

宽带路由器输入排队调度策略的综合研究^{*}

Study of Input-queued Router Working at its Scheduling Policy

刘宴兵¹ 李秉智¹ 幸云辉²

(重庆邮电学院计算机系 重庆400065)¹(北京邮电大学 北京100876)²

Abstract Currently there search on input-queued router is one of the most active research fields. The policy VOQ has the characteristics of improving throughput, satisfying QoS requirements and providing service fairly. If VOQ queueing policies are used with some cell scheduling policies, the 100% throughput can be achieved. This paper quantitatively evaluates and compares the performance of three scheduling policies: PIM, iSLIP and LPF.

Keywords Input-queued, Broadband router, Scheduling policy

1 引言

输入调度策略可以分成输入排队的调度策略和输出排队的调度策略。长期以来,人们一直认为输入排队的调度策略性能比较差,因而对输出排队的调度策略进行了大量的研究。但是,输出排队的调度策略要求输出端的接口速率是输入端的 N 倍(N 是端口数量),否则就会出现大量丢包的情况。随着输入端口速度的不断提高和输入端口数量的增多,输出排队的调度策略已经不能满足高速交叉开关的要求。分组在交换前被拆分成定长的数据包,数据包经过交换结构后在目的端口被组装成原分组,然后送往物理线路。目前的宽带路由器基本上采用输入缓冲的方式。这主要有以下两个方面的原因:第一,采用输入缓冲方式对存储器的速度要求不高。同输出缓冲方式和共享缓冲方式相比,输入缓冲方式对存储器的要求同交换结构的规模无关,从而可以使得路由器的规模较大。第二,在因特网上 TCP 数据所占的比例很大。TCP 数据具有突发的特点;并且 TCP 协议对丢包很敏感,一旦丢包发送速度就明显降低。为了保证 TCP 数据的业务质量,路由器内的存储容量应较大;而存储器的容量越大,其速度就越慢。这也要求采用输入缓冲方式以降低对存储器速度的要求。因此,研究人员又重新把注意力集中到了输入排队的调度策略,即大多数交换式路由器在交换时采用定长的信元^[1]。本文也正是基于以上背景展开综合研究的。

2 输入排队调度

输入排队调度是在交换结构的每条入线上设置缓冲器,如图1所示。

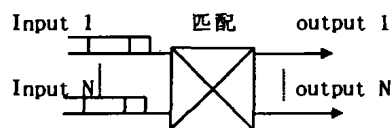


图1 输入缓冲

2.1 工作原理

输入缓冲一般采用简单的先进先出(FIFO)的排队规则。这样,当入线上每个时隙到来时,各个非空输入队列的队首信元进入竞争的仲裁,使去往不同出线的信元都能通过交换网

络而传送到出线,在竞争中失败的信元暂留在输入缓冲器中,等待下一轮(下一时隙)的竞争和传送。

2.2 排头阻塞现象

输入缓冲存在排头阻塞(HOL Blocking)现象。所谓 HOL 阻塞是指如果一个 FIFO 队列的队首信元竞争输出失败,其后继信元对应的输出端口即使是空闲的,该后继信元因受到队首信元的阻塞也不能输出的现象。

由于 HOL 阻塞,会使出线的传送效率降低。这里用吞吐率来表征信元通过的效率,可理解为在1个时隙内平均每条出线传送的信元数。显然,吞吐率最大为1。分析表明,在随机的均匀业务流模型下,当入线数 N 很大时,采用 FIFO 规则的输入缓冲方式的吞吐率为0.586。为解决线路排头阻塞现象问题,提出了一种称为虚拟输出队列(VOQ)的机制。其思路是:假设系统中有 N 个输入端口和 N 个输出端口,那么每个输入端口都有 N 个队列分别对应每个输出端口。这样,若一个输出端口发生阻塞对其他的输出端口不会产生影响。当然,在实际输出时,每个端口只对应于交叉开关中的一条线路,因此称之为虚拟输出队列。

3 基于 VOQ 输入排队调度策略模型及思想

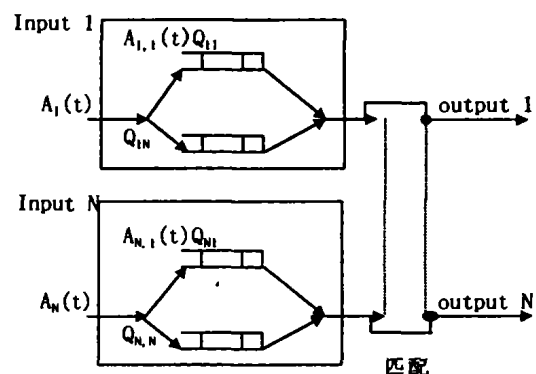


图2 VOQ 交换结构模型

图2是基于虚拟输出队列(VOQ)调度策略模型。图中 $A_i(t)$ 表示信元到达输入端口 i 的离散过程,任一信元时隙最多有一个信元到达输入端口 i ,此信元的目的端口确定为 $j(1 \leq j \leq N)$,该信元放入输入端口 i 的 j 缓冲队列 $Q(i,j)$ (其长度用

^{*} 青年教师科技基金项目(No. A2001-20)和信息产业部九五重点科技发展项目(No. 98048)。刘宴兵 讲师,硕士。李秉智 教授,博导。幸云辉 教授。

$L(i, j)$ 表示)。整个信元到达过程 $A(t) = \{A_i(t); 1 \leq i \leq N\}$ 必须是可接受的,也就是说每个端口输入和输出的总速率不超过其物理端口的带宽,即 $\sum_i \lambda_{i,j} < 1, \sum_j \lambda_{i,j} < 1$ 。

调度策略所寻求的结果就是在一个信元时隙内通过多次迭代到达输入和输出端口的最优匹配,从而使吞吐率接近 100%^[2]。

3.1 PIM 策略

PIM 策略基本思想是在信元时隙通过多步迭代在输入和输出端口之间匹配尽可能多的发送通道,并采用随机数防止仲裁的不公正性。在信元时隙开始,所有的输入和输出端口状态都标为空闲,每一步迭代之后都有一部分输入和输出端口被匹配上。每一步迭代中都只考虑未匹配上的端口,每次迭代信元含以下三个步骤:(1)请求:哪个输入端口 FIFO 非空,有信元要输出,该端口就向相应的输出端口提出发送请求;(2)授权:如果尚未匹配的输出口收到发送请求,它就在同时向它提出申请的多个输入端口中均匀地、随机地选择一个予以接受并对其进行授权;(3)接受:如果一个输入端口得到授权后(可能同时得到多个输出端口的授权),就从授权的端口中选择一个作为发送目标端口。

由于每次迭代只考虑上次迭代中未匹配的端口,因此迭代的结果不会和以前已经匹配的通道重复或冲突。策略的第二步:在多个请求的输入端口中随机地选择一个予以授权,可以确保所有的发送请求总能得到授权。

3.2 iSLIP 策略

和 PIM 策略一样,iSLIP 也是一种迭代策略,它克服了 PIM 策略的复杂和不公平性。每次迭代也有三个步骤:(1)请求:与 PIM 策略步骤 1 同;(2)授权:如果尚未匹配的输出口收到发送请求,它就按照循环优先级方式选择一个输入端口并通知该输入端口是否被许可。当且仅当输入端口接受该授权时,最高优先级指针 g_i 才加 1(以端口数 N 为模),指向获得授权端口的下一个端口。而刚被授权且接受授权的输入端口在下一时隙其优先级将为最低;(3)接受:当一个输入端口得到授权后(可能同时得到多个输出端口的授权),它就按照循环优先级方式选择一个,并使最高优先级指针 a_i 为 1(以端口数 N 为模)指向接受输出端口的下一个端口。

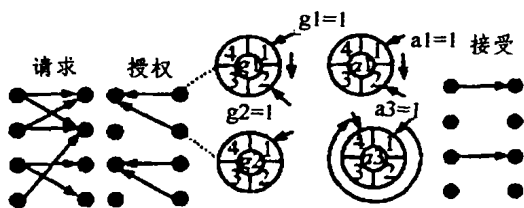


图3 iSLIP 策略的三个步骤

为了进一步理解策略的基本思想,下面根据数据结构中循环队列理论来阐述其原理。图3示出了一次迭代的三个步骤。图中输入端口1有信元分别指向输出端口1和2,输入端口2也有信元指向输出端口1和2,……。在信元时隙最后 a_1 和 a_3 分别指向了2和4,由于本次迭代输入端口1和输出端口1、输入端口3和输出端口3达成了匹配,而输入端口1虽然得到了输出端口2的授权,但它选择了输出端口1,所以指针 g_2 没有改变,待下次迭代进行匹配。可见,采用循环优先级指针不仅大大简化了实现的复杂度而且改善了公平性。此次建立起来的通道中的输入端口在下次匹配时将成为优先级最低的端口,而且当一个输入端口的请求未被满足时,它会在下次匹配时继续提

出该请求。所以,在输入端口的 N 个缓冲队列中,任何一个队列在 N^2 个时隙内必然得到发送机会。在我们项目中仿真模块实现授权的思想应用了这个基于循环队列的结构。

3.3 LPF 策略

LPF 策略是一种最大加权匹配策略,它考虑了各个输入端口缓冲队列长度以及信元竞争输出的拥塞程度。其加权值定义为:

$$w_{i,j}(n) = \begin{cases} R_i(n) + C_j(n), L_{i,j}(n) > 0 \\ 0, L_{i,j}(n) = 0 \end{cases}$$

式中 $L_{i,j}(n)$ 是第 n 时隙队列 $Q_{i,j}$ 的长度, $R_i(n) = \sum_j L_{i,j}(n)$ 表示输入端口 i 在第 n 时隙所有缓冲队列总和; $C_j(n) = \sum_i L_{i,j}(n)$ 表示所有要发送到输出端口 j 的信元缓冲队列长度总和。令 $S_{i,j}(n)$ 表示缓冲队列 $Q_{i,j}$ 得到发送的机会(得到发送机会, $S_{i,j}(n) = 1$; 否则, $S_{i,j}(n) = 0$)。LPF 策略就是寻求这样的匹配:在条件 $\sum_{i=1}^N S_{i,j}(n) \leq 1$ 和 $\sum_{j=1}^N S_{i,j}(n) \leq 1$ 下,使得 $\sum_{i,j} S_{i,j}(n) w_{i,j}(n)$ 最大。

4 调度策略综合评价

由于 PIM 策略采用随机数,所以实现比较复杂,而 iSLIP 策略仅需 $2N$ 个端口的调度器,而且每个调度器就是一个带优先级的 N 输入端编码器,所以实现比较简单。前面描述的 LPF 策略并不便于硬件实现,为了高速的硬件实现,必须将 LPF 策略变为迭代形式,但实现仍较复杂,具体综合评价如表1。

表1

策略	主要特点	吞吐率	非均匀业务量性能	实现复杂性
PIM	随机优先级最大匹配,多次迭代	单次迭代 63%, 多次迭代 100%	差	复杂
iSLIP	循环优先级最大匹配,多次迭代	100%	差	简单
LPF	最大加权匹配,可迭代实现	100%	优	复杂

结束语 本文对宽带路由器输入排队的调度策略进行了综合研究,这对当今宽带网络追求的改善吞吐量、保证 QoS 和公平服务等具有实际参考意义。我们认为研究输入队列调度策略不仅是设计高速 ATM 交换机的需要,而且对未来高速组播路由器的设计也有借鉴作用。在目前支持组播功能的高速路由器(如 BBN 公司的 50Gps 高速路由器 MGR)已经开始使用输入队列交换结构。我们相信,为了提高路由器的带宽,使用输入队列技术是未来高速路由器设计的一个趋势。

参考文献

- 1 Karol M. Input versus output queueing on a space division switch. IEEE Transactions on Communications, 1988, 35 (12): 1347~1356
- 2 Lee T. A modular architecture for very large packet switches. IEEE Transactions on Communication, 1990, 38(7): 1097~1106
- 3 黄立群,黄载禄. FQLP: ATM 网中一种新的实时业务调度算法. 电子学报, 2000, 28(4): 20~23
- 4 戴礼森,洪佩琳. 高速信元交换调度算法研究. 电子学报, 2000, 28(5): 96~98
- 5 刘宴兵,等. 基于三网合一解决方案的研究. 计算机科学, 2000, 27(11): 43