自适应路由算法 BNR 互连网络 计算机网络 ⑤

计算机科学 2000Vol. 27№. 12

BNR：最短路径无死锁全自适应路由算法的分析与设计工具*)

BNR: The Tool to Analyse and Design the Minimal Deadlock-Free Fully Adaptive Routing Algorithm

邓 波 杨晓东

（国防科技大学计算机研究所 长沙 410073）

Abstract In a massively parallel processors(MPP)system, a routing algorithm constitutes the primary factor influencing the performance of the interconnect network and MPP system. After analysising the characteristics of message routing in interconnection network, one new concept "the Best Network for Routing"(BNR)is proposed. Using it, we can analyse any minimal deadlock-free fully-adaptive routing algorithm(MDF^2A^2)proposed, and also can design two new MDF^2A^2: VBA and LCFAA. On this point, it gives guidelines to the interconnection network designers.

Keywords Routing vector, Routing direction, Best network for routing minimal deadlock-free fully-adaptive routing

1 引言

大规模并行计算机(MPP)系统性能的发挥极大程度上依赖于互连网络的通信性能[1]。互连网络采用的路由算法决定了消息在网络中如何选取路径，其性能对网络效率的发挥起着重要作用，根据允许选择路径的不同，路由算法有最短路径和非最短路径以及确定性和自适应性之分，自适应又有部分自适应和完全自适应之分。确定性算法唯一确定路径，算法控制和实现相对简单，但其选择的灵活性差，通道利用率低，网络易阻塞，从而导致网络效率不高。自适应路由算法对于一对源和目的结点，视网络的工作状态，可能有多条路径可选，灵活性好，网络的通道利用率高等优点。

随着人们对互连网络高带宽、低延迟的要求越来越迫切，采用自适应路由已经被认为是路由算法设计的必然趋热。大量自适应路由算法已经被提出，但它们或存在自适应性受限、或存在代价较大、或存在灵活性不够等缺点。本文提出"最佳寻径网"BNR(the Best Network for Routing)概念，据此分析了互连网络中最短路径无死锁全自适应路由(MDF^2A^2)算法，并利用BNR设计了两个最短路径无死锁全自适应路由算法VBA和LCFAA，给出了构造最短路径无死锁全自适应路由算法的设计方法，为网络算法设计人员提供指导。

本文提出的BNR概念适用于任何MPP互连网络，并且适用于任何当前采用的切换技术(包括包交换、虚穿透、虫孔路由等等)，故BNR有广泛的应用范围。基于虫孔路由切换技术的直接网络不仅是目前互连网络研究的主要方向之一，而且也是目前MPP系统实现所采用的主要互连网络之一，为叙述简单，后文是以虫孔路由切换技术下的直接无环k-ary n-cube网络为例。

2 "最佳寻径网"BNR

2.1 "寻径向量"R_V

消息在网络中寻径时，需要确定其在网络中的位置。对于无环k-ary n-cube网络来说，一般用$Msg=(m_{n-1},\cdots,m_1,m_0)$来标识寻径过程中消息所在结点的地址，Msg的变化反映了消息寻径的位置变化。由此，可把消息源结点记为$S=(s_{n-1},\cdots,s_1,s_0)$，目的结点记为$D=(d_{n-1},\cdots,d_1,d_0)$，消息在寻径过程中所经过的结点记为$C=(c_{n-1},\cdots,c_1,c_0)$。注意，此时C是可变的，它反映的是整个消息寻径过程中所经结点的地址变化；消息寻径其实就是在路由算法作用下，将消息从源结点经过变化的中间结点寻径最终到达目的结点的过程。

当我们将S、C、D分别看成n元向量$\vec{S}$、$\vec{C}$、$\vec{D}$时，则路由算法即是对源向量$\vec{S}$不断进行向量函数变换形成中间向量$\vec{C}$，最终变换为目的向量$\vec{D}$。

为更深入、准确描述网络的消息寻径行为，我们引

*)本课题得到国家"八六三"高技术和"九五"国防预研基金资助。邓 波 博士生，主要研究方向为高性能计算机体系结构、并行与分布处理及网络路由。杨晓东 教授，博士生导师，主要研究方向为高性能计算机体系结构、分布与并行处理及RAS技术。

入"寻径向量",定义 R_V=($\vec{D}$-$\vec{S}$)为消息在网络中寻径的"寻径向量"。若记 R_V=(r_{n-1},…,r_1,r_0),则 R_V=(r_{n-1},…,r_1,r_0)=(d_{n-1}-s_{n-1},…,d_1-s_1,d_0-s_0)。消息将根据"寻径向量"在网络中选择正确的寻径方向进行寻径,将 Dir(r_i)($0\leqslant i\leqslant n-1$)记为消息在第 i 维上的寻径方向,对于"寻径向量"R_V 的每个分量 r_i,若其值为正,则称消息在第 i 维上正向寻径,记 Dir(r_i)=1;反之其值为负,则称消息在第 i 维上负向寻径,记 Dir(r_i)=-1,若其值为零,则称消息在第 i 维上寻径终止,记 Dir(r_i)=0。对于给定消息的"寻径向量"R_V,它是 n 维空间中的一个向量,故由其确定的所有维上寻径方向是唯一的。

根据"寻径向量"R_V 定义易知"寻径向量"具有如下性质:

(1)由于网络每次只能向相邻结点传递消息,知道消息寻径时,通过路由算法一次变换作用只能修改"寻径向量"R_V 的一个分量 r_i($0\leqslant i\leqslant n-1$)。当网络为无环 k-ary n-cube 时,$r_i$ 被修改成 r_i-1 或 r_i+1;

(2)设计路由算法的根本目的就要使"寻径向量"R_V 通过路由算法作用最终变换成 n 元零向量(0,0,…,0);

(3)"寻径向量"R_V 的分量 r_i($0\leqslant i\leqslant n-1$)绝对值之和 $\sum_{i=0}^{n-1}|r_i|$ 是消息在网络中寻径所需的最少跳步数(hop),也即路由算法将"寻径向量"变换成 n 元零向量的最少变换次数;

(4)"寻径向量"R_V 所对应的寻径方向 Dir(r_i)($0\leqslant i\leqslant n-1$)由 R_V 唯一确定。

2.2 "最佳寻径网"BNR

网络在消息寻径时要达到最快传输的目标,必须要求能充分发挥其传输特性。对于在网络中寻径的消息来说,若"寻径向量"能经过最少变换次数变换成 n 元零向量,则网络就能以最快速度将消息传输至目的结点,根据"寻径向量"性质(3),网络可采用最短路径路由算法来实现"寻径向量"最少次数的零向量变换;若网络能提供给该消息到达目的结点的所有路径,则可使消息在网络中能最大限度地利用网络资源进行"寻径向量"向零向量的变换,故可以采用全自适应路由算法来作为网络充分发挥其最大性能的有效方法,此外,若路由算法使得网络中多个消息寻径相互占有其它消息需要使用的资源不能释放,进而阻塞网络其它多个申请被占资源的消息,造成网络死锁,使网络服务失效。因此在网络中给出的路由算法必须能解决死锁问题[2]。综上可知,设计最短路径无死锁全自适应路由算法(MDF^{2}A^2)成为 MPP 网络服务提供者需要考虑的关键内容之一。

为设计高效、易实现的最短路径无死锁全自适应路由算法,我们引入"最佳寻径网"BNR 概念,根据"最佳寻径网"BNR,从网络的拓扑结构出发,考虑网络采用的流控机制和切换技术,以及不同的应用目的,我们可设计出满足网络服务要求的高效、易实现最短路径无死锁全自适应路由算法,从而对网络服务提供者给出路由算法设计的具体指导。

下面对于一个即将在网络中寻径的消息我们构造一个网络,它是原来网络的一个子网,并且在这个网络中消息能按照上面所描述的以最短路径无死锁全自适应路由在网络中寻径传输。

设集合:

$F=\{(f(i))_{i\times n}\mid(0\leqslant i\leqslant n-1)\cap(f(i)=1\cup f(i)=-1)\}\forall\ \varphi\in F,\varphi=(f(n-1),f(n-2),\cdots,f(i),\cdots,f(0)),(0\leqslant i\leqslant n-1)$且 $f(i)\in\{1,-1\}$为已确定的值,我们在无环 k-ary n-cube 中根据 φ 构造一个子网,记为 $G(\varphi)$。网络 $G(\varphi)$构成原来网络的一个子网,并包含原来网络的所有结点。网络 $G(\varphi)$中每个结点的通道都是单向的,并且在第 i 维上通道方向就是 $G(\varphi)$中第 i 个分量 $f(i)$所示方向($f(i)$所示方向与 Dir(r_i)取值所示方向相同)。

下图即为 n=3 时,对给定 $\varphi=(f_2,f_1,f_0)$,4×3×3 无环网络对应的网络 $G_{(4\times3\times3)}$图示:(其中 $f(i)$表示"寻径向量"在第 i 维上对应的寻径方向)。

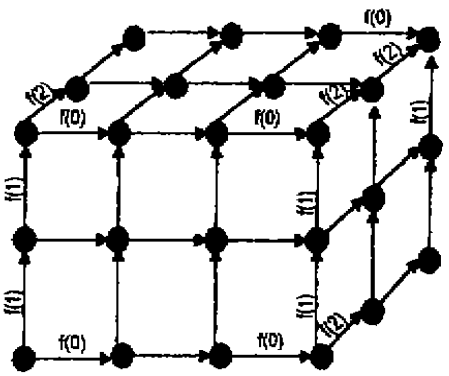

图 1　4×3×3 无环网对应的网络 $G_{(4\times3\times3)}$

从图 $G(\varphi)$的构造方法,我们知所有按照此方法构造的所有子网 $G(\varphi)$的集合其实就是构成了原来 k-ary n-cube 网络,即 $\bigcup_{\forall f\in F}G(f)$构成原来 k-ary n-cube 网络。

根据消息源结点 S 和目的结点 D 所对应的向量 $\vec{S}$、$\vec{D}$ 我们得到"寻径向量"R_V=(r_{n-1},…,r_1,r_0)=($\vec{D}$-$\vec{S}$)=(d_{n-1}-s_{n-1},…,d_1-s_1,d_0-s_0)。由"寻径向量"性质(4),对于一个给定的"寻径向量"R_V,存在唯一确定的寻径方向 Dir(r_i)($0\leqslant i\leqslant n-1$),若将给定 R_V 所确定的所有 Dir($r_i$)组成一个向量 Dir=($Dir'(r_i)$)$_{i\times n}$($0\leqslant i\leqslant n-1$),其中若 Dir($r_i$)=1 或 Dir($r_i$)=0 则 Dir'($r_i$)=1;若 Dir($r_i$)=-1 则 Dir'($r_i$)=-1。由此知 Dir∈F。由于对给定的消息,定义的 Dir 也被其唯一确定,

所以我们可根据任给的消息所确定的 Dir 在 F 中将其分类。任给消息 m，其确定的 Dir 记为 Dir_m，则 $Dir_m \in F$，由于集合 F 共有 2^n 个元素，故在无环 k-ary n-cube 网络中消息种类有且仅有 2^n 种，记 $m \in M_k$，$1 \leqslant k \leqslant 2^n$，$M_k$ 是 2^n 类消息中包含消息 m 的一类消息，易知 M_k 中所有消息确定的 Dir 都等于 Dir_m。

若将消息 m 放入由其确定的 Dir_m 所构造的子网 $G(Dir_m)$ 中寻径，我们知消息 m 在 $G(Dir_m)$ 中将可以按照最短路径无死锁全自适应路由寻径，则 $G(Dir_m)$ 就是要构造的，满足消息 m 按照上面所描述的以最短路径无死锁全自适应路由在网络中寻径传输的网络。

其实不仅对消息 m，而且对于 $m \in M_k$ 的 $M_k(1 \leqslant k \leqslant 2^n)$ 中所有的消息来说，原无环 k-ary n-cube 网络中能提供的所有最短路径全自适应路径都可以在网络 $G(Dir_m)$ 中找到，因此在 $G(Dir_m)$ 中再添加任何原无环 k-ary n-cube 网络中的物理通道对于 M_k 中所有的消息寻径都是冗余的。由于网络 $G(Dir_m)$ 是根据寻径方向进行构造的，若撤消网络 $G(Dir_m)$ 中某一结点的第 i 维物理通道，路由算法在该结点将无法对"寻径向量" R_V 进行第 i 维向量函数变换，使得 M_k 中的消息在该结点无法利用第 i 维通道向目的结点寻径，故在网络 $G(Dir_m)$ 撤消任一物理通道都将破坏 M_k 中的消息在网络 $G(Dir_m)$ 中进行全自适应寻径。又由于对于给定消息 m，其寻径方向已经确定，故消息在网络 $G(Dir_m)$ 中寻径根本不会形成环路，根据无死锁理论[2]，网络 $G(Dir_m)$ 是无死锁的。综上分析，我们知网络 $G(Dir_m)$ 是满足 M_k 中的所有消息进行最短路径无死锁全自适应路由的最小网络。我们就将这样的网络定义为给定消息 m 所属消息种类 $M_k(1 \leqslant k \leqslant 2^n)$ 的"最佳寻径网"BNR。

对于无环 k-ary n-cube 网络中的 BNR，易知 BNR 具有如下特征：

(1)对于给定的消息 m，它所对应的 BNR 是原网络中满足 $m \in M_k$ 的 $M_k(1 \leqslant k \leqslant 2^n)$ 中所有消息进行最短路径无死锁全自适应路由的最小子网；

(2)不同的 BNR 之间至少有一维寻径方向不同，即对于不同的 BNR_1 和 BNR_2，$\exists\ r_i \in R\_V_1$，$r'_i \in R\_V_2$，有 $Dir(r_i) \neq Dir(r'_i)$ $(0 \leqslant i \leqslant n-1)$；

(3)由于无环 k-ary n-cube 网络中消息种类有且仅有 2^n 种，故无环 k-ary n-cube 网络中存在且仅存在 2^n 个不同的 BNR。

3 用 BNR 判断路由算法的路由性质

根据 BNR，可快速判断出网络寻径所提供的路由算法是否是最短路径无死锁全自适应路由算法。要构成无环 k-ary n-cube 网络上最短路径无死锁全自适应路由算法，首先保证算法必须能提供 2^n 个不同的 BNR，使算法可以向消息提供网络所能提供的所有最短寻径路径；但如果算法提供 BNR 却无法保证消息在网络中可以占有网络所提供的最短寻径路径时，该算法仍不是最短路径无死锁全自适应路由算法，这样的算法还必须提供无死锁保证才能使消息在网络中占有网络所提供的最短寻径路径，综上分析，得到如下结论：

结论.1 当无环 k-ary n-cube 网络中一个无死锁的路由算法可以提供 2^n 个不同的 BNR 时，该算法是最短路径无死锁完全自适应路由算法。

根据结论 1 我们可以对目前已经提出的所有自适应路由算法分析其性质，判断其是否是最短路径全自适应路由算法。下面的两个实例根据结论 1 分析了两个典型路由算法的路由性质。

例 1.1 PAR(Planar_Adaptive Routing)算法[3]。由于它在维与维之间只能提供限定方向的寻径方向，使得平面自适应算法作用于无环 k-ary n-cube 网络中形成的 BNR 数小于 2^n，即对于某类消息，路由算法不能提供无环 k-ary n-cube 网络所能提供的所有最短寻径路径，所以平面自适应算法是部分自适应的最短路径无死锁路由算法。

例 1.2 PBFAA 算法[4]。由于其在虚拟子网 VIN1 中提供任意方向的寻径，使得该算法作用于无环 k-ary n-cube 网络中形成的 BNR 数可以达到 2^n，并且由于算法利用虚拟子网 VIN0 来避免死锁，所以 PBFAA 算法是最短路径无死锁全自适应路由算法。

除可以利用结论 1 判断路由算法的路由性质外，还可以利用结论 1 来构造新的最短路径无死锁全自适应路由 MDF^2A^2 算法。

4 用 BNR 设计 MDF^2A^2 算法

用 BNR 设计 MDF^2A^2 算法的关键在于所设计的算法中要提供消息寻径时对应的 BNR 网络寻径功能，使得消息按照所给的路由算法能在网络中按照各自对应的 BNR 进行寻径，从而保证消息能在网络中按照最短路径无死锁自适应路由进行寻径传输。下面提供的两个算法 VBA、LCFAA 均满足了这个要求。

4.1 基于虚拟 BNR 网的路由算法 VBA

VBA 算法的设计出发点即在网络中构造 2^n 个虚拟子网，使这些子网同 2^n 个 BNR 对应，从而保证了各类消息能在各自的 BNR 中按照最短路径无死锁全自适应路由寻径。VBA 算法描述如下：

1)算法在结点的每一物理通道上均设 2^{n-1} 条虚通道，分别用序号 $1,2,\cdots,2^{n-1}$ 标记，用 $VC_{dimension\ label\ direction}$ 来表示每个虚通道，其中 dimension 表示该虚通道沿哪一维传递消息；label 表示虚通道的序号；direction 可以为 +1(表示消息将沿正向传递)或 −1(表示消息将沿着负向传递)。

2)将网络划分成 2^n 个相互独立的虚拟子网，每个虚拟子网络都必须从每个结点的每维上选择一条虚通道，如此形成的虚拟子网即为 VBNR。

3)当一个消息出现在网络中时，根据它的"寻径向量"$R\_V=(r_{n-1},\cdots,r_1,r_0)$，得到其寻径方向 $Dir(r_i)$，从而确定了消息在网络中进行最短路径无死锁完全自适应时需要使用的虚通道，将消息放入由这些虚通道构成的 VBNR 中寻径。

根据结论 1，我们知 VBA 算法是 MDF^2A^2 算法。

4.2 利用 VNOF(虚网叠加)设计 MDF^2A^2 算法

如果不考虑 VBA 算法的实现代价，该算法将是满足 MDF^2A^2 的高效寻径算法，它能使消息在网络中利用 VBNR 提供的最大带宽进行传输。VBA 算法对每个 VBNR 的网络利用率将是最高的。但是，由于每个物理通道必须提供 2^{n-1} 个虚通道，对于无环 k-ary n-cube 网络，每个结点将必须提供 $2n\times 2^{n-1}$ 个虚通道。由于技术限制，当 n 较大时网络实现代价将是惊人的，故 VBA 算法在高维时不是一个满足 MDF^2A^2 的低代价、易实现路由算法。综合考虑算法的高效和易实现性，根据结论 1，结合已经提出的 VNOF(虚网叠加)[4] 框架，可以构造出一类易实现的高效 MDF^2A^2 算法。VNOF 的基本思想描述如下：

1)算法为每条物理通道设置 m 条虚通道，将网络分成 r 个虚拟子网，设为 VIN1，VIN2，…，VINr，r 个虚拟子网中的相应虚通道分时共享同一物理通道，不同虚拟子网使用每条物理通道不同序号的虚通道。

2)选定一个采用无死锁最短路径路由策略的 VIN1 子网，在该子网基础上叠加一或多个虚拟子网，在这些叠加的虚拟子网中根据需要可选择自适应的最短路径路由策略或其它路由策略。

我们构造 MDF^2A^2 基本方法如下：

方法 1 将给定的物理网络分为两类虚拟网络 V1 和 V2，V1 中的所有虚拟网络集合能提供 2^n 个不同的 BNR，V2 中的虚拟网络在设计时属于可选网络配置，当 V1 中的网络无法实现无死锁路由时，算法必须提供 V2 中的虚拟网络来实现无死锁路由。

根据结论 1 和 VNOF 构造框架性质，可知方法 1 所设计的路由算法是 MDF^2A^2 算法。

下面给出一个根据方法 1 设计的 MDF^2A^2 算法方案：LCFAA 算法。LCFAA 算法是迄今为止所见到的所需虚通道数最少的 MDF^2A^2 算法，对于无环 k-ary n-cube 网络，每个路由器仅需 3n-1 个虚通道。LCFAA 算法描述如下：

1)算法在结点的每一物理通道上均设一序号为 0 的虚通道，在第 1 至第 n−1 维的负向物理通道上增设一序号为 1 的虚通道。用 $VC_{dimension,label,direction}$ 来表示，其中 *dimension*、*label*、*direction* 的含义同前述。

2)将网络划分成两个虚拟子网：VIN0 和 VIN1。在 VIN0 中使用每条物理通道序号为 0 的虚通道；在 VIN1 中使用每条物理通道序号为 1 的虚通道。

3)在 VIN0 和 VIN1 中，均按最短路径完全自适应路径策略选择趋于目的结点的虚通道路由消息。

4)当一条消息的头到达某一结点时，如是目的结点，则消息被该结点接收，否则，执行下列操作：

a)若消息在前面传送中已使用过 VIN1 中虚通道，则：若 VIN1 中有可用虚通道，则将消息传向邻近结点；若 VIN1 中没有可用虚通道，则等待 VIN1 中有虚通道变为可用。

b)若消息在前面传送中未使用过 VIN1 中虚通道，则：若 VIN0 中有可用虚通道，则将消息传向邻近结点；若 VIN0 中没有可用虚通道，则：若 0 维上路由标志不等于 0 或存在路由标志大于 0 的维，则等待直至 VIN0 中有虚通道变为可用，再将消息传向邻近结点；若 0 维上路由标志等于 0 且所有其它维上路由标志均小于或等于 0，则：若 VIN1 中有可用虚通道，在 VIN1 中将消息传向邻近结点；若 VIN1 中没有可用虚通道，则等待 VIN1 中有虚通道变为可用。

由 LCFAA 算法性质，我们知 $VIN0\in V1$，$VIN1\in V2$。LCFAA 是一个低代价、易实现的高效 MDF^2A^2 算法，基于 LCFAA 算法我们设计了一个低代价完全自适应路由器[5]。

结语 本文提出的"最佳寻径网"BNR 概念适用于任何 MPP 互连网络，并且适用于任何当前采用的切换技术(包括包交换、虚穿透、虫孔路由等等)，故 BNR 有广泛的应用范围。本文以虫孔路由切换技术下的直接网络为例描述 BNR 概念，若要对其它网络或切换技术使用 BNR 概念分析网络路由算法，只需将直接网络替换为对应网络的描述并在所给的切换技术下进行相应修改即可。现在正分析 BNR 在其它网络(如各种静态、动态互连网络等)、其它切换技术(如包交换，虚穿透等)下的性质，最终达到能分析、设计 MPP 任意网络中最短路径无死锁全自适应路由算法的目标。

参 考 文 献

1 Duato J, Yalamanchili S, Ni L. Interconnnection networds·an engineering approach. IEEE Computer Society Press, 1997

2 Dally W J, Seitz C L. Deadlock-Free Message Routing in Multiprocessor Interconnection Network. IEEE Trans On Computers, 1987, 5(36): 258～264

3 Chien A A, Kim J H. Planar-adaptive routing·Low-cost adaptive networks for multiprocessors. In: Proc 19th Annual international symposium on Computer Architeches. 1992. 268～277

4 刘燕，杨晓东，等. 虚网叠加一构造自适应路由算法的有效框架，计算机研究与发展，1999(4)

5 刘燕，杨晓东，等. 一个低代价的完全自适应路由器设计. 电子学报，1998(11)