计算机科学1998 Vol. 25 №. 3

MPP 系统中自适应路由算法的分类及分析*)

Classification and Analysis of Adaptive Routing Algorithms in MPP Systems

刘 燕 王春艳 杨晓东

(国防科技大学计算机系 长沙410073)

摘 要 Interconnect network is an important part of massively parallel processors (MPP), and routing algorithm constitutes the primary factor influencing on the performance of it. In this paper, we discuss the routing algorithms for direct networks, and study the wormhole-routed adaptive algorithms in detail. Finally, we give some available points to design and evaluate new algorithms.

关键词 Massively parallel processors (MPP), Direct networks, Adaptive routing algorithms, Virtual channel

一个互连网络通常由拓扑结构、开关技术、流量控制和路由算法四方面来表征，其中路由算法之效率对网络性能起着很关键的作用。路由算法可分为确定性和自适应路由算法两种，确定性算法实现简单，已在很多商用 MPP 中得以实现；而自适应路由算法正成为新一代 MPP 系统的采用对象，如 Cray T-3E 系统采用了部分自适应路由算法。本文在直接网络的基础上，对自适应路由算法进行一个系统的分类，然后针对每类算法进行分析，并着重对采取虫孔路由开关技术的自适应路由算法进行研究和分析。

1 自适应路由算法及分类

自适应路由算法与预先唯一确定路径、不受网络状态影响的确定性路由算法(如 E-cube)不同，它对于一对源和目的结点，视网络的工作状态，可有多条选取的路径，因而有避免死锁、提高网络的带宽利用率和网络容错能力的好处；另外，自 Dally 提出虚通道概念[1]—多个逻辑通道分时共享同一物理通道—之后，采用虚通道的自适应路由算法实现起来更为经济和灵活，从而使自适应路由算法得到了极大的发展。一个好的路由算法应有三个特点：低通讯延迟、高网络吞吐率和 VLSI 工艺上的易实现性，近年来很多算法又将容错作为一项重要的指标，旨在寻求自适应性、性能和容错之间的最佳平衡。

下面给出自适应路由算法的一种分类(如图1所示)。在分类第一层，算法分为虫孔路由型和非虫孔路由型，其中前者是指在虫孔路由开关技术下采用的算法，而后者是在其它路由开关技术(如存储转发、线路交换、虚跨步、流水线路交换 PCS[8]等)下采用的路由算法。

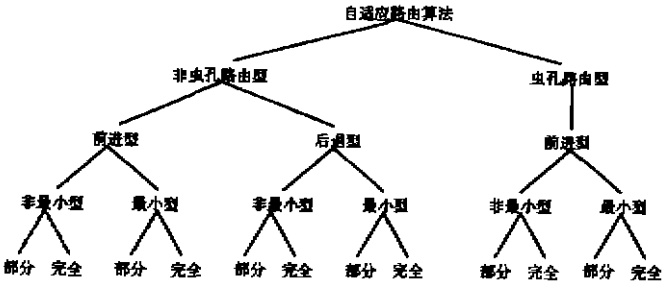

图1 自适应路由算法的一个分类

在分类第二层，算法分为前进型和后退型，前者指"向前"传送信息、无后退回溯能力，后者指算法系统搜索网络、需要时可后退回溯。虫孔路由型由于受其流水特性限制无法回溯，全部为前进型。

在分类的第三层，算法分为最小路由算法(路由总是选取从源到目的结点间最短的路径)和非最小路由算法(路由针对网络当前状态，允许信息沿一长路径进行传送)。

*)国家863高技术基金资助课题和九·五国防预研基金资助课题。刘燕 博士生，研究方向为高性能计算机体系结构，分布与并行处理。王春艳 硕士生，研究方向为 MPP 系统、互连通讯。杨晓东 教授，博士导师，长期从事巨型机和 MPP 系统理论研究。

在分类的最底层，算法分为完全自适应的和部分自适应的路由算法。前者可选取所有路径；而后者由于在路由上加以限制，仅有部分路径可选。

2 非虫孔路由型算法

在非虫孔路由算法中，前进型与后退型路由算法的区别在于当中间结点没有可选的通道空闲时采取的不同措施。前进型会等待、放弃或绕道到其它路径上去，而后退型本着"寻找其它路径比等待一条路径变为可用要好得多"的原则，会释放一些通道并后退回到先前的某结点处再在网络的其它部分寻找路径。在前进型的算法中，较为典型的有采用线路交换的基于已走 hop 数的算法、AI 算法、Idle 算法等[10]。该类算法类似于后面所述虫孔路由型算法，故在此不作讨论。在后退型算法中，有代表性的有采用最小路径进行深度优先搜索路由的 EPB 算法、采用最小和非最小路径进行深度优先搜索路由的 EMB 算法、将二者结合起来的 TPB-u 算法和采用 PCS 开关技术且路径中仅允许有小于等于 m 个非最小通道的 MB-m 算法等[9,10]。后退算法容错性能较好，但寻径头、延迟及附加成本都相对较大，且需解决活锁问题，复杂性较高，因此需在容错、性能和成本之间进行充分考虑后决定是否采用这种算法。这类算法一般适用于对容错要求较高的系统中。表1给出了对 k 元 n 方体网络，几种后退型路由算法相对于 E-cube 算法的寻径头大小比较。

表1 几种算法的寻径头比较

算法	类 别	寻径头大小
E-cube	确定性算法	$n\log_2 k$
EPB	完全自适应的最小后退算法	$n\log_2 k+1$
EMB	完全自适应的非最小后退算法	$n\log_2 k+1$
TPB-u	部分自适应的非最小后退算法	$n\log_2 k+1$
MB-m	部分自适应的非最小后退算法	$n\log_2 k+n\log_2 n+1$

3 虫孔路由型算法

虫孔路由(wormhole routing)是当今 MPP 系统中普遍采用的切换技术，在源结点处将要传送的消息报文划分成多个微片(flit)，消息头 flit 带路由信息，当头 flit 所需某通道空闲时，头 flit 经其向前传送，该通道被消息报文占用，后续数据 flit 以流水方式尾随头 flit 经其向前传送，直到尾 flit 经其传送后释放该通道；当头 flit 所需某通道被占用而受阻时，后续 flit 也被阻塞，存储在路径中各相应路由器的缓冲器中。由于虫孔路由开关技术的广泛采用，虫孔路由型算法是目前自适应路由算法中研究和采用得最多的一类，在直接网络中尤其如此，最有代表性的为 Linder 和 Harden 提出的2Pn 算法[1]，Glass 和 Li 提出的基于拐弯模型(Turn model)的算法[2]，Chien 和 Kim 提出的平面自适应路由算法[5]，Berman 和 Gravano 提出的 *-channel 算法[6]，Duato 提出的完全自适应路由算法[12]等。

3.1 部分自适应路由算法

部分自适应路由算法是实际实现中采用较多的自适应路由算法，它在硬件成本与自适应性间进行折衷，将自适应性限制在某部分，从而减少了避免死锁所需增加的硬件。

首先以静态限制路由自适应性的基于拐弯模型的算法[2]为例。拐弯模型不是基于增加物理或虚通道，而是基于分析网络中信息可转弯的方向和转弯所形成的环路，通过强制算法在路由消息过程中不出现某些转弯而阻止网络环路的出现，从而达到避免死锁的目的。以其中自适应性最好的负优先算法为例，在2维 mesh 结构中，设定西和南为负向、东和北为正向，该算法禁止从正向到负向的拐弯(即禁止北到西、东到南两个转弯)。当从源结点向目的结点传送信息时，算法要求首先在西和南方向上自适应地路由(如有必要)，再在东和北两个方向上自适应地路由，直至传向目的结点。该类算法是部分自适应的非最小路由算法。

基于拐弯模型的算法的突出优点是对硬件逻辑要求简单、成本较低，无需增加虚通道即可达到无死锁和部分自适应性。负优先算法对于某些非均匀流量模式性能很好，如对矩阵转置其性能优于确定性算法2倍。但该类算法由于偏重于某些通道，打破了流量的均衡性，易使网络过早进入饱和状态，如对于均匀流量模式其性能甚至还不如确定性路由算法。因此该类算法仅对某些非均匀流量模式的应用性能较好。为使流量均衡，很多研究者又在对该类算法进行改进，提高其在其它流量模式下的性能[4]。

平面自适应路由算法[5]代表了将路由的自适应性动态地限制在部分维上的部分自适应路由算法。以 n 维 mesh 结构上的算法为例，算法中为每个物理通道设3条虚通道，将整个网络分成 $A_0A_1-A_{n-1}$ 共 n 个自适应平面，其中每个平面 A_i 仅包含 d_i 和 d_{i+1} 两维。当从源向目的结点路由信息时，算法按从 A_0 到 A_{n-1} 的顺序进行路由，在每个平面 A_i 中，选择 d_i 和 d_{i+1} 维中任意趋近目标的通道进行自适应路由，直至某中间结点 d_i 维上的坐标等于目的结点在该维的坐标。在路由经过所有自适应平面后，信息到达其目的结点。该算法中任何时刻自适应性均限制在两维，从而达到了无死锁和降低网络成本的目的。这类算法仅需固定数目的虚通道，所需硬件成本和复杂性较低，且在流量分配上较拐弯模型均匀，对多数流量模式实际性能均较高，所以适用性较强，但其仍存在自

适应性受限的缺点，采用时尚需针对具体应用进行相应的改进。

3.2 完全自适应路由算法

3.2.1 *虚通道数大的完全自适应路由算法*。这类算法通过对每个物理通道采用大量的虚通道而达到无死锁和完全自适应的双重目的。以 n 维 mesh 结构上的2Pn 算法[1]为例。算法中为每个物理通道设 2^{n-1}条虚通道，每条物理通道的虚通道数可用一个 n 位二进制数来表示。当从源 $s=s_{n-1}s_{n-2}\cdots s_1s_0$向目的结点 $d=d_{n-1}d_{n-2}\cdots d_1d_0$传送信息时，算法为信息设置一个如下的标志 $t=t_{n-1}t_{n-2}\cdots t_1t_0$：

$$t_i=\begin{cases}1 & \text{if } s_i<d_i\\ 0 & \text{if } s_i>d_i\\ 0 \text{ or } 1 & \text{if } s_i=d_i\end{cases}$$

当消息从源结点前进了 i 步到达中间结点，则其占用该中间结点的序号为 t 的虚通道，将信息传送向目的结点。该算法是完全自适应的最小路由算法。

2Pn 类算法的突出优点是自适应度高，但所需的大量虚通道数使硬件增加，尤其当网络维数较大时，成本很大；其次，由于该算法每步都基于当前结点可用的局部信息来路由信息，故所选路径并不一定是全局最佳的，如在均匀和热点(hotspot)流量模式下，其性能甚至还不如 E-cube 算法。该类算法仅适用于低维小规模网络上某些流量模式非均匀的应用。有算法对其进行改进，如每步基于某类优先信息(如已走步数)来路由信息，使性能有所提高[7]。但巨大的虚通道需求量促使更多的研究者转入开发成本较低的自适应路由算法。

3.2.2 *虚通道数较少的完全自适应路由算法*。为以较少虚通道数达到完全自适应性，很多研究者对死锁的充分必要条件进行研究，其中最有代表性的为 Duato 提出的无死锁的充分必要条件和"通道依赖图是动态非循环的，则可避免死锁"的无死锁理论[11]，在这一新的理论指导下，人们又提出了大量新的算法。Berman 和 Gravano 等提出的'-channel 算法[6]是其中最为典型的一个，以 n 维 mesh 结构上算法为例，算法中为每个物理通道设置3条虚通道，并将虚通道分为带 * 号和不带 * 的。并约定：在带 * 号的通道上执行维序的确定性路由，在不带 * 号的通道上执行维序路由所不允许的路由。当从源向目的结点路由信息时，在中间结点处可选择任一空闲的不带 * 号通道，而仅可选择符合维序关系的空闲的带 * 号的通道，由此将信息传向目的结点。不带 * 号的通道提供了自适应性，带 * 号的通道避免了死锁，该算法为完全自适应的最小路由算法。

'-channel 代表了需要虚通道数较少的完全自适应路由算法，它在均匀和某些非均匀的(如位反)流量模式下性能都较好，尤其在均匀模式下其性能远优于确定性算法，这在自适应路由算法中是不多见的。但其开关成本和复杂性相对较高。

表2给出了 mesh 结构上几个典型虫孔路由算法每个路由器上所需虚通道数的一个比较。

表2

拓扑	确定性	部分自适应		完全自适应			
	维序路由	负优先[2]	平面自适应[5]	2Pn[1]	*-channel[6]	Opt-y[8]	Duato[12]
2Dmesh	4	4	6	6	8	6	8
3Dmesh	6	6	12	16	12	10	12
4Dmesh	8	8	18	40	16	14	16
nDmesh	2n	2n	6n−6	$(n+1)2^{n-1}$	4n	4n−2	4n

小结 通过对大量已有算法和资料进行分析，我们认为：1)虚通道和限制部分路由技术是避免死锁的有效方法；2)互连网络中最昂贵的资源为物理通道所需线路，其次为缓冲、开关和其它控制电路，虚通道虽不似物理通道那么昂贵，但其所必须的缓冲队列及相关的控制逻辑仍不可避免地带来额外成本，因而其数目选择上应受成本和复杂性限制，不宜过多；3)算法的选用和设计与具体应用密切相关(如对容错要求较高的应用，选择后退型算法较好，而对延迟要求较高的应用，选择虫孔路由型算法较好)；4)自适应性并非越高越好，这与具体应用的流量模式和算法的流量分配均衡性有关，对某些流量模式的应用，自适应性较高的算法性能并不一定优于自适应性较低的算法性能，有些自适应路由算法的性能甚至低于确定性路由算法，算法的性能与具体的消息流量模式有很大关系，自适应算法设计不当可能会导致硬件成本较高且性能较差。因此在为一个系统设计路由算法时，应充分考虑硬件成本和实际应用，选择设计性能价格比高、自适应性满足需要、实现尽量简单的无死锁路由算法，在成本、性能和实现复杂性上进行合理的折衷。

参 考 文 献

[1]D. H. Linder and J. C. Harden, An adaptive and fault-tolerant wormhole routing strategy for k-ary n-cubes, IEEE Trans. Computers, Jan. (40)1991

[2]C. J. Glass and L. M. Ni, The turn model for adaptive routing, in Proc. 19th Int. Symp. Comput. Arch, May 1992

超媒体系统 语义网络 信息语义 关键词(2)

计算机科学1998 Vol.25 №.3

一种基于语义网络的开放式超媒体系统结构

48-51

A Semantic Network Based Open Hypermedia System Architecture

TP391

李光亚 周学海 龚育昌 赵振西

(中国科学技术大学计算机系 合肥230027)

摘 要 In this paper, an open hypermedia system architecture used semantic networks as the underlying link storage mechanisms. The construct of content register table was proposed to support semantic-based hyperlinking to various kinds of information units. Semantic constraints were classified to describe the domain knowledge effectively. At last future work was suggested.

关键词 Semantic network, Content register table, Semantic constraint, Linking protocol, Open hypermedia system

现在的开放式超媒体系统,绝大多数只能有效地利用文本及图象两种媒体,未能充分利用文本、图形、声音、动画及视频等媒体来展示信息,因此需要提出一种有效的系统结构来支持各种媒体信息的集成。语义网络是由概念(结点)以及概念之间的语义关系(链)所组成的一种有向图形式的知识表示模型[8],它与超文本之间的类同关系在文[10]中就已被提出。虽然现在大多数超媒体系统都有一潜在的语义网络,但未能充分利用语义网络本身所固有的语义推理能力、应用领域的描述能力及其灵活的查询能力。对于开放式超媒体系统[6],语义网络更有其独特的优点:开放系统中潜在的应用领域多种多样,所集成的各应用系统也各不相同,这就需要一种机制来支持面向信息语义的链接,而不是某一特定应用的特定文档,该机制应该透明地支持不同应用文档的集成,而不必将文档数据从一种格式转化为另一种格式。真正的开放式超媒体系统它应该满足以下需求:

R1:它支持面向信息语义的链接,即可以链接到文档内容的任意元素,而不管具体文档的格式,只要

李光亚 博士生,主要研究超媒体系统和多媒体数据库。周学海 博士生,主要研究多媒体数据库和面向对象程序设计。
龚育昌 教授,主要研究数据库系统,CAI。赵振西 博士导师,主要研究数据库系统,面向对象和软件开发环境。

[3]C. J. Glass and L. M. Ni, Fault-tolerant wormhole routing in meshes without virtual channels, IEEE Trans. parallel and Distributed Systems, 7(6)1996

[4]J. H. Upadhyay et al., Efficient and balanced adaptive routing in two-dimensional meshes, prasant@iastate.edu

[5]A. A. Chien and J. H. Kim, Planar-adaptive routing: low cost adaptive networks for multiprocessors, in Proc. 19th Int. Symp. Comput. Arch, 1992

[6]P. E. Berman et al., Adaptive deadlock-and livelock-freerouting with all minimal paths in torus networks, in Proc. 4th Symp. on Parallel Algorithms and Architeches, 1992

[7]R. V. Boppana and S. Chalassai, A framework for designing deadlock-free wormhole routing algorithms, Same to [3], 7(2)1996

[8]L. Schwiebert, D. N. Jayasimha, Optimal fully adaptive minimal wormhole routing for meshes, J. of parallel and distributed computing, (27)1995

[9]P. T. Gaughan and S. Yalamanchili, A family of fault-tolerant routing protocols for direct multiprocessor networks, Same to [3], 6(5)1995

[10]P. T. Gaughan, Adaptive routing protocols for hypercube interconnection networks, IEEE Comput. Mag., (26)1993

[11]Jose Duato, A necessary and sufficient condition for deadlock-free adaptive routing in wormhole networks, Same to [3], May 1993

[12]Jose Duato and Pedro Lopez, Highly adaptive wormhole routing algorithms for n-dimensional torus, DIMACS serials in Discrete Mathematics and Theoretical Computer Science, (21), 1995

[13]W. J. Dally, Virtual-channel flow control, same to [3], 3(2)1992