计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
43(7):1225~1232, 2006

基于双轮转指针的输入与交叉点联合排队型(CICQ)交换结构调度算法

郑燕峰[1,2] 孙书韬[2] 贺思敏[1] 高 文[1,3]
[1](中国科学院计算技术研究所 北京 100080)
[2](中国科学院研究生院 北京 100049)
[3](北京大学信息科学技术学院 北京 100871)
(yfzheng@jdl.ac.cn)

A Dual Round-Robin Algorithm for Combined Input-Crosspoint-Queued Switches

Zheng Yanfeng[1,2], Sun Shutao[2], He Simin[1], and Gao Wen[1,3]
[1](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)
[2](*Graduate University of Chinese Academy of Sciences, Beijing* 100049)
[3](*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)

Abstract The CICQ switch fabric is an ideal solution to multi-terabit switch implementation owing to its nice distributed scheduling property. Round-robin algorithms have been extensively studied because of their simplicity for hardware implementation. It is known that round-robin algorithms provide high throughput under uniform traffic; however, the performance is degraded under nonuniform traffic. In this paper, the reason for the performance degradation of the existing round-robin algorithms is pointed out and then a class of dual round-robin algorithms is proposed. For the proposed algorithms, each input arbiter is associated with dual round-robin pointers named the primary pointer and the secondary pointer respectively. The input queue corresponding to the primary pointer has the highest priority being scheduled, and the decision for updating the primary pointer can be dynamically made relying on the input queue status. When the input queue corresponding to the primary pointer is blocked, other input queues can be uniformly scheduled according to the secondary pointer position. Simulations show that the dual round-robin algorithms can significantly improve the performance of the CICQ switch under nonuniform traffic.

Key words combined input-crosspoint-queued switch; virtual output queue; scheduling; throughput

摘 要 CICQ交换结构因具有良好的分布式调度特性而成为构建太比特(Tb/s)级以上交换机的一种理想选择.轮转型调度算法因硬件实现的简单性而得到广泛的研究,尽管此类型的调度算法在均匀流量下具有较高的吞吐率,然而在非均匀的流量下其性能则明显下降.指出了已有轮转型算法在非均匀流量下性能下降的原因,提出了一类基于双指针的轮转型调度算法,即每个输入调度器均有两个轮转指针(主指针和辅助指针).主指针对应的队列具有最高的调度优先级,算法可以根据各个队列的状态动态决定何时更新主指针,当主指针对应的队列被流控机制阻塞时,将根据辅助指针依次公平服务其他队列.实验结果表明,基于双指针的调度算法可以显著提高CICQ交换机在非均匀流量下的性能.

关键词 输入与交叉点联合排队型交换结构;虚拟输出队列;调度;吞吐率

中图法分类号 TP393.05

收稿日期:2005-06-02;**修回日期**:2005-11-09

基金项目:国家自然科学基金项目(69983008);国家"八六三"高技术研究发展计划基金项目(2001AA112100);中国科学院计算技术研究所基础研究基金项目(20056090)

1 引 言

交换结构作为路由器的核心组成部分,对提高路由器的交换容量起着关键性作用. 传统的输出排队型交换结构,由于需要较高的加速比,其可扩展性较差. 而输入排队型交换结构,即缓冲区设置在输入端口,由于仅要求存储器的带宽为线速即可,所以该交换结构具有较好的可扩展性.

对于输入排队型交叉开关内部无缓冲区(input-queued unbuffered crossbar)的交换结构,其调度方法一般归结为一种二分图的匹配问题,在每个调度时间片,输入端口与输出端口的匹配一般由一个集中式的调度方法生成[1,2],与输出排队型调度算法相比,集中式的调度方法具有较高的通信复杂度和计算复杂度. 为降低交换结构调度的复杂性,CICQ(combined input-crosspoint-queued)型交换结构得以提出,即不仅在每个输入端口设置缓冲区,而且还在每个交叉点(crosspoint)均设置了缓冲区,CICQ 交换结构由于采用分布式调度,从而具有更好的可扩展性,它也是当前交换结构研究领域的热点问题. 对于 CICQ 交换结构,现有算法大致可分为 3 类:

(1) 轮转型调度算法,即在每个输入与输出端口均采用轮转的方法进行调度. CIXB-1[3] 是采用该类算法的一种典型交换结构,对于均匀到达的通信量,CIXB-1 能够达到渐进 100% 的吞吐率,然而在非均匀的通信量下,CIXB-1 的吞吐率则明显下降. 为了提高轮转型算法在非均匀流量下的性能,文献[4]提出了一种基于帧的调度算法 RR-AF,它可以根据各个输入队列负载的大小来动态改变帧长,具有一定的自适应性. 但是 RR-AF 需要在每个输入队列及交叉点队列各设置两个计数器,由于统计帧长的计数器的值可以连续累加,而文献[4]并没有给出该计数器的上界.

(2) 基于权重的调度算法,通常采用输入队列的队长或者队首信元的排队延迟作为权重,在每个时隙,要进行多次比较,以找出具有最大权的队列进行调度,例如 LQF-RR[5] 和 OCF-OCF[6]. 由于需要在每个时隙内比较最大或最小值,随着线速的提高或者端口数的增加,此类算法的可扩展性会有较大的限制.

(3) 基于交叉点缓冲区状态的调度算法,例如 MCBF[7],对于每个输入调度器,它优先选择对应列交叉点缓冲区拥塞程度较小的输入队列;而每个输出调度器优先选择对应行交叉点缓冲区拥塞程度较大的交叉点队列. 这类算法的缺点仍然是需要多次比较,以寻找最大或最小值.

由于轮转型算法只需执行简单的轮转优先操作,具有硬件实现上的优势,尽管现有算法在均匀的流量下能够达到渐进 100% 的吞吐率,然而在非均匀的流量模式下,其性能明显下降. 对于非均匀的流量,由于负载多集中在少数几个队列上,如何提高这些队列的服务速率是提高 CICQ 交换结构性能的关键. 为此,本文提出了一类基于双指针的轮转型调度算法,其共同特征在于:每个输入调度器都分别有两个指针,其中主指针对应的队列具有最高的调度优先级,算法可以根据各个队列的状态来动态决定何时更新主指针,当主指针对应的队列被 CICQ 流控机制阻塞时,将根据辅助指针依次公平服务其他队列. 仿真实验表明,基于双指针的调度算法可以显著提高 CICQ 交换结构在非均匀流量下的性能.

2 CICQ 交换结构模型及工作原理

首先,做以下约定:①各个输入端口支持相同的线速率,以线速率传输一个信元所需的时间称为一个时隙;②加速比(speedup)定义为交换结构内部带宽与线速率的比值;③交换结构只处理定长分组,称为信元;④交换结构的规模均为 $N \times N$,即包括 N 个输入端口以及 N 个输出端口.

如图 1 所示,CICQ 交换结构具有以下特点:

(1) 分布式缓冲

在每个输入端口均采用 VOQ(virtual output queue)[8] 的队列组织方式,以避免队头阻塞,并且在每个交叉点设置了缓冲区 XPB. 在一个时隙内,每个交叉点缓冲区允许一次读操作和一次写操作,不需要内部加速比,所以具有良好的可扩展性.

(2) 分布式调度

为解决输入队列的竞争,每一个输入端口都设置一个调度器,以决定某一 VOQ 发送队首信元至相应交叉点缓冲区. 同样,为解决输出端口的竞争问题,每个输出端口也都设有一个调度器,以决定将某个交叉点缓冲区的信元发送至输出端口. 所以,CICQ 交换机共需要($2N$)个调度器,由于每个调度器都是异步独立运行的,所以具有较小的通信开销. 鉴于交叉点的队列容量较小,现有文献一般采用基于份额(credit)的流控机制[9],以避免该缓冲区溢出. 在一个时隙内,只有输入队列非空并且份额不为零

的 VOQ 才允许被调度到对应的交叉点缓冲区,记符合这种条件的输入队列为 EVOQ(eligible VOQ).

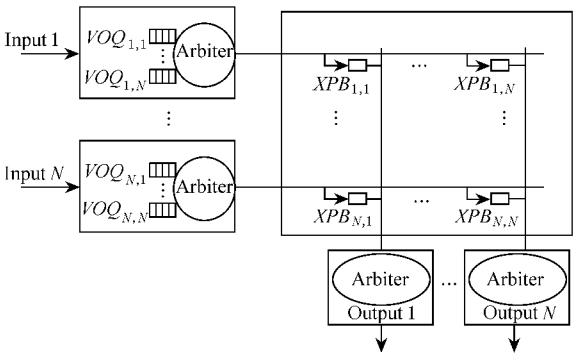

Fig. 1 An $N\times N$ CICQ switch architecture.

图 1 一个 $N\times N$ CICQ 交换结构

3 FD-RR 算法(full draining round-robin algorithm)

文献[10]指出:为了保持一个排队系统的稳定性,即队列及平均延迟均有界,则分组的到达速率应该小于其离开速率.

对于一个 $N\times N$ 的 CICQ 交换结构,考虑下面的一种非均匀流量模型,定义 H_VOQ 为具有最大负载的队列. 给定一个时隙 t,假设轮转指针指向 H_VOQ,并且其对应的交叉点队列已满,则 H_VOQ 被阻塞,现有轮转型算法[3,4]则会将指针移动到下一个非空并且不被流控机制阻塞的队列,则 H_VOQ 在随后的若干个时隙内具有较低的调度优先级,从而其离开速率下降;如果 H_VOQ 的平均到达速率大于其平均离开速率,则 H_VOQ 将会变得不稳定,从而导致吞吐率下降.

为了提高轮转型调度算法在非均匀流量模式下的性能,本文给出一种基于双指针的调度算法 FD-RR,与现有轮转型调度算法不同,每个输入端口调度器都有两个轮转指针(主指针和辅助指针);而每个输出端口调度器只需一个轮转指针.

现以上面的非均匀流量模型为例,介绍 FD-RR 的工作原理. 如果在第 t 个时隙,主指针指向 H_VOQ,而此时 H_VOQ 被流控机制阻塞,与现有同类算法[3,4]不同的是,FD-RR 的主指针仍将停留在 H_VOQ 位置直到该队列为空,其目的是使 H_VOQ 在接下来的若干个时隙($t+1,t+2,\cdots,t+k$)具有最高的调度优先级,从而提高信元的离开速率. 而当主指针指向的队列由于被阻塞而不能被调度时,为了公平服务其他队列,FD-RR 则根据辅助指针的位置按照轮转的方式依次服务其他队列.

3.1 FD-RR 算法

以输入端口 i($i=1,2,\cdots,N$)为例,描述输入调度器的工作过程:在任意一个时隙 t,不妨假设 P_VOQ 对应于主指针所指向的队列,首先检查 P_VOQ 队列是否为空:

(1) 队列为空

按照轮转的方式从 P_VOQ 的下一个位置寻找 EVOQ,如果找到,则把该 EVOQ 的队首信元发送至交叉点缓冲区,将主指针指向此队列,如果没找到 EVOQ,则主指针保持不变.

(2) 队列不为空

① 如果 P_VOQ 没有被流控阻塞,则其队首信元将会被发送到对应交叉点缓冲区,主指针保持不变;

② 如果 P_VOQ 被流控阻塞,则从辅助指针的位置按照轮转的方式寻找 EVOQ,如果找到,则该 EVOQ 的队首信元被发送至相应的交叉点缓冲区,并将辅助指针移向下一个位置;如果没有找到 EVOQ,则辅助指针保持不变.

根据主指针的更新规则可以看出,这是一种竭力服务的方式,即只要主指针指向的输入队列非空,主指针就保持不变,显然负载较大的队列获得服务的机会要多于负载较小的队列. 实验表明,这十分有助于提高 CICQ 交换结构在非均匀流量模型下的吞吐率. 由于流控的原因,主指针指向的队列可能被多次阻塞. 在这种情况下,FD-RR 将会根据辅助指针来服务其他队列,与主指针的更新方式不同,FD-RR 算法每根据辅助指针完成一次调度,就会将辅助指针从当前位置按照固定的轮转次序移动到下一个位置以服务其他队列,从而改善了算法 FD-RR 的公平性.

下面描述 FD-RR 输出调度器的工作过程:从轮转指针的当前位置开始,按照固定的轮转方式,寻找第一个非空的交叉点缓冲区,如果找到,则该队列的队首信元被发送至输出端口,轮转指针指向该队列的下一个轮转位置;如果没有找到非空的交叉点缓冲区,则轮转指针的位置保持不变.

3.2 FD-RR 算法仿真结果

(1) 交换结构模型

本文均采用 32×32 的 CICQ 交换结构①,并且

① 本文还进行了 4×4,8×8,16×16 下的仿真实验,由于所得结论与 32×32 相同,限于篇幅,仅给出 32×32 下的仿真结果.

设置 VOQ 的容量足够大,以保证没有信元丢失;每个交叉点 XPB 仅设置一个信元大小的缓冲区. 此外,假设各个输入端口具有相同的负载,用符号 $\rho(0<\rho<1)$ 来表示,符号 $\lambda_{i,j}$ 表示队列 $VOQ_{i,j}$ 的负载.

(2) 流量模型

均采用容许(admissible)的流量模型,即在任意时刻,输入端口与输出端口均不过载(overload),有 $\sum_i \lambda_{i,j} < 1, \sum_j \lambda_{i,j} < 1$.

均匀流量模型:

① Uniform Bernoulli i.i.d. traffic

信元的到达过程是 Bernoulli i.i.d. 的,并且每个 VOQ 具有相同的负载,即 $\lambda_{i,j} = \rho/N, \forall i, \forall j$.

② Uniform bursty traffic

该过程采用两态(on-off)马尔可夫模型来表示,当输入端口有突发流量到达时,处于 on 状态,其长度服从均值为 $l(l>1)$ 的几何分布;当输入端口空闲时,处于 off 状态,其长度服从均值为 $l(1-\rho)/\rho$ 的几何分布. 如果用随机变量 v 表示一个突发(burst)的目的输出端口,则其取值为输出端口 i 的概率为

$$P(v = i) = 1/N, i = 1,2,\cdots,N.$$

非均匀流量模型:

① Hot-spot traffic

信元的到达过程是 Bernoulli i.i.d. 的,该流量模型引入了一个参数 $w \in [0,1]$,每个 VOQ 的负载定义为:当 $i = j$ 时,$\lambda_{i,j} = \rho[w + (1-w)/N]$;当 $i \neq j$ 时,$\lambda_{i,j} = \rho(1-w)/N$.

② Log-diagonal traffic

信元的到达过程是 Bernoulli i.i.d. 的,并且满足 $\lambda_{i,j} = 2\lambda_{i,|j+1|}$ 以及 $\sum_i \lambda_{i,j} = \rho$,其中 $|j+1| = (j+1) \bmod N$.

③ Diagonal traffic

信元的到达过程是 Bernoulli i.i.d. 的,并且满足:当 $j = i$ 时,$\lambda_{i,j} = 2\rho/3$;当 $j = (i+1) \bmod N$,$\lambda_{i,j} = \rho/3$;在其他情况下,$\lambda_{i,j} = 0$. 根据上述定义,每个输入端口的负载仅分布在两个 VOQ 上,其余 VOQ 空闲.

(3) FD-RR 算法的吞吐率

吞吐率定义为能保证交换机稳定的最大输入负载,并采用以下办法来判断是否稳定:对于一个输入负载,分别测试以下 5 个时间区间:[0,200000],[0,400000],[0,600000],[0,800000],[0,1000000](每个时间单位为一个时隙),如果在某个输入负载下,VOQ 的平均队长在各个时间区间均增长,则该负载已经超过了交换机的吞吐率,否则交换机在该负载下是稳定的.

下面给出 FD-RR 算法在不同流量模型下的吞吐率,并与 CIXB-1[3],MCBF[7] 和 RR-AF[4] 算法做一比较.

① 均匀 Bernoulli i.i.d. 流量模型

FD-RR 算法的吞吐率为 99.5%,由文献[3,4,7]可知,CIXB-1,MCBF 和 RR-AF 同样具有接近 100% 的吞吐率,此处略去该模型下的比较.

② 非均匀容许流量模型

如图 2 所示,在 Hot-spot 流量模型下,对于参数 w 的不同取值,FD-RR 与 RR-AF 的吞吐率均高于 99%,而 CIXB-1 和 MCBF 的吞吐率则明显低于 FD-RR 与 RR-AF 的吞吐率. 对于 Log-diagonal 及 Diagonal 流量模型,如表 1 所示,FD-RR 的吞吐率均逼近 100%,明显高于其他几种算法.

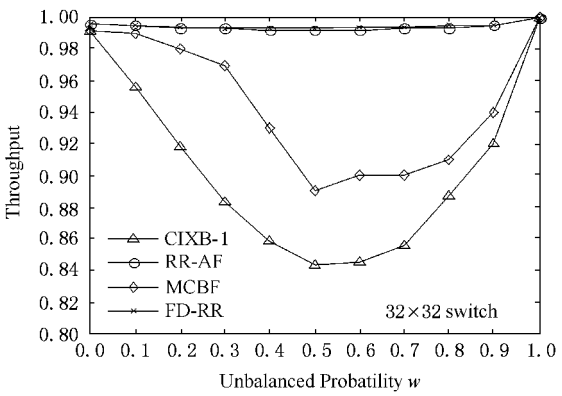

Fig. 2 The throughput under hot-spot traffic.

图 2 各算法在 Hot-spot 流量下的吞吐率

Table 1 The Throughput for Logdiagonal and Diagonal Traffic

表 1 各算法在 Log-diagonal 及 Diagonal 流量下的吞吐率

Traffic	CIXB-1	MCBF	RR-AF	FD-RR
Log-diagonal	0.863	0.932	0.941	0.994
Diagonal	0.87	0.872	0.912	0.993

4 基于份额的双指针轮转算法

尽管 FD-RR 在多种类型流量下具有逼近 100% 的吞吐率,但是该算法却存在"饥饿"问题. 其原因在于:主指针指向的队列有可能被无限制地服务. 为了避免该问题的发生,需要限制队列的最大服务次数. 为此,本文提出一种基于份额的双指针轮转算法 QD-RR(quantum-based round-robin algorithm).

易知,如果一个队列的平均服务速率小于其平均到达速率,则该队列的长度将会不断增加,所以队

长可以反映一个队列的拥塞程度. QD-RR 为每个 VOQ 都分配一个份额($Quantum$),该份额的初始值为零,当主指针从其他位置移向一个 EVOQ 时,为了避免主指针对于一个队列的服务过于"贪心",QD-RR 取该 EVOQ 队长的一半作为该队列的份额,并且主指针指向的队列每被调度一次,该队列的份额就减少一个单位,当份额为零时,则主指针就会被更新至其他位置以服务其他队列.

4.1 QD-RR 算法

以输入端口 i($i=1,2,\cdots,N$)为例,描述输入调度器的工作过程:在任意一个时隙 t,不妨假设 P_VOQ 对应于主指针所指向的队列,首先检查 P_VOQ 的份额是否为零:

(1) 份额为零

按照轮转的方式从 P_VOQ 的下一个位置寻找 EVOQ,如果找到,则该 EVOQ 的队首信元被发送至交叉点缓冲区,将主指针指向此队列,并且 EVOQ 的份额取值为 $\lfloor occupancy(\text{EVOQ},t)/2 \rfloor$,函数 $occupancy(\text{EVOQ},t)$ 返回该 EVOQ 在时隙 t 时的队长,如果没找到 EVOQ,则主指针保持不变.

2) 份额不为零

① 如果 P_VOQ 没有被流控阻塞,则其队首信元将会被发送到对应交叉点缓冲区,并且 P_VOQ 的份额减 1,主指针保持不变;

② 如果 P_VOQ 被流控阻塞,则从辅助指针的位置开始按照轮转的方式寻找 EVOQ,如果找到,则该 EVOQ 的队首信元被发送至相应的交叉点缓冲区,并将辅助指针移向下一个位置.如果没有找到 EVOQ,则辅助指针保持不变.

此外,QD-RR 输出端口调度器与 FD-RR 输出端口调度器的工作过程相同.

4.2 QD-RR 的一个实例

一个 6×6 的 CICQ 交换结构,如图 3(a)所示,在时隙 t 开始时,输入调度器 i 的主指针指向 $VOQ_{i,0}$,辅助指针指向 $VOQ_{i,3}$. 输入调度器 i 首先检查 $VOQ_{i,0}$的份额是否为 0.

(1) $Quantum=0$:从当前主指针的下一个位置($VOQ_{i,1}$)开始,寻找第 1 个 EVOQ,如果 $VOQ_{i,2}$满足调度条件,则 $VOQ_{i,2}$的队首信元被调度至 $XPB_{i,2}$队列,$VOQ_{i,2}$的份额取值为 $\lfloor occupancy(VOQ_{i,2},t)/2 \rfloor$,并且主指针指向 $VOQ_{i,2}$,如图3(b)所示.

(2) $Quantum>0$:如果 $VOQ_{i,0}$没有被流控阻塞,则 $VOQ_{i,0}$的队首信元被发送至交叉点队列 $XPB_{i,0}$,并且 $VOQ_{i,0}$的份额减少一个信元单位,主指针保持不变.如果 $VOQ_{i,0}$被流控阻塞,则从辅助指针的位置 $VOQ_{i,3}$开始寻找第 1 个 EVOQ,如果找到,则将该 EVOQ 的队首信元发送至对应的交叉点缓冲区,并更新辅助指针,如图 3(c)所示.

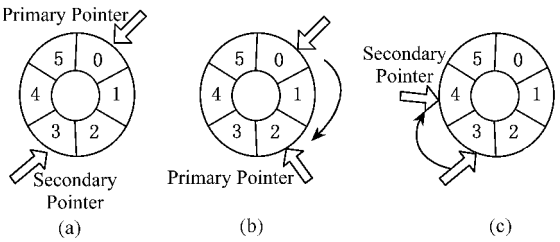

Fig. 3 An example of uptating the primary pointer and the secondary pointer. (a) The original position of primary and secondary pointer at time slot t; (b) Updating the primary pointer; and (c) Updating the secondary pointer.

图 3 一个主指针与辅助指针的更新示例.(a) 时隙 t 时主指针与辅助指针的初始位置;(b) 更新主指针;(c) 更新辅助指针

最后,总结一下 QD-RR 算法的特点:

(1) 根据主指针的更新规则,易知 VOQ 队列接受服务的次数与其获得的份额成正比,所以 QD-RR 是一种流量自适应的调度算法.

(2) 对于算法 QD-RR 的每个输入调度器,在每个时隙,它只需根据轮转指针(主指针或者辅助指针)的值进行简单的轮转优先操作,无需比较最大或最小值,所以它具有与 CIXB-1[3] 相同的时间复杂度.

(3) QD-RR 具有较低的通信开销,这是因为输入调度器设置在每个线卡内,可以直接获取本线卡内部队列的长度信息.

(4) 对于任意一个时隙,由于 VOQ 队长是有界的,所以主指针对应的 VOQ 所获得的份额也是有界的,当份额为零时,主指针就会被移动到其他位置,从而可以避免"饿死"问题的发生.

4.3 QD-RR 算法仿真结果

下面考察 QD-RR 算法的主要性能指标:吞吐率、平均延迟以及输出端口信元的平均突发长度.

(1) 吞吐率

① 均匀 Bernoulli i.i.d. 流量模型

QD-RR 算法的吞吐率为 99.6%,同样逼近 100%,故省略与相关算法的比较.

② 非均匀容许流量模型

先考察 Hot-spot 流量模型,如图 4 所示,对于 w 的不同取值,QD-RR 的吞吐率均大于 98%,明显高于 CIXB-1 以及 MCBF,而略低于 RR-AF 的吞吐率.对于 Log-diagonal 及 Diagonal 流量模型,如表 2 所示,QD-RR 算法的吞吐率均在 98% 以上,明显优于其他几种算法.

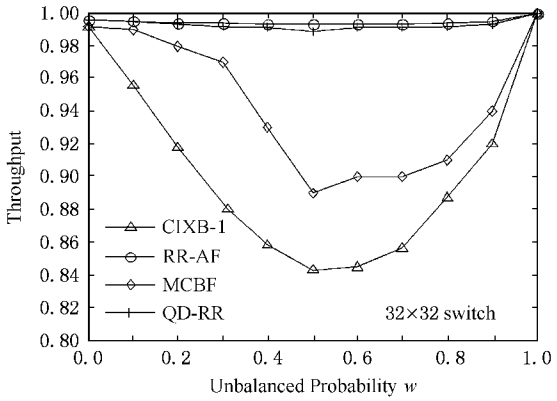

Fig. 4 The throughput under hot-spot traffic.

图 4 各算法在 Hot-spot 流量下的吞吐率

Table 2 The Throughput Under Log-diagonal and Diagonal Traffic

表 2 各算法在 Log-diagonal 及 Diagonal 流量模型下的吞吐率

Traffic	CIXB-1	MCBF	RR-AF	QD-RR
Logdiagonal	0.863	0.932	0.941	0.987
Diagonal	0.87	0.872	0.912	0.986

(2) 平均延迟

① 均匀流量模型

各算法无论是在均匀 Bernoulli i.i.d. 流量还是在均匀突发流量($l=16,32$)下,均具有十分相似的平均延迟,此处略去比较结果.

② 非均匀流量模型

各算法在 Hot-spot($w=0.5$)流量模型下的平均延迟性能如图 5 所示. 在相同的负载下,RR-AF 以及 QD-RR 算法的平均延迟明显低于 CIXB-1 和 MCBF 的平均延迟,而 RR-AF 和 QD-RR 的平均延迟非常接近. 对于 Log-diagonal 及 Diagonal 流量模型,根据图 6 与图 7 易知,QD-RR 的平均延迟显著低于其他几种算法. 此外,本文还比较了各算法在不同流量下的抖动情况,所得结论与平均延迟的结论相同,此处略去仿真结果.

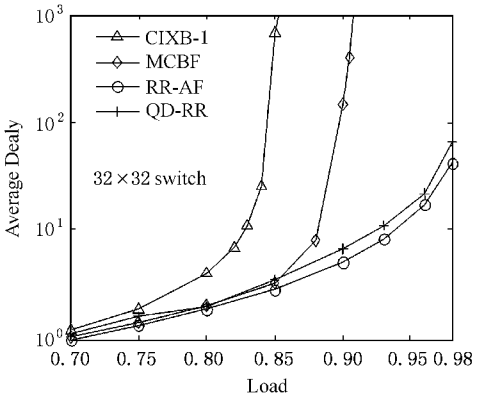

Fig. 5 The average delay under hot-spot($w=0.5$) traffic.

图 5 Hot-spot 流量下的平均延迟

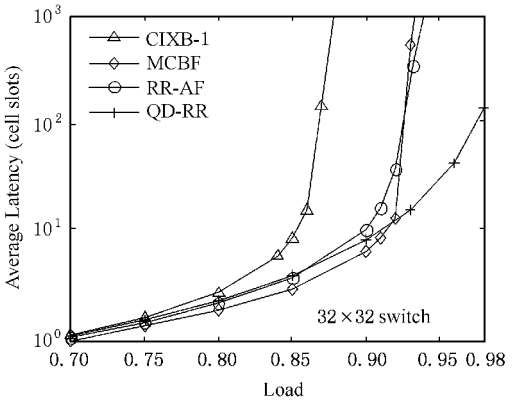

Fig. 6 The average delay under log-diaongal traffic.

图 6 Log-diaongal 流量下的平均延迟

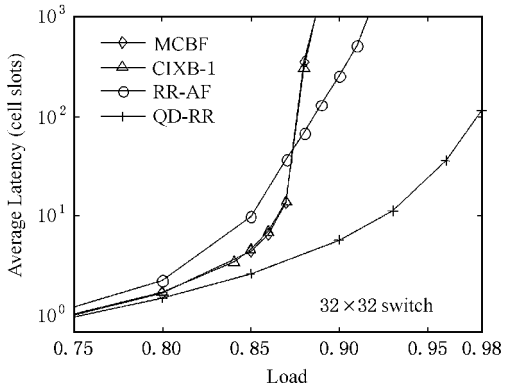

Fig. 7 The average delay under diagonal traffic.

图 7 Diagonal 流量下的平均延迟

(3) 输出端口信元的平均突发长度

由于突发的流量会对相邻路由器的性能产生不利影响[2],所以一个交换机应尽量减少其输出端口生成流量的突发性. 下面,以输出端口信元的平均突发长度为突发性的度量,来考察各调度算法在该项指标下的性能. 从图 8,9,10 可以看出,无论是均匀突发流量还是非均匀的流量模型,RR-AF 在输出端口信元突发的平均长度明显高于其他算法,这是因为 RR-AF 是一种基于帧的调度算法,其输出端口轮转指针需要等待整个帧发送完毕后,才更新其指针. 而对于 QD-RR,如果其输出端口的轮转指针所对应的交叉点缓冲区已经发送一个信元至输出端口,则该指针就会被移动到下一个位置,通过这种方式,来自不同输入端口的信元就可以交错分布在输出端口的信元序列中,从而有效减少了输出端口流量的突发长度.

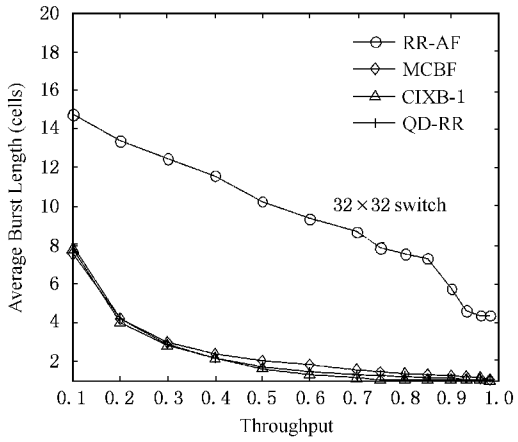

Fig. 8 The average burst length under uniform bursty ($l=16$) traffic.

图 8 均匀突发流量(均值为 16)到达下的输出端口的平均突发长度

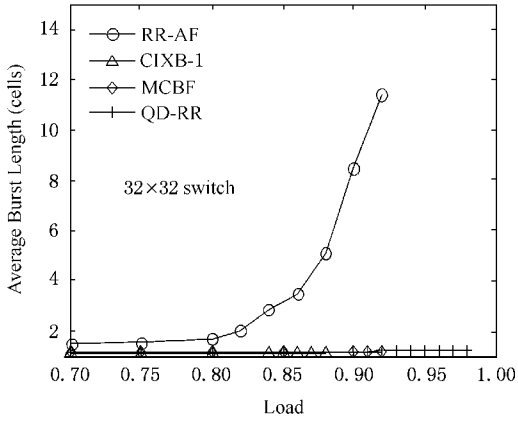

Fig. 9 The average burst length at under diagonal traffic.

图 9 log-diagonal 流量到达下的输出端口的平均突发长度

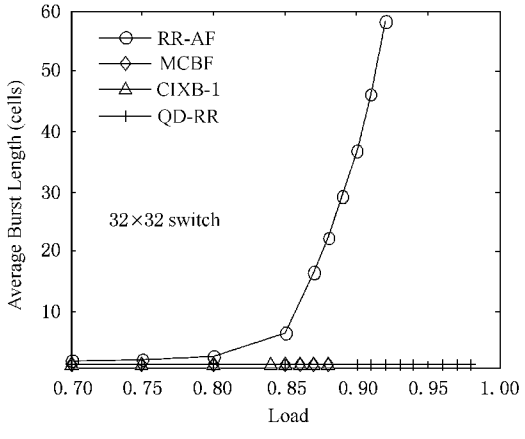

Fig. 10 The average burst length at under diagonal traffic.

图 10 Diagonal 流量到达下的输出端口的平均突发长度

最后,探讨 QD-RR 算法在构建太比特级交换机的可行性. 第 1,对于一个 $N\times N$ 的 CICQ 交换机,由于共有 N^2 个交叉点缓冲区,所以该缓冲区的大小是限制该结构可扩展性的一个瓶颈. QD-RR 仅要求每个交叉点缓冲区的大小为一个信元长度,从而在该缓冲区的需求上面达到了最小. 第 2,QD-RR 是一种分布式的轮转型调度算法,无需比较最大或最小值,无需内部加速比,易于硬件实现[2]. 第 3,在芯片的尺寸、I/O 针数以及功耗上面,现有的工艺水平对于构建太比特级的交换机是可行的[11].

5 结 论

为了提高轮转型调度算法在非均匀流量下的性能,本文提出了两种基于双指针的调度算法 FD-RR 和 QD-RR,尽管 FD-RR 在多种流量下具有逼近 100% 的吞吐率,然而却存在"饿死"问题. 为了进一步改进 FD-RR 算法的公平性和消除"饿死"问题,本文提出了一种基于份额的调度算法 QD-RR 算法,实验结果表明,QD-RR 在多种流量下均具有较好的性能.

参 考 文 献

1 T. E. Anderson, S. S. Owicki, J. B. Saxe, *et al*. High-speed switch scheduling for local area networks [J]. ACM Trans. Computer Systems, 1993, 11(4): 319~352

2 N. McKeown. The iSlip scheduling algorithm for input-queued switches [J]. IEEE/ACM Trans. Networking, 1999, 7(2): 188~201

3 R. Rojas-Cessa, E. Oki, Z. Jing, *et al*. CIXB-1: Combined input-one-cell-crosspoint buffered switch [C]. The Workshop on High Performance Switching and Routing, Dallas, USA, 2001

4 R. Rojas-Cessa. Round-robin selection with adaptable-size frame in a combined input-crosspoint buffered switch [J]. IEEE Communications Letters, 2003, 7(11): 555~557

5 T. Javidi, R. B. Magil, T. Hrabik. A high-throughput scheduling algorithm for a buffered crossbar switch fabric [C]. The 2001 IEEE ICC, Dallas, USA, 2001

6 M. Nabeshima. Performance evaluation of a combined input-and crosspoint-queued switch [J]. IEICE Trans. Communications, 2000, E83-B(3): 737~741

7 L. Mhamdi, M. Hamdi. MCBF: A high-performance scheduling algorithm for buffered crossbar switches [J]. IEEE Communications Letters, 2003, 7(9): 431~433

8 Y. Tamir, G. L. Frazier. Dynamically-allocated multi-queue buffer for VLSI communication switches [J]. IEEE Trans. Computers, 1992, 41(6): 725~737

9 H. T. Kung, R. Morris. Credit-based flow control for ATM networks [J]. IEEE Network, 1995, 9(2): 40~48

10 L. Kleinrock. Queueing Systems, Volume 2: Computer Applications [M]. New York: Wiley, 1976

11 F. Abel, C. Minkenberg, R. Luijten, *et al*. A four-terabit packet switch supporting long round-trip times [J]. IEEE Micro, 2003, 23(1): 10~24

Zheng Yanfeng, born in 1975. Ph. D. candidate. His main research interests include computer networks and multimedia communications.

郑燕峰,1975 年生,博士研究生,主要研究方向为计算机网络和多媒体通信.

Sun Shutao, born in 1967. Ph. D. candidate. His main research interests include computer networks and multimedia communications.

孙书韬,1967 年生,博士研究生,主要研究方向为计算机网络和多媒体通信.

He Simin, born in 1968. Ph. D. and associate professor, senior member of CCF. His main research interests include combinatorial optimization, real time scheduling, computer communications and bioinformatics.

贺思敏,1968 年生,博士,副研究员,中国计算机学会高级会员,主要研究方向为组合优化、实时调度、计算机通信与生物信息学.

Gao Wen, born in 1956. Ph. D., professor and Ph. D. supervisor, senior member of CCF. His main research interests include multimedia data compression, image processing, computer vision, multimodal interface, and artificial intelligence.

高文,1956 年生,博士,教授,博士生导师,中国计算机学会高级会员,主要研究方向为多媒体数据压缩、图像处理、计算机视觉、多模式接口和人工智能等.

Research Background

This work is partially supported by the Basic Research Program of the Institute of Computing Technology of the Chinese Academy of Sciences under the grant No. 20056090, the National Natural Science Foundation of China under the grant No. 69983008, and the National High Technology Development 863 Program of China under the grant No. 2001AA112100. Compared with an unbuffered crossbar, a combined input-crosspoint-queued (CICQ) switch is more attractive because of its distributed scheduling property. Among the different types of scheduling algorithms for CICQ switch, it is known that round-robin algorithms are the easiest to implement by hardware. Although the previously proposed round-robin algorithms achieve 100% throughput asymptotically under uniform traffic, these algorithms do not provide a satisfactory performance under nonuniform traffic. In this paper, we propose a class of dual round-robin algorithm for a CICQ switch with one-cell crosspoint buffers. With our algorithms, each input arbiter is associated with dual round-robin pointers. Unlike the previous round-robin algorithms, our algorithm has distinctive round-robin pointer updating rules which are powerful to cope with nonuniform traffic patterns. Extensive simulation results show that our algorithm achieves a satisfactory performance under both uniform and a broad class of nonuniform traffic patterns. Finally, we argue that the proposed algorithm is suitable to apply to implement multi-terabit capacity routers.