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

计算机科学丛书

可扩展并行计算

技术、结构与编程

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

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。

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 彬

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

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及工程系

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、单一系统映象、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。

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

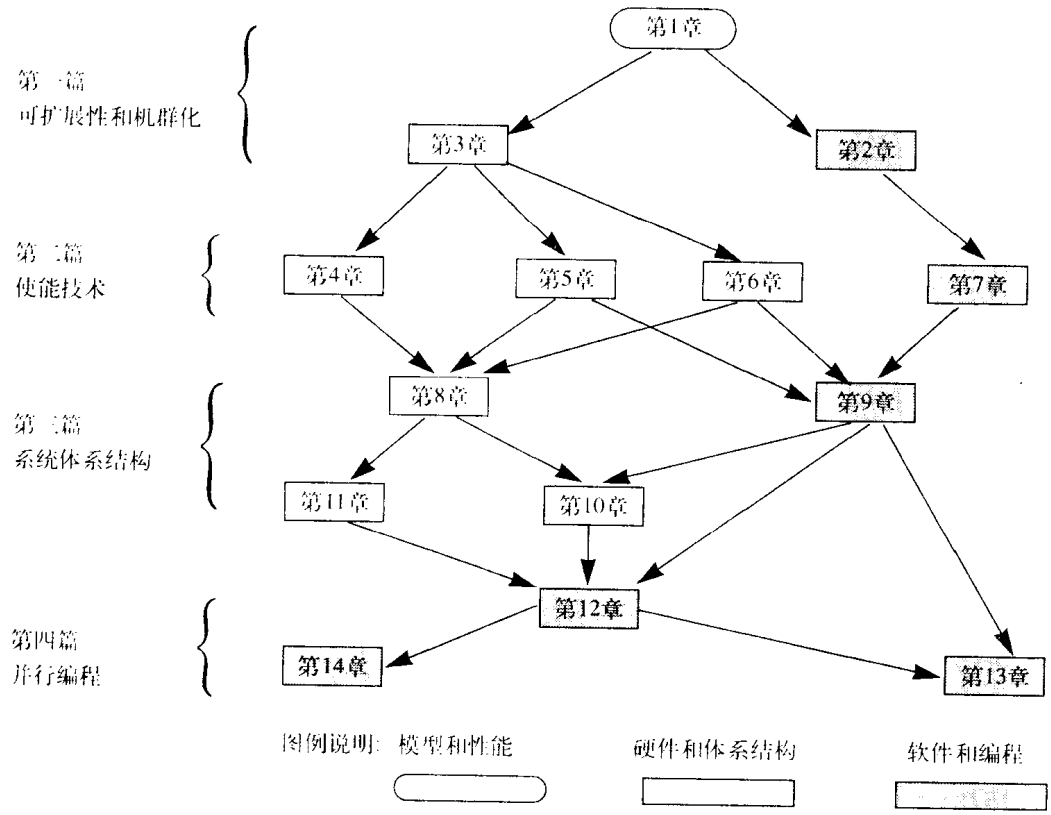
我们十分感谢您的反馈信息并希望您能乐于阅读本书。

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（带圆角的长方块），6章为硬件和体系结构（不带阴影长方块），而另6章则为软件和编程（带阴影长方块）。



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**：适合于计算机科学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

第二篇 使能技术

第三篇 系统体系结构

第四篇 并行编程

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　本章将建立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。