论上证明了 TSDM 结构在一定的控制机制下可以实现模拟输出排队.最后我们给出了该结构的工程简化实现方案,并通过仿真对其性能进行了验证.该方案既避免了输出排队结构的 N 倍加速问题,又不需像输入排队结构采用集中控制的调度算法,易于硬件实现,为高速路由器的调度设计提供了一个较好的选择方案.

参 考 文 献

1 Iyer S., Awadallah A., McKeown N.. Analysis of a packet switch with the memories running slower than the line rate. In: Proceedings of INFOCOM 2000, Tel-Aviv Israel, 2000, 2: 529～537.

2 Khotimsky D. A., Krishnan S.. Stability analysis of a parallel packet switch with bufferless input demultiplexors. In: Proceedings of ICC 2001, Helsinki, Finland, 2001, 213～217

3 Iyer S., Zhang R., McKeown N.. Routers with a single stage of buffering. In: Proceedings of SIGCOMM 2002, Pittsburgh, Pennsylvania, USA, 2002, 431～439

4 Adiseshu Hari, Parulkar Guru, Varghese George. A reliable and scalable striping protocol. In: Proceedings of ACM Sigcomm 1996, Stanford, CA, 1996, 131～141

5 Chiussi F., Khotimsky D., Krishnan S.. Generalized inverse multiplexing of switched ATM connections. In: Proceedings of Globecom 1998, Sydney, Australia, 1998, 268～276

6 Chiussi F., Khotimsky D., Krishnan S.. Advanced frame recovery in switched connection inverse multiplexing for ATM. In: Proceedings of ICATM 1999, Colmar, France, 1999, 173～182

7 Iyer Sundar. The Parallel Packet Switch Architecture [M. S. dissertation]. Stanford University, Stanford, 2000

8 Prakash A., Sharif S., Aziz A.. An $O(\log^2 N)$ parallel algorithm for output queuing. In: Proceedings of INFOCOM 2002, New York, USA, 2002, 127～136

9 Parekh A.. A generalized processor sharing approach to flow control in integrated services networks[Ph. D. dissertation], MIT, MA, 1992

10 Zhang L-X.. Virtual clock: A new traffic control algorithm for packet switching networks. In: Proceedings of ACM SIGCOMM' 90, Philadelphia PA, 1990, 19～29

YI Peng, born in 1977, Ph. D. candidate. His research interests include switch and schedule technology of routers.

WANG Peng, born in 1976, Ph. D. candidate. His research interests include high speed switching and system on chip technology.

GUO Yun-Fei, born in 1963, professor and Ph. D. supervisor. His research interests focus on the key technology of the highway network.

LI Yun-Tao, born in 1976, Ph. D. candidate. His research interests focus on key technologies in high-speed router.

Background

This work is part of our project, which is named Key Technology for Super High-speed Network Nodes. The goal of this project is to study and develop the state of the art communication technology, including routing, switching and network architecture, for next generation networks in China. With the project beginning in 2001, we focused on such fields as scalable switches, active networks and QoS routing and have made great man novel achievements. We have published over twenty research papers in domestic and overseas publications or conference proceedings. Moreover, some research results have been successfully applied into the implementation of Scalable Terabit IPv4/v6 Router. This paper focuses on scalable switches and priority sustained scheduling algorithms.