ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.11, November 2007, pp.2800−2809
DOI: 10.1360/jos182800

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

缓冲交叉开关交换结构性能分析*

孙书韬[1,3+], 贺思敏[2], 郑燕峰[2], 高 文[1,2]

[1](中国科学院 研究生院,北京 100049)

[2](中国科学院 计算技术研究所,北京 100080)

[3](中国传媒大学 计算机与软件学院,北京 100024)

Performance Analysis of a Buffered Crossbar Switch

SUN Shu-Tao[1,3+], HE Si-Min[2], ZHENG Yan-Feng[2], GAO Wen[1,2]

[1](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

[2](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[3](School of Computer and Software, Communication University of China, Beijing 100024, China)

+ Corresponding author: Phn: +86-10-65783396, Fax: +86-10-65783241, E-mail: stsun@cuc.edu.cn

Sun ST, He SM, Zheng YF, Gao W. Performance analysis of a buffered crossbar switch. *Journal of Software*, 2007,18(11):2800−2809. http://www.jos.org.cn/1000-9825/18/2800.htm

Abstract: This paper analyzes the performance of a buffered crossbar switch under bursty traffic. It derives the saturated throughput for a buffered crossbar switch with multiple queues at each input port by the proposed analytic model. The saturation throughput sharply decreases from 1 and converges to 0.5 with the increasing of average burst length, and it approaches 1 as the number of queues per input increases. The accuracy of the theoretic analysis is also investigated by extensive simulation. Results from this paper can be used as a guidance to design optimal buffered crossbar switches.

Key words: buffered crossbar switch; input queuing; scheduling; modeling; performance analysis

摘 要: 分析了一种缓冲交叉开关交换结构在突发流量到达下的性能.通过建立分析模型,给出了每个输入端口拥有单个或多个输入队列的缓冲交叉开关结构的饱和吞吐.结果显示,对于单输入队列结构而言,随着突发平均长度的增加,饱和吞吐迅速从 1 下降,并收敛于 0.5.随着每个输入端口输入队列数目的增加,饱和吞吐率逐渐接近 1.仿真实验验证了分析模型的准确性.该结果可以用于指导基于缓冲交叉开关的路由交换设备的优化设计.

关键词: 缓冲交叉开关交换结构;输入排队;调度;建模;性能分析

中图法分类号: TP393 **文献标识码**: A

输入排队技术广泛用于高速路由和交换设备.对于无缓冲的交叉开关结构的性能,人们进行了充分的研究.

* Supported by the National Natural Science Foundation of China under Grant Nos.69983008, 60773150, 90604029 (国家自然科学基金); the Knowledge Innovation Program of the Chinese Academy of Sciences under Grant No.KGCXZ-103 (中国科学院知识创新工程); the Basic Research Fund of Institute of Computing Technology, the Chinese Academy of Sciences, under Grant No.20056090 (中国科学院计算技术研究所基础研究基金)

Received 2005-10-13; Accepted 2006-08-21

文献[1]给出了采用单输入队列的无缓冲交叉开关交换结构的性能分析结果,在均匀分布的伯努利独立同分布流量下,对于一个$N\times N$的交叉开关交换结构,当N趋向于无穷大时,最大吞吐率为 58.6%.文献[2−4]分析了突发流量以及变长分组下单输入队列无缓冲交叉开关结构的性能.采用虚拟输出队列(virtual output queues,简称VOQs)结构[5,6]可以完全消除队头阻塞,取得了较高的吞吐率.文献[7,8]给出了VOQ结构的相关性能分析结果.有关具有虚拟输出队列的无缓冲交叉开关交换结构调度算法与性能的详细情况可参阅文献[9].对于具有多输入队列(每个输入端口输入队列数在 1 与N之间)的无缓冲交叉开关结构,文献[10,11]在假定逻辑上独立的子交换结构顺序调度的基础上给出了伯努利独立同分布的流量到达下的性能分析结果.

由于具有同时满足高速度与低成本要求的潜力,缓冲交叉开关交换结构近来引起了较多的关注.通过在交叉点加入少量的缓存,交换调度问题与无缓冲的交换结构相比有了很大的区别,可以大为简化.而增加少量交叉点缓存在当前的技术水平上是可以实现的.仿真和分析结果表明,带有虚拟输出队列的缓冲交叉开关结构可以达到较高的吞吐率[12−16].但是,目前关于非虚拟输出队列的缓冲交叉开关交换结构性能的研究还很少见,只有Lin等人[17]分析了一个典型的缓冲交叉开关结构的吞吐性能.这一典型结构的特征是:每个输入端口设置一个输入队列,每个交叉点设置一个分组大小的缓冲区;每一个时隙,当输入调度器发现输入队列队首分组要进入的交叉点缓冲区非空时,丢弃输入队列队首分组.他们通过丢弃阻塞的分组,使每一个时隙出现在输入队列队首的信元的目的端口具有马尔可夫性,从而建立起分析模型.结果显示,在均匀的伯努利独立同分布到达的流量下,总能达到超过 80%的吞吐率.特别地,当交换结构大小N趋向于无穷时,吞吐率趋向于 100%.基于对丢弃阻塞的分组情况下吞吐率的结论,他们推断,当N趋向于无穷时,即使不丢弃阻塞的分组,缓冲交叉开关结构也可以达到接近 100%的吞吐率.

然而,文献[17]的假设有两方面与实际情况存在较大差距.在实际情况下,一方面阻塞的分组通常需要缓冲并等待交换而不是丢弃;另一方面,到达流量通常表现出自相似特性,而不是伯努利独立同分布的.自相似的长程依赖和突发性对于排队系统性能具有很大的影响,传统的伯努利独立同分布流量模型无法体现流量的自相似特性.针对已有成果在缓冲交叉开关性能分析方面研究的空白,本文分析了在突发流量下,并且在交换机缓存阻塞的分组时,每个端口拥有 1 个或多个输入队列的缓冲交叉开关交换结构的吞吐性能,同时通过仿真验证了分析结论.

本文第 1 节给出交换结构与流量模型.第 2 节给出吞吐性能分析结果.第 3 节给出分析与仿真结果的比较分析.第 4 节为本文的研究结论.

1 交换体系结构与流量模型

1.1 体系结构

我们考虑一个 $N\times N$ 的缓冲交叉开关结构,每一个交叉点拥有一个分组大小的缓冲区,每个输入端口有 m 个先进先出的输入队列,编号为 0~(m−1),$1\le m<N$.输出端口分为 m 个不相交且数目相等的组,输入端口的每个输入队列对应且只对应一个输出组,也就是说,到对应输出组的分组只能进入这一输入队列等待调度.这样,一个$N\times N$的含有多输入队列的缓冲交叉开关交换结构就从逻辑上被分为 m 个 $N\times(N/m)$的子交换结构,我们将其编号为 0~(m−1).输入队列到输出组的一个很直接的对应方式是使每个输入端口中编号为 k 的输入队列对应编号为($im+k$)的所有输出端口,其中 $0\le i<\lfloor N/m\rfloor$.图 1 为一个采用上述分组及对应方案、每个输入端口有两个队列的 4×4 缓冲交叉开关结构示意图.

我们假定N为 2 的幂次,且为m的倍数;分组是定长的,时间被划分为时隙(slot);交换结构的加速比为 1,亦即,每一个时隙内,输入端口只能传送 1 个分组到交叉点缓冲区,输出端口也只能从交叉点缓冲区接收 1 个分组.到达输入端口i,目的为输出端口j的分组首先会被送入交叉点缓冲区$B[i,j]$,然后再从交叉点缓冲区传送到输出端口.调度分为两个阶段.第 1 阶段为输入调度,每一个输入端口独立、并行地检查交叉点缓冲区是否可以容纳队头分组,如果可以,则选定一个队头分组并传送到交叉点缓冲区.为了进行建模分析,与关于无缓冲交叉开关结构下多输入队列的研究[10,11]类似,我们假定交换结构的输入调度的检查是顺序进行的,每一个输入端口的调度

器从子交换结构 0 的输入队列开始,依次检查各个子交换结构输入队列的队首分组及其对应的目的交叉开关缓冲区,直到发现某一队列队首分组的目的交叉开关缓冲区为空为止,在此时隙传输这个分组.第 2 阶段为输出调度阶段,每一个输出端口独立、并行地检查其对应列的交叉点缓冲区中是否有分组,如果有,则从中随机选择一个输出到输出端口.特别地,在输入调度阶段开始传入交叉点缓冲区的分组可以直接参与输出调度,因而在后续分析中假定,在每个时隙开始时,分组被调度到交叉点缓冲区;在每个时隙结束时,到输出端口的分组被从交叉点缓冲区取走.

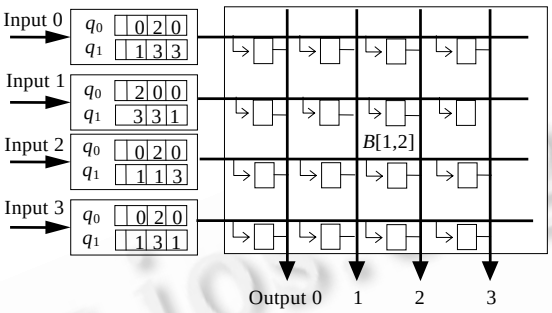


Fig.1 Architecture of buffered crossbar with multiple queues per input

图 1 多输入队列缓冲交叉开关体系结构

1.2 流量模型

在每一输入端口,分组到达过程表征为一个 2 态马尔可夫调制的伯努利过程的特例——ON-OFF模型(如图 2 所示)[18].在ON状态,每个时间片产生一个分组;而在OFF状态时不产生分组.模型从状态ON转移到状态OFF的概率为$(1-\alpha)$,从OFF状态转移到ON状态的概率为$(1-\beta)$.在ON状态时连续产生的分组称为一个突发(burst),一个突发内的所有分组都具有相同的目的端口.ON-OFF模型产生的突发的长度服从几何分布,均值$L=1/(1-\alpha)$.如果给定输入流量λ以及突发长度均值L,则可计算出α和β分别为$\alpha=1-1/L$和$\beta=(1-\lambda(2-\alpha))/(1-\lambda)$.同时,我们假定到一个子交换结构的流量分布是均匀的,即一个突发到对应子结构各个目的端口的概率均相等.

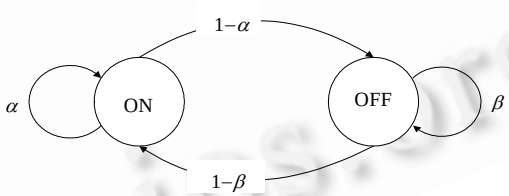


Fig.2 ON-OFF model

图 2 ON-OFF 模型

2 吞吐率分析

2.1 输入端口为单输入队列的缓冲交叉开关交换结构

首先,考察一个 $N\times N$ 的单输入队列缓冲交叉开关结构.一个 4×4 的单输入队列缓冲交叉开关结构如图 3 所示.我们假定饱和吞吐率为λ.所谓的饱和吞吐是指使输入队列总有积压(backlog)时输出端的吞吐率.对于这个 $N\times N$ 的输入排队缓冲交叉开关交换结构,在均匀分布的流模型下,所有的输入队列同时进入饱和状态,每个输入端口与输出端口的流量均为λ.

对于一个输出端口j,那些在第j列交叉点缓冲区中的分组构成了一个具有固定服务速率的队列,称为列队列Q_j.由于所有的列队列都是可互换的,具有相同的统计特性,因此在后续的讨论中只需研究一个代表性的列队列,称为标记列队列(tagged column queue).

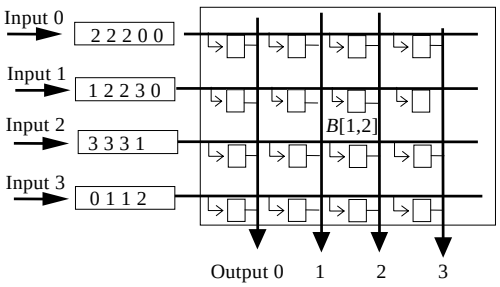


Fig.3 A 4×4 buffered crossbar switch with single queue per input

图 3 一个 4×4 的单输入队列缓冲交叉开关结构

下面分析如何为列队列建模.首先我们给出突发相遇的定义.这里的突发相遇专指从同一个输入端口到同一个输出端口的突发的行为.当一个突发的最后一个分组尚未从所在的交叉点缓冲区 $B[i,j]$中取走,而另一个从输入 i 到输出 j 的突发的第 1 个分组已经到达了输入队列队首时,我们说这两个突发相遇了.

从输出端来看,当从交叉点缓冲区取走一个分组之后,只要一个突发或几个相遇的突发没有结束,交叉点缓冲区立刻就会被突发中的后续分组填充.由ON-OFF流量模型可知,取走一个分组后,一个突发终止的概率为 $1/L$,则分组从交叉点缓冲区取走后,有后续分组填充的概率 $p'=1-1/L+p_m$,其中,p_m为两个突发相遇的概率.对于从同一输入到同一输出的两个突发,设s为前者离开输入队列进入交叉点缓冲区至后者到达输入队列队首的时间.当N趋向于无穷大时,$\forall T>0$,有概率$P(s<T)\to 0$;另外,$\forall \lambda<1$,列队列是稳定的,因而其队长是有界的,也就是说,一个分组在列队列中的停留时间是有界的.因此这两个突发相遇的概率趋向于 0.于是,当N足够大时,我们可以忽略突发相遇的影响,近似地认为两个从同一输入到同一输出的突发不会相遇.

另一方面,当 N 趋向于无穷大时,不同输入端口的输入队列变为互相独立的,在极限情况下,突发从各输入端口到达一个列队列的过程构成一个均值为 λ/L 的泊松过程.本文所说的突发到达列队列是指一个突发的第 1 个分组到达这个列队列的交叉点缓冲区.

基于以上讨论,可以将列队列建模为一个带反馈回路的 M/D/1 排队系统,当一个分组输出后,将以概率 $p=1-1/L$ 返回排队系统.分组进入 M/D/1 系统的速率为突发到达速率 λ/L.

设H^t标记列队列在第t个时隙内完成输入调度时的分组数目,则H^t的动态变化方程为

$$H^{t+1}=H^t-F^t\varepsilon(H^t)+A^{t+1} \tag{1}$$

其中,A^{t+1}是在$(t+1)$时隙到达标记的列队列的突发数;$\varepsilon(H^t)=\min(1,H^t)$;$F^t$以概率$p$取 0,以概率$(1-p)$取 1.采用文 献[19]Sec. 5.6 中提出的标准z变换分析技术,可以得到稳态时的队列长度分布的概率母函数,

$$H(z)=\frac{(1-z)(1-\lambda)}{(1-z)+lz-lze^{\frac{\lambda(1-z)}{l}}} \tag{2}$$

对 z 微分,并令 z 为 1,得到稳态队列长度的期望

$$E(H)=\frac{\lambda}{1-\lambda}-\frac{\lambda^2}{2L(1-\lambda)} \tag{3}$$

如果给定 $E(H)$的值,就可以通过式(3)解出λ,从而得到饱和吞吐率.幸而,通过对突发分组进入交叉点缓冲区过程的分析,我们的确可以得到 $E(H)$的值.

图 4 给出了一个含有 4 个分组的突发进入交叉点缓冲区的过程(突发长度为 4 个分组,e_i表示第i个分组进入的时刻).从前面的讨论可知,当N趋向于无穷时,两个从同一输入到同一输出的突发相遇的概率趋向于 0,也就是说,一个突发的首分组被阻塞的概率趋向于 0.因此,我们假定突发的首分组都会无阻塞地进入交叉点缓冲区.然而,对于其余的 3 个非首分组,它们不得不等到其前面的分组从交叉点缓冲区被取走后才能进入.这个突发进入交叉点缓冲区共经历 3 段等待时间加上最后一个分组进入占用的时隙.图中w_i是突发中第$(i+1)$个分组等待的时间,其时间跨度为从第i个分组进入交叉点缓冲区直到其离去的时间.等待时间W是一个随机变量.根据上文给出的流量模型,突发中平均非首分组的个数为$(L-1)$.假定系统已经以饱和吞吐率λ从某一个输入端口传送了K个

突发到某一交叉点缓冲区,则非首分组的平均等待时间为

$$E(W)=\frac{KL/\lambda-K}{K(L-1)}=\frac{L-\lambda}{\lambda\ (L-1)} \tag{4}$$

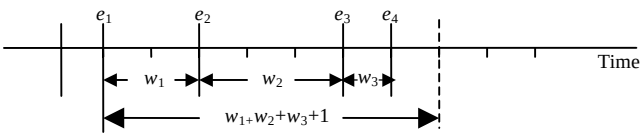


Fig.4 The entering process of a burst to a crosspoint buffer

图 4 一个突发进入交叉点缓冲区的过程

一个非首分组的等待时间实际上是其前驱在交叉点缓冲区所逗留的时间.因为从一个输入到一个输出的突发的到达是随机的,来自不同输入端口的突发具有完全相同的统计特性,并且几何分布具有无记忆性,对于长度大于 1 个分组的突发,非首分组的数量服从几何分布,分布参数与考虑所有突发时的突发长度分布参数相同.因而可以认为从一个输入到标记列队列的分组的等待时间是标记列队列中所有分组逗留时间的采样,标记列队列中平均分组逗留时间等于输入队列中分组的平均等待时间.通过排队系统的Little公式[20],有

$$E(H)=\lambda E(W) \tag{5}$$

将式(3)、式(4)代入式(5),得到

$$(3L-1)\lambda^2-4L^2\lambda+2L^2=0 \tag{6}$$

因为$\lambda\leq 1$,所以我们可以解得饱和吞吐率为

$$\lambda=\frac{2L^2-\sqrt{4L^2-2L^2(3L-1)}}{3L-1} \tag{7}$$

可以看出,当平均突发长度 L 趋向于 1 时,饱和吞吐率λ也趋向于 1;当 L 趋向于无穷大时,λ趋向于 0.5.

2.2 输入端口具有多输入队列的缓冲交叉开关结构

本节我们将把分析结论推广到每一个输入端口具有多个输入队列的情况.根据第 1 节的讨论可知,一个输出组的输入队列以及相应的交叉点缓冲区等在逻辑上构成了一个相对独立的子交换结构.为了叙述方便,我们在讨论有关输出组 k 的输出、列队列、对应的输入队列以及交叉点缓冲区时,一般将它们称作子交换结构 k 的对应要素.由于各子交换结构具有对称性,同种组成要素具有相同的特性,所以针对一个元素,如一个输出队列、一个列队列的讨论,都可以同样适用于其他同种要素.

设λ_k为子交换结构k的饱和吞吐率,这里,子交换结构k的饱和吞吐率是指其输出端口在输入队列处于饱和状态下的输出端口吞吐率.由于子交换结构k为一个$N\times(N/m)$的交换结构,因此其输入队列在饱和状态下的吞吐率为λ_k/m.

对于子交换结构 0,它的调度不受其他子交换结构的影响.其饱和吞吐率的分析与单数输入队列的情况类似.子交换结构 0 的列队列可以建模为一个带反馈回路的M/D/1 排队系统,当一个分组输出后,将以概率$p=1-1/L$返回排队系统.分组进入M/D/1 系统的速率为突发到达速率λ_0/L,则

$$E(H_0)=\frac{\lambda_0}{1-\lambda_0}-\frac{\lambda_0^2}{2L(1-\lambda_0)} \tag{8}$$

同样,输入队列中一个突发的分组以平均速率λ_0/m进入子交换结构 0 的交叉点缓冲区.通过分析突发进入子交换结构 0 交叉点缓冲区的过程,可以得到非首分组的平均等待时间,亦即分组在列队列中逗留的平均时间,

$$E(W_0)=\frac{L-\lambda_0/m}{(\lambda_0/m)(L-1)}=\frac{mL-\lambda_0}{\lambda_0(L-1)} \tag{9}$$

应用Little公式[20],有

$$E(H_0)=\lambda_0 E(W_0) \tag{10}$$

将式(8)、式(9)带入式(10),可得

$$(3L-1)\lambda_0^2-(2mL^2-2L^2)\lambda_0+2mL^2=0 \tag{11}$$

解得

$$\lambda_0=\frac{(m-1)L^2-\sqrt{(m-1)^2L^4-6mL^3+2mL^2}}{3L-1} \tag{12}$$

然而,其他子交换结构情况有所不同.例如,对于子交换结构 k,由于输入调度的检查过程是顺序的,如果它前面的子交换结构已经确定在一个时隙占用某一输入端口,输入一个分组到交叉点缓冲区,那么,即使交换结构 k 输入队列队首分组的目的交叉点缓冲区在此时隙是空的,子交换结构 k 也不能占用这一输入端口输入分组.因而公式(9)对于除了子交换结构 0 之外的其他子交换结构不再成立.幸而,我们仍然可以通过分析突发进入交叉点缓冲区的过程来估计对应列队列的平均长度.

图 5 是一个长度为 4 个分组的突发进入子交换结构k(k>0)的交叉点缓冲区的过程(k>0,突发长度为 4 个分组,e_i表示第i个分组进入的时刻).对于突发的第 1 个分组,在其到达队首之后,如果排在它前面的子交换结构的队列中的分组被选中进入交叉点缓冲区,则子交换结构k的突发的首分组不得不等待,即使其目的交叉点缓冲区为空.我们把从队头分组的目的交叉点缓冲区已经为空开始到队头分组实际进入交叉点缓冲区之间的时间段称为这个分组进入的竞争时间.竞争时间可以为 0.对于突发中的非首分组,当其前驱进入交叉点缓冲区后,它要等待其前驱从交叉点缓冲区移出后才有可能被调度进入交叉点缓冲区.把从其前驱进入交叉点缓冲区到被移出之间的时间段称为这一分组的等待时间.如果一个分组的前驱在一个时隙内进入并直接在此时隙末被移出交叉点缓冲区,则这一分组等待时间计为一个时隙.因而,每一个非首分组的进入时间由两个阶段组成,一个是等待时间,另一个是竞争时间.对于长度为l个分组的突发,它将经过l段竞争时间、(l–1)段等待时间,加上最后一个分组进入交叉点缓冲区花费的一个时隙后,完成进入交叉点缓冲区的全过程.

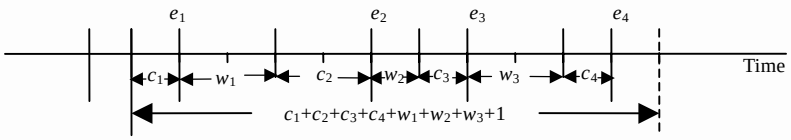


Fig.5 The entering process of a burst to a crosspoint buffer of subswitch k

图 5 一个突发进入子交换结构 k 的交叉点缓冲区的过程

现在考察子交换结构 1.假定在一个输入端口中,属于子交换结构 1 的一个输入队列已经传送了K个突发到某一个交叉点缓冲区.我们知道,饱和状态下,其传输速率为λ_1/m,因而传送这K个突发的总时间为$T^1=KL/(\lambda_1/m)$个时隙.在这些时隙内,子交换结构 0 的输入队列进行输入占用(不再检查并调度子交换结构 1 队列)的时隙所占比例为 $p_h^1=\lambda_0/m$.在剩下的那些时隙中,调度器检查子交换结构 1 队列队首分组的目的交叉点缓冲区是否为空,如果为空,则传送这个分组.因为我们知道,子交换结构 1 输入队列以速度λ_1/m传送分组进入交叉点缓冲区,所以,可以得到调度器检查子交换结构 1 输入队列队首分组的目的交叉点缓冲区并且发现为空的概率

$$p_{ae}^1=(\lambda_1/m)/(1-p_h^1)=(\lambda_1/m)/(1-\lambda_0/m).$$

据此可以推断,在子交换结构 0 队列传输分组而调度器没有检查子交换结构 1 中输入队列队首分组的目的交叉点缓冲区时,交叉点缓冲区为空的概率是 p_{nae}^1,有 $p_{nae}^1=p_{ae}^1$.注意到子交换结构 1 输入队列各个分组的竞争时间是其目的交叉点缓冲区为空,但调度器不检查调度这些分组,而是传输属于子交换结构 0 的分组的时间,因而,这 K 个突发经历的竞争时间的总长度为

$$T_c^1=T^1(\lambda_0/m)\ p_{nae}^1=\lambda_0KL/(m-\lambda_0) \tag{13}$$

则突发的非首分组的平均等待时间为

$$E(W_1)=\frac{T^1-T_c^1-K}{K(L-1)}=\frac{mL-\lambda_1}{\lambda_1(L-1)}-\frac{\lambda_0L}{(L-1)(m-\lambda_0)} \tag{14}$$

尽管对于子交换结构 1,每个时隙参与调度的输入队列的位置可能不同,但其参与调度的输入队列的数量

从统计观点来看是相同的.因而在此情况下,可以近似地采用一个带反馈回路的 M/D/1 排队系统来建模列队列的变化过程,则可以得到如下子交换结构 1 的列队列分组平均数量:

$$E(H_1)=\frac{\lambda_1}{1-\lambda_1}-\frac{\lambda_1^2}{2L(1-\lambda_1)} \tag{15}$$

应用Little公式[20],有

$$E(H_1)=\lambda_1 E(W_1) \tag{16}$$

将式(14)、式(15)代入式(16),可以得到:

$$\frac{\lambda_1}{1-\lambda_1}-\frac{\lambda_1^2}{2L(1-\lambda_1)}=\frac{mL-\lambda_1}{(L-1)}-\frac{\lambda_1\lambda_0 L}{(L-1)(m-\lambda_0)} \tag{17}$$

上式可化为如下关于λ_1的一元二次方程

$$a_1\lambda_1^2-b_1\lambda_1+c_1=0 \tag{18}$$

其中,$a_1=2\lambda_0L^2+3mL-3\lambda_0L-m+\lambda_0$;$b_1=2mL^2+2m^2L^2-2\lambda_0mL^2$;$c_1=2m^2L^2-2\lambda_0mL^2$.

显然,在利用式(12)得到λ_0后,可以根据一元二次方程的求解公式,从这个一元二次方程中解得λ_1的值.

同样地,假定一个输入端口中属于子交换结构k的一个输入队列已经传送K个突发到子交换结构k的一个交叉点缓冲区,其饱和传输速率为λ_k/m,则传送这K个突发的总时间为$T^k=KL/(\lambda_k/m)$时隙.我们知道,一个输入端口将以概率$p_h=\sigma_k/m(\sigma_k=\Sigma_i\lambda_i,0\le i\le(k-1))$优先被前$k$个子交换结构的输入队列占用.在其余的时隙中,子交换结构k的输入队列参加调度,检查队首分组目的交叉点缓冲区是否为空,结果为发现为空的概率是 $p_{ae}^k=(\sigma_k/m)/(1-\sigma_k/m)$.用此概率估计在子交换结构$k$因为前面的子交换结构占用输入端口而被禁止参与调度时,队首分组的目的交叉点缓冲区为空的概率,有 $p_{nae}^k=p_{ae}^k$.进而可以得到这 K 个突发中的分组经历的总竞争时间:

$$T_c^k=T^k(\sigma_k/m)\ p_{nae}^k=\sigma_kKL/(m-\sigma_k) \tag{19}$$

则非首分组的平均等待时间为

$$E(W_k)=\frac{T^k-T_c^k-K}{K(L-1)}=\frac{mL-\lambda_k}{\lambda_k(L-1)}-\frac{\sigma_kL}{(L-1)(m-\sigma_k)} \tag{20}$$

与子交换结构 1 类似,将关于带反馈回路的M/D/1 排队系统的结果及Little公式应用于子交换结构k,得到关于λ_k的一元二次方程:

$$a_k\lambda_k^2-b_k\lambda_k+c_k=0 \tag{21}$$

其中,$a_k=2\sigma_kL^2+3mL-3\sigma_kL-m+\sigma_k$;$b_k=2mL^2+2m^2L^2-2\sigma_kmL^2$;$c_k=2m^2L^2-2\sigma_kmL^2$.

可以从式(21)解出λ_k的值.从而整个交换结构总的饱和吞吐率为

$$\lambda_t=\sum_{k=0}^{m-1}\lambda_k/m \tag{22}$$

3 数值结果:分析与仿真

表 1 比较了分析结果与仿真结果得到的吞吐率的差别.其中,B 为平均突发长度,A 为分析结果,S 为仿真结果,kQ 为每个端口 k 个输入队列,ERR(%)为差异百分比,#为未仿真实验或计算.仿真结果收集自一个对1024×1024缓冲交叉开关交换结构的模拟.相对于仿真结果,当平均突发长度大于1.1个分组时,分析与仿真的差距小于 1.88%.但是,当平均突发长度进一步减小,接近无突发的 1 时,二者的差距快速增大.这是因为,对于一个1024×1024 的交换结构,当流量本身的突发性很低时,由于两个到同一个交叉点缓冲区的分组或突发相遇所导致的突发效应成为影响性能的主要因素,忽略其影响将导致较大的误差.如果交换结构的规模不断增大,如我们前面的讨论,两个到同一个交叉点缓冲区的分组相遇的概率趋向与 0,那么其影响可以忽略,吞吐率将上升.本文的理论分析结果表明,对于每一个输入端口只有一个队列的缓冲交叉开关结构,当突发长度等于 1 个分组,也就是伯努利独立同分布流量时,吞吐率渐近为 1,与文献[17]的推断一致,这也在某种程度上证明了我们的分析的正确性.

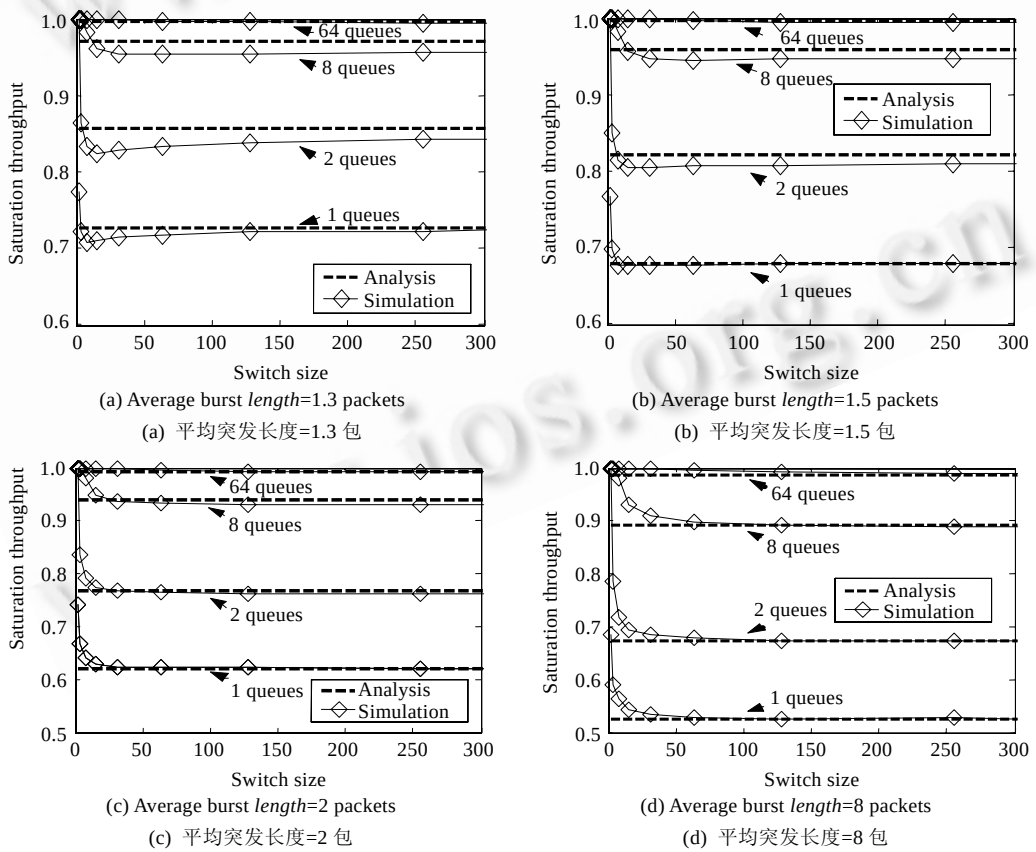
Table 1 Saturation throughputs of analysis and simulation

表 1 饱和吞吐分析与仿真结果

B	1	1.1	1.3	1.5	2	8	16	100	∞ (10 000 000)
A,1*Q*	1.0	0.817 9	0.726 3	0.679 6	0.620 2	0.524 7	0.511 8	0.501 2	0.500 0
S,1*Q*	0.927 7	0.814 6	0.725 6	0.679 4	0.620 3	0.524 5	0.512 5	0.503 3	#
ERR (%)	7.79	0.15	0.03	0.03	0.02	0.04	0.14	0.22	#
A,2*Q*	1.0	0.917 6	0.856 9	0.820 2	0.768 3	0.673 7	0.660 2	0.649 2	0.647 2
S,2*Q*	0.954 5	0.900 7	0.844 8	0.811 2	0.762 0	0.671 0	0.658 1	0.645 4	#
ERR (%)	4.76	1.88	1.49	1.11	0.83	0.40	0.32	0.59	#
A,8*Q*	1.0	0.986 1	0.970 8	0.959 2	0.939 2	0.891 3	0.883 2	0.876 3	0.875 0
S,8*Q*	0.982 4	0.972 5	0.958 2	0.948 0	0.930 2	0.886 0	0.879 7	0.874 5	#
ERR (%)	1.79	1.40	1.31	1.18	0.97	0.60	0.40	0.21	#
A,64*Q*	1.0	0.999 0	0.997 4	0.996 1	0.993 6	0.986 6	0.985 2	0.984 0	0.983 8
S,64*Q*	0.996 5	0.995 8	0.994 3	0.993 6	0.991 8	0.987 0	0.986 3	0.991 2	#
ERR (%)	0.35	0.32	0.30	0.25	0.18	0.16	0.11	0.73	#

另外,本文的分析结果表明,在单输入队列情况下,如果平均突发长度从 1 个分组增长为 2 个分组,吞吐率迅速从 1 降为 0.62.尽管很难为交叉点具有多个分组缓冲的交换结构建立一个分析模型,但是从本文讨论的典型结构的分析结论推断,一个单队列的缓冲交叉开关结构,如果想取得较好的吞吐性能,大于输入流量平均突发长度的交叉点缓冲区则是十分必要的.如果一个交换结构在每个输入端口设置 64 个队列,即使交叉点缓冲只有 1 个分组,突发长度很大,也能达到接近 100%的吞吐率.

图 6 给出了不同输入流量突发长度下的仿真与分析结果随着 N 增长而收敛的情况.平均突发长度分别为 1.3,1.5,2 和 8 个分组.在每一突发长度下,又分为端口队列个数为 1,2,8,64 这 4 种情况.



(a) Average burst *length*=1.3 packets (a) 平均突发长度=1.3 包

(b) Average burst *length*=1.5 packets (b) 平均突发长度=1.5 包

(c) Average burst *length*=2 packets (c) 平均突发长度=2 包

(d) Average burst *length*=8 packets (d) 平均突发长度=8 包

Fig.6 Saturation throughput with switch size under burst traffic

图 6 突发流量下饱和吞吐率随交换结构大小变化

从图 6(a)可以看出,当流量平均突发长度为 1.3 时,N 大于 32,仿真得到的饱和吞吐变得基本恒定;虽然在队列个数为 2 和 8 时有少许误差,但与分析结果还是十分接近的.图 6(d)给出了在输入流量平均突发长度为 8 个分组时,结果随着 N 的增长而收敛的情况.此时,仿真结果得到的饱和吞吐随着 N 的增大收敛于分析结果的速度比平均突发长度为 1.3,1.5 和 2 个分组时要慢一些.当 N 大于 64 时,仿真得到的饱和吞吐变得基本恒定,并且十分准确地收敛于分析结果.从这 4 个不同平均突发长度流量下的评估结果总体来看,突发长度较大时,收敛速度随着流量平均突发的增长而有所降低,但仿真结果的收敛点与分析结果的差距将减小.

4 结 论

本文给出了一个典型的缓冲交叉开关结构在突发流量下的渐进饱和吞吐率的分析模型.分析结果显示,在单输入队列情况下,饱和吞吐率随着平均突发长度的增加而迅速减小,从 1 趋近于 0.5.当每个输入端口输入队列数量增加时,系统的饱和吞吐率也随之增加;如果每一个输入端口设置 64 个输入队列,就能取得接近于 1 的吞吐率.分析结论表明,对于缓冲交叉开关结构,如果要得到满意的吞吐性能,一定量的交叉点缓冲区或输入队列是必要的.本文的结果可以使我们更好地了解缓冲交叉开关交换结构的性能特性,为基于缓冲交叉开关的路由交换设备的优化设计提供理论支持.

References:

[1] Karol MJ, Hluchyj MG, Morgan SP. Input versus output queuing on a space-division packet switch. IEEE Trans. on Communications, 1987,35(12):1347−1356.

[2] Jacob L, Kumar A. Saturation throughput analysis of an input queuing ATM switch with multiclass bursty traffic. IEEE Trans. on Communication, 1995,43(2-4):757−761.

[3] Li SQ. Performance of a non-blocking space-division packet switch with correlated input traffic. In: Proc. of the 1989 GLOBECOM Conf. New York: IEEE Communication Society Press, 1989. 1754−1763.

[4] Manjunath D, Sikdar D. Variable length packet switches: Delay analysis of crossbar switches under Poisson and self similar traffic. In: Proc. of the IEEE INFOCOM 2000. New York: IEEE Communication Society Press, 2000. 1055−1064.

[5] Anderson T, Owicki S, Saxe J, Thacker C. High speed switch scheduling for local area networks. ACM Trans. on Computer Systems, 1993,11(4):319−352.

[6] McKeown N. Scheduling algorithms for input-queued switches [Ph.D. Thesis]. Berkeley: University of California at Berkeley, 1995.

[7] Nong G, Hamdi M, Muppala JK. Analytical analysis of ATM switches with multiple input queues with bursty traffic. In: Proc. of the 1999 GLOBECOM Conf. New York: IEEE Communication Society Press, 1999. 1222−1226.

[8] Nong G, Muppala JK, Hamdi M. Analytical analysis of nonblocking ATM switches with multiple input queues. IEEE/ACM Trans. on Networking, 1999,7(1):60−74.

[9] Pang B, Gao W, He SM. A survey on input-queued scheduling algorithms in high-speed IP routers. Journal of Software, 2003, 14(5):1011−1022 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1011.htm

[10] Yeung KL, Hai S. Throughput analysis for input-buffered ATM switches with multiple FIFO queues per input port. Electronics Letters, 1997,33(19):1604−1606.

[11] Kim H, Kim K. Performance analysis of the multiple input-queued packet switch with the restricted rule. IEEE/ACM Trans. on Networking, 2003,11(3):478−487.

[12] Rojas-Cessa R, Oki E, Chao HJ. CIXOB-*k*: Combined input crosspoint-output buffered packet switch. In: Proc. of the IEEE 2001 GLOBECOM. New York: IEEE Communication Society Press, 2001. 2654−2660.

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

[14] He SM, Sun ST, Zhao W, Zheng YF, Gao W. Smooth switching problem in buffered crossbar switches. Special Issue of Performance Evaluation Review, 2005,33(1):386−387.

[15] Javidi T, Magill R, Hrabik T. A high-throughput scheduling algorithm for a buffered crossbar switch fabric. In: Neuvo Y, ed. Proc. of the Int'l Conf. Communications. Helsinki: IEEE Communication Society Press, 2001. 1586−1591.

[16] Chrysos N, Katevenis M. Weighted fairness in buffered crossbar scheduling. In: Proc. of the IEEE HPSR 2003. New York: IEEE Communication Society Press, 2003. 17−22.

[17] Lin MJ, Mckeown N. The throughput of a buffered crossbar switch. IEEE Communications Letters, 2005,9(5):465−467.

[18] Adas A. Traffic models in broadband networks. IEEE Communication Magazine, 1997,35(7):82−89.

[19] Kleinrock L. Queuing Systems, Vol.1. New York: John Wiley & Sons, 1975.

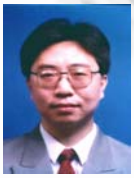
[20] Little JD. A proof of the queuing formula $L=\lambda W$. Operations Research, 1961,9(3):383−387.

附中文参考文献:

[9] 庞斌,高文,贺思敏.高速 IP 路由器中输入排队调度算法综述.软件学报,2003,14(5):1011−1022. http://www.jos.org.cn/1000-9825/14/1011.htm

孙书韬(1967－),男,辽宁朝阳人,博士,工程师,主要研究领域为计算机网络,多媒体通信,信息系统.

郑燕峰(1975－),男,博士,工程师,主要研究领域为计算机网络,多媒体通信.



贺思敏(1968－),男,博士,副研究员,CCF 高级会员,主要研究领域为组合优化,实时调度,计算机通信,生物信息学.

高文(1956－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为多媒体数据压缩,图像处理,计算机视觉,多模式接口,人工智能.