超级递归基准互连网络性能分析*)

Performance Analysis of Super Recursive Baseline Interconnection Networks

侯国峰　杨愚鲁

(南开大学计算机科学与技术系　天津300071)

Abstract　Based on Delta network, Baseline network and current VLSI technique development level, a new family of MINs using 8×8 switches which is called Super Recursive Baseline Interconnection Network (SRB) has been presented. This paper makes network performance analysis, simulation and comparison, and proves that SRB has superior qualities in network pass rates, bandwidths and performance/cost ratios, etc. Therefore, SRB is proved to be a family of Multistage Interconnection Network simple in routing, superior in performance, and easy in expanding.

Keywords　Interconnection Network, Network pass rate, Bandwidth, Performance/cost ratio

1　引言

在并行处理领域,研究并行机中多处理器连接的方式(即互连网络)是一个很重要的课题。互连网络是MPP的核心部分,拓扑结构、寻径控制和流控策略是其要素。为了降低互连网络的代价、提高其传输性能和可伸缩能力,研究人员已经提出了许多种互连网络,其中Delta网络和基准网络是较早提出的总体性质较好的互连网络[1,2],它们已被用于许多种并行机中处理器连接的拓扑结构(如BBNTC-2000,IBM RP3)。Delta网络具有较高的频带和性能价格比,但可扩展性差。基准网络使用2×2交换开关,具有简单的寻径控制和较好的可伸缩性等性质,但硬件代价较大。从集成电路技术角度,系统规模的增大使得许多互连网络结构难以实现[3,4],系统的组装受限于组装单元的管脚数(边界面积)和布局面积,这种组装技术是互连网络结构的最终决定因素[5],网络结构只能在此前提下有效地开发系统组装的特点,才能获取尽可能高的性能。即使在当今的集成电路技术条件下,解决LSI芯片上的引脚数目限制问题仍然是困难的,研究表明,基于当前VLSI技术水平,8×8交叉开关具有最优的性能价格比。使用16×16的交叉开关的系统也可能实现,但此时大多数情况下需要使用多路复用器和多路分配器,因此系统性能价格较低。8×8交叉开关成为目前应用最广泛的开关元件[6]。基于当前集成电路技术发展水平,在Delta网络和基准网络的构造思想基础上,一种使用8×8交换开关的新型动态互连网络被提出[7],称为超级递归基准互连网络(Super Recursive Baseline Interconnection Network, SRB)。本文介绍了它的基本性质,SRB具有简单的寻径控制方法,能够实现多种置换并且具有优良的网络规模可扩展性;并对SRB进行了性能分析与比较,证明它具有较高的频带和网络通过率,比现有的Delta网络、基准网络和交叉开关网络等具有更高的性能价格比,并且它适应于当前VLSI技术的发展水平,具有很好的应用前景。

2　SRB的定义及其拓扑结构

SRB是一种使用8×8的交叉开关的多级阻塞动态互连网络。每个交叉开关内部构造是一个8×8 Crossbar结构,从交叉开关的任何一个输入端可以输出到它的任一个输出端。SRB网络的规模为N×N,其中$N=8^n(n=2,3,\cdots)$。对于一个SRB网络的命名规则是从左到右,开关段依次为$0,1,\cdots,\log_8N-1$,链路级依次为$0,1,\cdots,\log_8N$。开关段中的每个开关位置用8进制表达式$p_np_{n-1}\cdots p_1$来表示,其中$n=\log_8N-1$,用$(p_np_{n-1}\cdots p_1)_i$表示段i中的一个交换开关;而交叉开关之间的每条链路则用$p_np_{n-1}\cdots p_0$来表示,用$(p_np_{n-1}\cdots p_0)_i$表示级i中的一条链路。如果链路是从交换开关的最上端端口引出,则$p_0=0$;否则按引出端口从上到下的顺序p_0依次取值1,2,…,7。

*)本文系教育部高等学校骨干教师资助计划和天津市自然科学基金(重点)(编号:003800111)资助项目。侯国峰　硕士研究生,主要研究方向为并行体系结构、计算机网络、人工智能,杨愚鲁　教授,博士生导师,主要研究方向为并行体系结构、互连网络、人工智能。

其拓扑描述规则定义为：$\beta^k[(p_np_{n-1}\cdots p_0)_i]=(p_n\cdots p_{n-i+1}\,k\,p_{n-i-1}\cdots p_2)_{i+1}$，通过链路$(p_np_{n-1}\cdots p_1k)_{i+1}$，$0\leqslant i<n$，$0\leqslant k<n$。用文字表示就是：$\beta^k$ 通过链路$(p_np_{n-1}\cdots p_1k)_{i+1}$，$0\leqslant i<n$，$0\leqslant k<n$，将段 i 中的交叉开关$(p_np_{n-1}\cdots p_1)_i$ 连接到段 i+1的交叉开关$(p_n\cdots p_{n-i+1}\,k\,p_{n-i-1}\cdots p_2)_{i+1}$。

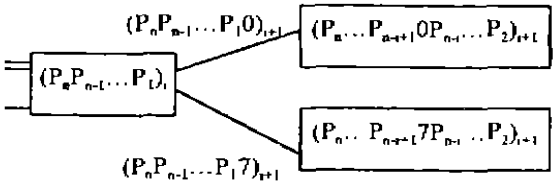

图1　拓扑描述规则的物理意义

SRB 的结构可以用如图2所示方法递归生成。它表示的是第一次迭代，将 N×N 网络分为第一级 N×N 的开关模块和第二级的8个(N/8)×(N/8)网络子块 $C_i(i=0,1,\cdots,7)$。第二次迭代再将 $C_i(i=0,1,\cdots,7)$ 各分为(N/8)×(N/8)的开关模块和(N/64)×(N/64)的网络子块。这样递归下去，直到分为 N/8个8×8的子块为止。总的迭代次数为 $\log_8N-1$，得到的 SRB 网络的开关段数是 $\log_8N$，链路级数是 $\log_8N+1$。由于 SRB 的递归型结构，可以方便地利用小规模 SRB 构造更大规模 SRB，因此 SRB 具有优良的网络规模可扩展性。

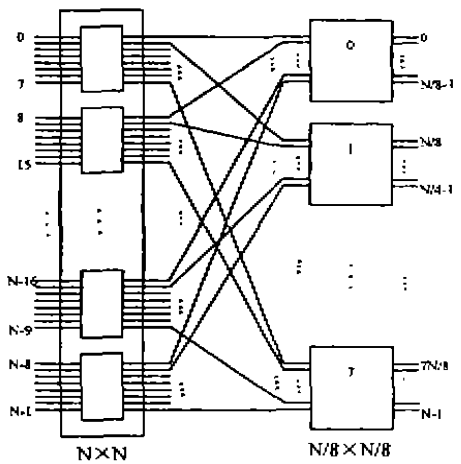

图2　产生 SRB 网络的递归过程

3　SRB 的寻径控制方法

SRB 网络采用单元控制方式，即寻径控制机构有单独设置每个交换开关状态的能力。寻径算法使用目的标记寻径算法[8]，使用该算法 SRB 可以建立任何源端到目的端的一条连接。记 $S_nS_{n-1}\cdots S_1S_0$ 为源端链路的8进制标记，$D_nD_{n-1}\cdots D_1D_0$ 为目的端链路的8进制标记($n=\log_8N-1$)。寻径算法规定 $c=D_{n-i}$ 作为第 i 段交叉开关的寻径控制信号，来控制源端到目的端所经过的第 i 段交叉开关的状态，使输出链路从该交叉开关自上而下顺序地从第 c+1个输出端口引出。因此，第 i 段($i=0,1,\cdots,n$)交叉开关实现的是链路 $p_np_{n-1}\cdots p_1p_0$ 到 $p_np_{n-1}\cdots p_1D_{n-i}$ 的变换，而第 i+1级($i=0,1,\cdots,n-1$)混洗连接实现的是 $p_np_{n-1}\cdots p_1D_{n-i}$ 到 $p_n\cdots p_{n-i+1}D_{n-i}p_{n-i-1}\cdots p_1$ 的变换，因此当源端链路经过网络到达目的端时，链路标记必然已经经过 n+1次上述的变换变为 $D_nD_{n-1}\cdots D_1D_0$。所以，它一定可以正确到达目的端；反之，也可以从目的端到源端反向进行寻径。SRB 具有的简单的目的寻径控制方法使网络寻径代价较低，具有优良的寻径效率。

通过分析可知，SRB 具有其它多级阻塞网所不具备的优点，它能容许通过位序颠倒置换，这对实现 FFT 很有利；并且经二次通过能实现任意置换[9]。它不能实现恒等置换，+1 mod N 置换等，但这些问题可以通过网络重构加以解决。

4　SRB 的性能分析

互连网络接受随机访问请求的概率称为网络通过率，它和频带、性能价格比三者均是评价多级互连网络性能的重要参数[10]，本节主要从这三方面对 SRB 网络进行性能分析。在分析计算之前，先作几点假设[9]：(1)每台处理机发出的访问请求是随机的、独立的，这些请求均匀分布于所有的存储模块。(2)每台处理机发出访问请求的平均速率为一个周期内 m 次。这里，一个周期是指网络的通过时间、访问存储器的时间以及返回处理机的时间之和、不论访问存储器的操作是读还是写。显然，$m\leqslant1$。(3)每个周期内所产生的各个访问请求都同时发出。每台处理机每个周期发出请求的次数最多为1。(4)凡被阻塞掉(未被接受)的请求即不再考虑，下一个周期发出的是与这个阻塞请求无关的新的请求。

上述几点假设中，假设(2)与 SRB 网络的具体实现有关；假设(2)、(3)旨在说明每台处理机每个周期内产生的请求概率为 m；假设(4)是为简化分析而作的。实际上被拒绝的请求在下个周期内是要重发的，这样使问题分析起来比较复杂。但模拟研究表明，如不作假设(4)，则网络接受请求的概率只是稍低一点而已，故这样假设下所得到的分析结果也是可靠的[11]。

令 P_A 为一个请求能被网络接受的概率，m 是每台处理机产生请求的平均概率，则 N×N SRB 网络的频带为 mNP_A，它表示网络在一个周期内可以接受的请求数。针对 SRB 网络，讨论 $N=8^n$ 的情况。

假定一个8×8模块 M 上有 i 个请求，用 $P(0|i)$ 表

示在M一个输出端上没有响应的概率。由于M的8个输出端被请求访问的概率是一样的，因此$P(0|0)=1$，$P(0|1)=7/8$，$P(0|2)=49/64$，…，$P(0|8)=(7/8)^8$。再用$P(1|1)$表示在M一个输出端上响应请求的概率，则$P(1|1)=1-P(0|1)$。这样，M可以接受的请求数为$E(1)=8P(1|1)$，现在来讨论网络中某一级h的情况。令$q_h(k)$为k个请求到达h级一个模块上的概率，$P_A(h)$为一个到达h级的请求能被接受的概率，则$P_A(h)=$(h级上一个模块可接受的请求数)/(到达h级一个模块上的请求数)，由$E(1)$的表达式及$q_h(k)$的定义，可得$P_A(h)=(E(1)q_h(1)+E(2)q_h(2)+\cdots+E(8)q_h(8))/(q_h(1)+2q_h(2)+\cdots+8q_h(8))$，其中$q_h(k)$可以用下列递归方法求得：假定某一级上的模块$M_1$有$i_1$个请求，模块$M_2$有$i_2$个请求，…，模块$M_8$有$i_8$个请求。用$r(k|i_1,i_2,\cdots,i_8)$表示有k个请求可以到达下一级的模块A的概率（A的上一级模块中有且仅有模块M_1，M_2，…，M_8具有与模块A直接相连的链路），则

$$r(0|i_1,i_2,\cdots,i_8)=P(0|i_1)P(0|i_2)\cdots P(0|i_8),$$

$$r(1|i_1,i_2,\cdots,i_8)=P(1|i_1)P(0|i_2)\cdots P(0|i_8)+P(0|i_1)P(1|i_2)\cdots P(0|i_8)+\cdots+P(0|i_1)P(0|i_2)\cdots P(1|i_8),$$

$$\cdots,$$

$$r(8|i_1,i_2,\cdots,i_8)=P(1|i_1)P(1|i_2)\cdots P(1|i_8)$$

因此，$q_{h+1}(k)$可以表达为$q_{h+1}(k)=\sum_{i_1+i_2+\cdots+i_8=k} r(k|i_1,i_2,\cdots,i_8)q_h(i_1)q_h(i_2)\cdots q_h(i_8)$，至于q的起始值，可以通过每台处理机产生请求的平均速率m求出。在第0级一个8×8模块上的请求分布可以写为$q_0(0)=(1-m)^8$，$q_0(1)=8m(1-m)^7$，$q_0(2)=28m^2(1-m)^6$，…，$q_0(8)=m^8$。

所以，一个请求可以为$8^n\times 8^n$ SRB网络所能接受的概率$P_A=P_n(0)P_n(1)P_n(2)\cdots P_n(n-1)$，网络在一个周期内可以接受的访问请求数，即它的频带$BW=mNP_A$。

根据上述分析，可以对SRB网络与现有的交叉开关网络、Delta网络和基准网络的性能作一比较，结果如图3～图6所示。

图3中，对N×N交叉开关网络来说，P_A在$N\geqslant 1$的所有整数时才有值；对基准网络和$2^n\times 2^n$ Delta网络而言，只有在$N=2^n(n\geqslant 1)$时才有定义；对SRB网络而言，也只有在$N=8^n(n\geqslant 1)$时才有定义。但图中的P_A为了便于观察它随N的变化趋势，画成了平滑的曲线。SRB网络，Delta网络和基准网络的P_A随N增加而下降，交叉开关网络在N逐渐增大时，其P_A接近一常数，在同等规模条件下，SRB的P_A低于交叉开关网络的P_A，但明显高于Delta网络和基准网络的P_A。

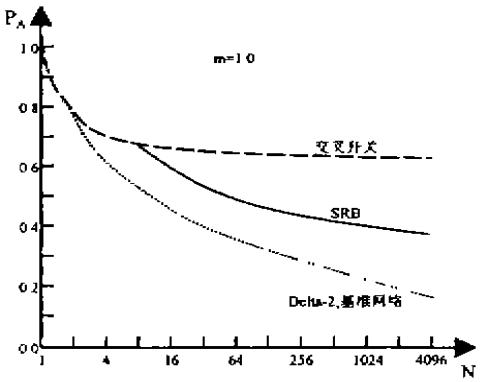

图3 SRB与其它网络$P_A(N)$关系的比较

图4说明，这几种网络的频带均随N的增加而增加。但当存储器的访问时间大于网络的延时时间时，它们的频带相差不十分大。图5表示的是N较大时P_A与m的关系，随着请求的平均速率增加，P_A会逐渐下降，SRB在这两个方面的性能分析中均优于Delta网络和基准网络。

图6表示的是性能价格比的曲线，性能在这里是指频带，价格则取决于网络中每个开关结点所含的电路。对交叉开关，其数量级为$O(N^2)$；对Delta-2网络和基准网络，其数量级为$O(N\log_2 N)$；对SRB网络，其数量级为$O(N\log_2 N)$。取1×1交叉开关的性能价格比为1，则图中纵坐标所代表的性能价格比都是相对于1×1的交叉开关而言的。结果表明：当$N\geqslant 64$时，SRB网络的性能价格比明显优于交叉开关、Delta网络和基准网络。采用SRB网络作为多处理机的互连网络，比上述其它各种网络具有更大的吸引力。

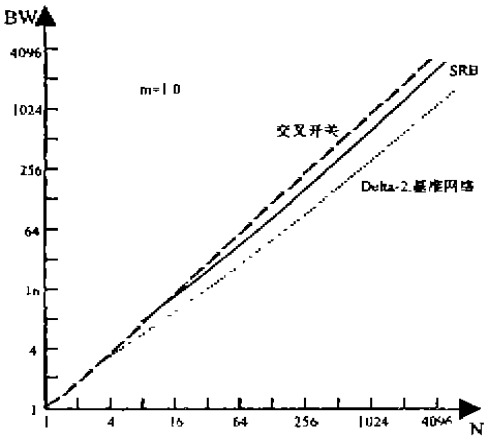

图4 SRB与其它网络BW(N)关系的比较

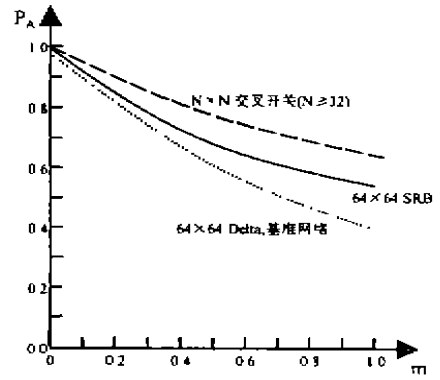

图5 SRB 与其它网络 $P_A(m)$ 关系的比较

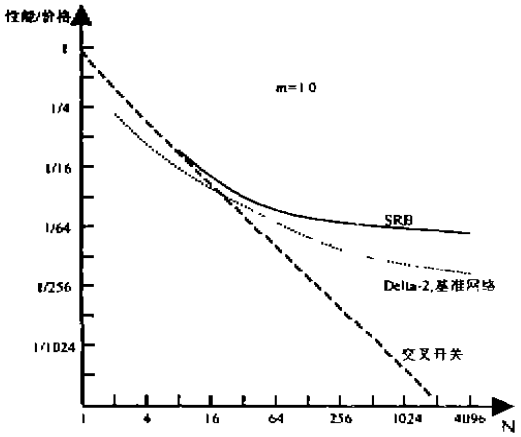

图6 SRB 和其它网络性能价格比的比较

结论 基于 Delta 网络、基准网络和当前 VLSI 技术的发展水平，一种使用8×8交叉开关的新型多级动态互连网络——超级递归基准互连网络(SRB)被提出。SRB 具有简单的寻径控制方法，能够实现多种置换并具有优良的网络规模可扩展性。更进一步地，SRB 具有较高的频带和网络通过率，比 Delta 网络、基准网络和交叉开关网络等具有更优良的性能价格比；并且 SRB 适应于当前 VLSI 技术的发展水平，具有良好的应用前景。

参 考 文 献

1 Dias D M, Jump J R. Analysis and Simulation of Buffered Delta Networks. IEEE Transactions on Computers, 1981, C-30(4): 273～282

2 Wu Chuan-Lin, Feng Tse-Yun. On a Class of Multistage Interconnection Networks. IEEE Transactions on Computers, 1980, C-29(8): 694～702

3 Engels M, Lauwereins R, Peperstraete J. The influence of technology on the choice of a multiprocessor Interconnection Network. In: Proc. of the Second Workshop on Parallel and Distributed Processing. Sofia, Bulgaria, Mar. 1990. 91～110

4 侯国峰，李涛，杨愚鲁. 动态互连网络研究. 见：2000年中国计算机学会计算机体系结构学术年会论文集. 哈尔滨，2000: 123～128

5 Hwang K. Advanced Computer Architecture. Parallelism, Scalability, Programmability. USA. McGraw-Hill, 1993

6 Amano H. Survey report on high performance switches. [The annual report from Advanced Information Technology Center(Survey report on high end computing technology)]. Tokyo, Japan: Advanced Information Technology Center, Apr. 2000. 55～75

7 Hou Guofeng, Yang Yulu. Super Recursive Baselines: A Family of New Interconnection Networks with High Performance/Cost Ratios. In: Proc. of the Fifth Int Symposium on Parallel Architectures, Algorithms and Networks. Dallas, Texas, USA, IEEE Computer Society Press, 2000. 260～265

8 Lawrie D H. Access and Alignment of Data in an Array Processor. IEEE Transactions on Computers, 1975, C-24(12): 1145～1155

9 王鼎兴，陈国良. 互连网络结构分析. 北京：科学出版社，1990

10 郑纬民，汤志忠. 计算机体系结构(第二版). 北京：清华大学出版社，1998

11 Patel J H. Performance of Processor-Memory Interconnections for Multiprocessors. IEEE Transactions on Computers, 1981, C-30(10): 771～780