一种新的输入排队 crossbar 的公平调度算法

彭来献[1] ,田 畅[1] ,郑少仁[2]
(1. 解放军理工大学通信工程学院 ,江苏南京 210007 ;2. 南京邮电学院 ,江苏南京 210003)

摘 要: 本文提出一种新的输入排队 crossbar 调度算法 ——WMFS(weighted max-min fair scheduling) 算法 ,可以为各个竞争的流提供近似的归一化公平服务 ,保证了带宽分配的加权 max-min 公平性. 另外 ,仿真结果表明在均匀业务流到达情况下 ,WMFS 具有 100 %的吞吐量和良好的时延性能.

关键词: QoS; 输入排队; 加权 max-min 公平性; 归一化公平服务
中图分类号: TP393. 05 **文献标识码**: A **文章编号**: 0372-2112 (2004) 12A-048-04

A Novel Fair Scheduling Algorithm for Input-Queued Crossbars

PENG Lai-xian[1] ,TIAN Chang[1] ,ZHENG Shao-ren[2]
(1. *Institute of Communication Engineering , PLA University of Science and Technology , Nanjing , Jiangsu* 210007 , *China* ;
2. *Nanjing Univ. of Posts and Telecommunications , Nanjing , Jiangsu* 210003 , *China*)

Abstract: We proposed a weighted max-min fair scheduling (WMFS) algorithm ,that provides approximately normalized fair service guarantees and weighted max-min fair bandwidth allocation distribution among the contending flows ,for input-queued crossbars. Furthermore ,results from simulation show that the algorithm is able to achieve asymptotically 100 % throughput and low cell latency under uniform traffic.

Key words: quality of service ;input queuing ;weighted max-min fairness ;normalized fair service

1 引言

为了适应高速的网络环境和提高数据传输效率 ,网络交换设备逐渐摒弃传统的输出排队交换结构 ,通常采用基于输入排队 crossbar 的定长交换结构[1~3]. 此外 , Internet 中多媒体业务的增加和新业务的涌现迫使交换机/路由器必须具有 QoS(Quality of Service) 保证能力. 其中 ,带宽分配的公平性是一个重要的指标. 所谓带宽分配的公平性是指各流得到的带宽与其预约带宽成正比例[4]. 队列调度算法是提供公平性和其他 QoS 保证能力的主要手段.

在输出排队交换结构中 ,目前存在多种公平的调度算法[5] (如 WFQ) ,它们的目标是逼近理想化模型 GPS (Generalized Processing Sharing) [6] ,并提供带宽分配的公平性和其他 QoS 保证能力. 然而在输入排队 crossbar 中 ,分组在输入端缓存 ,到达一个输出端的流分布在不同的输入端. 为了实现 QoS 保证 ,分组调度算法需要了解分布在各个输入端口的队列状态 ,致使这些公平调度算法难以直接用于输入排队的队列调度. 因此 ,将分组调度与 crossbar 调度相结合 ,实现带宽分配的公平性和 QoS 保证也就成为路由器技术发展的一个关键技术问题.

一直以来 ,输入排队 crossbar 调度算法的研究主要围绕高速、高吞吐量、硬件易实现等方面 ,很少考虑带宽分配的公平性. 本文主要研究输入排队 crossbar 的公平调度问题 ,提出了加权 max-min 公平调度 (WMFS : Weighted Max-min Fair Scheduling) 算法并对 WMFS 进行了性能分析. 结果表明 ,WMFS 不仅能够保证带宽分配的公平性 ,而且还具有高吞吐量、低时延等良好性能.

2 问题描述

2.1 输入排队 crossbar 交换结构及其调度

图 1 表示了一个 $N \times N$ 的基于输入排队 crossbar 的交换结构. 处理的数据单元为固定长度的信元. 为了方便叙述和分析 ,我们假设所有输入/输出端的速率相同 ,crossbar 交换一个信元的时间间隔称为一个时隙. 为避免由于 HOL 阻塞 (Head

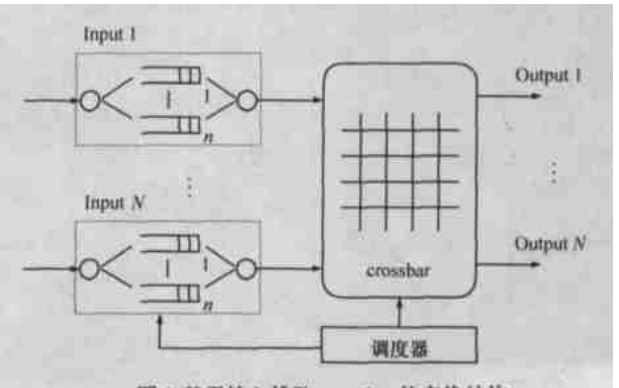

图 1 基于输入排队 crossbar的交换结构

收稿日期:2004-10-18;修回日期:2004-11-19
基金项目:国家"863"支持课题(No. 2001AA121063)

of Line Blocking)而导致的交换结构最大吞吐量下降问题[9],一般采用虚拟输出排队技术来消除 HOL 阻塞,即在输入队列中为每个流维护一个独立的 FIFO 队列.如果每个输入端都有 n 个流,那么 N 个输入端总共需要维护 $N \times n$ 个 FIFO 队列.

本文只考虑单播数据流.用 λ_i 表示输入端 i 的负载,指每个时隙内平均到达的信元数,要求 $\lambda_i \leq 1$.如果到达的信元平均分布在各个流中,则称业务流是均匀的,否则是非均匀的.

输入排队 crossbar 的调度算法主要解决信元在输入/输出端的竞争,即保证在一个时隙内每个输入端口至多发送一个信元,每个输出端口至多接受一个信元.因此,输入排队 crossbar 的调度问题通常也被描述成二部图匹配问题,输入排队 crossbar 调度算法有时也称为匹配算法.调度算法由图 1 中的调度器负责执行.

2.2 现有调度算法举例

目前,输入排队 crossbar 调度算法的设计主要考虑高吞吐量和高速,其中 PIM[7]、iSLIP[8]最具代表性.然而,这些算法没有考虑带宽分配的公平性.WPIM[10]是对 PIM 算法的改进,改善了带宽分配的公平性.WPIM 在一个帧(时隙的整数倍)内为每个流预约时隙数,使用 PIM 作为调度核心,$\log(N)$ 次迭代收敛.在一个帧内,当流得到服务的时隙数超过预约时,则屏蔽该流发出的请求,否则不屏蔽.因此,在超过预约之前,WPIM 实际就是 PIM,并不能保证带宽分配的公平性.

3 WMFS 算法描述

本节首先介绍加权 max-min 公平性和归一化服务概念,然后描述 WMFS 算法.

max-min 公平性最初用于流量控制中[11],保证所有的流分配的带宽在不超过各自负载的情况下尽可能均等,那些负载小于预约的流将产生多余的带宽,这些带宽将平均分配给那些负载大于预约的流.如果将某个流分配的带宽与预约带宽的比值称为"归一化带宽",那么加权 max-min 公平性就是"归一化带宽"的 max-min 公平性,保证分配的带宽与预约带宽成正比例,多余的带宽按比例分配.

在输出排队公平调度算法中,通常需要维护一个系统虚时间 $V(\cdot)$,并且根据预约带宽记录每个流的开始服务虚时间 $S_i(\cdot)$和结束服务虚时间 $F_i(\cdot)$.如果 $V(\cdot)$用比特作为单位,那么 $V(t)$则表示在时刻 t 各个流应得到的"归一化公平服务",$S_i(t)$则表示流 i 在时刻 t 得到的"归一化服务".GPS 在任何时刻 t 保证 $S_i(t)$等于 $V(t)$,也保证在任何时刻带宽分配都是加权 max-min 公平的.因此,公平调度算法设计的目标是尽量减少两者之间的差距.也就是说,$S_i(t)$越接近 $V(t)$,算法公平性越好.

本文提出的 WMFS 基本思想就是利用归一化服务解决输入排队 crossbar 的输入/输出端竞争,尽力缩小 $S_i(t)$和 $V(t)$的差距.同时,我们也使用两者差距的大小衡量算法公平性的优劣.

WMFS 采用迭代的策略改善吞吐量和信元时延性能,根据文献[7,8,10]的分析,我们规定 WMFS 一次执行过程也包含 $\log(N)$次迭代.WMFS 运行前所有输入/输出端均是未匹配的,每次迭代只考虑当前未匹配的输入/输出端.WMFS 每次迭代只包含如下两个步骤:

Step1:**许可**.每个未匹配的输出端从各个流中选择归一化服务最小的流,并向该流所在的输入端发送许可信号.

Step2:**接受**.如果一个未匹配的输入端接收到多个许可信号,则从中选择归一化服务与归一化公平服务差距最大的流,并向相关输出端发出接受信号,从而建立一个输入/输出连接.

4 性能分析

本节主要分析 WMFS 的吞吐量、时延和公平性等性能.在同样的参数设置下,输出排队的交换网络具有最优的性能.因此,我们将输出排队的性能作为其他算法性能的参考.本文仿真模型是 $N \times N$(N = 16 或 4)的输入排队 crossbar,每个输入端只有 N(即 $n = N$)个流,并且每个流分别到达不同的输出端.仿真的长度均为 100000 个时隙.

4.1 时延性能

仿真模型采用一个 16 ×16 的 crossbar,为了便于与 iSLIP 比较,假设各个流的预约带宽相同.图2(a)比较了在均匀业

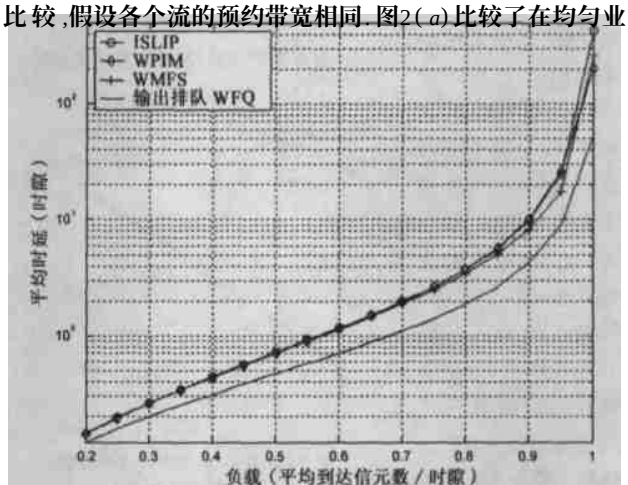

图 2(a)　WMFS、iSLIP 和 WPIM 在均匀业务流到达情况下的平均时延性能比较

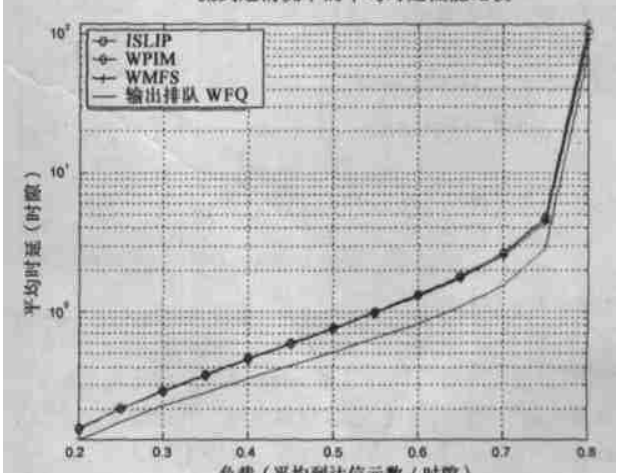

图 2(b)　WMFS、iSLIP 和 WPIM 在非均匀业务流到达情况下的平均时延性能比较

务流到达情况下 WMFS、iSLIP 和 WPIM(迭代次数均为 4)的平均时延性能.图中可以看出,当负载较低时,三者平均时延性能几乎相同;当负载较高(>0.8)时,WMFS 的性能甚至要优于 iSLIP,原因是 WMFS 更加公平地调度信元,缓解了突发对时延性能地影响,当负载越大,这个优势越明显.

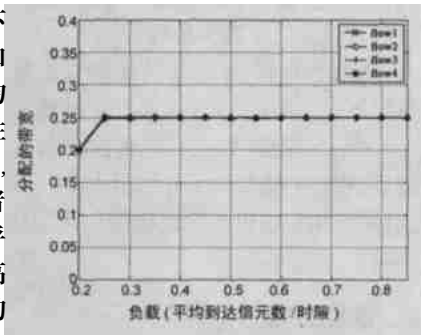

图 3 使用 iSLIP 调度的带宽分配

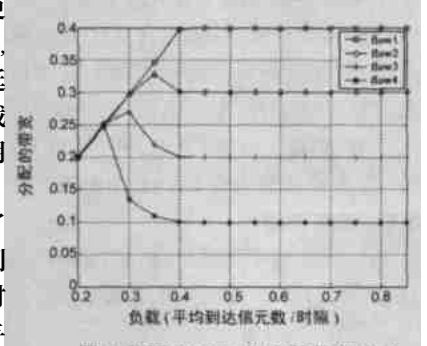

图 4 使用 WMFS 调度的带宽分配

图 2(*b*)比较了在非均匀业务流到达情况下的平均时延性能.这里采用与文献[10]中相同的非均匀业务流到达模式.假设 4 个端口连接服务器,其余 12 个连接客户机,每个客户机向每个服务器发送 10%的流量,其余均匀地发送到其他客户机.每个服务器向所有客户机均匀地发送 95%的流量,其余 5%则均匀地发送给其他服务器.所有服务器和客户机的发送负载均相同.在这种情况下,三者平均时延几乎相等.

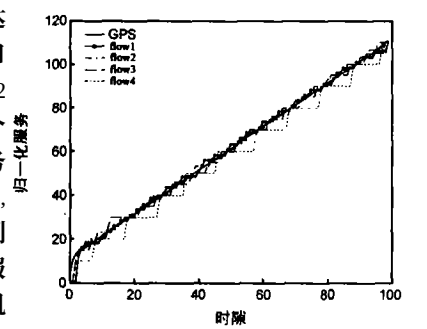

图 5 (*a*) 在时隙 [0,99] 内 WMFS 提供的归一化服务

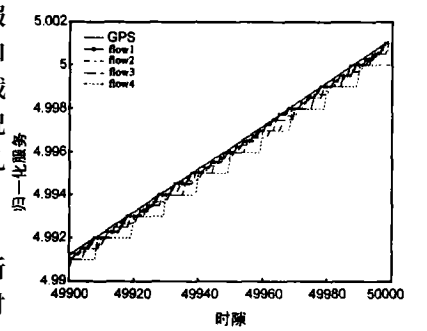

图 5 (*b*) 在时隙 [49900,49999] 内 WMFS 提供的归一化服务

4.2 带宽分配

下面我们分析当输出端超载时 WMFS 带宽分配的公平性.为了方便绘图,仿真模型采用一个 4×4 的 crossbar.每个输出端维护 4 个分别来自不同输入端的流,假设输出端 1 维护的 4 个流分别称为 flow1、flow2、flow3 和 flow4,它们分别来自输入端 1、2、3 和 4,预约带宽分别为 0.4、0.3、0.2 和 0.1,并且负载 p 相同.其他的流的负载均为 0.05.由于要求每个输入端 $\lambda_i \leq 1$,所以 flow1～4 最大负载为 0.85.当 flow1～4 负载在[0,0.85]中变化时,图 3、4 分别表示使用 iSLIP 和 WMFS 调度时 flow1～4 所获得的带宽.

从图 3 中可以看出,当 $p \leq 0.25$ 时,4 个流获得的带宽相同;当 $p > 0.25$ 时,iSLIP 仍然均等地为各个流分配带宽,这违背了加权 max-min 公平性.另一方面,图 4 展示了当 $p > 0.25$ 时,WMFS 根据预约带宽为每个流分配不同的带宽.当 $p \geq 0.4$ 时,每个流只获得预约的带宽.当 $0.25 < p < 0.4$ 时,WMFS 分配带宽是加权 max-min 公平的,例如当 $p = 0.3$ 时,flow1～4 获得的带宽分别为 0.3,0.3,0.267 和 0.133.因此,WMFS 在任何时刻分配带宽都是加权 max-min 公平的.

4.3 归一化服务

为深入研究带宽分配的公平性,我们分析 WMFS 为各个流提供的归一化服务,并与 WPIM 进行比较.仿真模型采与 4.2 节中的相同.当 $p = 0.5$ 时,分别使用 WMFS 和 WPIM 调度,WPIM 帧长为 1000 个时隙,我们记录下在时隙[0,99]和 flow1～4[49900,49999]内获得的归一化服务和归一化公平服务,如图 5、6 所示.

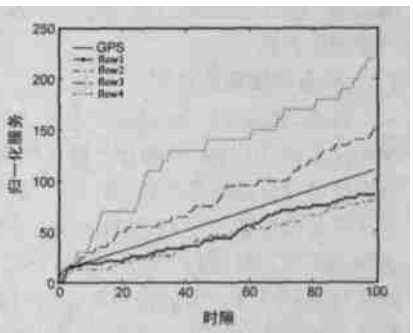

图 6 (a) 在时隙 [0,99] 内 WPIM 提供的归一化服务

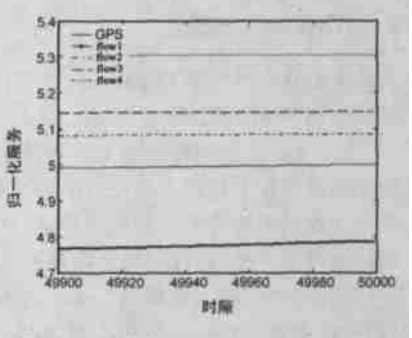

图 6 (*b*) 在时隙 [49900,49999] 内 WPIM 提供的归一化服务

从这些图中明显地看到,WMFS 为各个流提供的归一化服务非常接近理想的公平服务,它们之间的差距也很小.但是使用 WPIM 调度时,各个流得到的归一化服务和公平服务之间的差距变化剧烈.仿真结果表明:WMFS 可以提供近似的归一化公平服务,也就是说,在分配带宽时具有更好的公平性.

5 结束语

本文提出了一种输入排队 crossbar 的公平调度算法 WMFS,WMFS 具有带宽分配的加权 max-min 公平性,并且可以提供近似的归一化公平服务保证.仿真结果表明 WMFS 与其它调度算法相比,在拥有良好的公平性的同时还具有高吞吐量、低时延等性能.本文首次采用归一化服务来衡量公平性的好坏,可以更加精细地分析算法带宽分配的公平性.

由于 WMFS 需要维护归一化公平服务,当流的数目 n 很大时,WMFS 的开销将会限制其在高速交换环境中的应用,这是 WMFS 的缺陷.下一步,我们计划使用 SCFQ[12] 算法来克服

这个缺陷,并给出实用的、硬件易实现的公平调度策略.

参考文献:

[1] N Mckeown, et al. The Tiny Tera: A packet switch core[J]. IEEE Micro Magazine, 1997, 17(1): 26 - 33.

[2] Cisco Inc. Cisco 12000 series ——Internet Router. Product Overview [Z]. http://www.cisco.com, Oct 2001.

[3] Partridge C, et al. A 50-Gb/s IP router[J]. IEEE Trans. on Networking, 1998, 6(3): 237 - 248.

[4] A Demers, S Keshav, S Shenker. Analysis and simulation of a fair queueing algorithm[J]. Journal of Internetworking Research and Experience, 1990, 1(1): 3 - 26.

[5] H Zhang. Service displines for guranteed performance service in packet-switching networks[J]. Proc of IEEE, 1995, 83(10): 1374 - 1396.

[6] A Parekh, R Gallager. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The single node case [J]. IEEE/ACM Trans. on Networking, 1993, 1(3): 344 - 357.

[7] T E Anderson, S S Owicki, J B Saxe, C P Thacker. High speed switch scheduling for local area networks[J]. ACM Transactions on Computer Systems, 1993, 11(4): 319 - 352.

[8] N Mckeown. Scheduling algorithm for input-queued cell switches[D]. Ph D Thesis, UC Berkeley, May 1995.

[9] M J Karol, M Hluchyj, S Morgan. Input versus output queuing on a space-division packet switch [J]. IEEE Trans On Comm, 1987, 35 (12): 1347 - 1356.

[10] D Stiliadis, A Varma. Providing bandwidth grarantees in an input-buffered crossbar switch [A]. Proc of INFOCOM '95 [C]. Boston, MA, USA, April 1995. 960 - 968.

[11] D Bertsekas, R Gallager. Data networks (second ed.) [M]. Englewood Cliffs, NJ, USA: Prentice Hall, 1992.

[12] S Golestani. A self-clocked fair queueing scheme for broadband applications[A]. Proc. of INFOCOM '94[C]. Toronto CA USA: June 1994. 636 - 646.

作者简介:

彭来献　男,1978 年 3 月出生于安徽阜阳,1999 年于解放军通信工程学院获工学学士学位,2004 年 6 月在该院获得通信与信息系统专业博士学位,主要研究领域为高速路由器体系结构和高速交换结构.

田　畅　男,1963 年 2 月出生于山东青岛,博士,副教授,中国电子学会高级会员,主要从事宽带交换技术、网络安全和无线分组网的研究,在国内外有关刊物、会议发表论文 40 余篇.