国 防 科 技 大 学 学 报
第 20卷 第 3期 JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY Vol 20 No. 3 1998

Banyan网的一类非阻塞特性及其在 ATM 交换机设计中的应用*

赵东升 卢锡城 周兴铭

(国防科技大学计算机系 长沙 410073)

摘 要 Banyan网具有结构简单、自寻址、硬件复杂度低等优点, 因而被许多交换系统和多机系统作为基本互连网络, 但其内阻塞特性限制了它的性能。对某些特定输入模式, Banyan网是非阻塞的。本文讨论 Banyan网的一类非阻塞特性, 说明它们在 ATM 交换机设计中的应用—— 通过数据分布提高 ATM 交换机的负载均衡能力, 用更小的硬件代价实现多目广播功能。

关键词 Banyan网, 非阻塞特性, 数据分布, 多目广播

分类号 TP393

A Class of Nonblocking Property of Banyan Networks and Its Applications in ATM Switch Design

Zhao Dongsheng Lu Xicheng Zhou Xingming

(Department of computer, NUDT, Changsha, 410073)

Abstract With the property of simple structure, self-routing and low hardware complexity, Banyan networks has been taken as interconnection network by many multi-processor systems and switching systems. However, its blocking property limits its performance and wide use. It is known that Banyan network is nonblocking for some special input/output patterns. In this paper, a class of nonblocking property of Banyan network is discussed, then some useful examples are demonstrated through application to ATM switch design: to promote ability of load sharing through data distribution and to implement multicast ability using lower hardware cost.

Keywords Banyan networks, nonblocking property, data distribution, multicast

Banyan网是一种自选路由的多级互连网络, 具有结构简单、自寻址、硬件复杂度低等优点, 因而被许多交换系统和多机系统作为基本互连网络。然而, Banyan网是有内阻塞的, 这限制了它的性能和应用效果。但是, 对某些特定的输入输出模式, Banyan网具有非内阻塞的性质。文献 [2]指出: 如果输入分组占据连续的输入端口, 且它们的目的地址单调, 则 Banyan网无内阻塞。这一性质已被广泛使用, 最典型的例子是 Batcher-Banyan交换结构 [2]。文献 [3]针对逆向 Banyan网 (Banyan的镜像), 证明了如果输入分组具有模 N 连续的目的地址, 则它非阻塞。由于 Banyan网和逆向 Banyan网互为镜像, 容易得到 Banyan网的另一非阻塞条件: 如果输入分组以模 N 方式占据连续的输入端口, 且它们的目的地址单调, 则不会发生内阻塞。本文讨论 Banyan网和逆向 Banyan网的这一类非阻塞性质, 着重说明它们在 ATM 交换机设计中的应用。

1 Banyan网的一类无阻塞特性

$N \times N$ Banyan网是具有 $\log_2 N$ 级的多级互连网, 每级有 $N/2$ 个 2×2 的交换开关, 数据只能逐级通过网络。Banyan网中, 任意输入到输出间只存在一条唯一的路径。如果对 Banyan网中每个开关的输入从上到下编号为 0和 1, 则每条输入和输出间的路径均可通过边的编号队列表示出来, 它们联在一起即表

* 国防预研项目资助
199?年 9月 15日收稿
第一作者: 赵东升, 男, 1970年生, 博士生

示目的地址的编号。令 $n=\log_2 N$, 目的地址二进制表示为 $d_1 d_2 \cdots d_n$, 则第 i级的交换开关仅需要目的地址的第 i 位 d_i: 如果 d_i 为 0, 则送上面的输出端; 如果 d_i 为 1, 则送下面的输出端。逆向 Banyan 网是 Banyan 网的"镜像", 即将输入端和输出端对换, 级的编号从 n 到 1进行, 其路由选择算法与 Banyan网一样。Banyan网和逆向 Banyan网都是自选路由的, 数据分组在每个时钟周期前进一级, 经 $\log_2 N$ 步到达目的端口。

本文讨论内部阻塞问题, 即具有不同目的地址的分组竞争同一条边的情况。所谓无内阻塞网就是指对所有输入和输出之间的任何一对一映射, 网络内部都无冲突。Banyan网和逆向 Banyan网是有内阻塞的网络, 但对特定的输入输出模式, 数据分组不会在内部发生冲突, 所有的分组可在 $\mathrm{Log}_2 N$ 步到达目的地址。下面以定理的形式给出它们的一类非阻塞条件, 为叙述方便, 定义有分组出现的输入端口为活动输入。

定理 1 Banyan网是非阻塞的, 如果其活动输入 $x_1, x_2, \cdots, x_k$ 和相应的目的地址 $y_1, y_2, \cdots, y_k$ 满足以下条件:

(1)输出单调: $y_1 < y_2 < \cdots < y_k$ (或 $y_1 > y_2 > \cdots > y_k$);

(2)输入集中: $x_1, x_2, \cdots, x_k$ 占据连续的输入端口。

证明略 (见文献 [5]Beckmann的证明)。

如果交换开关有数据复制能力, 则 Banyan 网可用于一对多寻址 (多目广播)。现在考虑一对多寻址问题, 设输入 x 有目的地址集合 $Y=\{y_1, y_2, \cdots, y_k\}$, $k \leqslant N$. 定义 $Y_1 < Y_2$为对 $\forall y \in Y_1$和 $\forall y' \in Y_2$, 有 $y < y'$. 根据定理 1, 以下推论成立。

推论 1 广播 Banyan网是非阻塞的, 如果其活动输入 $x_1, x_2, \cdots, x_k$ 和相应的目的地址集合 $Y_1, Y_2, \cdots, Y_k$ 满足以下条件:

(1)输出单调: $Y_1 < Y_2 < \cdots < Y_k$ (或 $Y_1 > Y_2 > \cdots > Y_k$);

(2)输入集中: $x_1, x_2, \cdots, x_k$ 占据连续的输入端口。

证明 任选 $y_1 \in Y_1, y_2 \in Y_2, \cdots, y_k \in Y_k$, 由定理 1可知, 各路径 $<x_1, y_1>$, $<x_2, y_2>, \cdots, <x_k, y_k>$ 互不冲突。命题得证。

文献 [3]中给出了逆向 Banyan网的一个非阻塞条件, 即下面的定理 2

定理 2 逆向 Banyan网是非阻塞的, 如果其活动输入 $x_1, x_2, \cdots, x_k$ 和相应的目的地址 $y_1, y_2, \cdots, y_k$ 满足以下条件:

(1)输入单调: $x_1 < x_2 < \cdots < x_k$ (或 $x_1 > x_2 > \cdots > x_k$);

(2)输出模 N 集中: $y_1, y_2, \cdots, y_k$ 以模 N 方式占据连续的输出端口。

因为 Banyan网和逆向 Banyan网互为镜像, 将二者的输入输出对换, 则可得到 Banyan网的又一非阻塞条件。

定理 3 Banyan网是非阻塞的, 如果其活动输入 $x_1, x_2, \cdots, x_k$和相应的目的地址 $y_1, y_2, \cdots, y_k$满足以下条件:

(1)输出单调: $y_1 < y_2 < \cdots < y_k$ (或 $y_1 > y_2 > \cdots > y_k$);

(2)输入模 N 集中: $x_1, x_2, \cdots, x_k$以模 N 方式占据连续的输入端口。

推论 2 广播 Banyan网是非阻塞的, 如果其活动输入 $x_1, x_2, \cdots, x_k$ 和相应的目的地址集合 $Y_1, Y_2, \cdots, Y_k$满足以下条件:

(1)输出单调: $Y_1 < Y_2 < \cdots < Y_k$ (或 $Y_1 > Y_2 > \cdots > Y_k$);

(2)输入模 N 集中: $x_1, x_2, \cdots, x_k$ 以模 N 方式占据连续的输入端口。

2 ATM 交换机的输入负载平衡

ATM 交换机中的缓冲方法主要有输入排队和输出排队[6], 输出排队系统在吞吐率、时延等方面都优于输入排队系统, 但其硬件复杂度高。输入排队系统硬件代价低, 但由于存在输出冲突, 性能受到限制。为提高输入排队系统的性能, 可采用加速因子 (Speed- up)或输出扩展等方法[7], 在输入端口和输出

端口都设缓冲器,交换结构成为混合排队系统。一般的混合排队系统中,每个输入端口的缓冲器是独占的。在非均衡输入业务模式和突发业务模式下,输入端口不能得到充分利用,降低了交换机的吞吐率。因为缓冲区不能共享,即使轻载端口的缓冲区还有空闲,重载端口也可能由于缓冲区溢出而丢弃信元(即定长分组)。

我们可以利用逆向 Banyan网的非阻塞特性实现共享的输入队列来解决这个问题,如图1所示。共享输入队列由一个数据分布网络(逆向 Banyan网)和一个 N 体交叉存储器组成,用两个指针 TOP和 BOTTOM 实现全局 FIFO;N 体交叉存储器的每一个体对应交换机的一个输入端口。每个交换机周期,K 个可能到达的输入信元通过逆向 Banyan网被寻址到从 BOTTOM 开始模 N 连续的存储器地址中,$0 \leqslant K \leqslant N$,地址的生成可用并行前缀算法。由定理2可知,信元不会在网络中阻塞,而共享存储器以 N 体交叉方式组织,相邻地址的存储单元位于不同的体中,所以 K 个到达的信元可以并行写入存储器。同时,交换机构每个周期最多可从共享队列中读出以 TOP开始的模 N 连续的 N 个信元。输入信元以循环方式均匀地分布到交换机构的 N 个输入端口(每个体中的信元数之差不超过1),解决了负载不均的问题。由于共享效应,信元丢失率得以降低。这里,为控制并行处理,所有硬件只需以输入链路的速度运行,复杂度低,易于实现。共享输入排队加上小加速比的交换结构,性能可以接近输出排队系统,但硬件代价低得多。

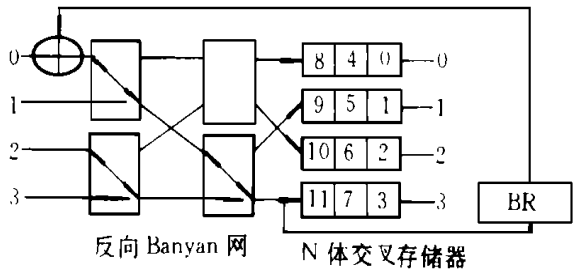

图1 共享输入队列的实现

3 多目广播交换机的设计

点-多点通信常用多目广播(multicast)和广播(broadcast)来实现,广播可看作多目广播的特例(一到全部的多目广播)。直接在多级互连网上实现多目广播是一个困难的问题。多目广播 ATM 交换机一般用一个拷贝网络复制信元,拷贝网络的输出作为点-点交换网的输入,由点-点交换网完成最终的信元寻址[5],其结构如图2所示。

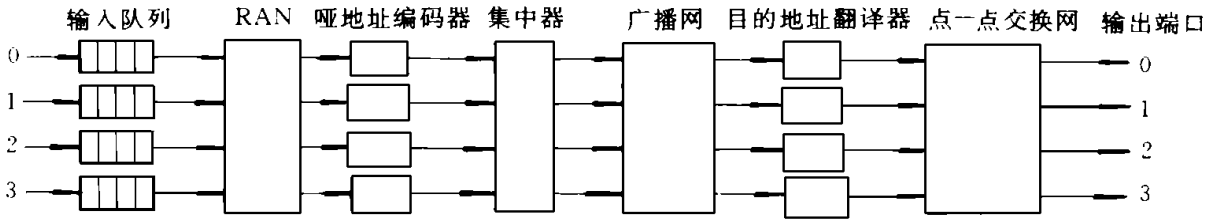

图2 拷贝网一般结构(4×4为例)

游动加法器网络(Running adder network,简写为 RAN)对各输入的拷贝数计算前缀和,生成各输入信元在拷贝网中的输出地址范围。例如,输入端口0的信元有三个目的地址,则生成的临时地址范围是[0,2]。集中器将输入信元送到拷贝网从0开始的连续输入端口,以满足 Banyan网数据广播的无阻塞条件(推论1),集中器用逆向 Banyan网实现。拷贝网将输入信元广播到由临时地址范围指定的连续输出端口,这是无阻塞的。复制后的信元由交换网送到最终的目的端口。在一个周期中,拷贝网络至多能输出 N 个信元,当输入信元的拷贝请求数之和超过 N 时就会发生输入溢出问题。因为 RAN 每次都从输入端口0开始计数,当溢出发生时总是上端的输入信元优先得到服务,导致不公平性。文献[8]用循环 RAN 解决输入公平性问题,循环 RAN 每次从上一周期中第一个未能得到服务的输入端口开始计数。但是,集中器的输入从任意输入端口开始计数不满足逆向 Banyan网的非阻塞条件。文献[8]使用两个平行的

Banyan 网或“膨胀度”为 2的逆向 Banyan 网来解决这个问题, 文献 [9]中给出了其非阻塞性质.

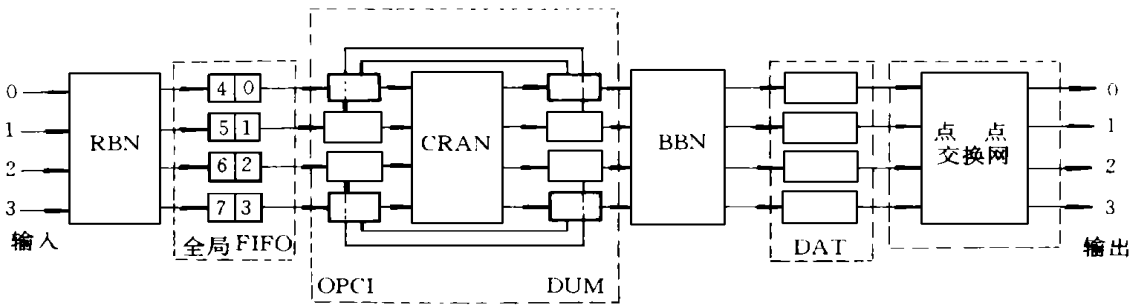

图 3 新的拷贝网体系结构

我们利用推论 2的结论, 提出一种改进的结构, 可用更小的硬件代价解决这个问题并能得到更高的性能[10]。所提结构如图 3所示, 它由共享输入队列 (由逆向 Banyan网和 N 体交叉存储器组成)、模 N 循环 RAN (CRAN)、广播 Banyan网 (BBN)构成。在每一周期, 输入信元首先被送到共享输入队列的模 N 连续地址单元 (从全局 FIFO 的队尾开始算起)。模 N 循环 RAN 每次从共享队列的全局队首位置开始对拷贝请求计数, 选定 K 个允许读出的信元, 这 K 个信元的拷贝请求数之和不超过 N. 被选定的信元由拷贝网络读出, 广播到拷贝网从 0开始的连续输出端口, 由推论 2可知, 这一广播寻址是非阻塞的。因为共享输入队列在逻辑上是一个全局的 FIFO, 所以输入信元以循环方式公平地写入和读出, 解决了拷贝网的输入公平性问题。这一结构同时解决了非平衡输入业务模式和突发业务模式下的输入负载不均问题, 使拷贝网的输出端口最大限度地得到利用。与 [8]中的方案相比, 这里用一般的逆向 Banyan网代替了“膨胀度”为 2的逆向 Banyan网, 降低了硬件代价。另外, 使用共享输入队列还可以减少存储器需求。我们在文 [10]中给出了这一方案的具体控制算法和性能评价, 表明它有更高的性能。

4 结论

本文讨论了 Banyan网的一类非阻塞特性, 利用它们解决 ATM 交换系统中的输入负载不均问题和广播网络的输入公平性问题。这些问题的解决方法都是基于并行处理和分布控制的, 硬件代价小, 具有良好的可伸缩性, 适用于构造大规模的交换系统。

参考文献

1 huang A, Knauer S. Starlite A wideband digital switch. In: Proc GLOBECOM'84, Nov. 1984

2 Narashima M. The batcher-banyan self-routing network: universality and simplification. IEEE Trans on Commun, 1988, 36: 1175~1178

3 Kim H S, Garcia A L. Nonblocking property of reverse banyan networks. IEEE Trans Commun, 1992, 40(3)

4 Bianchini R P, Kim H S. The tera project: a hybrid queueing ATM switch architecture. IEEE JSAC, 1995, 13(4)

5 Lee T T. Nonblocking copy networks for multicast packet switch. IEEE JSAC, 1988, 6

6 Liew S C. Performance of various input-buffered and output-buffered ATM switch design principles under bursty traffic: simulation study. IEEE Trans Commun, 1994, 42(2)

7 Oie Y et al. Effect of speedup in nonblocking packet switch. In: Proc ICC'89, 1989

8 Byun J W, Lee T T. The design and analysis of an ATM multicast switch with adaptive traffic controller. IEEE Trans on Networking, 1994, 2(3)

9 Lee T T, Liew S C. Broadband packet switches based on dilated interconnection networks. IEEE Trans Commun, 1994, 42

10 赵东升, 周兴铭, 卢锡城. 基于共享输入队列和自适应交通控制器的无阻塞 ATM 拷贝网. 通信学报 (已录用, 待发表)