第 35 卷 第 8 期 计 算 机 工 程 2009 年 4 月
Vol.35 No.8 Computer Engineering April 2009
·网络与通信· 文章编号：1000—3428(2009)08—0131—03 文献标识码：A 中图分类号：TP393

基于并行存储机制的交换结构及其调度算法

郑德任[1]，汪斌强[1]，扈红超[1]，李 挥[2]
（1. 国家数字交换系统工程技术研究中心，郑州 450002；2. 北京大学深圳研究生院，深圳 518055）

摘 要：针对目前多数交换机制可扩展性差、实现复杂度大的问题，基于并行存储机制构建高性能交换结构 PSS，采用流模型证明在不加速的情况下 PSS 交换结构对满足大数定律的可容许到达业务能够实现 100%的吞吐量，在该结构的基础上提出简单优先双轮询算法 SPDRR。仿真结果表明，应用 SPDRR 算法的 PSS 交换结构能够获得很好的性能。

关键词：交换结构；调度算法；双轮询；优先级

Switching Architecture Based on Parallel Storage Scheme and Its Scheduling Algorithm

ZHENG De-ren[1], WANG Bin-qiang[1], HU Hong-chao[1], LI Hui[2]
(1. National Digital Switching System Engineering & Technological R & D Center, Zhengzhou 450002;
2. Shenzhen Graduate School, Peking University, Shenzhen 518055)

【Abstract】 In order to improve the performance of the presented schemes with low scalability and great complexity, this paper builds a high-performance switching architecture based on parallel storages scheme named Parallel Storage Scheme(PSS). With the flow model techniques, it proves that PSS switch can achieve a throughput of 100% without speedup to the arbitrary admissible traffic that satisfies the Strong Law of Large Number(SLLN). Simple Priority Double Round Robin(SPSDRR) algorithm is presented based on PSS. Simulation results indicate that PSS switch with SPDRR algorithm can obtain high performance.

【Key words】 switching architecture; scheduling algorithm; double round robin; priority

1 概述

现在的交换机不但要支持越来越快的速率，而且必须保证不同混合业务源的服务质量(QoS)。输出排队(OQ)交换结构虽然在提供服务质量保障方面极具优势，通过简单的调度机制就可获得高吞吐量和良好的时延性能[1]，但 OQ 交换结构 N 倍加速问题限制其大容量构建。输入排队(IQ)交换结构虽只需工作在线路速率(与 N 无关)[2]，可以实现大容量的构建，并且采用虚拟输出排队(VOQ)用以解决队头的阻塞，但 IQ 交换结构采用复杂的集中式调度算法，如极大权重匹配算法能提供 100%吞吐量，但其算法复杂度为 $O(N^3\mathrm{lb}N)$[3]。联合输入输出排队(CIOQ)交换结构虽然已被证明在加速比为 2 时能够完全模拟 OQ 交换结构，但调度机制需要集中考虑输入调度与输出调度，其集中式匹配算法过于复杂、不易实现[4]。

本文基于分布式系统设计思想构建了新的交换结构——PSS(Parallel Storage Scheme)。此结构不但能对满足大数定律的可容许到达业务实现 100%的吞吐量，而且调度算法简单，因此，既能获得好的性能，又易于工程实现。

2 PSS 交换结构

2.1 PSS 交换结构简介

PSS 交换结构由输入、输出和存储 3 个模块组成。3 个模块分别由 N 个同构的 $1\times N$ 交换单元、N 个同构的 $N\times 1$ 交换单元和 $N\times N$ 个并行端口分布式存储单元组成。输入交换模块的 $1\times N$ 交换单元用作输入端口的分路器，输出交换模块的 $N\times 1$ 交换单元用作输出端口的合路器。所有的交换单元都相互独立，因此，易于实现。

图 1 为一个 $N\times N$ PSS 交换结构模型。并行端口分布式存储单元 S_{ij} (i,j=1, 2,…,N)与输入端口 i 输出端口 j 相连，从输入端口 i 到输出端口 j 的信元被写进 S_{ij} (i,j=1, 2,…,N)，此存储单元中的信元通过调度机制被送入输出端口 j。由于 PSS 交换结构是分布式的，可以采用分布式调度算法，因此调度机制能够获得 QoS 保证。尽管 PSS 交换结构能直接处理变长包，但为了简化分析，本文假设经过此结构的信元大小固定。

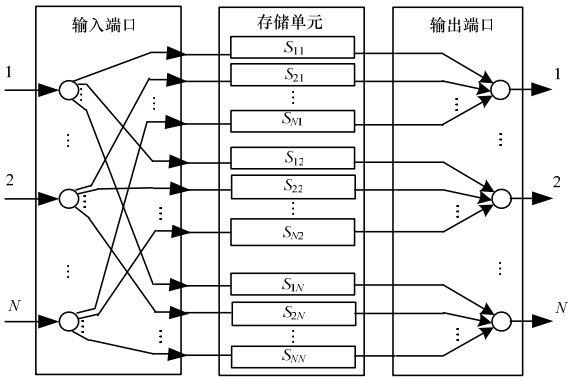

图 1 $N\times N$ 的 PSS 交换结构

基金项目：国家自然科学基金资助项目(60572042)；国家“863”计划基金资助项目(2008AA01Z214)
作者简介：郑德任(1983－)，男，硕士研究生，主研方向：高性能路由器交换结构与调度算法；汪斌强，教授、博士生导师；扈红超，博士研究生；李 挥，副教授、博士
收稿日期：2008-08-02 **E-mail：**zhengderen2004@163.com

2.2 PSS 吞吐量分析

本文采用流模型技术[5]证明 PSS 交换结构对任意可允许到达业务能够获得 100%的吞吐量。令 $A_{ij}(n)$表示到第 n 时隙所有到达输入端口为 i、目的端口为 j 的信元个数。假设到达过程$\{A_{ij}(\cdot), i, j = 1,2,\cdots, N\}$以概率 1 满足强大数定律，即

$$\lim_{n\to\infty}\frac{A_{ij}(n)}{n}=\lambda_{ij},\ i,j=1,2,\cdots,N \tag{1}$$

其中，λ_{ij} 是到达 VOQ_{ij} 的速率。如果 λ_{ij} 满足式(2)，则到达业务$\{A_{ij}(\cdot), i, j=1,2,\cdots, N\}$被称作可容许的。

$$\sum_i\lambda_{ij}\leqslant 1 \ \text{ and } \ \sum_j\lambda_{ij}\leqslant 1 \tag{2}$$

令 $D_{ij}(n)$表示到第 n 时隙所有离开 S_{ij} 的信元个数。文献[5]证明：如果 $D_{ij}(n)$以概率 1 满足式(3)，则交换结构为速率稳定的，可以获得 100%的吞吐量。

$$\lim_{n\to\infty}\frac{D_{ij}(n)}{n}=\lambda_{ij},\ \ i,j=1,2,\cdots,N \tag{3}$$

定理 1 在加速因子为 2 时，采用任意尽职工作型调度算法的 PSS 交换结构对所有满足强大数定律的可容许到达业务能获得 100%的吞吐量。

证明 令 $Z_{ij}(t)$表示 t 时刻缓存于 S_{ij} 中的流的总量。根据文献[5]的引理 1 可知，要证明流量模型是弱稳定的，只需证明对于任意 $Z(t)>0$，$Z'(t)\leqslant 0$，其中，$Z'(t)$为 $Z(t)$的导数。即证明当 PSS 交换结构队列中有堵塞时，可容许到达业务缓存队列中流的总量是非递增的。在时隙 n 对任意 $Z_{ij}(n)>0$，由于调度算法是尽职工作型，因此在时隙 n 结束时必定有一个信元被调度离开 PSS 交换结构。令 $L_j(t)=\sum_i Z_{ij}(t)$ 为 t 时刻所有输出端口为 j 的流总量，则

$$L_j(n+1)-L_j(n)\leqslant\sum_i(A_{ij}(n+1)-A_{ij}(n))-1 \tag{4}$$

结合式(2)，应用流模型的求极限运算可得

$$L'_j(t)=\sum_i Z'_{ij}(t)\leqslant\sum_i\lambda_{ij}-1\leqslant 0 \tag{5}$$

由于

$$Z(t)=\sum_{ij}Z_{ij}(t)$$

$$Z'(t)=\sum_{ij}Z'_{ij}(t)=\sum_j L'_j(t)\leqslant 0$$

流模型是弱稳定的，且利用工作保持型调度算法的 PSS 交换结构是速率稳定的，因此利用尽职工作型调度算法的 PSS 交换结构不加速即可对满足大数定理的任意可容许到达业务实现 100%的吞吐量。

3 SPDRP 调度算法

由于带有优先级调度策略的调度算法实现太复杂，在工程设计中不易实现，因此本文从简化设计、易于工程实现的角度出发，提出了基于优先级双轮询调度算法 SPDRR (Simple Priority Double Round Robin)。

为了给不同优先级的业务提供不同的服务，每个端口分布式存储单元逻辑上被分成 k 个不同输出优先级的虚拟优先级队列。PSS 交换结构维护的优先级个数决定 k 的大小。定义 VPQ_{ijk} 为 S_{ij} 中优先级为 k 的虚拟优先队列。在每个时隙，SPDRR 调度算法实现 2 级判决调度：在第 1 级，SPDRR 调度算法根据所有优先级间带宽的分配决定哪个优先业务被服务；在第 2 级，SPDRR 调度算法从所有具有相同优先级和相同输出端口的 VPQ 中用简单的轮询机制选择一个 VPQ 调度输出。

图 2 给出在 PSS 交换结构的输出端 j 上 SPDRR 调度算法的实现机制。K 个计数器分配给 k 种不同优先级服务业务，通过优先级轮询(PRR)给输出端口 j 分配输出带宽。每个时隙通过为不同优先级水平的计数器分配不同的输出带宽来获得不同优先级的业务。由于具有相同优先级的业务可能分布在不同存储单元的 N 个 VPQ 中，因此在每个优先级中用一个简单的轮询(SRR)判决器决定服务哪个 VPQ。每个 SRR 判决器维护一组指向相应 VPQ 的指针寄存器(VPR)和一个被服务的 VPQ 的服务指针寄存器(SPR)。每个时隙调度更新后，SRR 判决器的服务指针指向轮询机制中下一个非空的 VPQ。

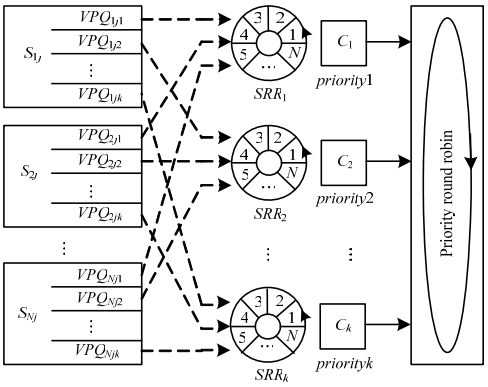

图 2 输出端口 j 的 SPDRR 调度机制

SPDRR 调度算法具体描述如下：

(1)起始状态将所有优先级业务的带宽计数器初始化为 $C_i=0$, $i=1, 2,\cdots,k$。

(2)每次轮询到一个优先级为 i 的业务时，为对应的带宽计数器 C_i 分配输出带宽 N_i，且令 $C_i=N_i+C_i$，其中，N_i 由优先级为 i 的业务决定。

(3)轮询到计数器 C_i 输出时，如果 C_i 大于一个包长 L 且优先级为 i 的 VPQ 非空，则轮询选择一个优先级为 i 的 VPQ 服务且令 $C_i=C_i-L$，直到该优先级对应的所有 VPQ 为空或者该计数器的值小于 L，L 是信元的长度。

(4)轮询到计数器 C_i 输出时，如果 C_i 小于一个包长 L，则将该轮应发的字节数储蓄起来留到下一轮使用，并跳到下一个优先级业务。

(5)轮询到计数器 C_i 输出时，如果优先级为 i 的 VPQ 都为空，则清空计数器 C_i 且跳到下一个优先级业务。

本文算法特点如下：

(1)算法通过对计数器分配不同的带宽实现区分优先级的调度。

(2)由于算法的第 1 级区分业务的轮询调度仅仅根据带宽计数器大小决定优先级，因此第 1 级的轮询调度与第 2 级区分业务带宽的轮询调度同样是最简单的轮询调度，算法复杂度低。又因为所有输出端口调度程序都能独立并行地工作，所以此机制在工程设计中易于实现。

(3)不仅同一优先级的业务通过简单轮询获得公平调度，而且区分优先级的业务在一次轮询后将剩余的小于信元长度的包累积到下次该优先级业务调度，不会造成某一优先级业务小负载的过度丢弃，因此，不同优先级的业务之间也能够实现公平调度。

4 性能仿真

本文采用 NS-II 仿真器对采用 SPDRR 调度算法的 PSS 交换结构进行性能仿真。为了便于与其他类型的交换结构比较，本文同时仿真了采用 iSLIP 调度算法的 IQ 交换结构(IQ-iSLIP)、采用 WFQ 调度算法的 OQ 交换结构(OQ-WFQ)、采用 LQF-RR 调度算法的 CICQ 交换结构(CICQ-LQF-RR)。

所有的交换结构采用 16×16 的大小，所有到达输入端口的业务有 4 级优先级。仿真器工作在时隙粒度且仿真时间为 100 000 时隙，信元大小为 64 Byte。

图 3 给出了不同交换结构在到达业务为 uniform 分布($\lambda_{ij}=\rho/N, \forall\, i, j$)条件下的平均时延特性。可以看到，所有交换结构受业务负载影响较小、性能较好。图 4 给出了不同交换结构在到达业务为 non-uniform 分布($\lambda_{ii}=0.8\rho$, $\lambda_{i|i+1|}=0.2\rho$)条件下的平均时延特性。当负载过重时，IQ-iSLIP 机制的时延性能明显下降，其他机制性能较好。在所有业务形式和不同负载下，PSS-SPDRR 机制的平均时延性能与 OQ-WFQ 机制几乎相同。这从另一个方面证明了 PSS 交换结构能够获得 100%吞吐量。

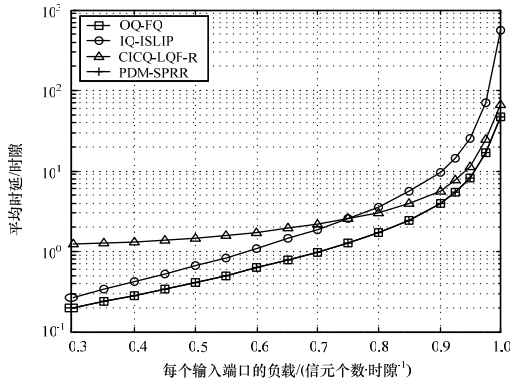

图 3　业务源为 uniform 分布的时延性能

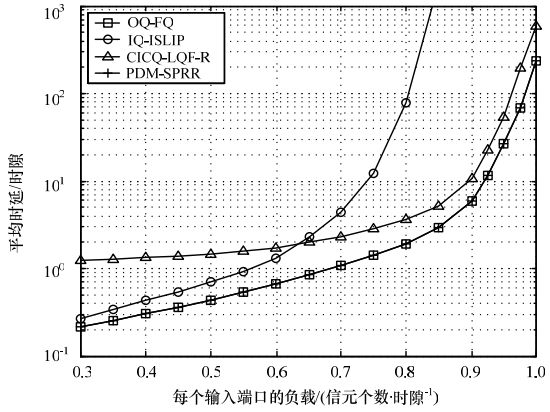

图 4　业务源为 non-uniform 分布的时延性能

为了验证 PSS-SPDRR 交换机制获得区分优先级服务的能力，测试并比较了每个不同优先级业务流的平均时延特性。由于 IQ-iSLIP 机制和 CICQ-LQF-RR 机制不能直接实现优先级调度，因此只能比较 OQ-WFQ 机制和 PSS-SPDRR 机制之间优先级业务流的平均时延。假设到达输出端口的业务根据它们的优先级被聚合为 4 种流，这些流根据优先级从低到高定义为 flow1, flow2, flow3, flow4。它们约定的带宽分别为 0.1, 0.2, 0.3, 0.4 且每个输出端口的整个带宽为 1。所有这些流具有相同的负荷，定义为 ρ。图 5 给出了在到达业务为 uniform 分布($\lambda_{ij}=\rho/N, \forall\, i, j$)的条件下，在 PSS-SPDRR 机制和 OQ-WFQ 机制下优先流的平均时延特性。图 6 给出了在到达业务为 non-uniform 分布($\lambda_{ii}=0.8\rho$, $\lambda_{i|i+1|}=0.2\rho$)的条件下，在 PSS-SPDRR 机制和 OQ-WFQ 机制下优先流的平均时延特性。2 种机制都能针对不同业务水平提供不同的服务。事实上，2 种分布情况下优先业务流的平均时延非常相似，共同的特点是对于高优先级的业务流能获得较好的平均时延特性。

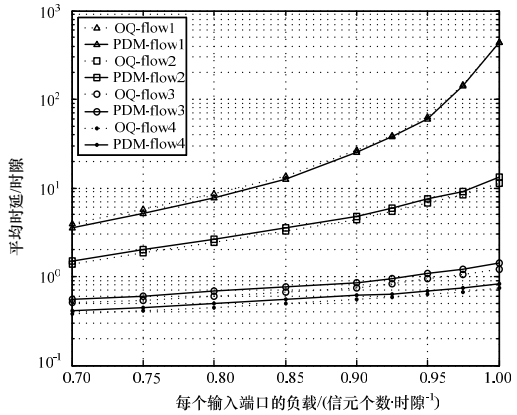

图 5　区分服务的业务流为 uniform 分布的时延性能

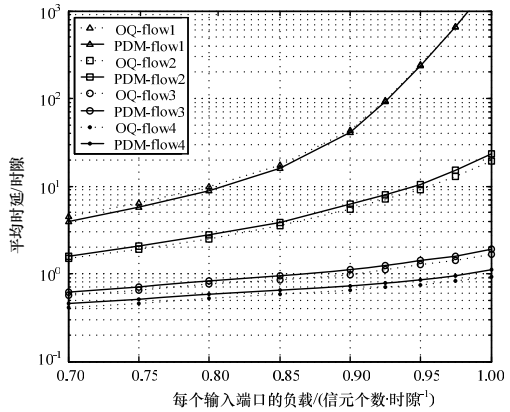

图 6　区分服务的业务流为 non-uniform 分布的时延性能

5　结束语

本文基于分布系统设计思想，用并行端口分布存储机制构建了一个新的 PSS 交换结构。通过流模型技术证明了 PSS 交换结构对满足大数定律的可容许到达业务能够获得 100%的吞吐量，无需加速即可实现不同优先级业务的 QoS 保障。由于 PSS 交换结构只需工作在线路速率，因此能够在高速网络中扩展实现。为易于在工程设计中实现，基于 PSS 交换结构提出了 SPDRR 调度算法。该算法比传统优先级调度算法的调度门限更多，需要的判决时间却更少，并且采用了简单的优先级双轮询机制，易于工程实现。仿真结果表明，应用 SPDRR 调度算法的 PSS 交换结构能够获得很好的性能。

参考文献

[1] Kesidis G, McKeown N. Output-buffer ATM Packet Switching for Integrated-services Communication Networks[C]//Proc. of IEEE ICC'97. Montreal, Canada: IEEE Press, 1997.

[2] McKeown N. Scheduling Algorithms for Input-queued Cell Switches[D]. Berkeley, USA: University of California, 1995.

[3] McKeown N, Anantharam V, Walrand J. Achieving 100% Throughput in an Input-queued Switch[C]//Proc. of IEEE INFOCOM'96. [S. l.]: IEEE Press, 1996.

[4] Chuang Shang-Tse, Goel A, McKeown N, et al. Matching Output Queueing with a Combined Input Output Queued Switch[J]. IEEE J. of Selected Areas in Communications, 1999, 17(6): 1030-1039.

[5] Dai Jim, Prabhakar B. The Throughput of Data Switches with and Without Speedup[C]//Proc. of IEEE INFOCOM'00. [S. l.]: IEEE Press, 2000.

编辑　张　帆