死锁恢复的多维交换结构容错路由算法

许 都，宋 雷，王 宏

(电子科技大学宽带光纤传输与通信网技术教育部重点实验室 成都 610054)

【摘要】在高性能路由器中采用多维交换结构是解决可扩展性的一种方法。在实现这种交换结构时，随着节点数目的增多，交换结构出现故障的概率也随之增加。该文在mesh/torus结构上提出了一种基于死锁恢复策略的容错路由算法MMAR。基于各非故障节点周围链路的状态，MMAR能容错任意形状的故障模型且所需虚拟通道数少。通过在凹形区域表面节点中设置该凹形区域内节点位置信息表，该算法能避免消息进入与其目的节点无关的凹形区域以使绕道路径最短。该文给出了在256个节点的二维torus中的仿真结果，验证了算法的有效性。

关 键 词 死锁恢复； 故障模型； 容错路由算法； 多维交换结构

中图分类号 TP302.8 文献标识码 A

Deadlock Recovery-Based Fault Tolerant Routing Algorithm for Multi-Dimensional Switching Fabric

XU Du, SONG Lei, and WANG Hong

(Key Laboratory of Ministry of Education for Broadband Optical Fiber Transmission and Communication Networks, University of Electronics Science and Technology of China Chengdu 610054)

Abstract Scalable switching fabrics can be used to implement high performance routers by employing multi-dimensional switching fabrics. But the fault probability of switching fabric also increases with the increase of components. A novel fault-tolerant algorithm on the mesh/torus, named as minimal misrouted adaptive routing (MMAR), is proposed based on deadlock recovery mechanism. According to the status of links around each fault-free node, MMAR can accommodate arbitrary shaped fault models using minimal number of virtual channels. When encountering concave fault models, through setting the position tables for the nodes within the concave regions at the surface nodes of each concave region, MMAR can minimize the length of the misrouted paths by avoiding routing the message into the irrespective holes. Performance results of a simulation on torus with 256 nodes are also given.

Key words deadlock recovery; fault model; fault-tolerant routing algorithm; multi-dimensional switching fabric

分布式体系结构的多维交换结构具有控制简单、便于实现分布式管理、能并行处理消息等优势，逐渐受到了高性能路由器设计者的青睐，其中mesh和torus两种最为典型。多维交换结构的中心问题是路由算法、切换技术和流控技术[2]。虫孔路由(wormhole routing)[2]是多维交换结构中普遍采用的切换技术。该技术在源节点处将要传送的消息报文划分成多个微片(flit)。消息头微片(header flit, HFL)携带路由信息先行，后续数据微片以流水方式尾随其向前传送。虚通道流控[2]是多维交换结构中普遍采用的流控方式，当其与一些特殊的仲裁策略相结合时能有效提高网络的性能。

随着多维交换结构规模的扩大，系统出现多个节点故障或节点之间链路故障的可能性也随之增加。因此，设计较好的容错路由策略使得系统在发生故障的情况下仍能实现有效的路由具有重要意义。目前对多维交换网络中容错路由机制进行了大量研究，提出了很多算法[6-8]。但它们或有影响了正常网络的性能、或有容错的故障模型单一、或有所需虚拟通道数较多、或有需使多个非故障节点丧失(disable)功能[2]等缺点。基于死锁[2]恢复机制与真完全自适应路由(true fully adaptive routing)算法[5]，本文提出了一种自适应强的容错路由算法 MMAR。基于对各节点周围链路状态的考查，消息在未遇到故

收稿日期：2007－05－29；修回日期：2008－02－28
基金项目：国家自然科学基金(60372011)
作者简介：许 都(1968－)，男，副教授，主要从事通信网络理论及应用方面的研究.

障时将按照 TFAR 路由规则进行传送；一旦遇到故障将采取适当自适应路由策略绕过当前故障区域完成路由。本文通过在凹形区域的表面节点设置关于凹形区域内节点位置信息表，以防止目的节点不在凹形内的消息路由进入当前的凹形区域，以缩短绕道路径长度。MMAR 算法所需虚拟通道数少，能容错任意形状的故障模型, 且不需使任何非故障节点丧失其功能。

1　故障模型及相关技术

1.1　故障模型

故障模型是故障节点和故障链路的集合。它是容错路由算法的基础，其特点为：(1) 模型边界上仅有非故障节点和链路；(2) 内部仅有故障节点或链路；(3) 每个故障仅且必属于一个故障模型。

在形成故障模型时，每个非故障节点通过合适的检测机制周期性地检测自身和与其相连的各条链路的状态，并发送状态信号给邻居节点。当一个节点发生故障，它的邻居节点通过与该故障节点相连的链路可推断出，所以链路故障与节点故障在本质上是一致的。上述过程经反复迭代至没有新的节点或链路被设为故障后，连接环绕故障区域的无故障部件形成故障环(fault ring，f-ring)，便可得到各种形状的故障模型。如图1所示，该网络存在两个故障模型，其中一个为矩形 M_1，另一个为多边形 M_2(包含了凹形和凸形等区域)。

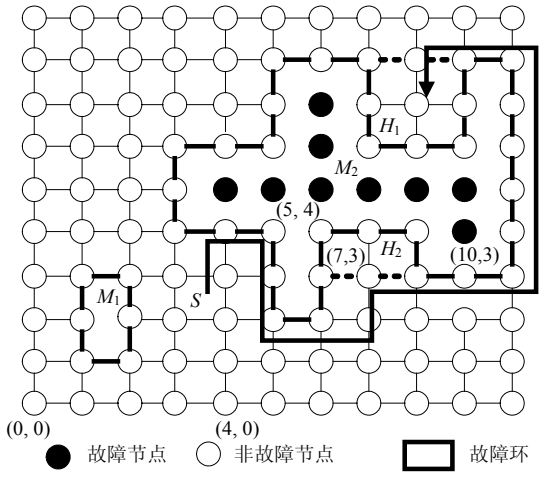

图1　故障模型示意图(32×32 torus的部分结构)

1.2　基本定义

定义 1　在 n 维的mesh/torus网络中(k_i 为第 i 维节点数)每个节点都有其唯一的地址表示：$A(\delta_0,\delta_1,\cdots,\delta_i,\cdots,\delta_{n-1})(0\leqslant\delta_i\leqslant k_i-1)$。在torus网络中，源节点与目的节点的距离向量可以表示为 $R(\Delta\delta_0,\Delta\delta_1,\cdots,\Delta\delta_{n-1})$且在每一维上相对距离 $\Delta\delta_i$ (或正或负)将满足：

$$\lfloor(k_i-1)/2\rfloor\leqslant|\Delta\delta_i|\leqslant\lceil(k_i-1)/2\rceil$$

定义 2　文献[1]将故障模型的凹形区域定义为孔。在二维mesh/torus结构中，孔通过分布式算法得以形成。图1中，H_1 表示一个孔，其中(7, 6)、(8, 6)、(9, 6)为孔底部节点，(7, 8)、(8, 8)、(9, 8)为孔的表面节点，且孔内所有节点均为非故障节点，它们能继续产生、接收、转发数据包。

定义 3　在二维mesh/torus网络中，孔表面节点将被连起来。被连部分与部分原故障环(即不在孔内的部分)连起来构成新的故障环，称为虚故障环(vf-ring)。

1.3　TFAR与死锁恢复机制

TFAR具有最短路径和最强自适应性，其选择任意趋近于目的节点的链路为消息的输出通道，并在已选定的物理链路中任意选择虚拟通道来传送该消息。当网络负载低于饱和点时，TFAR带来极低的死锁概率[2]；当负载超过饱和点时，其带来的死锁概率也较低。

多维交换结构中，采用虫孔方式的切换技术由于各消息相互竞争容易导致死锁。文献[4]指出，当所采用的路由算法自适应性较强时，网络中发生死锁的概率会很低。在低死锁概率的网络中，采用死锁恢复技术可充分提高资源利用率，从而提高整个结构的吞吐率。在死锁恢复机制中，本文采取了FC3D技术[10]来检测死锁。它是一种分布式死锁检测机制，先找到几个阻塞分组所构成的依赖树(tree)。如果根节点阻塞而且其请求的资源为阻塞树中的一个，则出现死锁。死锁恢复技术采用流行的前进型死锁恢复方式DISHA[9]，该技术采用简单硬件实现机制就能从被检测到的死锁中恢复出来。每个节点都配备了一个额外的中央缓冲器或死锁缓冲器。在系统范围内，这些缓冲器集中在一起形成一条无死锁恢复通路，该通路可以看作是一个节点所有物理维共享的浮动虚通道。如果发现死锁，消息就交换到无死锁通路，沿着通往目的节点的路径自适应路由。

2　MMAR算法

2.1　支持容错的一些特殊节点功能

本文以torus网络为基础介绍MMAR算法。为了使MMAR能工作，网络节点需作如下准备工作：

(1) 全网中每个非故障节点检测其周围的链路

状态(是否为故障)，并建立相应链路状态表(link status table，LST)。图1中，在 S 节点建立的LST包含4项(即 $X+$、$Y+$、$X-$、$Y-$方向的链路各为一项)，各表项分别表示其周围链路的状态。显然S节点LST的4个表项均为非故障；而在节点(5,4)，经检测发现 $X+$、$Y+$方向的链路为故障，因此其LST表项将标明 $X+$、$Y+$方向的链路故障，其余两条链路正常。

(2) 通过合适的分布式算法构造故障环。当有凹形故障区域存在时，构建孔标记为 H_i。图1网络中存在 H_1 和 H_2 两个孔，每个表面节点根据构建孔时所得到的孔内部所有节点的位置信息建立位置表(position table，PT)。 在图1中表面节点(7,3)的PT包含 6项，各项分别记录节点(6,3)、(7,3)、(8,3)、(6,4)、(7,4)、(8,4)的位置信息。与节点(7,3)同在一个孔的表面节点(6,3)、(8,3)将记录与其相同的位置信息。

(3) 各孔的表面节点在原LST的基础上构建两个链路状态表：LST1和LST2，其中LST1与LST保持一致；在LST2中，孔内并沿孔的反方向的链路被强制设为故障链路。这样消息可以根据其目的节点与当前孔相对位置关系选择适当的链路状态表以查询各链路状态。图1中节点(7,3)的LST1与原LST一致即4个表项均为链路正常；而LST2中的 $Y+$ 方向链路被强制设为故障。

(4) 将各孔表面节点连接后与不在孔内的原故障环部分连起来形成虚故障环(图1中用虚线表示)。

2.2 MMAR算法流程

算法中，n_s、n_d、n_c 分别表示为源节点、目的节点和当前节点。当消息开始注入网络时被标记为正常消息(normal)。未遇到故障(即当前节点不在故障环或虚故障环上)时，MMAR的行为与TFAR一致。消息任意选择趋近其目的节点的链路作为输出通道且在已选定的输出链路中任意选择一条虚拟通道进行路由。当遇到故障时，消息将按以下步骤路由离开当前节点：

1) 计算 n_c 到 n_d 的相对距离 R。根据 R 值，如果消息到达 n_d，则将该消息送出交换结构，路由过程结束；如果没有到达 n_d，路由转为步骤2)。

2) 得出消息在 n_c 的可能输出链路(possible output link，POL)即趋近其目的节点的链路，并判断 n_c 是否为孔表面节点。如果确为表面节点，则路由过程转为步骤4)；否则，路由转为步骤3)。

3) HFL检查链路状态表(LST、LST1或LST2)中关于POL的各表项，将会出现以下3种情形之一：

(1) 所有POL均非故障，则HFL在这些通道中任意选择一条作为输出链路。

(2) 部分POL为故障或某条链路在上一跳已经被路由，则HFL从剩下的POL中任意选择一条作为输出链路。在图2中，情形(1)与情形(2)合并为一类即消息选择任意一条可用的POL为其输出链路。

(3) 所有的POL不可用(链路故障或在上一跳已经被使用)，则该消息类型变为绕道消息(misrouted)。它将沿着故障环(虚故障环)绕行，这里仍有一个判决，即HFL是否在上一跳已在故障环上(虚故障环)绕行。如果上一跳已在环上绕行，消息继续保持该方向沿故障环(虚故障环)路由；否则消息将任意选择一条沿故障环(虚故障环)的链路来路由。

4) HFL检查PT各表项并与其目的节点进行对比，以确定其目的节点是否在当前孔中，然后路由转至步骤3)。如果目的节点在当前孔中，在步骤3)中HFL检查链路状态表LST1，否则检查LST2。检查完LST2后，消息将沿着虚故障环(而非故障环)路由以避免消息进入该孔。

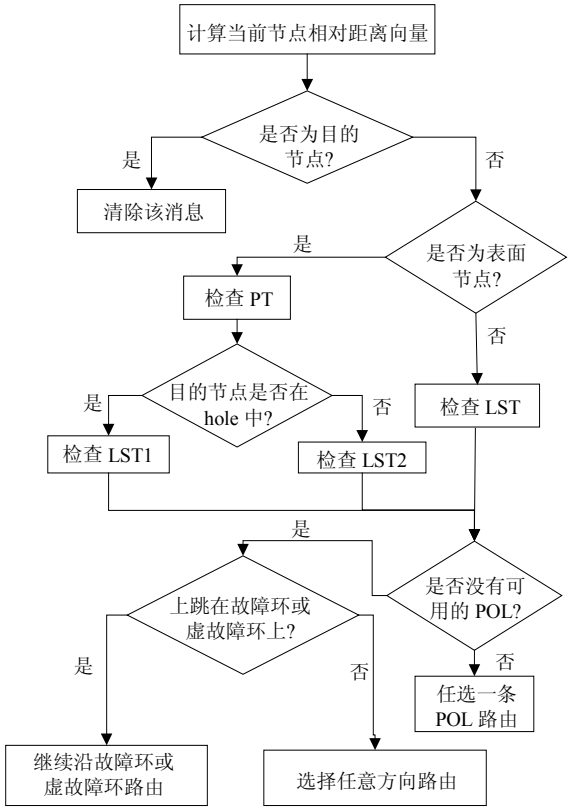

图2 MMAR算法流程图

图1描述了在32×32 torus网络中，一消息从源节点(4,3)到目的节点(8,7)的路由全过程。在源节点，HFL任选链路(4,3)→(4,4)到达节点(4,4)。在该节点HFL检查LST后发现只有一条POL(4,4)→(5,4)为非

故障，故选择。在节点(5,4)所有POL均不可用，消息将沿故障环绕行。由于其上一跳已在故障环上，所以消息路由的方向将保持不变。消息经过节点(5,3)到达节点(5,2)。因为在节点(5,2)处已经有一条可用POL，则消息类型重新记为正常且HFL选择该链路作为输出链路。在节点(6,2)两条POL均可用，其中一条已不在故障环上。根据路由步骤3)的情形(1)，消息可以路由到节点(7,2)并离开当前的故障环。在表面节点(7,3)，HFL检查PT后发现其目的节点不在该孔后检查LST2。消息将沿着唯一的非故障链路(7,3)→(8,3)离开该节点到达节点(8,3)。根据路由规则经过类似的若干跳后，消息可到达 H_2 的表面节点(9,8)。消息检查PT后发现在当前孔中。通过检查LST1，消息经由节点(8,8)最终到达目的节点(8,7)。

3　仿真结果与分析

本文使用OPNET搭建仿真模型，采用的包长为5 flit且流量模式为均匀模式(即消息以相等的概率送往交换结构中的每一节点)。本文考虑了两个重要的性能参数：信息的平均延迟和网络的吞吐率。吞吐率按标准化带宽的百分比来度量。标准化带宽很容易通过网络等分面均匀随机流量的50%导出[2]。如果一个N节点的网络的等分带宽为B bit/s，则网络中每个节点最大可以注入$2B/N$ bit/s的流量。仿真模型中采用虚拟通道方式来复用物理链路。每条虚拟通道拥有各自输入存储区和输出存储区，每个存储区的长度为1 flit。每条物理链路中的虚拟通道条数为2(死锁恢复虚拟通道除外)。

以16×16 torus为例观察MMAR的容错性能。以网络中故障节点的数目为参数，本文取了5种情况进行比较。第一种为正常网络即网络中没有任何故障。其他4种的故障节点数分别为18(占全网节点数的7.03%)、27(10.5%)、39(15.2%)和47(18.4%)。

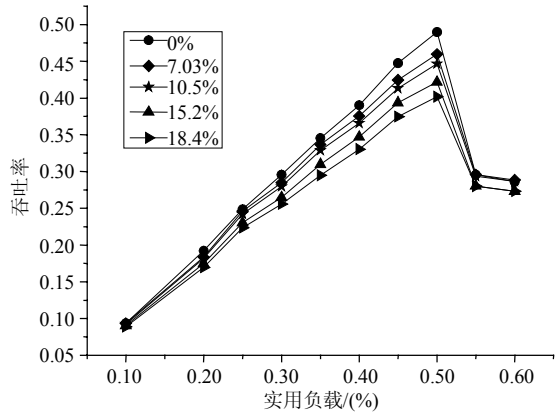

图3　各故障节点数情形下负载-吞吐率曲线图

通过图3和图4的性能曲线，可看出MMAR在严重故障情况下仍能继续工作。在不同故障节点数情形下，网络流量实用负载饱和点均在0.5左右。虽然在故障情况下网络的性能有所降低但幅度并不大。吞吐率方面，MMAR能容错任意故障模型，不使任何非故障节点丧失功能以保证每对非故障节点之间都能互相通信,且只有故障节点失去收发信息功能,使系统保持良好的吞吐率性能。平均信息延迟方面，大部分消息能按其最短路径到达目的节点；MMAR采取避免消息进入与其目的节点无关的孔以缩短绕道路径的策略，这样信息平均延迟变化能保持相对稳定。但随着故障节点数目的增多，网络中信息平均延迟还是随之有所增加。这是由于故障节点数目的增多导致了单个故障模型范围的扩大，部分消息必须绕更长路径来到达目的节点。

当网络中无故障时，整个网络中的MMAR与TFAR行为一致。此时，网络的吞吐率达到最大值，信息的平均延迟降为最低。

可见，MMAR在多维交换结构中体现出了良好的容错性能。在实际中，交换结构中多个节点同时发生故障的概率极低。因此，MMAR为mesh/torus交换结构中容错的好算法。

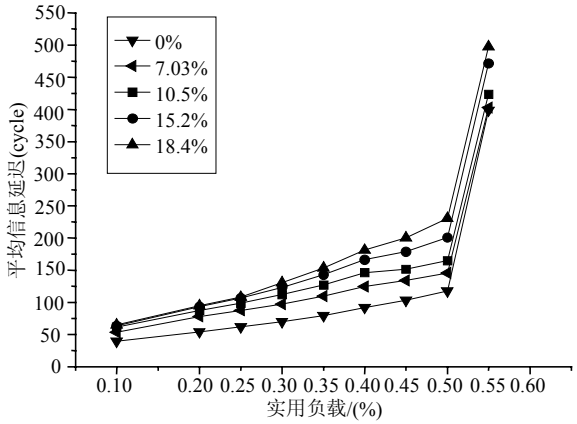

图4　各故障节点数情形下负载-平均时延曲线图

4　结 束 语

本文提出了一种在多维交换结构中使用的基于死锁恢复策略的容错路由算法MMAR。算法通过在各非故障节点中建立链路状态表容错任意形状的故障模型，以确保网络中任何两个非故障节点正常通信；通过收集局部故障信息并在凹形区域表面节点建立关于凹形区域内各节点的位置信息表，缩短了故障消息的绕道路径，降低了平均信息延迟。

(下转第 854 页)

次重传基本上可以保持传输的正确，因此采用星座重排和不采用星座重排的区别不大。而本文提出的SBV-PR-HARQ方案无论在信噪比低还是信噪比高的时候比原有方案的性能均有较大的提高。

4 结 论

传统的星座重排混合ARQ方案只是单独地强调各个比特软值的平均，而没有考虑可靠性低的比特是造成数据包的错误的主要原因，本文提出了一种基于比特软值的部分重传混合ARQ方案，理论分析和计算机表明，该方案在减少重传比特的情况下，依然能较好地平均比特间的软值，从而提高系统的性能。

参 考 文 献

[1] WENGERTER C, GOLITSCHEK A, EDLER V E. Advanced hybrid ARQ technique employing a signal constellation rearrangement[C]//VTC 2002-Fall. Vancouver, Canada: IEEE, 2002.

[2] SHEA J M. Reliability-based hybrid ARQ[J]. IEE Electronics Letters, 2002, 38: 644-645.

[3] VISOTSKY E, SUN Y, TRIPATHI V, et al. Reliability-based incremental redundancy with convolutional codes[J]. IEEE Transactions on Communications, 2005, 53: 987-997.

[4] ROONGTA A B, MOON J-W, SHEA J M. Reliability-based hybrid ARQ as an adaptive response to jamming[J]. IEEE Journal on Selected Areas in Communication, 2005, 23(5): 1045-1055.

[5] BERROU C, GLAVIEUX A, THITIMAJSHIMA P. Near Shannon limit error-correcting coding and decoding: Turbo-codes[C]//Proc ICC ’93. Geneva, Switzerland: [s.n.], 1993.

[6] PYNDIAH R, PIEART A, GLAVIEUX A. Performance of block Turbo coded 16-QAM and 64-QAM modulations[C]// IEEE GLOBECOM ’95. New York: IEEE, 1995.

[7] GOFF L S, GLAVIEUX A, BERROU C. Turbo-codes and high spectral efficiency modulation[C]//IEEE SUPERCOMM/ICC ’94. New Orleans: [s.n.], 1994.

[8] GU Xin-yu, LI W Y, NIU Kai, et al. A universal soft output algorithm for M-QAM demapper[C]//Proceedings of ICCC’2004. Beijing: ICCC, 2004.

[9] GU Xin-yu, WANG Yi-chen, YU Xiao-bo, et al. Advanced hybrid ARQ technique employing a signal constellation rearrangement based on 64-QAM[J]. Journal of Electronics & Information Technology, 2005, 27(11): 1686-1690.

编 辑 熊思亮

(上接第847页)

参 考 文 献

[1] GU Hua-xi, SHEN Hong, LIU Zeng-ji, et al. A new routing method to tolerate both convex and concave faulty regions in mesh/torus networks[C]//Proceedings of the 6th International Conference on Parallel and Distributed Computing, Applications and Technologies. Dalian China: IEEE Computer Society, 2005.

[2] DUATO J, YALAMANCHILI S, NI L. Interconnection networks: an engineering approach [M]. (revised edition). San Francisco: Morgan Kaufmann, 2002.

[3] MARTINEZ J M, LOPEZ P, DUATO J. A cost-effective approach to deadlock handling in wormhole networks [J]. IEEE Transactions on Parallel and Distributed Systems, 2001, 12(7): 716-729.

[4] PINKSTON T M. On deadlocks in interconnection networks[C]// Proceedings of the Int’l Symposium on Computer Architecture. Colorado USA: ACM Press, 1997: 38-49.

[5] BAYDAL E, LOPEZ P, DUATO J. A family of mechanisms for congestion control in wormhole networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2005, 16(9): 772-784.

[6] HO C T, STOCKMEYER L. A new approach to fault-tolerant wormhole routing for mesh-connected parallel computers[J]. IEEE Trans on Computers, 2004, 53(4): 427-439.

[7] GOMEZ M E, NORDBOTTEN N A, FLICH J, et al. A routing methodology for achieving fault tolerance in direct networks [J]. IEEE Transaction on Computers, 2006, 55(4): 400-415.

[8] CHALASANI S, BOPPANA R V. Communication in multi-computers with non-convex faults[J]. IEEE Transactions on Computers, 1997, 46(5): 616-622.

[9] KHONSARI A, FARAHANI A. Disha: a performance model of a true fully adaptive routing algorithm in k-ary n-cubes[C]//Proceedings of the 10th IEEE International Symposium on Modeling, Analysis, Simulation of Computer and Telecommunication Systems. Texas USA: IEEE Computer Society, 2002: 183-190.

[10] RUBIO J M, LOPEZ P, DUATO J. FC3D: flow control-based distributed deadlock detection mechanism for true fully adaptive routing in wormhole networks[J]. IEEE Transactions on Parallel and Distributed systems, 2003, 14(8): 765-778.

编 辑 张 俊