一种新型大容量路由器交换网络中的高效路由策略*)

杨君刚[1,2] 邱智亮[1] 刘增基[1] 李红卫[2]

(西安电子科技大学综合业务网国家重点实验室 西安 710071)[1] (西安通信学院 西安 710106)[2]

摘 要 路由算法对交换网络性能具有很大的影响。本文针对一种新型的大容量路由器交换网络拓扑—XD(Cross-Direct)网络[1,2]的特点,提出了一类和应用于传统直连网络中的基于简单维序路由算法具有相同网络性能的路由算法—对角矢量映射法(DVM)。该类算法分为两种,文中对这两种算法进行了详细描述和性能分析,给出了它们各自的应用场合。

关键词 交换网络,直接连接网络,XD网络,对角矢量映射法,维序路由算法

An Efficient Routing Scheme in a New Switching Fabric of Large Capacity Router

YANG Jun-Gang[1,2] QIU Zhi-Liang[1] LIU Zeng-Ji[1] LI Hong-Wei[2]

(National Key Lab. of Integrated Service Networks, Xidian Univ., Xi'an 710071)[1] (Xi'an Communication Institution, Xi'an 710106)[2]

Abstract Routing algorithm plays a very important role in obtaining good performance of switching fabric. XD (Cross-Direct) network is a new topology used as switching fabric in large capacity routers[1,2]. Diagonal Vector Mapping (DVM) scheme is proposed in this paper, which is based on the characteristics of XD network. It possesses similar performance as the simple dimension-ordered routing used in traditional direct interconnection networks. Two kinds of algorithms are proposed based on DVM scheme. The algorithms are described in details, with the applications and the performance analysis presented.

Keywords Switch fabric, Direct interconnect networks, XD network, Diagonal Vector Mapping (DVM) scheme, Dimension-ordered routing

1 引言

大容量路由器是实现高速骨干网并决定其性能的关键。要实现大容量(数T比特级)路由器,一方面要提高端口速率(从OC-48到OC-192以至OC-768),另一方面要增加端口数目(几十个直到几百个)。将信息从输入端口转发到输出端口的交换网络是任何路由器设计的核心。传统的路由器是利用背板总线(如Cisco公司12000系列[1])或Crossbar(如Juniper M160路由器[2])来完成交换的。然而,总线结构不具备良好的可扩展性,整个路由器的交换容量受限于总线带宽;而Crossbar的成本与端口数的平方成正比,也不能经济的扩展到大容量。所以,大容量路由器的交换网络必须另作选择。

为了满足网络运营商对灵活性、可靠性和高性能的需求,一个大容量路由器交换网络需要具有分布式管理(Distributed Management)能力、高的可扩展性(Scalable)和强的容错(Fault Tolerance)能力。直接连接网络(direct interconnect networks)[3]是实现大容量路由器交换网络的一种理想方式,具有以下优点:(1)便于实现分布式管理,因为在直接连接网络中,各个网络结点都是相对独立的交换实体,因此其网络管理具有先天性的分布特性;(2)具有良好的路径多样性(Path Diversity)。在多维直接连接网络中,任意一对网络结点间都有多条(和网络维数成正比)路径,因此该网络具有良好的容错性;(3)具有良好的可扩展性。由于直接连接网络的各个结点都是相对独立的交换实体,因此,网络扩展非常简单,同时,网络维数越少,扩展的颗粒度越小,扩展成本越低。从网络性价比的角度,目前在大容量路由器中常用的直接连接网络拓扑有3D Torus和超立方体网络,如Avici公司的TSR40采用3D Torus结构[3],而Pluris公司的TeraPlex20系列产品采用超立方体结构[4]。但是这两种网络都是多维结构在可实现性和网络的可扩展性方面都不能令人满意,对此我们提出了一种具有良好可实现性和网络可扩展的直接连接交换网络拓扑结构—XD(Cross-Direct)网络[5,6],证明了该网络具有和传统直接连接网络同样的对称性(是一种Cayley图)[5];同时,由于该网络是一种两维的平面结构,便于工程实现和进行小颗粒度的扩展[6]。然而,具有良好的网络拓扑并不能保证网络具有良好的性能,网络的路由算法对交换网络的性能具有至关重要的影响。本文首先对XD网络的结构进行简单介绍,在此基础上提出了一类适用于XD网络的路由算法—对角矢量映射法(DVM:Diagonal Vector Mapping scheme),该类算法分为两种,在下文中分别以算法1和算法2表示,文中对这两种算法进行了详细描述,对其性能进行了分析,说明了其各自的应用场合。

2 XD网络拓扑介绍[5,6]

XD网络采用二维平面结构,每一维的结点个数可以相等也可以不相等,两维结点数相等的XD网络称为规则XD网络,两维结点数不相等的XD网络称为不规则XD网络。网络中每一个结点有6条链路和相邻结点相连(网络的度为

*)基金项目:国家863项目资助(2002AA103062);ISN国家重点实验室开放课题资助(ISN7-03)。杨君刚 讲师,博士研究生,主要研究方向:大容量路由和交换技术以及下一代光传送网络关键技术研究。

6)。如果用坐标(x,y)来表示结点在一个$k_x \times k_y$ XD网络中的位置,那么与该结点相连的邻结点是$(x \pm 1 \bmod k_x, y \pm 1 \bmod k_y)$和$(x, y \pm 1 \bmod k_y)$6个结点。图1和图2分别表示的是一个规则的5×5和一个不规则的6×5的XD网络拓扑示意图。为了以后论述方便,定义在图1和图2中沿$(x, y \pm 1 \bmod k_y)$方向的链路为直向链路,沿$(x \pm 1 \bmod k_x, y \pm 1 \bmod k_y)$方向的链路为斜向链路。

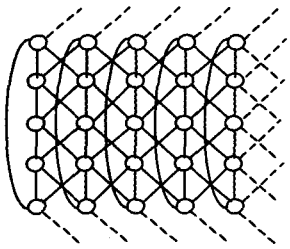


图1 $k_x=k_y=5$的规则XD网络示意图

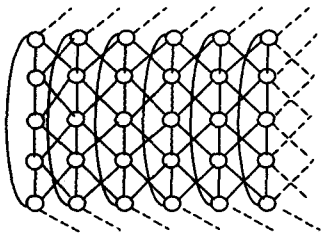


图2 $k_x=6, k_y=5$不规则XD网络示意图

3 XD网络中的路由算法－对角矢量映射法(DVM)

为了寻找适合于XD网络的路由算法,首先对该网络特征进行分析。在文[5]中,我们证明了XD网络所对应的图为Cayley图,因此,XD网络是一种结点对称网络。因此,对XD网络路由的讨论可以针对网络中的一个结点来讨论。由图1和图2 XD网络的结构图可以看出XD网络不是一种链路对称网络。斜向链路对网络连通性的贡献明显强于直向方向链路,因此在路由算法中应尽量避免过分依赖斜向链路,应该尽量向直向方向链路分担网络业务。下面,结合XD网络的特点,首先将在传统直连网络中简单、高效的路由算法－维序路由算法(dimension-ordered routing)进行改造应用到XD网络,提出算法1;在此基础上,针对该算法对网络直向链路利用率不高的缺点,提出了算法2。由于算法1和算法2是结合了XD网络的特点,在实现中需要网络坐标的变换和旋转,将这两种算法统称为对角矢量映射法(DVM)。最后,对算法的性能进行了分析。

3.1 算法1

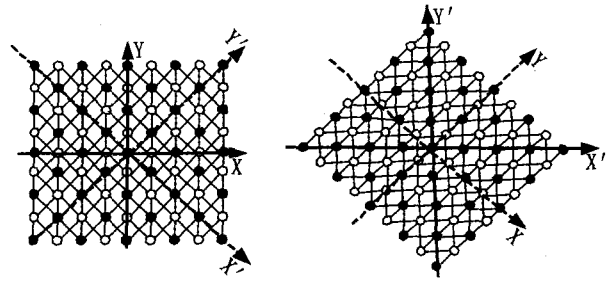
维序路由算法是传统直接连接网络中一种经典的路由算法,具有无死锁、是网络最短路由算法、实现简单等优点[3,7,8],算法1是结合XD网络的特性,将维序路由进行改造应用到其中的路由算法,这种算法具有维序路由的良好特性。下面对算法1进行描述。

(1)坐标变换:由于XD网络是结点对称网络[5],每一个结点在网络中的地位是等同的,因此,总可以通过坐标变换把信息的源结点变为坐标原点,这样便于实现和讨论。所以,该算法的第一步通过坐标变换将信息的源结点变为坐标原点,并求出信息的目的结点和源结点坐标相对位置。在此,假设信息的目的结点相对源结点的坐标为(x,y)。

(2)网络结点坐标映射(坐标旋转):完成坐标变换后,将信息的目的结点分为两大类。一类是目的结点坐标$|x+y|$为偶数的结点(如图3(a)中的深色结点),另一类是目的结点坐标$|x+y|$为奇数的结点(如图3(a)中的浅色结点)。将坐标系X-Y顺时针方向旋转45°,可以得到如图3(b)所示的坐标系X'-Y',在坐标系X-Y中的两类结点对应到坐标系X'-Y'中也对应分为两类(分别为图3(b)中的深色和浅色结点)。XD网络可以看作这两类结点组成的部分2Dtorus网络的叠加,而两个网络之间的连接是由结点在X-Y坐标系中的直方向(对应到X'-Y'—坐标系为斜方向)链路完成的。在其中每一个网络的内部结点间的通信可以采用类似于一般2D Torus中的维序路由算法[7];而两个网络结点之间的通信,必须经过两步来完成,第一步按单个网络内的路由算法到达和目的结点最接近的结点,第二步通过两个网络间的连接链路到达目的结点,第二步只需要一跳。下面对算法过程进行严格的数学描述。

假设要寻找从结点(x_1, y_1)到结点(x_2, y_2)的路径,令$x=x_2-x_1, y=y_2-y_1$,如果$|x+y|$为偶数,则令:

$$\begin{cases} x'=\dfrac{x+y}{2} \\ y'=\dfrac{y-x}{2} \end{cases} \tag{1}$$



(a) (b)

图3

如果$x' \neq 0$,则按x'方向走$|x'|$跳(此处x'方向分正方向和负方向,和x'值的正负相对应,对应到XY坐标系中x'的正方向和负方向分别和矢量(1,1)和(−1,−1)相对应),然后,如果$y' \neq 0$,再按y'方向走$|y'|$跳(y'方向也分正负,和y'值的正负相对应,对应到XY坐标系中y'的正方向和负方向分别和矢量(−1,1)和(1,−1)相对应),这样就可以到达目的结点。

如果$|x+y|$为奇数,若$y \geqslant 0$,则令:

$$\begin{cases} x'=\dfrac{x+y-1}{2} \\ y'=\dfrac{y-x-1}{2} \end{cases} \tag{2}$$

按照和$|x+y|$为偶数时同样的走法走完$|x'|+|y'|$跳,最后,沿坐XY标系中矢量(0,1)走一跳,这样便可以到达目的结点;若$y<0$,则令:

$$\begin{cases} x'=\dfrac{x+y+1}{2} \\ y'=\dfrac{y-x+1}{2} \end{cases} \tag{3}$$

按照和$|x+y|$为偶数时同样的走法走完$|x'|+|y'|$跳,最后,沿XY坐标系中矢量(0,−1)走一跳,这样便可以到达目的结点。从上面的算法描述,可以看出路由算法1非常简单,便于

实现。

3.2 算法2

虽然算法1非常简单易行，但是仔细观察图3就会发现，按这种路由算法，在直向链路上的链路利用率很低。如果假设在XD网络中信息传输的平均距离为 $\bar{d}$（XD网络的平均距离计算见文[5]），那么，直向链路利用的概率仅为斜向链路上的 $\frac{1}{2\bar{d}}$，直向链路利用率不高，会大大加重斜向链路的负担，造成网络的不稳定。因此，需要对算法1进行改进，这样引入了算法2。算法2和算法1的思路大体相同，只是提高了直向链路上的链路利用率。算法2可以描述如下：

按照和算法1相同的方法完成坐标转换，将信息的源点变换为坐标原点，然后将 X-Y 坐标系按顺时针方向旋转45°，按照公式(1)得到坐标系 X'-Y'。在坐标系 X'-Y' 中，源结点和第二或第四象限的结点通信时，依然按照算法1来完成寻路。如果源结点和第一或第三象限的结点通信时，按以下算法寻路：

如果 $|y'|\geqslant|x'|$ 且 $y'\geqslant 0$，先沿 y' 正方向走 $|y'|-|x'|$ 跳，再沿坐标系 X'-Y' 中矢量 $(\frac{1}{2},\frac{1}{2})$（对应 X-Y 坐标系中的矢量(0,1)）走 $2\times|x'|$ 跳，这样就可以到达目的结点；如果 $y'<0$，则先沿 y' 负方向走 $|y'|-|x'|$ 跳，再沿坐标系 X'-Y' 中矢量 $(-\frac{1}{2},-\frac{1}{2})$（对应 X-Y 坐标系中的矢量(0,−1)）走 $2\times|x'|$ 跳，这样就可以到达目的结点。

如果 $|y'|<|x'|$ 且 $x'\geqslant 0$，先沿 x' 正方向走 $|x'|-|y'|$ 跳，再沿坐标系 X'-Y' 中矢量 $(\frac{1}{2},\frac{1}{2})$（对应 X-Y 坐标系中的矢量(0,1)）方向走 $2\times|y'|$ 跳，这样就可以到达目的结点；如果 $x'<0$，则先沿 x' 负方向走 $|x'|-|y'|$ 跳，再沿坐标系 X'-Y' 中矢量 $(-\frac{1}{2},-\frac{1}{2})$（对应 X-Y 坐标系中的矢量(0,−1)）走 $2\times|y'|$ 跳，这样就可以到达目的结点。

3.3 算法性能分析

下面对上述算法性能进行分析。最短路由算法是一类可以很好利用网络资源的路由算法，下面证明按以上所述路由算法所寻找出来的路径是XD网络的最短路径。

定理1 按照路由算法1所选的路径是XD网络的最短路径。

证明：可以把一个XD网络看作是经过坐标转换和旋转后，在 X'-Y' 坐标系中两个网络的叠加，这两个网络的每一对相应结点间都有链路相连，这两个网络都可以看作 X'-Y' 坐标系中一个2D Torus网络的一部分（如图3(b)所示）。由于在一个2D Torus中维序路由是网络的最短路由[7,8]，在对称的部分2D Torus网络中这个结论也是成立的。所以，如果信息是在这两个网络中的一个网络内部传递时，按照路由算法1肯定是最短路径。如信息是在两个网络结点之间传递，假设由路由算法1决定的路径长度为 d，由于两类网络之间必须再经过一跳才能到达目的结点，所以 $d=|x'|+|y'|+1$，由于 $|x'|+|y'|$ 为最短路径，所以 d 也为最短路径。证毕

下面证明算法2也是XD网络的最短路由算法。

定理2 按路由算法2所选的路径也是XD网络中的最短路径。

证明：当目的结点在坐标系 X'-Y' 的第二或第四象限时，算法2和算法1寻路方法完全相同，由定理1可知，算法2是最短路由算法。当目的结点在坐标系 X'-Y' 的第一或第三象限时，按照算法2源结点到达结点 (x',y') 的路径长度 $d=|x'|+|y'|$，如果该结点是 X-Y 坐标系中 $|x+y|$ 为偶数的结点和算法1的路径长度是相同的，根据定理1，该路径也是最短路径。如果目的结点是 X-Y 坐标系中 $|x+y|$ 为奇数的结点，由于在算法1中，对这类结点的坐标旋转变换是按(2)或(3)式来完成的，这里仅讨论按(2)式转换的情况（按(3)式转换的情况类似）。算法2中目的结点在坐标系 X'-Y' 的第一或第三象限时，坐标旋转变化按(1)式，假设 $y>x>0$（其他情况按类似方法）。按算法1的路由长度为 $d=|x'|+|y'|+1=\frac{x+(y-1)}{2}+\frac{(y-1)-x}{2}+1=y$；

按算法2的路由长度为 $d'=|x'|+|y'|=\frac{x+y}{2}+\frac{y-x}{2}=y$，可见 $d=d'$。按照定理1，在这种情况下该路径也是最短路径。综上所述，算法2完成的寻路是XD网络的最短路径。证毕

算法2虽然比算法1稍微复杂，但是它可以提高网络直向链路的利用率，减少网络斜向链路的负担，对于保证网络的稳定性具有重要意义；从网络性能来讲，算法1和算法2都是衍生于维序路由算法，因此，具有维序路由算法良好的性能。

结论 交换网络拓扑结构的选择是大容量路由器设计中至关重要的一步。XD网络是一种很适合于大容量路由器（T比特以上级）的交换网络拓扑结构，本文针对XD网络的拓扑特点，提出了在其上简单、高效的路由算法一对角矢量映射法（DVM）。该算法包括算法1和算法2两种。算法1是网络最短路由算法，实现简单，便于在进行网络性能分析中使用，但是对直向链路利用率不高；算法2虽然比算法1稍微复杂，但是不仅具有和算法1同样好的网络性能，同时提高了网络直向链路的利用率，分担了网络斜向链路的负担，更利用网络性能的稳定，因此，在实际设计中通常采用算法2。

参考文献

1 Cisco system. Cisco 12016 Gigabit Switch Router, data sheet, 2001. http://www.cisco.com

2 Juniper networks. Juniper M160, White paper. 2000. http://www.juniper.net/

3 Dally W J. Scalable Switching Fabrics for Internet Routers [WhitePaper]. Avici Systems, Inc, 1997

4 TeraPlex -Architecture Overview [WhiterPaper]. Pluris, Inc, 2001

5 杨君刚，邱智亮，刘增基. 一种新型大容量路由器交换网络结构的研究[J]. 西安电子科技大学学报，2003，7：89～95

6 邱智亮，陈震，杨君刚，等. 一种大容量可扩展分组交换网络结构[P]. 中华人民共和国发明专利，ZL 03 114464.0. 2004，9

7 Dally W J. Performance Analysis of k-ary n-cube Interconnection Networks [J]. IEEE Transactions on Computers, 1990, C-39(6): 775～785

8 Sullivian H, Bashkow T R. A large scale, homogeneous, fully distributed parallel machine [A]. In: Proc. 4th Annu. Symp. Comput. Architecture, Mar. 1977. 105～117