

计 算 机 科 学 丛 书

可扩展并行计算 技术、结构与编程

Scalable Parallel Computing
Technology, Architecture, Programming

黄 铠 徐志伟 著

陆鑫达 曾国荪 邓倩妮 等译
陆鑫达 校



机械工业出版社
China Machine Press

TP338.6
01

计算机科学丛书

3587/A2

可扩展并行计算

技术、结构与编程

黄 钊 徐志伟 著

陆鑫达 曾国荪 邓倩妮 等译

~~陆鑫达~~ 校



机械工业出版社
China Machine Press

并行处理机是计算机设计的未来。而实现高效并行处理机的一个重要问题是构成程序模型和表示应用所用语言基础的计算机模型。本书是唯一涉及对高效并行处理理解的三个支撑点：计算模型、基础选择及程序设计范例所有方面的一本教材，是本科生、研究生及各类计算机设计人员必读的教科书。

Kai Hwang & Zhiwei Xu: Scalable Parallel Computing Technology, Architecture, Programming.

Original edition copyright © 1998 by McGraw-Hill. All rights reserved.

Chinese edition copyright © 2000 by China Machine Press. All rights reserved.

本书中文简体字版由作者授权机械工业出版社独家出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

版权所有，侵权必究。

本书版权登记号：图字：01-2000-0123

图书在版编目(CIP)数据

可扩展并行计算技术、结构与编程 / (美)黄铠(Hwang, K.)、徐志伟著；陆鑫达等译.
—北京：机械工业出版社，2000.5

(计算机科学丛书)

书名原文：Scalable Parallel Computing Technology, Architecture, Programming

ISBN 7-111-07580-3

I. 可… II. ①黄… ②徐… ③陆… III. 并行计算机-基本知识 IV. TP338.6

中国版本图书馆CIP数据核字(2000)第10353号

机械工业出版社(北京市西城区百万庄大街22号 邮政编码100037)

责任编辑：周 桦

北京第二外国语学院印刷厂印刷·新华书店北京发行所发行

2000年5月第1版第1次印刷

787mm×1092mm 1/16·34.25印张

印数：0 001-5 000册

定价：49.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

敬 献

献给我的家庭，感谢他(她)们的关爱和支持

黄铠

献给我的中学老师张大华女士，感谢她在一个动荡的年代里教给了我科学的宁静

徐志伟

译者序

本书是关于可扩展并行处理机的一本很有特色的教材，取材新颖。从软硬件技术，计算机体系结构以及计算模型、编程技术和环境等方面来论述可扩展并行计算是本书的一大特色。本书的另一特色是有较强的模块性，从而可灵活地组合成适合不同专业和不同对象的有关计算机的课程。由于本书中还介绍和剖析了当今正在从事的大多数并行处理机的研究和实现，因而对从事相关研究工作的读者来讲也是一本很有价值的参考书。我们有幸翻译此书，以飨读者，希望读者能从中受益。

参加本书翻译工作的有陆鑫达、曾国荪、邓倩妮、于洋、张玉珩和王景村。陆鑫达教授负责和组织了全书的翻译，并翻译了序、前言、目录、第1章至第4章、第12章至第14章以及Web资源表目录。曾国荪副教授翻译了第6章及第7章。邓倩妮讲师翻译了第8章及第9章。第5章由于洋翻译，第11章由张玉珩翻译，第10章由王景村翻译。译稿全文由陆鑫达教授作了校对。

在本书的翻译过程中，承蒙本书作者之一徐志伟博士拨冗到沪，及时告知原著中的错漏之处和一些修改之处，使本书的翻译能更为正确，在此谨向他深表谢意。此外还要感谢机械工业出版社华章公司的策划、编辑人员，是他们的不懈努力才使本译著得以顺利出版。

由于翻译时间较为仓促，翻译水平有限及新术语的不断涌现，翻译中的不妥之处，还望读者不吝批评指正。

陆鑫达

上海交通大学计算机科学与工程系

2000年2月16日

序

并行处理机是计算机设计的未来。业已证明技术可重复性是促使并行处理机实现不断增长的动力。另一个主要理由是重要应用领域对性能增长的需求。然而建造高效、可扩展并行处理机也已被证实是异常困难的。

数十年来，似乎大型（ $n > 100$ 处理机）并行处理机只需再稍作努力和研究就可有效地加以实现，然而这一点尚未被证实。在程序中很难找到必要的并行性，而当找到这种并行性时，它又不能转换成正比于参加运行的处理单元数的程序执行加速比。当 n 超过100时，这一点表现得尤为明显。

实现高效并行处理机的一个重要问题是构成程序模型和表示应用所用语言基础的计算机模型。过去所开发的模型能有效地适用于单处理机，因为我们通晓顺序动作规范的语言概念，所以它们很容易映射到串行计算和程序设计模型上。但至今已经证实，企图将这些表示重新映射到并行模型上是相当低效的。

支持对高效并行处理理解的三个支撑点是：计算模型、基础选择（underlying alternative）以及程序设计范例。本书是唯一地涉及到上述所有三个方面的一本教材。性能指标为理解计算的基本模型提供了定量基础。

黄、徐两位教授综合地分析了处理机硬件模型、互连网络和一系列广域网，从而给出了硬件当前发展状况的全貌。它们涉及的范围从硬件的指令级并行性（ILP）扩展至由工作站网（NOW）可获取的并行性形式。他们以一体化的计算模型、硬件以及程序设计模型的系统观点，述评了当今正在从事的大多数并行处理机的研究与实践。

这种硬件和软件集成方式使本书很有价值，因为只有通过从应用到系统映射以及到处理机构成映射的理解，我们方能期待构造有效的可扩展并行处理机系统的连贯进展。

Michael J. Flynn

斯坦福大学

前 言

一个数字的世界

本书内容包括多处理机、多计算机及基于网络的机群平台的可扩展体系结构和并行编程。数字技术已造就了计算机工业。现在数字技术正在以另一个冲击波对远程通信和信息工业施以根本影响。在高度自动化的社会中，将一切转换为数字便是未来成功的关键。

各种技术的杂交要求新一代计算机能适应于可扩展、并行和分布计算。计算机和信息技术中的这些变化已促使计算机专业人员去研究本教科书中所提供的材料。其最终目的是要为迎接21世纪的新挑战做好准备。

本书概要

本书由四篇14章组成，对以下四方面进行论述：原理、技术、体系结构和编程。

- 第一篇中的3章涉及可扩展计算机平台和模型、并行编程基础以及并行性能指标。
- 第二篇对商品微处理器、分布式高速缓存和存储器体系结构、开关式互连、千兆位网和通信作了述评。
- 第三篇将涉及对称式多处理机（SMP）和高速缓存一致性、非均匀存储器访问（CC-NUMA）机、工作站机群（COW）以及大规模并行处理机（MPP）。
- 第四篇介绍并行语言、侧重Unix编程环境的编程模型、消息传递、数据并行以及PVM、MPI、Fortran 90和HPF在可扩展计算机上的使用。

本书中90%以上的论题内容是在近5年来新技术进展和研究开发的基础上重新写成的，它不是黄铠教授前几本著作的修改本。下面重点介绍本书的独特之处。

热点芯片和互连

本书将评述构造可扩展多处理机和多计算机机群的商品微处理器和热点芯片。将研究分布式高速缓存/存储器和千兆位网以及时延隐藏机制。尤其是我们将研究多处理机总线和纵横交叉开关、SAN（系统局网）以及如千兆位以太网那样的LAN（局域网）、SCI（可扩展一致性接口）以及ATM（异步传送模式）网络。

可扩展平台和机群

本书将重点介绍可扩展体系结构、快速通信机制、时延隐藏、分布式共享存储器、高速缓存一致性协议以及存储器一致性协议。我们将涉及有更高可用性的软件扩展、单一系统映像、故障恢复系统以及计算机机群中的作业管理。

个例研究包括HP/Convex Exemplar、Cray T3D/T3E、IBM SP2、Digital Trucluster、Microsoft Wolfpack、Sun Ultra Enterprise 10000、SGI Origin 2000、Sequent NUMA-Q及Intel/Sandia ASCI Option Red。我们还将讨论在斯坦福Dash、伯克莱NOW、普林斯顿

SHRIMP以及Rice TreadMarks各机器系统中所获得的教训。

并行软件环境

本书中有将近一半以上的篇幅介绍软件工具和并行编程系统。对于共享存储器方法，我们将研究ANSI X3H5、Pthreads、SGI Power C、OpenMP以及C++语言。我们还将通过研究Solaris MC和LSF（负载共享设施）了解可用性、单一系统映像和机群作业管理。

对于并行编程，我们将研究数据并行、消息传递、共享存储器以及隐式范例。我们还要研究做显式并行性表示用的MPI（消息传递接口）、PVM（并行虚拟机）、Fortran 90和HPF（高性能Fortran）以及做隐式并行性表示用的语言和编译器。

基于基准测试程序的评价

本书得益于我们对6台可扩展计算机开展基准程序测试的经验。这6台计算机是Maui高性能计算机中心的SP2，San Diego超级计算机中心的T3D/T3E和Paragon，Cray Eagan数据中心的T3D以及香港大学的SP2、SGI服务器和Pearl机群。

用第一手的基准程序测试结果对各种机器平台中的集合MPI通信进行了评估。根据NASA并行NAS和USC/HKU STAP基准程序测试结果，我们揭示了有关体系结构的含义。对这些基准程序性能测试结果进行了评估，并同时对机器规模和问题规模可扩展性做了分析。

Web资源

在瞬息万变的世界中，任何计算机书籍在几年后就会变得陈旧。通过选用实际的论题和讨论可持续几代计算机系统的基础问题，我们努力使本书有较长的生命周期。所给出的实例和定量数据均出自实际设计或基准程序的测试实验。

在本书的最后我们编译了大量的，可连接包括学术、商业和政府各部门的计算机公司、研究科研项目、信息技术中心以及主要应用组在内的几千个主页。在香港大学还保留有该表的在线总站。若要访问我们的主页，请参阅致教师/读者。

致谢

我们感谢本领域的6位顶级专家对我们草稿的专业评述。是他们的建议使我们的手稿得以修改成目前的形式和内容。我们感谢Choming Wang为本书所制订的索引以及王卓立博士维护在香港大学的本书的Web所给予的帮助。

与Dharma Agrawal, Jean-Loup Baer, Gordon Bell, David Culler, Jack Dongarra, Michael Flynn, Ian Foster, Jeffrey Fox, Mark Franklin, Wolfgang Giloi, Allan Gottlieb, Anoop Gupta, John Hennessy, Ken Kennedy, Duncan Lawrie, Charles Leiserson, Kai Li, Guojie Li, Lionel Ni, David Patterson, Gregory Pfister, John Rice, Sartaj Sahni, Chuck Seitz, Bruce Shriver, H.J.Siegel, Burton Smith, Daniel Tabak, H.C.Torng以及Ben Wah等人的思想交流令人鼓舞。

我们感谢McGraw-Hill图书公司Eric Munson、Lynn Cox以及Besty Jones各位编辑的支持。并对Richard DeVitto、Francis Owen以及Nina Kreiden所做的出版工作和香港大学Polly Leung的协助表示衷心感谢。

VIII

感谢在撰写本书过程中，麻省理工学院林肯实验室、香港研究批准委员会以及香港大学所给予的研究经费的支持。尤其要提到的是由香港大学所提供的精良设备和环境，使本书的作者得以非常愉快地从事本书的写作。

反馈信息地址

所有技术接触、提议、校误或信息交换，敬请读者和大学教师用以下E-mail与任一作者进行联系：

kaihwang@usc.edu zxu@ncic.ac.cn

我们十分感谢您的反馈信息并希望您能乐于阅读本书。

黄 铠

徐志伟

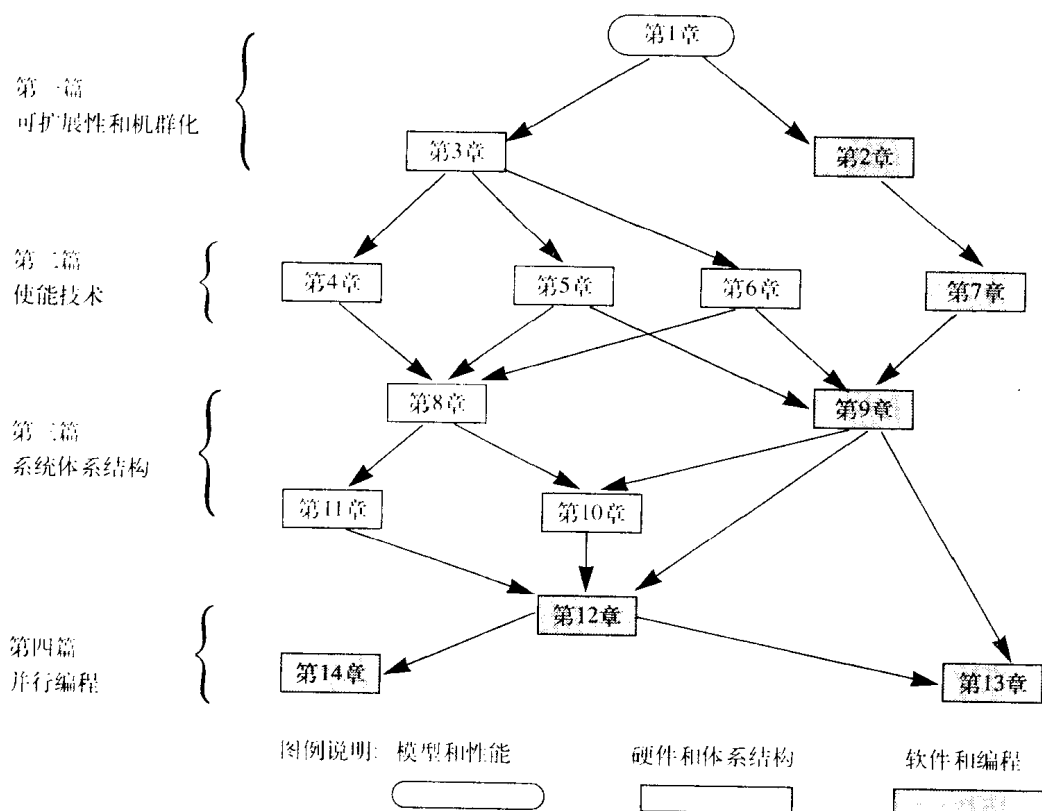
1997年11月15日

香港

致教师/读者

本书是专为学院/大学程度的计算机科学或计算机工程课程的课堂教学编写的标准教科书。适合的课程包括：计算机体系结构、并行处理、分布计算、并发编程、基于网络计算和计算机工程等。

阅读流程图 该流程图表明了本书中14章间的逻辑流程。在该图的左边标出了四篇的标题。本书中有两章是理论和模型化（带圆角的长方形），6章为硬件和体系结构（不带阴影长方形），而另6章则为软件和编程（带阴影长方形）。



课程设置

下面建议5种课程设置，授课均可采用本书，每门课持续一学期共45个学时：

- **计算机体系结构**：适合于电气和计算机工程专业及偏重硬件的学生，应包括第1、3~6和8~11各章内容。
- **并行编程**：适合于计算机专业及偏重编程和软件的学生，应包括第1、2、7~10和12~14各章内容。
- **并行处理**：适合于计算机科学/工程和电气工程专业硬、软件复合型的学生，应包括第1、2、4~6、8~10和12各章内容。

- **分布式计算**: 适合于计算机科学和电气工程复合专业的学生, 内容应包括适合于一学期使用的第1, 2, 5~7, 9~10和13~14各章。
- **计算机工程**: 是计算机技术和软件系统设计的高等课程。应包括第1, 2, 4~7和8~10各章内容。

所有学生应从第1章的内容开始阅读。工程类学生可多阅读些图中左边部分的有关技术和体系结构的章节内容, 而偏重软件的学生则应多阅读流程图右边子树中的章节内容。从逻辑关系来讲, 本书的阅读应自上而下按箭头指示的方向循序进行。

硬件和体系结构的6章内容形成了计算机体系结构或计算机工程的课程核心, 而软件和编程的6章内容则形成了并行编程课程的核心。并行处理课程包括了所有可扩展并行系统但较偏重于共享存储器多重处理。分布计算课程则较侧重于消息传递系统和基于网络的机群计算。

教师手册

采用本书的教师可获得教师手册。该手册提供了家庭作业习题的选择解答、视图作品、某些采样测试以及有关学期课程设计题目的建议。

教师可写信至: 机械工业出版社华章公司市场部: 北京百万庄南街1号 邮政编码 100037(Tel: 68326677-2829)传真68311602提交书面请求以获取该教师手册。

Web站点访问

作为新的参考书, 本书也可供系统设计人员、学术研究人员、应用程序员、系统分析员、资源管理员、解答提供人员以及一般的计算机专业人员自学使用。为防止内容变得陈旧, 敬请读者访问我们的Web站以获取已更新的WWW连接。

<http://www.cs.hku.hk/~kaihwang/book98.html>

该Web站将动态地更新。如果您希望您的机构或您的科研项目加入到连接表中, 请用以下E-mail与香港大学的C.L.Wang(王卓立)博士联系: clwang@cs.hku.hk

Web地址表

作为一本随时更新的参考书, 本书也考虑了一般系统设计人员、应用程序设计员、系统分析员、资源管理员、解答提供人员和计算机专业人员的自学需要。为避免落伍, 请读者访问我们时时更新的Web地址。

<http://www.cs.hku.hk/~kaihwng/book98.html>

此Web地址表动态更新。如果你想把你的组织或项目加在表上, 请与香港大学的C.L.Wang联系, Email地址: cl.wang@usc.edu。

目 录

译者序		
序		
前言		
致教师/读者		
 第一篇 可扩展性和机群化		
第1章 可扩展计算机平台和模型	2	
1.1 计算机体系结构演变	3	
1.1.1 计算机代	3	
1.1.2 可扩展计算机体系结构	3	
1.1.3 计算机系统体系结构发展的趋势	5	
1.2 可扩展性范围	6	
1.2.1 资源可扩展性	6	
1.2.2 应用可扩展性	7	
1.2.3 技术可扩展性	7	
1.3 并行计算机模型	8	
1.3.1 语义属性	9	
1.3.2 性能属性	11	
1.3.3 抽象机模型	12	
1.3.4 物理机模型	16	
1.4 机群化的基本概念	19	
1.4.1 机群特征	19	
1.4.2 体系结构的比较	20	
1.4.3 机群的优越性和难点	21	
1.5 可扩展设计原理	23	
1.5.1 独立原理	24	
1.5.2 平衡设计原理	25	
1.5.3 可扩展性设计	28	
1.6 参考文献注释和习题	30	
第2章 并行编程基础	33	
2.1 并行编程综述	33	
2.1.1 并行编程缘何艰难	33	
2.1.2 并行编程环境	35	
2.1.3 并行编程方法	36	
2.2 进程、任务和线程	38	
2.2.1 抽象进程的定义	38	
2.2.2 执行方式	40	
2.2.3 地址空间	40	
2.2.4 进程现场	42	
2.2.5 进程描述符	42	
2.2.6 进程控制	43	
2.2.7 进程的变异	45	
2.3 并行性问题	46	
2.3.1 进程中的同构性	46	
2.3.2 静态和动态并行性	48	
2.3.3 进程编组	48	
2.3.4 分配问题	49	
2.4 交互/通信问题	50	
2.4.1 交互操作	50	
2.4.2 交互方式	52	
2.4.3 交互模式	53	
2.4.4 合作和竞争交互	54	
2.5 并行程序中的语义问题	55	
2.5.1 程序的终止	55	
2.5.2 程序的确定性	55	
2.6 参考文献注释和习题	56	
第3章 性能指标和基准程序	59	
3.1 系统和应用的基准程序	59	
3.1.1 微基准程序	60	
3.1.2 并行计算的基准程序	62	
3.1.3 商业和TPC基准程序	64	
3.1.4 SPEC基准程序系列	65	
3.2 性能与成本	66	
3.2.1 执行时间和吞吐率	67	
3.2.2 利用率和成本有效性	68	
3.3 基本性能指标	70	

3.3.1 工作负载和速度指标	70
3.3.2 防止对顺序性能误解的说明	72
3.4 并行计算机性能	73
3.4.1 计算特征	73
3.4.2 并行性和交互开销	75
3.4.3 开销定量化	76
3.5 并程序性能	82
3.5.1 性能指标	82
3.5.2 基准程序中的可用并行性	85
3.6 可扩展性和加速比分析	86
3.6.1 Amdahl定律: 固定问题规模	87
3.6.2 Gustafson定律: 固定时间	88
3.6.3 Sun和Ni定律: 存储器受限	90
3.6.4 等性能模型	93
3.7 参考文献注释和习题	95

第二篇 使能技术

第4章 微处理器构件	100
4.1 系统发展趋向	100
4.1.1 硬件进展	100
4.1.2 软件进展	102
4.1.3 应用进展	103
4.2 处理器设计原理	105
4.2.1 指令流水线基理	105
4.2.2 从CISC到RISC及进一步延伸	108
4.2.3 体系结构性能的增强方法	111
4.3 微处理器体系结构系列	112
4.3.1 主要的体系结构系列	112
4.3.2 超标量和超流水处理器	113
4.3.3 嵌入式微处理器	116
4.4 微处理器的实例研究	117
4.4.1 Digital的Alpha 21164微处理器	117
4.4.2 Intel高能奔腾处理器	120
4.5 后RISC、多媒体和VLIW	124
4.5.1 后RISC处理器特征	124
4.5.2 多媒体扩展	126
4.5.3 VLIW体系结构	129
4.6 微处理器的未来	130
4.6.1 硬件发展趋向和物理极限	130

4.6.2 未来的工作负载和挑战	131
4.6.3 未来微处理器的体系结构	132
4.7 参考文献注释和习题	134
第5章 分布式存储器和时延容忍	137
5.1 层次存储器技术	137
5.1.1 存储部件特性	137
5.1.2 存储器层次性质	139
5.1.3 存储器容量的规划	141
5.2 高速缓存一致性协议	142
5.2.1 高速缓存一致性问题	143
5.2.2 监听一致性协议	144
5.2.3 MESI监听协议	145
5.3 共享存储器一致性	148
5.3.1 存储器事件排序	148
5.3.2 存储器一致性模型	150
5.3.3 非严格的存储器模型	151
5.4 分布式高速缓存/主存体系结构	153
5.4.1 NORMA、NUMA、COMA 和DSM模型	153
5.4.2 基于目录的一致性协议	158
5.4.3 斯坦福Dash多处理机	159
5.4.4 Dash中基于目录的协议	161
5.5 时延容忍技术	163
5.5.1 时延的避免、减小和隐藏	163
5.5.2 分布式一致性高速缓存	164
5.5.3 数据预取策略	165
5.5.4 非严格的存储器一致性的效果	167
5.6 多线程时延隐藏	167
5.6.1 多线程处理机模型	167
5.6.2 现场切换策略	169
5.6.3 组合时延隐藏机制	172
5.7 参考文献注释和习题	173
第6章 系统互连和千兆位网络	178
6.1 互连网络基础	178
6.1.1 互连环境	178
6.1.2 网络部件	180
6.1.3 网络特征	181
6.1.4 网络性能指标	182
6.2 网络拓扑结构和性质	183

6.2.1 拓扑结构和功能性质	183
6.2.2 路由方案和功能	184
6.2.3 网络拓扑结构	187
6.3 总线、纵横交叉开关和多级开关	191
6.3.1 多处理机总线	191
6.3.2 纵横交叉开关	193
6.3.3 多级互连网络	195
6.3.4 开关互连比较	197
6.4 千兆位网络技术	199
6.4.1 光纤通道和FDDI环	199
6.4.2 快速以太网和千兆位以太网	201
6.4.3 构造SAN/LAN的Myrinet网	203
6.4.4 HiPPI和超级HiPPI	204
6.5 ATM交换器和网络	207
6.5.1 ATM技术	207
6.5.2 ATM网络接口	208
6.5.3 ATM四层体系结构	209
6.5.4 ATM互连网络连接性能	211
6.6 可扩展的一致性接口	212
6.6.1 SCI互连	213
6.6.2 实现问题	214
6.6.3 SCI一致性协议	216
6.7 网络技术比较	217
6.7.1 标准网络及其发展前景	217
6.7.2 网络性能和应用	218
6.8 参考文献注释和习题	219
第7章 线程化、同步和通信	223
7.1 软件多线程化	223
7.1.1 线程概念	224
7.1.2 线程管理	225
7.1.3 线程同步	226
7.2 同步机制	227
7.2.1 原子性和互斥	227
7.2.2 高级同步结构	230
7.2.3 低级同步原语	234
7.2.4 快速锁机制	237
7.3 TCP/IP通信协议组	239
7.3.1 TCP/IP协议组的特性	239
7.3.2 UDP、TCP和IP	241

7.3.3 Sockets接口	244
7.4 快速和有效通信	245
7.4.1 通信中的关键问题	246
7.4.2 LogP通信模型	250
7.4.3 低级通信支持	251
7.4.4 通信算法	257
7.5 参考文献注释和习题	259

第三篇 系统体系结构

第8章 对称多处理机和CC-NUMA 多处理机	264
8.1 SMP和CC-NUMA技术	264
8.1.1 多处理机体系结构	264
8.1.2 商品化SMP服务器	268
8.1.3 英特尔SHV服务器电路板	269
8.2 SUN Ultra Enterprise 10000系统	270
8.2.1 Ultra-E10000系统	271
8.2.2 系统电路板的结构	272
8.2.3 可扩展性和可用性支持	273
8.2.4 动态域和性能	274
8.3 HP/Convex Exemplar X-Class	275
8.3.1 Exemplar X系统的体系结构	275
8.3.2 Exemplar软件环境	277
8.4 Sequent NUMA-Q 2000	278
8.4.1 NUMA-Q 2000的体系结构	278
8.4.2 NUMA-Q的软件环境	281
8.4.3 NUMA-Q的性能	282
8.5 SGI/Cray Origin 2000超级服务器	284
8.5.1 Origin 2000系列的设计目标	284
8.5.2 Origin 2000的体系结构	285
8.5.3 Cellular IRIX环境	290
8.5.4 Origin 2000的性能	293
8.6 CC-NUMA 体系结构的比较	294
8.7 参考文献注释和习题	296
第9章 机群化和可用性支持	298
9.1 构造机群的挑战性	298
9.1.1 机群的分类	298
9.1.2 机群的体系结构	299

9.1.3 机群设计要考虑的问题	300	10.5 Berkeley NOW研究项目	356
9.2 对机群可用性的支持	302	10.5.1 适合快速通信的主动消息	357
9.2.1 可用性概念	302	10.5.2 适合于全局资源管理的GLUnix	360
9.2.2 可用性技术	304	10.5.3 xFS无服务器网络文件系统	361
9.2.3 检查点和故障恢复	308	10.6 TreadMarks: 用软件实现的	
9.3 对单一系统映像的支持	311	DSM机群	366
9.3.1 单一系统映像层	311	10.6.1 边界条件	366
9.3.2 单一入口和单一文件层次	312	10.6.2 DSM中的用户接口	367
9.3.3 单一I/O、网络化和存储空间	316	10.6.3 实现问题	368
9.4 Solaris MC中的单一系统映像	317	10.7 参考文献注释和习题	369
9.4.1 全局文件系统	318	第11章 MPP的体系结构和性能	371
9.4.2 全局进程管理	318	11.1 MPP技术概论	371
9.4.3 单一I/O系统映像	319	11.1.1 MPP特性和要点	371
9.5 机群的作业管理	320	11.1.2 MPP系统概述	373
9.5.1 作业管理系统	320	11.2 Cray T3E系统	374
9.5.2 作业管理系统综述	324	11.2.1 T3E的体系结构	374
9.5.3 负载共享工具(LSF)	326	11.2.2 T3E的系统软件	376
9.6 参考文献注释和习题	331	11.3 新一代ASCI/MPP系统	376
第10章 服务器和 workstation 机群	334	11.3.1 ASCI可扩展设计策略	377
10.1 机群产品和研究项目	334	11.3.2 硬件和软件需求	378
10.1.1 支持机群产品的潮流	334	11.3.3 定约的ASCI/MPP平台	379
10.1.2 SMP服务器机群	336	11.4 Intel/Sandia ASCI Option Red	380
10.1.3 机群研究项目	336	11.4.1 Option Red的体系结构	380
10.2 面向NT机群的微软Wolfpack	338	11.4.2 Option Red的系统软件	382
10.2.1 微软Wolfpack配置	338	11.5 并行NAS基准程序测试结果	384
10.2.2 热备份多服务器机群	339	11.5.1 NAS并行基准测试程序	384
10.2.3 主动可用性机群	339	11.5.2 超步结构和颗粒度	385
10.2.4 容错多服务器机群	341	11.5.3 主存、I/O和通信	386
10.3 IBM SP系统	341	11.6 MPI和STAP基准程序测试结果	387
10.3.1 设计目标和策略	341	11.6.1 MPI性能测试	388
10.3.2 SP2系统的体系结构	343	11.6.2 MPI时延和总计(聚集)带宽	389
10.3.3 I/O和网络互连	345	11.6.3 MPP的STAP基准程序测试	
10.3.4 SP系统软件	347	评估	391
10.3.5 SP2及其前景	349	11.6.4 MPP体系结构的含义	394
10.4 Digital TruCluster	350	11.7 参考文献注释和习题	396
10.4.1 TruCluster体系结构	350	第四篇 并行编程	
10.4.2 存储器通道互连	352	第12章 并行范例和编程模型	399
10.4.3 TruCluster编程	354	12.1 范例和可编程性	400
10.4.4 TruCluster系统软件	356		

12.1.1 算法范例	400	13.2.4 集合MPI通信	445
12.1.2 可编程性问题	402	13.2.5 MPI-2扩展	448
12.1.3 并行编程举例	402	13.3 并行虚拟机	450
12.2 并行编程模型	405	13.3.1 虚拟机结构	451
12.2.1 蕴式并行性	405	13.3.2 PVM中的进程管理	453
12.2.2 显式并行模型	407	13.3.3 用PVM进行通信	455
12.2.3 四种模型的比较	410	13.4 参考文献注释和习题	458
12.2.4 其他并行编程模型	412	第14章 数据并行编程	462
12.3 共享存储器编程	413	14.1 数据并行模型	462
12.3.1 ANSI X3H5共享存储器模型	413	14.2 Fortran 90方法	462
12.3.2 POSIX线程模型	416	14.2.1 并行数组操作	463
12.3.3 OpenMP标准	417	14.2.2 Fortran 90中的本征函数	464
12.3.4 SGI Power C模型	420	14.3 高性能Fortran	466
12.3.5 C++ 一种结构化的并行C语言	422	14.3.1 对数据并行性的支持	466
12.4 参考文献注释和习题	426	14.3.2 HPF中的数据映射	468
第13章 消息传递编程	429	14.3.3 对Fortran 90和HPF的总结	472
13.1 消息传递范例	429	14.4 其他的数据并行方法	474
13.1.1 消息传递库	429	14.4.1 Fortran 95和Fortran 2001	475
13.1.2 消息传递方式	430	14.4.2 PC++和Nesl方法	477
13.2 消息传递接口	432	14.5 参考文献注释和习题	480
13.2.1 MPI消息	434	参考文献	482
13.2.2 MPI中的消息信封	439	Web地址表	510
13.2.3 点对点通信	443		

第一篇 可扩展性和机群化

1.1 目的

本篇介绍有关并行和分布计算系统可扩展性和可编程性的原理。目的在于为读者学习以后各章打下必要的基础。

第1章 本章将建方可扩展计算机平台的模型。我们首先对可扩展性概念、按资源、应用和技术三个正交轴加以定义。然后我们描述三个抽象机模型——PRAM、BSP和阶段并行模型以及5种物理机模型——PVP、SMP、MPP、COW和MPP系统的特征。可扩展性设计原理将通过举例加以介绍，目的在于使系统设计能独立于技术、体系结构、算法、语言、应用和所使用的网络环境。有关平衡设计、过度设计和向后兼容性的基本概念也将在本章加以叙述。

第2章 这一章专门叙述可扩展并行计算机的编程。除了叙述有关进程、任务、线程和环境的基本概念外，还覆盖了并行性管理、进程交互、程序语义、算法范例和软件可移植性中的所有重要问题。

将介绍下列四种并行编程模型，有关这些模型的细节将在第四篇中介绍。

- 并行化编译器模型
- 数据并行模型
- 消息传递模型
- 共享存储器模型

第3章 这一章将涉及基本性能的基准测试程序和指标。其目的是识别有关可扩展性能的属性。我们首先对并行基准测试程序组进行综合介绍，然后详尽叙述兼顾性能和成本的折衷方法。此外还将指明如何防止对顺序程序执行的误解。

对有关并行性管理和软件交互作用的开销进行了定量分析。对颗粒度大小、可用并行性、并行性能指标、Amdahl定律、Gustafson定律、Sun和Ni定律以及各种等性能模型都用具有说明性的基准测试程序测试结果进行了定量分析。

1.2 读者注意

第1章必须首先阅读。这对前言中所建议的所有四种可能的课程设置均为如此。

第2章必须在侧重软件的第7、9、12、13和14章前阅读。对于侧重硬件的读者，在第一次阅读时可以跳过这几章。

第3章内容将对理解出现在第4、5、6、8、10和11各章中涉及性能的内容颇有帮助。

将本书作为导论课的主修计算机科学和电气工程混合专业的学生，在第一次阅读时可跳过第3章。

然而对于侧重研究的学生来讲，只要他所选择的研究题目与系统性能有关，则将发现第3章的内容特别有用。