一种递归构造的合成 BANYAN 网络

任开新,顾乃杰,潘 伟,刘 刚
(中国科学技术计算机科学技术系,安徽合肥 230027)

摘 要: 该文提出了一种新的多路径多级互连网络——递归构造的合成 BANYAN 网络,网络由若干级 3 ×3 的开关组成. 通过增加中间链路,解决了在已有的由 Seo 和 Feng 提出的合成 BANYAN 网上不能实现所有置换的问题. 该网络无需复杂的数值计算,通过二进制操作就可以很容易的产生路由标志,得到更多的路径,从而大大提高了路由成功率和容错能力. 该文中还给出了路由算法,并提出通过设置标识开关性能的标志位,使在路由时选取正确的路由标志,提前避开不起作用的开关,达到"预容错"的目的.

关键词: 多级互连网络; 合成 BANYAN 网; 递归构造; 路由标志; 路由算法
中图分类号: TP393.02 **文献标识码**: A **文章编号**: 0372-2112 (2003) 02-0228-04

The Recursively Constructed Composite Banyan Network

REN Kai-xin, GU Nai-jie, PAN Wei, LIU Gang
(Department of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China)

Abstract: A new multipath multistage interconnection network called the recursively constructed composite banyan network is proposed. The basic building blocks of the network are 3 ×3 switching elements. The advantage of this network is to resolve the problem that the composite banyan network proposed by Seo and Feng can't perform arbitrary permutation by adding middle stages. The network can easily generate a primary routing tag and more alternate tags through simple binary operations, which increase the degree of fault-tolerance. A permutation routing algorithm suitable for the network is introduced. The major feature of the network is pre-fault-tolerance. By setting a flag bit to identify whether the status of one witch is usable or not, the network can choose appropriate routing tag to avoid the switch that can't work in advance.

Key words: interconnection networks; composite banyan network; recursive construction; routing tags; routing algorithm

1 引言

多级互连网在并行系统中有着很重要的地位,并行系统的性能同多级互连网的时延与吞吐量密切相关. BANYAN 网络由 $\log_2 N$ 级的开关元件(SE)组成,写与 Cube、Omega 等网络都是拓扑等价的[1~4]. 这类网络的显著特征就是唯一路径,它们都是阻塞网. 改进的方法有:(1)增加硬件代价,例如增加额外级数、链接,改变开关的尺寸例如 3 ×3 或 4 ×4;(2)通过路由算法寻找其他路径来增加网络容错能力,避免错误路径. 近年来很多研究者在 BANYAN 网的基础上提出了新的结构,Mayez 将每个输入多路输出选择器的输入端都以二叉树的形式扩充,以较调换硬件代价减少路径冲突[5],Seo 和 Feng 将正反两个 BANYAN 垒加构造了合成 BANYAN 网[6],在此基础上 Seo 和 Lee 又提出了扩充合成 BANYAN 网[7],将合成 BANYAN 网的构造元件改进为 4 ×4,它们均是通过在输入输出之间提供多路径来增加网络的容错.

一个较好的容错网络应该在结构上具有正规性,并且在错误或冲突发生的时候容易重新路由. 本文提出了一种新型的多路径网络结构,称为递归构造的合成 BANYAN 网,对以往合成 BANYAN 网[6,7]不能实现任意置换的缺陷进行了改进. 该网络的基本构造元件是 3 ×3 的开关,通过主次路由标志来获取任意一对源地址和目的地址之间的更多的分离路径来大大提高网络的容错能力.

2 合成 BANYAN 网介绍

Seo 和 Feng 提出的 $N \times N$ 合成 BANYAN 网[6](图 1)具有 $\log_2 N$ 级,每级由 $N/2$ 个 3 ×3 的 SE 组成. 相邻两级 i 和 $i+1$ 之间的链接 ℓ_i 是根据连接准则将反 BANYAN 网的链接叠加到 BANYAN 网上得到的. 合成 BANYAN 网中的 SE 大小都相同,第一级输入的第二个端口和最后一级的第二个输出端口没有用到,它们实际上是按照 2 ×3 和 3 ×2 的 SE 工作的. 每个 SE 均有三个输出(最后一级除外),每位都有三个端口值,$t_i \in \{0, 1, \bar{1}\}$,其中 0 表示正反 BANYAN 网络均有的连线,对应 $K_i = K_{i+1}$ 的情况;$\bar{1}$ 表示正 BANYAN 网络的连线,对应 $K_{i+1} = K_i \sim 2^i$ 的情况;1 表示反 BANYAN 网络的连线,对应 $K_{i+1} = K_i \sim 2^{n-2-i}$ 的情况.

收稿日期:2001-12-24;修回日期:2002-09-10

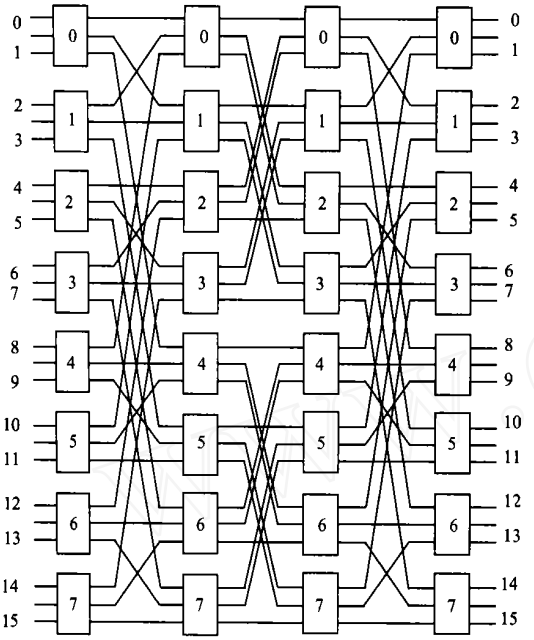

图 1　16 ×16 合成 BANYAN 网络

在合成 BANYAN 网中,主路由标志是由源和目的地址的前 $n-1$ 位按位异或得到的,取得主路由标志以后,可以按照以下算法求得三种情况对应的次路由标志:

CASE1:对于位置对称而值不等于 0 的位,当对应位的值均为 1($\bar{1}$)时,则将对应位的值均换为 $\bar{1}$(1)可得到次路由标志;

CASE2:对于位置对称而值相反的位,则将对应位的值为为 1 和 $\bar{1}$,$\bar{1}$ 和 1,0 和 0 的路由标志都是次路由标志;

CASE3:对于路由标志长度为奇数而中间级恰为 0 的情况例外,该中间级上不能作任何操作.

设 $n=\log_2 N$,当 n 为偶数的时候,在中间级的链路上对标记为"0"的 8 条链路(见图 1 中间级之间的加重线)产生竞争的可能性非常大. 根据 Sue 和 Feng 提出的求次路由标志的算法 CASE3,无法对中间位做任何修改,不能求得不占用中间级"0"链路的次路由标志.

3　递归构造的合成 BANYAN 网(RCBN)

在合成 BANYAN[5]网中,对于映射 $\{0,1,2,\cdots,N\}\to\{0,1,2,\cdots,N\}$(其中 $n=\log_2 N$ 为偶数)的置换,主路由标志以及由它得到的次路由标志的中间标志位总是 0,无法改变 N 个输入竞争中间级中 $N/2$ 条"0"链路的状态. 由此我们可以得出以下的推论:

推论 1　$N\times N$(其中 $n=\log_2 N$ 为偶数)的合成 BANYAN 网络,若主路由标志的中间位(第 $\lfloor(n-1)/2\rfloor$ 位)为"0"的个数大于 $N/2$,则在中间级链路上发生的竞争冲突或硬件故障是不可排除的,因此无法完成所有置换.

证明　当 $n=\log_2 N$ 为偶数时,主路由标志的中间位对应着偶数级的合成 BANYAN 网中中间一级的"0"链接,该链接的个数为 $N/2$. 所以,当中间位值为"0"的主路由标志多于 $N/2$ 时,对中间级"0"链接的要求超过了网络本身结构所能提供的链接数,又由次路由标志的算法 CASE3 得知,此时无法对中间位做任何修改,因此无法完成所有置换.

上面的推论表明:并不是所有的合成 BANYAN 网络都能实现所有的置换,特别是在级数为偶数级时,在中间链路发生竞争的可能性较大,这是目前在合成 BANYAN 网中至今未能解决的关键问题.

为了解决上述问题,本文提出了一种新型网络——递归构造的合成 BANYAN 网,基本组成元件是 3 ×3 的开关,级数为 $2\log_2 N-3$ 级,递归的最小单位为 8 ×8 的合成 BANYAN 网络,比 Benes 网[7]级数($2\log_2 N-1$)少两级.

定义 1　设 N 为 RCBN 的输入端口数(其中第一级中每个 SE 只有 2 个输入端起作用),$k=N/2$ 为每一级的 SE 个数. $N\times N$ 的 RCBN 就是第一级的链路以及最后一级的链路遵循合成 BANYAN 网络连接准则,而中间链接着上、下两个子网,这两个子网均为 $N/2\times N/2$ 的 RCBN,依次类推,RCBN 中最小的递归单位为 8 ×8 的合成 BANYAN 网络.

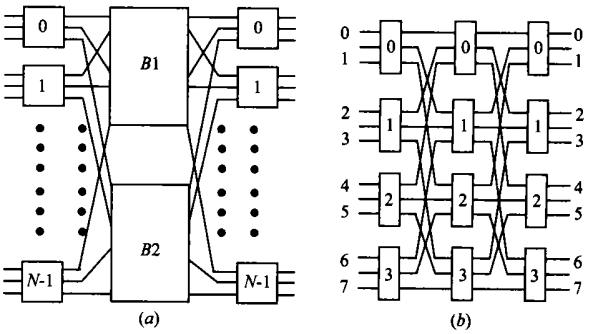

图 2　(a) $N\times N$ RCBN 的递归构造,B_1 和 B_2 都是 $N/2\times N/2$ RCBN;(b) 8 ×8 的合成 BANYAN 网是 RCBN 的最小递归单位

定义 2　对于第 k 级的第 i 个 SE($0\le k\le l$,$0\le i\le N/2-1$,其中 l 为链路的级数)来说,有:

如果 $0\le k<\frac{l}{2}$,当输出标签值分别为

$$\begin{cases}1\\ 0\text{时},K_{i+1}=\begin{cases}K_i\sim 2^{l-k-1}\\ K_i\\ K_i\sim 2^{0}\end{cases};\\ \bar{1}\end{cases}$$

如果 $\frac{l}{2}\le k<l$,当输出标签值分别为

$$\begin{cases}1\\ 0\text{时},K_{i+1}=\begin{cases}K_i\sim 2^{0}\\ K_i\\ K_i\sim 2^{k-l/2-1}\end{cases};\\ \bar{1}\end{cases}$$

从上述构造方法得知,网络的总级数为 $2\log_2 N-3$ 级,每级有 $N/2$ 个 SE,器件数为 $N(2\log_2 N-3)/2$ 个.

4　RCBN 中的多路径

本节将具体给出在 RCBN 中获取主、次路由标志的策略. 主路由标志可以通过对源和目的的二进制地址进行简单操作得到,一旦主路由标志确定下来,则次路由标志就可以很容易

的确定.

4.1 主路由标志的获取

由于网络以递归方式构造,所以主路由标志要按递归的层数逐层求得.

设主路由标志为 $P=(p_{l-1}p_{l-2}\cdots p_0)$,其中 l 为 RCBN 的中间链路级数,$l=2n-4$,其中 $n=\log_2 N$. 已知源和目的地址,将其表示为二进制的形式:$s=(s_{n-1}s_{n-2}\cdots s_1s_0)$ 和 $d=(d_{n-1}d_{n-2}\cdots d_1d_0)$. $Temp_0$ 是由对源和目的地址进行按位异或操作得到的:$Temp_0=(t_{n-1}t_{n-2}\cdots t_1t_0)=s\sim d=((s_{n-1}\oplus d_{n-1})\cdots(s_1\oplus d_1)(s_0\oplus d_0))$,其中:$0\le i\le n-1$,$t_i=s_i\oplus d_i$. 舍去最后一位,得到 $Temp_1=(t_{n-1}t_{n-2}\cdots t_1)$,令 $p_{l-1}=t_{n-1}$,$p_0=t_1$;考虑第二层 RCBN,将 s 中的 s_{n-1} 与 t_{n-1} 异或,d 中的 d_1 与 t_1 异或,去除 s 和 d 中的最高位,可以得到二次地址 $s^1=(s_{n-2}s_{n-3}\cdots s_0)$ 和 $d^1=(d_{n-2}d_{n-3}\cdots d_1d_0)$,再将 s^1 和 d^1 按位异或,舍去最后一位后取前后两个标志位作为第二层的主路由标志位;以此类推,照上述方法直至求至最内一层 8×8 合成 BANYAN 网即可.

16×16 的 RCBN 中,当源和目的地址分别为 $s=(0000)$ 和 $d=(1111)$ 时,得到的 $Temp_0=(1111)$,舍去最后一位为 (111),则递归的最外层主路由标志位为 $p_3=1$ 和 $p_0=1$;将 s 中的 s_{n-1} 与 t_{n-1} 异或,将 d 中的 d_1 与 t_1 异或,相应得到 (1000) 和 (1101),去除最高位得到二次地址 $s^1=(000)$ 和 $d^1=(101)$,进行按位异或得到 (101),去除最后一位得到 $p_2=1$ 和 $p_1=0$,即为最内层合成 BANYAN 网络的主路由标志位. 最后得到的 16×16 的 RCBN 的主路由标志为 ——$P=(1101)$.

4.2 次路由标志的确定

一旦主路由标志已经求出,则可相应求得次路由标志.

定义 3 设路由标志为 $P'=(p'_{l-1}p'_{l-2}\cdots p'_0)$,则标志位中的两个位 i 和 j 互为对称位当且仅当满足条件 $j=l-i-1$,其中 $0\le i\le l/2-1$,$l/2\le j\le l-1$.

每对对称位构成了路由标志 $P'(p'_{l-1}p'_{l-2}\cdots p'_0)$ 中的一层,也就是一层递归的链路. 路由标志共有 $l/2$ 层,每一层都可以用下面的等式表示:

$$P'=\underbrace{p'_{l-1}p'_{l-2}\cdots p'_{l/2}}_{level\frac{l}{2}}\underbrace{p'_{l/2-1}\cdots p'_1p'_0}_{level1}$$

RCBN 的路由标志位也继承了合成 BANYAN 网的路由标志特性:

如果路由标志 $P'=(p'_{l-1}p'_{l-2}\cdots p'_0)$ 中的两个位 i 和 j 互为对称位(这里假定 $i<j$),则:

CASE 1. $p'_i=p'_j=0$,则将 p'_i 和 p'_j 的值换为 1 和 $\bar{1}$ 或者换为 $\bar{1}$ 和 1 均为次路由标志,反之亦然;

CASE 2. $p'_i=1$ 并且 $p'_j=0$,则将 p'_i 和 p'_j 值换为 0 和 $\bar{1}$ 也同样为次路由标志,反之亦然;

CASE 3. $p'_i=p'_j=1$,则将 p'_i 和 p'_j 的值换为 $\bar{1}$ 也同样为次路由标志,反之亦然.

对于 $N\times N$ 的 RCBN,如果将每级的 SE 从上到下划分、每两个划分成一对的形式(共有 $N/4$ 对),则相临两级的 SE 对之间不但都有输出标签为 0 的链路连接,而且都有交叉链路连接,即满足图 3 所示.

表 1 对称位的标志位设置

主路由标志中的相应对称位	一对对称位的值			
	00	01	10	11
对应位的次路由标志设置	$\bar{1}1$ $1\bar{1}$	$\bar{1}0$	$0\bar{1}$	$\bar{1}\bar{1}$

由上述特征我们可以得出以下结论:

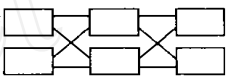

图 3 RCBN 中每对 SE 之间的必有联系

推论 2 给定路由标志位中的两个位 i 和 j,若 $|i-j|=1$,则:

CASE 4. 若满足均在标志位的高位半部且 $p'_i=p'_j=0$,则 p'_i 和 p'_j 的值设为 $\bar{1}$ 也是次路由标志,反之亦然;如果 $p'_i=\bar{1}$ 并且 $p'_j=0$,则将 p'_i 和 p'_j 的值对换也为次路由标志,反之亦然;

CASE 5. 若满足均在标志位的低位半部且 $p'_i=p'_j=0$,则 p'_i 和 p'_j 的值设为 1 也是次路由标志,反之亦然;如果 $p'_i=1$ 并且 $p'_j=0$,则将 p'_i 和 p'_j 的值对换也为次路由标志,反之亦然;

定理 1 在 $N\times N$ RCBN 中,在任意一对输入端和输出端之间,至少存在 $2^{(\log_2 N-1)/2}$ 条分离(disjoint)路径.

证明:由 RCBN 的路由标志特性 CASE 2、CASE 3,可以得知每层 RCBN 对应的路由标志至少有两种,即对应的分离路径至少有两条,而递归总层次为 $(\log_2 N-1)/2$,所以在任意一对输入端和输出端之间至少存在 $2^{(\log_2 N-1)/2}$ 条分离路径. $N>8$ 时,由特性 CASE 4、CASE 5 可知,分离路径数大于 $2^{(\log_2 N-1)/2}$(如图 4),特别的,当 $N=8$ 时,有 2 条分离路径.

推论 3 在 $N\times N$ RCBN 中,如果主路由标志 $P=0$,即输入端口号和输出端口号相同,则总的路径数至少为 $3^{(\log_2 N-1)/2}$ 条.

证明:由 RCBN 的路由标志特性 CASE 1 可知,当输入端口号和输出端口号相同时,每层对应的路由标志有三种,即对应的分离路径有三条,而递归层次为 $(\log_2 N-1)/2$,所以总的路径数至少为 $3^{(\log_2 N-1)/2}$.

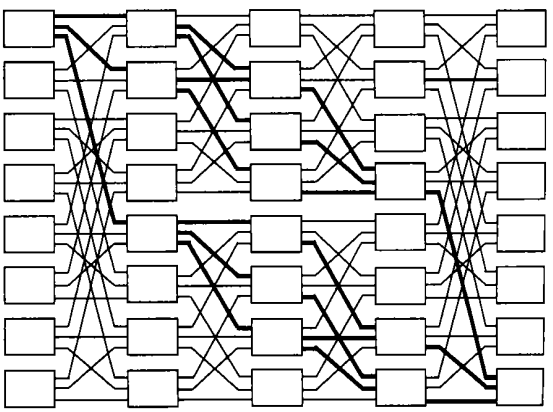

图 4 16×16 RCBN 中的多路径(源节点 0000 →目的节点 1111)

从上面的结论可知,RCBN 与已有的合成 BANYAN 网相比,可以在任意一对源和目的节点之间得到更多的路径,大大

提高了路由的成功率.

5　RCBN 的路由算法及容错性讨论

由于网络和路由标志的逐层递归构造特性,所以根据目的地址 s 和路由标志 P',可以利用 RCBN 的拓扑描述求得到达下一个递归层的 N_{SE}(N_{SE}为一级中 SE 的序号,$0 \le N_{SE} \le N/2-1$);统计每个 SE 即将接收到的路由信息数目,若大于 3 个,则需要选择其他的路由标志;若某个 SE 上即将接收到的路由信息(即路由标志)在该 SE 所在级无冲突(即相应的标志位不同),则确定该层对应的路由标志位后,再依次类推下一个递归层;若某个 SE 即将接收到的路由信息在该 SE 所在级发生冲突,则需要修改路由标志,直至符合无冲突条件. 在每一个递归层都通过算法确定出部分路由标志——即该层对应的两个标志位,然后根据确定下来的两个标志位行进至下一个递归层直至最小的递归单位,即在确定最外层的路由标志后,问题转化为在下面的两个子网 B_1 和 B_2 上进行路由,依此类推,直至求到最小的 8×8 子网,由上可知,该算法的递归深度为 $(\log_2 N-1)/2$.

从递归合成 BANYAN 网络的构造和次路由标志位序列的获得来看,该网络比较其他网络增添了大量的次路由标志序列,一对源地址和目的地址之间的可用路径大大增多,这样就极大限度的提高了网络的容错性.

一般网络的故障通常会发生在链路故障和 SE 故障两个方面,而对于最开始一级和最后一级的 SE 来说,如果发生硬件故障,是无法解决的. 如果中间级的 SE 发生故障,则同该 SE 相连的输入输出共六条链路均会发生失效,通常情况下,若一级中发生故障的 SE 数目高于 $N/6$,则该级失效的链路数也将高于 $N/2$,导致该级可用链路数少于可供路由的最少链路数 N,此时将无法正确完成 $N \times N$ 的全置换. 其他的情况均可以为其选择次路径来进行重路由.

根据上面的路由算法来看,该网络在确定每个递归层的路由标志位时,首先要求出到达的下一级的 N_{SE}(即到达的 SE 序号),如果在每个 SE 上设置一个状态位 mark 用来标识该开关的状态是否良好,当 mark 标志显示即将到达的 SE 状态不正常时,就可以提前修改路由标志,使其路由后不会到达该 SE,这样可以最大程度的避免 SE 故障对路由的影响,并且可以使信息尽量一次性通过网络路由成功而无需进行重路由,达到"预容错"的目的.

6　结论

本文介绍了一种新型的递归构造的合成 BANYAN 网,网络由 $2\log_2 N-3$ 级的 3×3 的 SE 递归构造,元器件总数为 $N(2\log_2 N-3)/2$. 该网络可以有效和快速的在任意一对源和目的地址之间确定至少 $2^{(\log_2 N-1)/2}$ 条分离路径. 一旦信息在网络传递中发生了开关和链路错误,可以选择其他的次路由标志,重新进行路由. 该网络具有多路径、重路由能力,除了具有非常强大的容错性以外,通过设置一个标志位 mark 来标识开关是否工作良好,达到"预容错"的目的.

参考文献:

[1] C L Wu, T Feng. On a class of multistage interconnection networks [J]. IEEE Trans, 1980, 29(8): 694 - 702.

[2] D H Lawrie. Access and alignment of data in an array processor [J]. IEEE Trans, 1975, 24(12): 1145 - 1155.

[3] T Feng. Data manipulating functions in parallel processors and implementations [J]. IEEE Trans, 1974, 23(3): 309 - 318.

[4] K E Batcher. The flip network in STARAN [A]. Proc. Int'l Conf. Parallel Processing [C]. 1976. 67 - 71.

[5] Seung Woo, et al. The composite banyan network [J]. IEEE Trans, 1995, 6 - 10(10): 1043 - 1054.

[6] Mayez A, et al. Evaluation of pipelined dilated banyan switch architectures for ATM networks [J]. IEEE Trans, 1999, 7(10): 724 - 740.

[7] Hyoung Il Lee, et al. The augmented composite banyan network [A]. Madras India: 1998 5th International Conference on High Performance Computing [C]. 1998. 285 - 292.

[8] V Benes. Mathematical, Theory of Connecting Networks [M]. New York: Academic Press, 1965.

作者简介:

任开新　女,1977 年出生于辽宁海城,中国科学技术大学计算机科学与技术系学士,毕业后留校任教,研究方向:多级互连网络,并行体系结构,并行计算. E-mail: renkx@ustc.edu.cn.

顾乃杰　男,1961 年出生于江苏南通,中国科学技术大学计算机科学技术系副教授,研究方向:并行分布式算法、并行分布式计算中的通讯问题. E-mail: gunj@ustc.edu.cn.