基于有限缓存输入队列交换机的缓存管理策略

白小明,邱桃荣
BAI Xiao-ming,QIU Tao-rong

南昌大学 信息工程学院,南昌 330031
College of Information Engineering,Nanchang University,Nanchang 330031,China
E-mail:baixm2004@163.com

BAI Xiao-ming,QIU Tao-rong.Buffer memory management strategy for input queuing switches of limited buffer space. Computer Engineering and Applications,2008,44(11):139-141.

Abstract: This paper addresses scheduling and memory management in input queued switches having fnite buffer with the objective of improving the performance in terms of throughput and average delay.Most of the prior works on scheduling related to input queued switches assume infinite buffer space.In practice,buffer space being a finite resource,special memory management scheme becomes essential.

Key words: buffer memory management;switches;scheduling

摘 要:介绍了一种基于有限缓存的输入队列交换机的集成调度和缓存管理策略(ISMM),该方案在吞吐量和平均延迟两项性能指标上有了很大改善。以前的许多关于输入队列交换机的调度方案都是在无限缓存的前提下完成的,但实际上,缓存是一种有限的资源,因此缓存管理方案就非常重要。

关键词:缓存管理;交换开关;调度
文章编号:1002-8331(2008)11-0139-03 **文献标识码:**A **中图分类号:**TP301

1 引言

输入队列结构交换机的分组缓存在输入端,存储器可以工作在链路速率,可扩展性大大优于输出队列结构。但由于分组缓存在输入端所造成的 HOL(Head Of Line)阻塞现象导致输入队列交换机采用 FIFO 调度时的吞吐率只能达到 58.6%[1]。近几年,随着链路速率和存储器速率矛盾的日益突出,输入队列交换机引起了研究人员的普遍重视,输入队列开关交换结构是高速交换机中最流行的结构。

交换机中的信元调度和对共享存储交换开关的缓冲管理是两个非常值得研究的问题,一直以来,所有的关于输入队列交换机的文献都认为输入队列长度是无限的,所以没有提出一种缓存管理方案,而且对调度算法和缓存管理都是分开研究的。最近,文献[2]的作者提出了一种考虑缓存是有限资源的基于最大权匹配(MWM)的缓存管理方案(BCT),然而,此方案的复杂性是其实现的最大瓶颈。

本文将针对具有有限缓存的输入队列交换机提出一种集成的调度和缓存管理方案,它提供了一种相对具有无限缓存系统有更好的时间延迟特性和更大吞吐量的解决方法。文中理论分析了其性能,并进行了仿真实验。

2 基于 ISLIP 的 ISMM 方案(ISLIP-MM)

2.1 网络模型

考虑一个有 N 输入和 N 输出的输入队列交换机,在输入端口处的缓存大小分别是 $B1,B2,\cdots,Bn$。到达的信元经过决策判定它的目的输出端口,从输入端口 i 到达由输出端口 j 离开的信元存储在 $VOQ(i,j)$中,在输入端口 i 的所有等待的 VOQ 总长度小于端口的缓冲 Bi。为了使算法尽量简单,假定它等于缓冲区的大小,$Bi=B$,到达的固定长度的分组或者信元允许把时间分割成离散的时间片,且在每个时间片里调度发生一次,假定信元在每个时间片的起始时到达并且在每个时间片的结束时离开。

2.2 ISMM

缓存管理问题是对那些到达同一输入口却拥有不同输出口的信元决定如何共享缓存,这部分工作是在每个信元到达时由一个允许性测试完成的,它决定是接收还是拒绝此信元,然而这个测试仅仅在输入缓存占有率达到一个阀值时才有效。如果没有达到这个阀值,那么所有到达的信元都是可以接收的。这个测试考虑了目前的缓存状态和开关状态来时做出的决定更加有效。缓存状态包括了在每个虚拟输出队列(VOQ)中信元的数目,但是在一个输入端口 i 可以获得的信息是每个 VOQ 那个输入口的信元数目。

开关的状态取决于在交换机中实现的特定调度算法,在开关状态中可以获得两个重要的特性。第一个特性是输入口 i 没有得到输出端口 j 准许信号的最小时间片数目,这个特性取决于其它输入口上有相同输出端口 j 的信元的数目(除了输入口

基金资助:江西省科技厅工业攻关项目(No.2005112)。
作者简介:白小明(1966-),男,副教授,研究方向为计算机网络及其应用;邱桃荣(1964-),男,博士,教授,研究方向为人工智能及其应用。
收稿日期:2007-08-23 **修回日期:**2007-11-23

i),它和那些输入端口的信元总数是没有关系的。在请求阶段,每个输入端向任一个带有信元缓存的输出端口发送一个请求,所以每个输出端口也拥有每个输入端口以那个输出端口为目的输出的信元数目。因为使用的是决定型算法,所以输出端冲突的数目可以由输出仲裁单元计算出来。

这些信息需要从输出端口传送到不同的输入端口,在任一个迭代匹配算法中除了请求、准许、接受的三个步骤外加入另一个步骤,称为反馈阶段。输出端口发送包含竞争同一端口的输入端口数目的反馈信号给输入端口,这要求总数为 $N2$ 条消息从输出仲裁到输入仲裁。在每个时间片传送这些信息减慢了开关的工作速度。需要附加大量的硬件在每个时间片计算出数值,然而网络负载流量和突发模式下的信元是紧密相关的,因为在连续的时间片里队列长度几乎没有任何改变。因此以同一输出端口为目的端口的输入端口数目也是变化很小的。在 ISMM 算法中,输出仲裁仅仅给那些请求信号被准许的输入端口发送反馈信号,通过仿真发现它和全反馈方案相比,可以减少很多信息流量,而在性能上却没有多少损失。

第二是任一个输入端口 i 没有收到任一个输出端口 j 的接收信号的最少时间片数目。这个特性取决于输入 i 的信元到达其它输出端口(除了端口 j)的数目,称之为输入端冲突(比如属于同一个输入端口的 VOQs 相互间竞争准许输出信号)。这个特性的所有相关信息需要当场计算出来。在时间片 t_0 时输入 i 到输出 j 的冲突总数是输入端冲突和输出端冲突数目之和。

然后,检查在每个输入端口处所有的 VOQs 共享的缓存空间,输入 i 到输出 j 的信元丢弃率和 VOQ(i,j)所占有的总缓存大小也是成比例的。这个结论有两个用途:(1)可以保持不同链路间的公平;(2)减少平均延迟时间。

ISMM 方案通过在缓存占有率到达阀值后丢弃高延迟的信元使交换机可以继续接收低延迟的信元。

2.3 ISMM 和 ISLIP

ISLIP 的开关状态由所有输出端口的允许指针和所有输入端口的接受指针,但是在输入端口 i 可以获得的信息是每个虚拟输出队列中输入端口 i 信元的数目和接受指针的位置。

由于 ISLIP 是一种轮循调度算法,只有输入端口 i 和输出端口 j间的信元被允许和接受后,其它输入端口到输出端口 j 的信元才能得到下一个指针位置。当在时间片 t_0 输出端口 j 发送允许请求信号到输入端口 i,反馈的数值可以有下面的公式决定:

$$output\text{-}congestion_{ij}(t_0)=\sum_{k=i+1}^{K_{j-1}} r_{kj}(t_0),$$

$$r_{kj}=\begin{cases}1 & \text{当输入 } k \text{ 到输出 } j \text{ 有请求}\\ 0 & \text{其它}\end{cases}$$

输出端冲突值表明了 t_0 时刻开始输入端 i 的请求没有被输出端 j 准许的时间片数目,这个值是低范围的,因为可能很多输入口有在 t_0 时刻以后到输出端口的信元,而不是在 t_0 时刻,这个值在每次输入 i 的请求被输出 j 准许后更新,所以输入口 i 维持 N 条这种带输出端冲突的队列,每条恰好对应一个输出端口,且它在每个不同的时间片(当输入 i 的请求被相应得输出端口 j 准许时)里更新。

输入端口 i 发送往输出端口 j 的信元一旦被准许和接受后,只有当输入口 i 到达其它端口的请求被准许和接受后才能进行调整。在时间 t_0 处,输入端冲突的值可以有下面的公式决定:

$$input\text{-}congestion_{ij}(t_0)=\sum_{k=a_i}^{j} l_{kj}(t_0),\ l_{kj}=\begin{cases}1 & \text{当 } l_k>0\\ 0 & \text{其它}\end{cases}$$

最坏的情况是输入端口必须等到一个输出端口给它发送一个准许信号并且接受它的时间片数目是 $N2$,当所有输入端的虚拟输出队列非空时会发生上述情况。

3 仿真与结果

任一种资源分配方案都是基于它的效率和公平性来判定的。定义 Peff 为系统吞吐率和系统响应时间的比值,下面仿真不同的负载流量下评价 ISLIP-MM 的性能并且和 ISLIP,MWM 和 BCT 算法的结果进行对比。

3.1 算法 ISLIP-MM 和 ISLIP 的比较

图 1 表明了两者在突发均匀流量模型下,信元长度为 32 和 64 时的平均延迟和吞吐率,突发长度反映了每个忙周期的平均时间。ISLIP-MM 算法在不牺牲吞吐率的前提下使平均延迟时间减少 10%~15%。然而当负载大于 90%时,各算法延迟时间的差异非常小。这是因为,在重负载的条件下 ISLIP 要分割成更多的时间单元所有延迟时间由调度算法单独决定。在突发均匀流量模型下,信元到达时可以由开关和缓存的状态来决

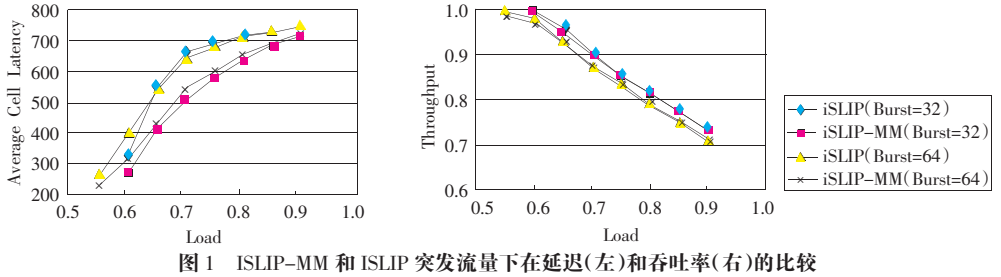

图 1 ISLIP-MM 和 ISLIP 突发流量下在延迟(左)和吞吐率(右)的比较

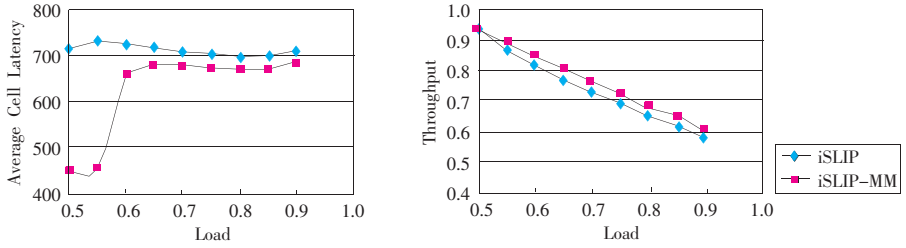

图 2 ISLIP-MM 和 ISLIP 非均匀负载流量下延迟(左)和吞吐率(右)的比较

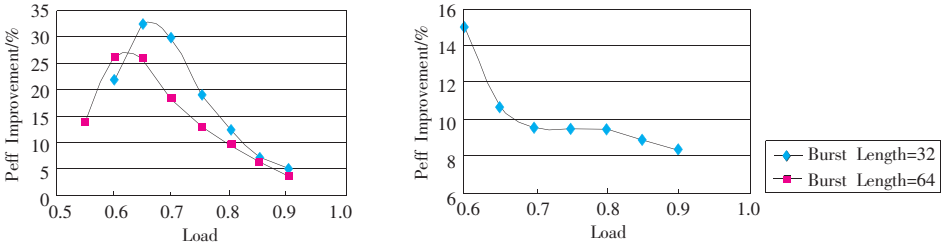

图 3 ISLIP-MM 在突发流量(左)和非均匀流量(右)性能效率相比 ISLIP 的改善率

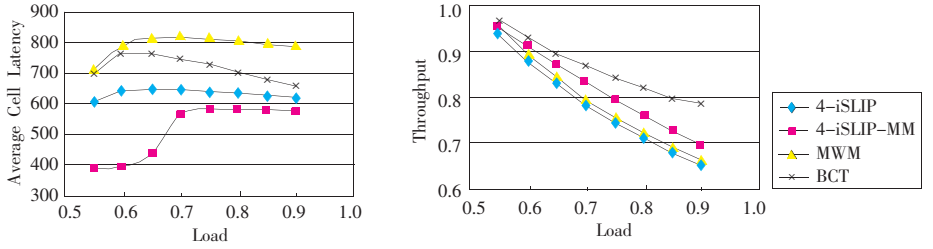

图 4 4-ISLIP、4-ISLIP-MM、MWM 和 BCT 在非均匀流量负载下的延迟时间(左)和吞吐率(右)的比较

定是否丢弃。

图 2 表明了两者在非均匀流量模型下的平均延迟和吞吐率。ISLIP-MM 算法在延迟时间和吞吐率上较 ISLIP 算法的性能有 5% 的提高,这说明在输入端流量扰动越大时 ISLIP-MM 算法将获得越好的性能。

图 3 表明了 ISLIP-MM 算法在性能效率上相比 ISLIP 的改善率大致在 10%~35%之间 。在低负载流量下,Peff 先增加到一个比较高的值后然后开始下降, 这是由调度算法本省决定的。低负载情况下没有很多的信元丢弃所有算法在吞吐率和延迟上也没有多少差异,而在重负载情况下,ISLIP 算法被分成更多的时间片,所以延迟时间和吞吐率更多是由调度算法本身而不是缓存管理方案来决定的。

3.2 ISLIP-MM,MWM 和 BCT 的比较

图 4 表明了 4-ISLIP,4-ISLIP-MM,MWM 和 BCT 在非均匀流量负载下的平均延迟时间和吞吐率,前面的数字 4 表示每个调度周期的准许-接收阶段的 4 次迭代。在负载流量变化很大时,4-ISLIP-MM 不论是在信号延迟还是在吞吐率上都比 MWM 性能好的多。 而 BCT 算法相比 4-ISLIP-MM 算法有着更高的吞吐率但是延迟时间太长,但是由于算法本身的高度复杂性,它实际上是不可能实现的。

4 结论

本文主要介绍了有限缓存输入队列交换机中一种叫做 ISMM 的存储器管理方案来改善平均延迟和吞吐率两个性能指标。这个方案在决定是否接收还是丢弃一个信元是同时考虑了开关的状态和缓存的状态。所提出的算法实现起来非常简单且对交换机工作速度没有多大的影响。仿真结果表明上述方案在系统延迟和吞吐率方面相比 ISLIP (滑动迭代轮循匹配),MWM(最大权匹配)算法特别是在非均匀负载流量(负载流量变化很大)情况下有很大改善,在吞吐率方面和 BCT(终端平衡阻塞算法)算法差不多。

参考文献:

[1] Karol M,Hluchyj M,Morgan S.Input versus output queuing on a space division switch[J].IEEE Trans on Communications,2003,35(12):1347-1356.

[2] Sarkar S.Optimum scheduling and memory management in input queued switches with finite buffer space[J].IEEE INFOCOM,2003:792-799.

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

[4] Marsan M A,Bianco A.Packet scheduling in input-queued cell-based switches[C]//IEEE INFOCOM'2001,Alaska,USA,2001.

[5] Wu Jun,Chen Qing,Luo Junzhou.A round-robin scheduling algorithm by iterating between slots for input-queued switches[J].Journal of Software,2005,16(3):375-383.

[6] Wu Jun,Chen Qing,Luo Junzhou.A scheme and behavior analysis of duplicated ports switches with line rate buffers[J].Journal of Software,2003,14(12):2060-2067.

[7] Lin M,McKeown N.The throughput of a buffered crossbar switch[J].IEEE Communications Letters,2005,5:465-467.

[8] Yoshigoe K,Christensen K J.An evolution to crossbar switches with virtual output queuing and buffered cross points[J].IEEE Network,2003,17(5):48-56.

[9] Rojas-Cessa R,Oki E,Jing Z,et al.CIXB-1:combined input-one-cell-crosspoint buffered switch[C]//IEEE HPSR'01,Dallas,Texas,USA,2001.

[10] Rojas-Cessa R.Round-robin with adaptable-size-frame arbitration for input-crosspoint buffered switches[C]//2004 IEEE Int'l Conf on Communications,Paris,France,2004.

[11] Yoshigoe K,Christensen K,Jacob A.The RR/RR CICQ switch:Hardware design for 10-Gbps link speed[C]//The 2003 IEEE Int'l Performance,Computing,and Communications Conference,Phoenix,USA,2003.

[12] Gunther N J,Christensen K J,Yoshigoe K.Characterization of the burst stabilization protocol for the RR/RR CICQ switch[C]//The 28th Annual IEEE Int'l Conf on LCN'03,Bonn,Germany,2003.

[13] Mhamdi L,Hamdi M.MCBF:a high-performance scheduling algorithm for buffered crossbar switches[J].IEEE Communications Letters,2003,7(9):451-453.

[14] 李勇,罗军舟,吴俊.一种交叉点小缓存 CICQ 交换机高性能调度算法[J].计算机研究与发展,2006,43(12):2033-2040.

[15] 王重钢,隆克平,龚向阳,等.分组交换网络中队列调度算法的研究及其展望[J].电子学报,2001,29(4):553-559.