2000 年 10 月
第 2 卷第 10 期

中国工程科学
Engineering Science

Oct.2000
Vol.2 No.10

研究报告

关于大规模并行处理机系统可扩展性设计

卢锡城
（国防科学技术大学，长沙 410073）

[摘要] 大规模并行处理系统旨在满足国防和国民经济许多重要应用领域对高性能计算能力的需求。长期以来，结构上的可扩性和编程上的友好性一直是并行计算机系统设计中追求的重要而又互相矛盾的两个目标。文章结合研究实践，对大规模并行处理机系统（MPP）可扩展性设计的若干问题进行探讨。

[关键词] 体系结构；大规模并行处理机（MPP）；对称多处理机（SMP）；基于 Cache 一致性的非一致存储访问（CC－NUMA）；群机（cluster）；超结点

1 前言

大规模并行处理机系统（MPP）已非传统意义上的 MPP，已成为泛指一切结构上可扩展性好的并行计算机系统。结构上的可扩性和编程上的友好性一直是并行计算机系统设计中追求的十分重要而又互相矛盾的两个目标。全系统只有一个操作系统副本，软件采用“共享存储”编程模式的并行计算机系统，编程上较方便，但由于受到硬、软件及工艺技术上的制约，系统规模的可扩展性受限，一般全系统 CPU 数不超过 512 个。每个结点（或超结点）上运行各自的操作系统，结点间采用“消息传递”编程模式的 MPP 系统，可扩展性好，但并行编程却较困难。在目前技术条件下，要实现规模大到上千乃至上万个 CPU 的并行计算机系统，后者仍是唯一的技术途径。本文结合工作实践，就 MPP 系统可扩展性设计的若干问题进行探讨。

2 关于超结点（Hypernode）设计

超结点是指 MPP 系统中的结点是由多个 CPU 组成的支持“共享存储”编程模式的多处理机系统。

2.1 超结点的结构

目前超结点的结构大体分二类：a. 对称多处理机（SMP）结构；b. 基于 Cache 一致性的非一致存储访问（CC－NUMA）结构；表 1 列出了二者的主要特点。

表 1 SMP 和 CC－NUMA 的主要特点

Table 1 Main features of SMP and CC－NUMA

结构	SMP	CC－NUMA
并行模型	UMA	NUMA
物理存储	共享	分布
逻辑存储	共享	共享
可扩充性（CPU 数）	≤64	≤512
可编程性	好	好

两种结构中，CC－NUMA 可扩展性比 SMP 好的特点，使之成为当前技术条件下实现大超结点的主要技术选择。CC－NUMA 结构亦称为 S^2MP（可扩展共享存储多处理机）结构。

2.2 CC－NUMA 超结点结构

CC－NUMA 超结点的可扩展性，硬件上受到 CC 机制实现的制约，软件上受到 NUMA 特性的

[收稿日期] 2000－05－19；修回日期：2000－07－03
[基金项目] 国家自然科学基金资助项目（69933030）
[作者简介] 卢锡城（1946－），男，江苏靖江市人，中国工程院院士，国防科学技术大学教授，博士生导师

制约。CC－NUMA 超结点结构见图 1。

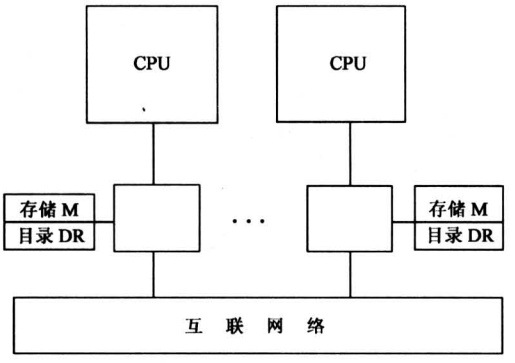

图 1 CC－NUMA 结构

Fig.1 Structure of CC－NUMA

存储器物理上分布的 CC－NUMA 结构要支持有效的“共享存储”编程模式，关键是要做到平均访存时间可与 SMP 相比，这一性能是靠用硬件实现的高速、低延时互联网络及 Cache 一致性机制保证（CC）来实现的。例如，采用 CC－NUMA 结构的 SGI 公司产品 SN1，其局存访问时间为 115 ns，远存访问时间为 180 ns ＋50 ns／hop（跳步）。

CC－NUMA 结构的超结点规模，硬件上受到 CC 保证机制，软件上受到 NUMA 特性的制约。由于超结点规模的扩展将引起目录表规模的增大和 Cache 一致性机制实现的复杂性增加，硬件开销就成为制约因素。为克服 NUMA 特性的负面影响，系统软件必须尽可能保证数据的局部性，如采用页迁移、数组私有化等技术。根据一些典型测试程序测试，采用局部化技术，通信量可降低 20％到一个量级。随着超结点规模增大，软件高效局部性优化就越困难。软件优化是一项正在发展中的技术，软件优化技术一直滞后于硬件实现技术。

2.3 超结点规模受结点间通信实现技术的制约

图 2 给出 1 个 2 结点并行计算最简单的模型。

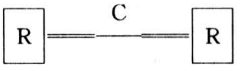

图 2 两结点并行模型

Fig.2 Parallel model with two nodes

设两个结点上并行运行任务的计算时间为 t_R，因并行而引入的通信等开销时间为 t_C，设通信与计算时间不重叠，则制约并行效率的主要因素是 t_R/t_C 的值。t_R/t_C 值越大，并行效率越好。提高 t_R/t_C 值可以从系统硬、软件设计、并行算法、并行程序设计等各个方面努力。为保持系统的均衡，结点计算能力增强，相应结点间通信带宽也应增强，而实现高通信带宽的技术难度及成本限制了超结点规模的扩展。

3 关于层次结构设计

MPP 系统硬件上可分多层是由类型不同的互联网互联，但这里所讨论的层次式结构是指含共享存储和分布存储两种不同层次的结构设计。

3.1 层次式体系结构是 MPP 系统发展的趋势

1）因共享存储 SMP 结构的并行计算机系统规模受限，基于分布存储的 MPP 结构获得了较快的发展。一个自然的延展就是以 SMP 系统为结点组成 MPP 系统。这就形成了硬件上的二层结构，即结点内共享存储、结点间分布存储。

2）随着微电子技术的发展，多处理机芯片即将问世，芯片本身就可较好地支持直接互联成 SMP 或 CC－NUMA 结构。图 3 所示的 IBM 公司正致力开发的 Power 4 微处理器，每个芯片内含两个主频大于 1 GHZ 的 CPU，并可方便地直接互联成多处理机系统。图 4 所示是 Compaq 正在开发的 Alpha 21364 微处理机芯片及以其构成的多机系统。这些为层次结构的硬件实现提供强有力的支持。

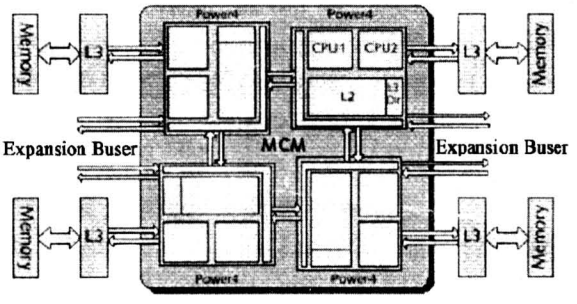

图 3 基于 POWER 4 构成的共享存储多处理机系统

Fig.3 SMP system based on Power 4

3）从软件编程角度，所谓层次结构是指存在两种编程模式，即超结点内共享存储，超结点间消息传递编程模式。一些应用问题本身的物理特性存在两种粒度不同的并行性，层次结构可以有效地映象到两种不同的编程模式上。

3.2 层次结构可扩展性好，但编程困难

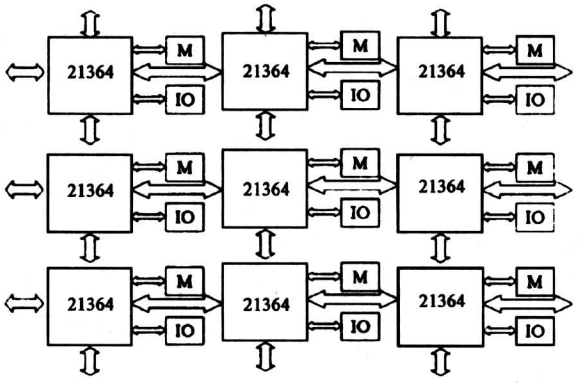

图 4　Alpha 21364 构成共享存储多处理机系统

Fig.4　SMP system based on Alpha 21364

层次结构减少了每一层上互联的结点数，既可简化互联网的硬件结构，又可在每一层上提供较好的扩展能力。但是两种编程模式给用户编程带来较大困难。给用户提供自动、高效，对层次透明的编程支持是人们不断努力的目标。

4　关于系统平衡性设计

系统设计的平衡性直接影响系统的可扩展性，不同应用领域对并行计算机系统各部分能力的平衡要求不同。如美国 ASCI 计划，根据核模拟应用问题特点，提出系统各部分能力应按如下比例扩展：

1 TFlops 计算性能	1 TB 主存
50 TB 磁盘	16 TB/s Cache 带宽
3 TB/s 主存带宽	0.1 TB/s (I/O) 带宽
10 GB/s 磁盘带宽	1 GB/s 归档带宽
10 PB 文档存储空间	

4.1　影响系统性能的主要因素

4.1.1　*MPP 系统中，输入/输出带宽往往是用户业务流程中的瓶颈*　目前 CPU 计算性能已超过 1GFlops，互联网每个互联方向的通信性能超过 1 Gb/s，存储带宽也超过 5 GB/s。预计未来 1～3 年内，CPU 峰值性能将达 3～4 GFlops，每个互联网方向通信可达 10 Gb/s，存储带宽可大于 10 GB/s。而 I/O 技术的发展则远远落后，I/O 系统的可扩展能力往往成为制约系统整体性能的瓶颈。

4.1.2　*全局操作是通信技术中制约系统可扩展性的重要因素*　设 N 是参加计算的 CPU 数，在实现一次全局操作的通信中，最少需要 lb N 个操作步骤，如 Broadcast（广播），而实现 Gather / Scatter（收集/散布）这样的全局操作，则需要 $N-1$ 个操作步骤才能完成。规模越大，一次全局操作的开销越大，全局操作开销制约了系统的可扩展性。

4.1.3　*高存储带宽是系统平衡的基础*　从体系结构角度看，存储子系统是现代并行计算机系统的中心，CPU、输入/输出、机间通信均要共享存储带宽。因此高存储带宽是系统平衡设计的基础。

4.2　系统各层次间的和谐是平衡性设计的难题

一个实际问题映像到并行计算系统，可以分为三层模型、两级映射，见图 5。如何协调三层模型，两级映射的关系，以使系统设计整体和谐，是并行计算机设计中具有挑战性的课题：

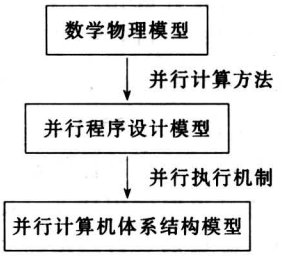

图 5　实际问题到并行计算系统映像

Fig.5　Mapping from the real problem to the parallel computer system

1）面向应用作针对性设计，是提高 MPP 系统实用性能的重要技术途径。不同应用问题的数学物理模型具有不同的并行计算特征，例如，数值天气预报全球谱模式呈现一层并行性，核物理问题则具有粒度不同的两层次并行特征。因此一台大规模并行机系统很难做到高效地通用。据美国有关部门统计，ASCI 计划的并行计算机，能利用 50 %以上系统能力的题目约占 5 %。面向应用作针对性设计，是提高并行计算机实用性能的重要技术途径。

2）并行程序设计是从并行算法到生成高效、可扩展性好的并行程序过程中关键步骤，性能优劣依赖于并行程序设计人员对并行体系结构的了解。优化数据的划分和分布是减少通信量、实现高可扩的关键。对含千万个 CPU 的高端系统，系统很难自动实现优化的数据划分和分布，必须靠并行程序设计人员结合并行体系结构的特点进行设计。

3）硬件实现接口原语、系统软件支持多种优化策略可提高性能、价格比。应用程序编程人员给出指导策略信息，可较合理地协调各层关系。如在 CC－NUMA 系统中，处理机调度问题的实现可以是：硬件实现同步等基本原语，系统软件实现自调

度（SS）、块自调度（CSS）、指导性自调度（GSS）等多种策略。并根据应用问题的特点，通过编译指导方式选择不同策略，从而较有效地解决处理机调度、负载平衡等问题。软、硬件技术结合，可降低硬件实现的复杂度，提高系统的性价比。

5 关于对称结构设计

MPP系统中结点按职能分工可分为计算结点（PN）、输入/输出结点（ION）、服务结点（SN）、控制结点（CN）等。对称结构设计是指硬件上把职能不同的结点均设计成相同的结构。硬件对称和软件可裁减能增强系统可扩性并可降低系统软件开销。随着硬件设备和器件成本的下降和功能的增强（如存储芯片、硬盘等），MPP系统诸结点按对称结构设计已成趋势。结构对称的硬件和配置可裁减的软件，使系统可根据应用特点（计算密集型、数据密集型或通信密集型）对结构作不同配置，以提供计算、I/O等能力的灵活升级、扩展。

6 值得重视的群机（cluster）技术

90年代以来，随着高性能工作站、服务器及高速网络互联技术的发展，cluster正成为超级计算系统中的一支劲旅，实际上，MPP与cluster间的界限已变得模糊。今天提供高性能计算的cluster已不再是传统的工作站加局域网（LAN）组成的工作站网络（NOW）或者工作站群（COW），而是采用专用、同构结点和专门设计的高速紧耦合互联网组成的并行计算机系统。其重要特点与传统MPP系统的区别，是各部件均遵照国际标准接口设计的商用产品，其主要优势是可扩展性好，造价低。表2列出了Compaq公司基于cluster结构发展超级计算机的计划。现Alpha Server SC系统已做到点点通信延时：driver级 $<3\ \mu s$，MPI级$<5.5\ \mu s$。1999年底，国际上超级计算机排行榜TOP500前10名中cluster结构占4席，见表3。

表2 Compaq MPP系统发展之路

Table 2 Way of Compaq 's HP System

年代	主频/MHz	CPU数（SMP）	节点数	TFlops
1999	667	4	128	0.8
2000	>700	32	128	≈7
2001	>1 000	64	256	≈30
2002	>1 200	64	256	≈40
2003	≈1 500	64	256	≈100

表3 TOP500前10名（1999年底发布）

Table 3 The top ten of TOP500 (Late of 1999)

Rank	Manufacturer 厂商	Computer 名称	Rmax /GFlops	Country	Year	Area of Installation	# Proc	Rpeak /GFlops
1	Intel	ASCI Red	2 379.6	USA	1999	Research	9 632	3 207
2	IBM	ASCI Blue-Pacific SST IBM SP 604e	2 144	USA	1999	Research Energy	5 808	3 868
3	SGI	ASCI Blue Mountain	1 608	USA	1998	Research	6 144	3 072
4	Cray/SGI	T3E1200	891.5	USA	1998	Classified	1 084	1 300.8
5	Hitachi	SR8000/128	873.6	Japan	1999	Academic	128	1 024
6	Cray/SGI	T3E900	815.1	USA	1997	Classified	1 324	1 191.6
7	SGI	ORIGIN2000/ 250MH	690.9	USA	1999	Research	2 048	1 024
8	Cray/SGI	T3E900	675.7	USA	1999	Research Weather	1 084	975.6
9	Cray/SGI	T3E1200	671.2	Germany	1999	Research Weather	812	974.4
10	IBM	SP Power3 222 MH	558.13	USA	1999	Research	960	852.4

cluster技术促进了并行计算技术的普及，大量并行系统软件、并行中件软件、并行应用软件等自由软件应运而生。许多提高性能、方便应用的并行技术，如单一系统映像（SSI）技术、分布共享存储（DSM）技术，资源管理与调度（RMS）技术等蓬勃发展，cluster技术给MPP技术发展注入了

新的活力。可以预计，随着并行、串行源同步通信技术、光互连技术、高性能 I/O 标准（如，NGIO）、存储总线标准及 SSI 等技术的发展，cluster 技术将呈现更强的竞争力。

7 后记

人们把“并行计算机系统”定义为“并行算法”和“并行体系结构”的结合，它说明了并行计算机系统与应用领域的紧密相关性。算法设计人员与系统设计师之间广泛、深入的交互，在系统设计阶段有利于指导强化系统设计能力和成本的重点投向，以提高系统的性价比。在系统定型后，可以帮助编制出结合系统特点的高效并行程序。用户与系统研制方紧密配合，应用软件开发与并行计算机系统研制同步进行的“并发工程（Concurrent Engineering）”模式已被实践证明是行之有效的模式。要进一步加强的是更多更宽地从应用问题算法入手，结合应用全过程分析模拟信息流的特点，设计出整体性能均衡、高效的 MPP 系统。体系结构的创新源于对算法的深入分析。

Issues on the Scalability in Designing a Massively Parallel Processor

Lu Xicheng

(*National University of Defense Technology*, *Changsha* 410073, *China*)

[**Abstract**] The massively parallel processor (MPP) has been designed to meet the requirements for the high performance computing in many application fields of both national defense and economy. The structural scalability and the friendly programming are the two important and conflicting goals in designing a MPP system. Based on practice, the issues on the scalable design of MPPs are discussed in this paper.

[**Key words**] architecture; MPP; SMP; CC－NUMA; cluster; hypernode

(Cont. from p.97)

Research on operation behavior of multiple arch dam

Gu Chongshi, Li Xuehong

(*College of Water Conservancy and Hydropower Engineering*, *Hohai University*, *Nanjing* 210098, *China*)

[**Abstract**] The multiple arch dam is sensitively affected by its environment because of its thin structure. Based on the observation data, the operation behaviors of Meishan and Foziling multiple arch dams are comparatively analyzed with many kinds of models and analysis methods. Some unfavorable loading conditions are obtained for the deformation, stability, strength, crack, etc.

[**Key words**] multiple arch dam; operation behavior; unfavorable loading condition; comparative analysis

* * *

更　正

《中国工程科学》2000 年第 7 期第 68 页，续表 2 第 1 栏的 Ru Rh Pd Ag 系 Os Ir Pt Au 之误。特此更正，并向作者、读者致歉。