计算机科学2003Vol. 30№. 9

局部子立方体连通的超立方体网络容错路由算法和概率分析研究*)

肖晓麟　王国军　陈建二

(中南大学信息科学与工程学院　长沙410083)

Research on Fault Tolerant Routing Algorithms and Probabilistic Analysis in Locally Subcube-Connected Hypercube Networks

XIAO Xiao-Lin　WANG Guo-Jun　CHEN Jian-Er

(College of Information Science & Engineering, Central South University, Changsha, 410083)

Abstract　In this paper, first we analyze and give opinions of fault tolerant routing and probabilistic analysis. Then, on the basis of locally subcube-connected hypercube networks, we put forward some ideas to develop efficient fault tolerant routing algorithms and powerful probabilistic analysis techniques to study fault tolerant models and the corresponding routing algorithms, which is of great importance to the research of parallel computer interconnection networks.

Keywords　Interconnection networks, Hypercube networks, Network fault tolerance, Local-subcube-connectivity, Probabilistic analysis

1　引言

并行计算机是现在高性能计算领域的杰出代表。并行计算机体系结构由数据传输网络和多处理机一起组成，其核心是通信体系结构，通信体系结构的核心则是并行计算机互联网络。随着并行计算机互联网络和VLSI技术的迅速发展，系统中的并行处理机越来越多，仍采用简单的互连结构，已不能满足需求，于是人们提出了用超级的互连拓扑结构来进行设计。国内外现已对各种主要的并行计算机互联网络拓扑结构进行了大量研究，并对其中的一些拓扑结构已研制出了相应的商用和研究用的并行计算机系统。在这些商用和实验性多处理机系统中，通常采用直接互联网络连接系统中各个结点(也称为处理机)，其中最典型的网络拓扑结构就是超立方体型。超立方体网络拓扑结构之所以能引起研究者对其在不同方面的研究兴趣是由于它具有正规性、对称性、可靠性、强容错性、直径短、可嵌入性和网络通信能力的可扩展性等优点，这也使其深受实践者的欢迎。但是，随着并行计算机系统和互联网络规模的不断扩大，系统中出现结点故障或链路(也称为边)故障的可能性也随之增加，即网络中出错的可能性在增大。因而研究具有大量错误结点的网络容错模型和容错路由算法并进行概率分析具有非常重要的意义。尽管国内外现已对超立方体网络从网络容错模型、容错路由算法和概率分析等方面进行了大量的研究，但是现有的研究未能充分挖掘超立方体网络的容错能力，这方面还有许多研究可做，这对如何建立可靠的并行计算机系统很有帮助。

2　研究基础

2.1　相关概念

一个n维超立方体网络H_n由2^n个结点和$n\cdot 2^{n-1}$条边构成，每个结点可表示成一n位二进制地址串$b_1b_2\cdots b_n$，两两相邻的结点的地址必须只有一位不同，即这两点之间有一条边相连。每一个长度为n－k的二进制串$b_1b_2\cdots b_{n-k}$对应于H_n中的一个具有2^k个结点的k维子立方体H_k。2个n维系统可以构成一n+1维系统，方法是对其中一n维超立方体的各结点地址前加上一位"0"，而另一n维超立方体各结点地址前加上一位"1"。反之，一个n维超立方体可由完全相同的两个n－1维子立方体合成，即将两个n－1维子立方体中具有相同编号的结点连接起来，然后将其中一个子立方体的编号均加上2^{n-1}。依照这种递归构造方法，可用子立方体构造任意多维的超立方体。图1表示的是一个4维超立方体网络，它中间隐含了3维子超立方体，同时也可以看作是由内外两个3维子超立方体构成。用隐含子结构或递归的方法来看待n维超立方体网络结构时显得更容易理解。

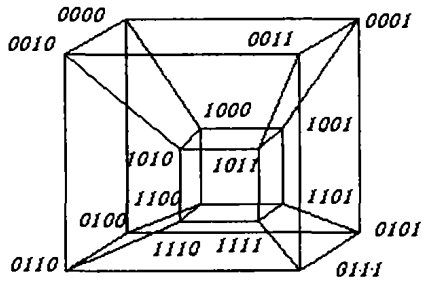


图1　隐含子结构的4维超立方体网络

下面是文[1,2]中对局部连通性的两个概念的定义。

定义1(局部k维子立方体连通性)　如果n维超立方体H_n中每一个k维子立方体H_k中的错误结点数少于正确结点

*)本文得到国家自然科学基金项目(编号:69928201)资助。肖晓麟　硕士研究生，主要研究兴趣包括网络容错、网络路由和网络安全等。王国军　副教授，博士，从事计算机网络、路由算法、容错性、软件工程等领域的研究。陈建二　博士，教授，博士生导师，从事计算机网络、计算机优化和计算机图形学等相关领域的研究工作。

数且所有 H_k 中的正确结点组成一个连通图，则称 H_n 是局部 k 维子立方体连通的。

定义2(局部子立方体连通性) 如果 n 维超立方体 H_n 中对任一个 k 维子立方体 H_k($k \geqslant 1$)，存在一个包含 H_k 的 h 维子立方体 H_h(这里 $h \geqslant k$)，且 H_h 是局部 h 维子立方体连通的，则称 H_n 是局部子立方体连通的。

2.2 超立方体网络容错模型的研究

一些研究者在结点错误模型或/和链路错误模型的基础上，提出了超立方体网络上的多种网络容错模型。现有的容错路由算法大都是基于 forbidden faulty set[3]概念的网络容错模型和 k-safe 容错模型[4]的。还有一些容错模型的概念被提出。尽管这些概念和模型在容错性、简单性、是否需要网络的全局状态信息等方面各有特点，但都不允许错误结点数大于 O(n)，容错能力相当有限。后来，王国军等人提出了两种容错模型：局部 k 维子立方体连通性模型和局部子立方体连通性模型[1,2]，它们的容错性与超立方体网络的结点数成比例，即可容许大量的错误结点而仍能确保正确结点的连通性。

2.3 超立方体网络容错路由算法的研究

早期的研究者在设计路由算法时，一般不考虑网络容错情况，如文[5]中提出的基于海明距离的三种确定性寻径算法。随着网络规模的增大，网络中结点和/或边出错的可能性也变大。因此，在研究超立方体网络路由算法时，容错能力被当做一个重要的性能指标考虑进来。研究者已对超立方体网络上单播(Unicast)、广播(Broadcast)、并行(Parallel)、多播(Multicast)、选播(Anycast)的路由算法进行了大量的研究。其中，Bao 等基于随机分布的 Byzantine 错误类型设计的两个所有结点到所有结点间的广播容错路由算法[6]。这些容错算法都是基于结点和/或边容错模型的，容错能力都很小。后来，王国军等人提出了基于局部连通性网络的单播容错路由算法[1]和并行容错路由算法，并针对 shouting 广播通信模式[7]设计了广播容错路由算法。这些算法都是基于局部信息的，只需结点知道其邻结点的状态，而无需知道整个网络信息。另外，不少研究人员还对近年提出的超立方体的变种如广义超立方体、扭立方体和超级交叉立方体的结构和路由算法进行了研究。

2.4 超立方体网络容错模型和容错路由算法的概率分析研究

根据网络容错模型对错误分布情况的分类，人们提出了两类最常使用的错误分布模型，即界限模型(bounded model)和概率模型(probabilistic model)。在界限模型中，要求给错误结点和/或边设定上界值，并且假定以最坏情况进行考虑。在概率模型中，错误的发生是随机的并且是相互独立的，同时应给定一个错误概率。这方面已有一些研究成果[6,8,9]，但是这些方法多研究一些极端的不大可能的情形，低估了超立方体网络的容错能力。直到2001年，王国军等人才首次研究了在给定结点错误概率时的基于局部连通性概念的容错模型和容错路由算法的容错性概率，并证明了超立方体网络能够容许相当多的错误结点而仍能确保整个网络的连通性[2]。

3 研究内容及其实现

3.1 研究内容

实践经验和结果表明，超立方体网络的容错性是相当强的，现有研究大都未能充分展示超立方体网络的容错能力。后来，王国军等人提出了局部连通性的概念[1,2]。局部连通的超立方体网络能够容许大量的、与网络中结点总数成比例的错误结点而仍能确保整个网络的全局连通性。局部连通性模型充分挖掘出了超立方体网络的容错能力。尽管他们的研究取得了突破性的进展，但同样也存在不足：他们并未对局部 k 维子立方体连通性模型和局部子立方体连通性模型进行深入的对比分析，特别是未能深刻揭示两种容错模型各自的适用场合；对基于局部 k 维子立方体容错模型的多播和选播等容错路由问题没有进行深入的研究，对局部子立方体容错模型只研究了单播容错路由算法；对超立方体网络进行的容错性和效率测试，大都是基于均匀结点错误分布的，没有考虑到结点错误分布不均匀的情况，更没有用随机分布，而结点非均匀错误分布模型下超立方体网络的容错路由应具有更高的现实意义和研究价值；在概率分析研究中也仍假定 n 维超立方体中结点具有均匀和独立的错误概率。上述问题在相关的研究中都是值得注意的方面。

经过我们的研究，局部子立方体连通性模型是比局部 k 维子立方体连通性模型更为一般、更具有普遍意义的模型，虽然基于这种模型进行路由算法的研究和概率分析的研究难度更大，但一旦成功，对并行计算机互联网络的研究将具有更重大的意义。所以本文的研究目标和方法是沿着另一条主线进行的，即基于局部子立方体连通性容错模型的各种容错路由算法及其概率分析研究。因此，我们要解决的关键问题有[10]：局部 k 维子立方体连通性容错模型和局部子立方体连通性容错模型的对比研究；基于局部子立方体连通性容错模型的各种高效容错路由算法的研究；局部子立方体连通性容错模型和容错路由算法的容错性的概率分析研究。文[1,2]中局部 k 维子立方体连通性模型主要针对整个网络中结点错误分布比较均匀的情况，而局部子立方体连通性容错模型主要针对整个网络中结点错误可能分布很不均匀的情况。这是我们现在对两种模型的初步理解。为了使基于两种模型的容错路由算法和概率分析更具有针对性，我们还要对上述两种模型继续进行深入的对比研究。

我们所考虑的结点错误分布不均匀的理想情况是结点错误随机分布，这就牵涉到了随机算法的问题。由于计算机的完全确定性，现有的许多随机算法都是伪随机的。好的伪随机数生成器创建的序列的属性与许多真正随机数的序列的某些属性是一样的。这些属性的实现决定于伪随机数生成器的开始值和模数。另外，其安全性还和硬件有关，目前大多数机器使用的32位的伪随机数生成器易被野蛮攻击手段解破。最佳的办法是选用128位的伪随机数生成器并用好的物理度量来生成随机数或伪随机数发生器的种子。现在已有的伪随机算法虽多，但为了得到结点错误随机分布的理想状态，并兼顾采用超立方体拓扑结构的并行计算机互连网络的安全性，所以我们也有必要研究伪随机算法的设计以及开始值和模数的选取，这有助于建立可靠的并行计算机系统。

3.2 实现结点错误概率的不均匀分布的方法

为了实现具有大量错误结点的、结点错误分布可能不均匀的满足局部子立方体连通性条件的超立方体网络上的高效容错路由算法；并对局部子立方体连通性网络容错模型和容错路由算法的容错性进行概率分析。首要考虑的就是如何合理地实现结点错误概率的不均匀分布。

由于 n 维超立方体具有非常规范的几何结构，对于顶点分类的分类超平面也应具有非常规范的几何结构；而 boole 函数的前向网络实现总是希望网络对输入数据的容错能力最

强，即分割 n 维超立方体任两不同类顶点的分割超平面应经过这两顶点连线的中点。所以有的研究者依据 n 维超立方体的构造方法以及 n 维超立方体的记数性质-boole 函数的非可线性可分性，提出了对 n 维超立方体顶点进行容错分类的容错分类复杂度的概念，并给出了容错分类复杂度为1的 boole 函数的基本模式[11]。根据这个思想，对属于不同错误分类的结点给定不同的错误概率，就可简单地实现错误概率不均匀分布，也具有一定的实际意义。

一些研究者对超立方体上自适应(无死锁)虫洞(wormhole)路由技术进行了研究[12~16]，其中涉及到对热点(hot spot)问题以及通道消息到达率等相关计算的研究。虽然对路由算法是否死锁问题的研究主要是基于早期的不考虑网络容错情况的，但是根据热点的定义，将超立方体中某个或多个结点作为热点时，由于距离热点近的通道(channel)可以获得更高的消息到达率(traffic rate)，而超立方体中的各结点与热点的距离远近不同，因此热点是能和非均匀分布(nonuniform distribute)网络发生联系的，所以这在我们的研究中是很值得注意的。但是他们的研究多是基于网络中只有一个热点(single hot-spot)的情形的，但是扩展到多个热点(multiple hot-spot)应该是可行的，也是很容易的。在只考虑一个热点时，由于超立方体的对称性，其中任意一点做热点都可以。对于扩展到一个超立方体中多个热点的情形要分两种情况来考虑：第一种基于均匀分布(uniform distribute)，可以考虑每一个 k 维子立方体一个热点，这样，根据定义，一个 n 维超立方体将被分为 m($m\geqslant 1$，并且 $m=h-k$，$n\geqslant k$)个 k 维子立方体。由于划分的子立方体间不可能交叉，因此这个 n 维超立方体中共有 m 个热点，当且仅当 $n=k$ 时，只存在唯一热点。另一种是基于非均匀分布的，这就回到了最理想的情况，找到某种有意义的随机错误分布模型，把它映射到 n 维超立方体结构上。在这个方面，文[7]曾做过一定研究，但未具体解决错误结点怎样实现随机分布的问题。以上处理的意义在于：由于超立方体或某子立方体中各点距离热点的不同，消息在各通道或结点上的等待时间或阻塞时间也不同，消息一旦长久阻塞甚至出现死锁时路由就进行不下去了，这时会有可能认为该路由通路在该结点上出了错误，因为热点的存在，所以各点出现错误概率也是不同的。应该怎样扩展到超立方体网络中具有多个热点的情形、到底应该有多少个热点、哪些结点应该是热点以及最终怎样和均匀/非均匀的容错分布联系起来等都是要通过研究解决的难题。至于超立方体网络中应该有多少个热点、哪些结点是热点、怎样和非均匀容错分布联系起来等都是其中的难点。文[15,16]中还特别强调了通道消息到达率的概念。在基于消息通信的多处理机系统中，通道消息到达率影响消息通信开销，通道消息到达率越小，结点越有可能阻塞。因此，我们可以试探性地这样做：利用超立方体网络中各点不同的通道消息达到率，结合随机算法随机生成源结点、目的结点以及路径，得出各点的错误概率。将网络中每个结点都处理一次，则可得到各点的错误概率。显然，用这种方法得到的错误概率是不均匀分布的，且具有一定的实际意义，因为由于热点的存在使得通道消息到达率和结点错误概率间也存在一定的联系。

在通过并行化来加速光线跟踪算法的研究中，并行化有两种空间分割的方法，其中之一就是将计算分布到各处理机上的图形空间分割。在这方面的研究中，文[17]的研究者提出了一种研究方法，即在超立方体互连拓扑结构上把形成图像的3D 空间子分成不相交的矩形子空间，并把一个子空间内的计算和一个物体数据指定给一个单处理机。这种方法能获得很好的数据相关性。另外，Macdonald 和 Booth[18]采用的空间分解模式考察了两个空间的子分试探法。其中的表面试探法指出表面区域与光线与物体相交的概率成正比，由此得出最优分解平面是位于物体中线与空间中线之间，有效地减少了需要找出分解平面的搜索范围。虽然以上方法存在一些缺陷，但我们仍然可以将结合利用他们的思想，进行在超立方体互连拓扑下有效的并行子分，即把一个给定的空间图形自适应地分解成矩形区域，并把这些合成区域映射到多处理机的结点处理机上。利用有效的结构来加速标识分解的平面，并在并行空间分解的同时完成区域和物体映射到结点处理机上。对这些经分解映射后不同区域中的结点可以分别取不同错误概率就实现了错误概率的不均匀分布。同时这个思想对研究如何将平面中的概率统计中某种合适的随机变量分布，如泊松分布，作为结点的不均匀错误分布映射到 n 维空间中的超立方体网络上去也有借鉴作用。前者可联系随机算法，生成 n 个数的序列，用于对分成的 n 块区域的错误概率。

在以上方法中，涉及到根据超立方体的结构来对点、面或空间划分为不同错误概率的若干区域时，最好都能联系随机算法，随机生成若干随机数的序列来作为错误概率的数值。

结论 本文以文[1,2]提出的局部子立方体连通性容错模型为基础，对超立方体网络容错路由算法和概率分析进行了比较全面的研究，并创造性地提出了一些研究方法。以前，人们只有这样一个印象，即超立方体网络容错模型和容错路由算法的容错性"应该"是很强的。但对局部子立方体连通的超立方体网络容错模型和容错路由算法的容错性到底有多强，还没有人使用概率分析的方法严格证明过。由于局部子立方体连通性容错模型中的子立方体的概念，将一个大的超立方体物理上划分为若干个小的子立方体，而对这些小的子立方体进行概率分析是容易做到的，所以借助上述研究方法并结合概率分析的方法最终是可能得出严格的概率意义的理论分析结果来证明超立方体网络容错模型和容错路由算法的容错性的。

参 考 文 献

1 王国军，陈建二，陈松乔. 具有大量错误结点的超立方体网络中的高效路由算法的设计与讨论. 计算机学报，2001，24(9)：909～916

2 Chen Jianer, Wang Guojin, Chen Songqiao. Locally Subcube-connected hypercube networks: Theoretical analysis and experimental results. IEEE Transactions on Computers, 2002, 51(5): 530～540

3 Esfahanian A H. Generalized Measures of Fault Tolerance with Application to N-Cube Networks. IEEE Transactions on Computers, 1989, 38(11): 1586～1591

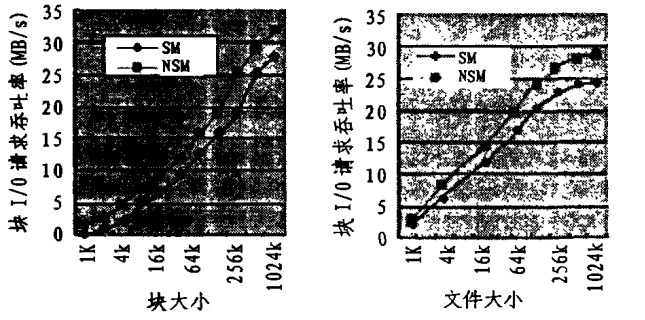
4 Latifi S, Hedge M, Naraghi-Pour M. Conditional Connectivity Measures for Large Multiprocessor Systems. IEEE Transactions on Computers, 1994, 43(2): 218～222

5 陈国龙，张德运，王小东. 基于 Hamming 距离的超立方体网络的寻径算法. 小型微型计算机系统，1999，20(6)：422～424

6 Bao F, Igarashi Y, Katano K. Broadcasting in Hypercubes with Randomly Distributed Byzantine Faults. In: 9th Intl. Workshop on Distributed Algorithms, WDAG'95, Lecture Notes in Computer Science 972, Springer, 1995. 215～229

(下转第110页)

的情况，第二次测试 iSCSI 块客户和 USN 服务器加安全模块的情况，第三次测试文件客户和 USN 服务器不加安全模块的情况，第四次测试文件客户和 USN 服务器加安全模块的情况。其块 I/O 吞吐率和文件 I/O 吞吐率显示如图7所示。从显示的结果看，块 I/O 客户加载安全模块并选择加密存储时比不加载安全模块慢14%-22%；而文件 I/O 客户加载安全模块并选择加密存储时，比不加载安全模块慢15%-25%。在这两种情况下，如加载安全模块，但选择非加密存储时，对整个系统的性能影响更小。



(a) 块 I/O 请求吞吐率 (MB/s)　(b) 文件 I/O 请求吞吐率 (MB/s)

注：图中 SM 表示加安全模块，NSM 表示没有加安全模块。

图7　USN 顺序访问性能

结论　目前，在国际上，NAS 和 SAN 技术是两种比较成熟的技术，而寻求 NAS 技术和 SAN 技术融合构建 USN 则是一种全新的技术。我们在国家自然科学基金项目（统一存储网络的设计、建构和实验研究，编号：69873017）和高等学校骨干教师资助项目（存储区域网智能调度和管理技术研究）的资助下，用211工程提供的设备构造了一个简单的 USN。在此基础下，我们结合我们所设计的 iNAS 安全系统[7]，并基于 NAS 的 USN 的特点，进一步设计出了一套 USN 安全算法。通过实验验证，该安全算法对整个 USN 系统性能影响不大，与文[5]相比该系统通过用户口令认证，可防止拒绝服务攻击，并且既可面向块 I/O 请求，又可面向文件 I/O 请求。与文[7]相比，我们对客户端文件 I/O 请求安全模块进行了改进，使安全系统对文件 I/O 请求客户端的性能影响更小；在 USN 服务器端，我们通过增加 HMAC 认证和 I/O 写并行算法模块，使安全系统对 USN 服务器性能影响更小。

参考文献

1　IBM redbook：IP Storage Networking：IBM NAS & iSCSI Solutions

2　Brocade whitepaper. comparing the Storage Area Network and Network Attached Storage

3　Krawczyk H, Bellare M, Canetti R. HMAC: Keyed-Hashing for Message Authentication. IETF Network working Group RFC2104, Feb. 1997

4　Reid J. Plugging the Hole on Host-Based Authentication. Computers and Security, 1996. 661～671

5　Miller E, Long D, Freeman W, Reed B. Strong Security for Distributed File Systems. IEEE Micro, 2001, 20(1): 34～40

6　Thadani M, khalidi Y A. An Efficient Zero-Copy I/O, Framework for UNIX. SUN Microsystems Laboratories, Inc

7　韩德志，等. 基于 iSCSI 协议的附网存储安全系统的研究与设计. 小型微型机系统（已录用）

（上接第102页）

7　Santoro N, Widmayer P. Distributed Function Evaluation in the Presence of Transmission Faults. In: Proc. Intl. Symposium on Algorithm, SIGAL'90, Lecture Notes in Computer Science 450, Springer, Berlin, 1990. 358～369

8　Najjar W, Gaudiot J L. Network Resilience: A Measure of Network Fault Tolerance. IEEE Transactions on Computers, 1990, 39(2): 174～181

9　Chen M S, Shin K G. Depth-First Approach for Fault-Tolerant Routing in Hypercube Multicomputers. IEEE Transactions on parallel and Distributed Systems, 1990, 1(2): 152～159

10　王国军，陈建二，张祖平. 局部子立方体连通的超立方体网络容错路由算法和概率分析研究：[湖南省自然科学基金报告]. 2002

11　Zhang Junying, Xu Jin, Bao Zheng. Tolerantly Linear Separability of Boolean Functions and its Numbering. In: 1996 Intl. Conf. on Signal Processing Proc. 1996. 1433～1436

12　Ould-Khaoua M, Sarbazi-Azad H. An Analytical Model of Adaptive Wormhole Routing in Hypercubes in the Presence of Hot Spot Traffic. IEEE Transactions on Parallel and Distributed Systems, 2001, 12(3): 283～292

13　Sarbazi-Azad H, Ould-Khaoua M, Mackenzie L M. An Analytical Model of Fully-Adaptive Wormhole-Routed k-Ary n-Cubes in the Presence of Hot Spot Traffic. In: 14th Intl. Parallel and Distributed Processing Symposium (IPDPS'00), 2000. 605～610

14　dandamudi S P, Eager D L. Hot-Spot Contention in Binary Hypercube Networks. IEEE Transactions on Computers, 1992, 41(2): 239～244

15　Dally W J, Aoke H. Deadlock-free Adaptive Routing in Multicomputer networks Using Virtual Channel. IEEE Transactions on parallel and Distributed Systems, 1993, 4: 466～475

16　Abraham S, Padmanabham K. Performance of the direct Binary n-cube network for multiprocessors. IEEE Transactions on Computers, 1989 (7): 1000～1011

17　Ozguc B, Isler V, Aykanat C. Subdivision of 3D Space Based on the Graph Partitioning for Parallel Ray Tracing. In: Proc. of the 2nd Eurographics Workshop on Rendering, Barcelona, 1991

18　MacDonald J D, Booth K S. Heuristics for Ray Tracing Using Space Subdivision. The Visual Computer, 1990, 6(3): 153～166