国外著名高等院校
信息科学与技术优秀教材

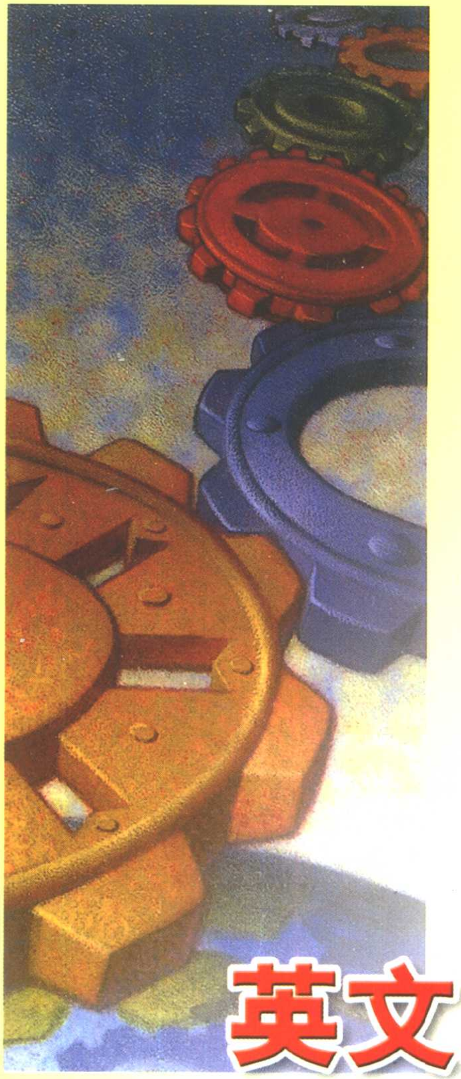
PTR
PH

高性能集群计算：结构与系统（第一卷）

High Performance Cluster Computing: Architectures and Systems

VOLUME 1

Rajkumar Buyya



英文版

人民邮电出版社
POSTS & TELECOMMUNICATIONS PRESS

国外著名高等院校信息科学与技术优秀教

高性能集群计算：结构与系统

（第一卷）

（英文版）

High Performance Cluster Computing: Architectures and Systems, Volume 1

Rajkumar Buyya

人民邮电出版社

图书在版编目（CIP）数据

高性能集群计算. 第1卷，结构与系统 / （美）布亚（Buyya, R.）著.
—北京：人民邮电出版社，2002.8
国外著名高等院校信息科学与技术优秀教材
ISBN 7-115-10346-1

I. 高… II. 布… III. 集群—计算—高等学校—教材—英文 IV. TP301.6

中国版本图书馆 CIP 数据核字（2002）第 049560 号

版权声明

国外著名高等院校信息科学与技术优秀教材

高性能集群计算：结构与系统（第一卷）（英文版）

◆ 著 Rajkumar Buyya
责任编辑 李 际

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街 14 号
邮编 100061 电子函件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
读者热线 010-67180876
北京汉魂图文设计有限公司制作
北京朝阳展望印刷厂印刷
新华书店总店北京发行所经销

◆ 开本：800×1000 1/16
印张：55.25
字数：1233 千字 2002 年 8 月第 1 版
印数：1-3 000 册 2002 年 8 月北京第 1 次印刷

著作权合同登记 图字：01－2002－3751 号

ISBN 7-115-10346-1/TP・2905

定价：75.00 元

本书如有印装质量问题，请与本社联系 电话：(010)67129223

内 容 提 要

本书是一本覆盖面非常广的专著，内容包含了有关集群计算的体系结构、网络、协议和I/O、进程调度、资源共享和负载平衡，以及目前典型的集群系统剖析。其中每章都是由该研究领域的国际最知名的专家撰写，因而具有非常高的学术价值和学术指导意义。

本书聚集了高性能集群计算领域中 100 多位资深的从业者所做出的贡献。实质上，对该领域中每一个与系统相关的关键问题本书都提供了最新的信息。在高性能并行计算领域中，无论您是一位开发者、研究者、管理员、教师、学生，还是一个管理者，本书都是一本难得的经典书籍。

出版说明

2001 年，教育部印发了《关于“十五”期间普通高等教育教材建设与改革的意见》。该文件明确指出，“九五”期间原国家教委在“抓好重点教材，全面提高质量”方针指导下，调动了各方面的积极性，产生了一大批具有改革特色的新教材。然而随着科学技术的飞速发展，目前高校教材建设工作仍滞后于教学改革的实践，一些教材内容陈旧，不能满足按新的专业目录修订的教学计划和课程设置的需要。为此该文件明确强调，要加强国外教材的引进工作。当前，引进的重点是信息科学与技术和生物科学与技术两大学科的教材。要根据专业（课程）建设的需要，通过深入调查、专家论证，引进国外优秀教材。要注意引进教材的系统配套，加强对引进教材的宣传，促进引进教材的使用和推广。

邓小平同志早在 1977 年就明确指出：“要引进外国教材，吸收外国教材中有益的东西。”随着我国加入 WTO，信息产业的国际竞争将日趋激烈，我们必须尽快培养出大批具有国际竞争能力的高水平信息技术人才。教材是一个很关键的问题，国外的一些优秀教材不但内容新，而且还提供了很多新的研究方法和思考方式。引进国外原版教材，可以促进我国教学水平的提高，提高学生的英语水平和学习能力，保证我们培养出的学生具有国际水准。

为了贯彻中央“科教兴国”的方针，配合国内高等教育教材建设的需要，人民邮电出版社约请有关专家反复论证，与国外知名的教材出版公司合作，陆续引进一些信息科学与技术优秀教材。第一批教材针对计算机专业的主干核心课程，是国外著名高等院校所采用的教材，教材的作者都是在相关领域享有盛名的专家教授。这些教材内容新，反映了计算机科学技术的最新发展，对全面提高我国信息科学与技术的教学水平必将起到巨大的推动作用。

出版国外著名高等院校信息科学与技术优秀教材的工作将是一个长期的、坚持不懈的过程，我社网站（www.ptpress.com.cn）上介绍了我们陆续推出的图书的详细情况，敬请关注。希望广大教师和学生将使用中的意见和建议及时反馈给我们，我们将根据您的反馈不断改进我们的工作，推出更多更好的引进版信息科学与技术教材。

人民邮电出版社
2001 年 12 月

序　　言

集群技术就是采用通用的计算机硬件部件和常用的（甚至是免费的）软件，来构造高性能的计算平台和服务平台，解决大规模科学计算、工程领域和商业应用。集群技术的发展得益于计算机领域许多关键技术的飞速发展和突破，包括廉价且高性能的微处理芯片、高速网络以及用于高性能分布式计算的标准软件。当然集群技术的兴起也是科学计算和商业应用领域对计算能力的迫切需求。目前集群技术可广泛地应用于包括超级计算、高可靠应用、超级网络服务器、电子商务、高性能数据库应用等在内的许多领域。美国国会曾经制定过一个限制高性能计算出口的规定，但由于集群技术的飞速发展和广泛应用，使得这个规定形同虚设，现在采用集群技术非常容易构造每秒运行几百万亿次、几千万亿次，甚至几万万亿次的超级计算机。目前世界上几乎所有的国家，尤其是发展中国家，都采用集群技术作为高性能计算的平台。

同时集群技术也为高等院校在计算机专业普及并行处理知识，提供了极其便利的平台。多年来，许多高等院校的计算机专业由于缺少教学和实验平台，使得学生在学习并行处理或并行程序设计课程的时候，往往只有书本上的概念，而缺乏感性的认识。集群技术为这些高等院校提供了一个廉价而具体的实验平台，学校不必因为需要提供实验平台而投资上百万元的资金去购买超级计算机。

Rajkumar Buyya 博士编写的这本《高性能集群计算》第一卷，是一本覆盖面非常广泛的学术专著，所写内容包含了有关集群计算的体系结构、网络、协议和 I/O、进程调度、资源共享和负载平衡，以及目前典型的集群系统剖析。其中每个章节都是由该研究领域的国际最知名的专家撰写，因而具有非常高的学术价值和学术指导意义。

我作为 Rajkumar Buyya 博士的朋友和合作伙伴，在该书发行前便得到 Rajkumar Buyya 博士赠送的样书，随后我在美国南加州大学和华中科技大学采用该书作为研究生并行处理课程的教材，取得了非常好的效果。我还与 Rajkumar Buyya 博士一起撰写了有关在高等院校进行集群计算课程教学的学术论文，并发表在第一届 IEEE/ACM 集群计算和网格计算国际会议 (CCGrid 2001)的论文集上。有关该书各章节的教学幻灯片（Powerpoint 文件），可以在 Rajkumar Buyya 博士的个人网页上得到（http://www.cs.mu.oz.au/~raj/）。同时 Rajkumar Buyya 博士的个人网站上还提供了大量最新的有关集群计算的资料、演讲稿、

文章和软件信息。相信这本书以及相关的资料，能够给每位对集群计算有兴趣的研究人员和学生带来丰富的理论和实践知识。

金海

教授　博导
华中科技大学计算机学院
2002 年 6 月 20 日

Preface

The initial idea leading to cluster[1] computing was developed in the 1960s by IBM as a way of linking large mainframes to provide a cost-effective form of commercial parallelism. During those days, IBM's HASP (Houston Automatic Spooling Priority) system and its successor, JES (Job Entry System), provided a way of distributing work to a user-constructed mainframe cluster. IBM still supports clustering of mainframes through their Parallel Sysplex system, which allows the hardware, operating system, middleware, and system management software to provide dramatic performance and cost improvements while permitting large mainframe users to continue to run their existing applications.

However, cluster computing did not gain momentum until three trends converged in the 1980s: high performance microprocessors, high-speed networks, and standard tools for high performance distributed computing. A possible fourth trend is the increased need of computing power for computational science and commercial applications coupled with the high cost and low accessibility of traditional supercomputers. These building blocks are also known as killer-microprocessors, killer-networks, killer-tools, and killer-applications, respectively. The recent advances in these technologies and their availability as cheap and commodity components are making clusters or networks of computers (PCs, workstations, and SMPs) an appealing vehicle for cost-effective parallel computing. Clusters, built using commodity-off-the-shelf (COTS) hardware components as well as free, or commonly used, software, are playing a major role in redefining the concept of supercomputing.

The trend in parallel computing is to move away from specialized traditional supercomputing platforms, such as the Cray/SGI T3E, to cheaper and general purpose systems consisting of loosely coupled components built up from single or multiprocessor PCs or workstations. This approach has a number of advantages, including being able to build a platform for a given budget which is suitable for a large class of applications and workloads.

This book is motivated by the fact that parallel computing on a network of computers using commodity components has received increased attention recently, and noticeable progress towards usable systems has been made. A number of re-

[1] Cluster is a collection of interconnected computers working together as a single system.

searchers in academia and industry have been active in this field of research. Although research in this area is still in its early stage, promising results have been demonstrated by experimental systems built in academic and industrial laboratories. There is a need for better understanding of what cluster computing can offer, how cluster computers can be constructed, and what the impacts of clustering on high performance computing will be.

Though a significant number of research articles have been published in various conference proceedings and journals, the results are scattered in many places, are hard to obtain, and are difficult to understand, especially for beginners. This book, the first of its kind, gathers in one place the current and comprehensive technical coverage of the field and presents it in a tutorial form. The book's coverage reflects the state-of-the-art in high-level architecture, design, and development, and points out possible directions for further research and development.

Organization

This book is a collection of chapters written by leading scientists active in the area of parallel computing using networked computers. The primary purpose of the book is to provide an authoritative overview of this field's state-of-the-art. The emphasis is on the following aspects of cluster computing:

- Requirements, Issues, and Services
- System Area Networks, Communication Protocols, and High Performance I/O Techniques
- Resource Management, Scheduling, Load Balancing, and System Availability
- Possible Models for Cluster-based Parallel Systems
- Programming Models and Environments
- Algorithms and Applications of Clusters

The work on High Performance Cluster Computing appears in two volumes:

- Volume 1: Systems and Architectures
- Volume 2: Programming and Applications

This book, Volume 1, consists of 36 chapters, which are grouped into the following four parts:

- Part I: Requirements and General Issues
- Part II: Networking, Protocols, and I/O
- Part III: Process Scheduling, Load Sharing, and Balancing

- Part IV: Representative Cluster Systems

Part I focuses on cluster computing requirements and issues related to components, single system image, high performance, high availability, scalability, deployment, administration, and wide-area computing. Part II covers system area networks, light-weight communication protocols, and I/O. Part III discusses techniques and algorithms of process scheduling, migration, and load balancing along with representative systems. Part IV covers system architectures of some of the popular academic and commercial cluster-based systems such as Beowulf and SP/2.

Readership

The book is primarily written for graduate students and researchers interested in the area of parallel and distributed computing. However, it is also suitable for practitioners in industry and government laboratories.

The interdisciplinary nature of the book is likely to appeal to a wide audience. They will find this book to be a valuable source of information on recent advances and future directions of parallel computation using networked computers. This is the first book addressing various technological aspects of cluster computing indepth, and we expect that the book will be an informative and useful reference in this new and fast growing research area.

The organization of this book makes it particularly useful for graduate courses. It can be used as a text for a research-oriented or seminar-based advanced graduate course. Graduate students will find the material covered by this book to be stimulating and inspiring. Using this book, they can identify interesting and important research topics for their Master's and Ph.D. work. It can also serve as a supplementary book for regular courses, taught in Computer Science, Computer Engineering, Electrical Engineering, and Computational Science and Informatics Departments, including:

- Advanced Computer Architecture
- Advances in Networking Technologies
- High Performance Distributed Computing
- Distributed and Concurrent Systems
- High Performance Computing
- Parallel Computing
- Networked Computing
- Trends in Distributed Operating Systems
- Cluster Computing and their Architecture.

Cluster Computing Resources on the Web

The various software systems discussed in this book are freely available for download through the Internet. Please visit this book's website,

- http://www.phptr.com/ptrbooks/ptr_0130137847.html

for pointers/links to further information on downloading Educational Resources, Cluster Computing Environments, and Cluster Management Systems.

Acknowledgments

First and foremost, I am grateful to all the contributing authors for their time, effort, and understanding during the preparation of the book.

I thank Albert Zomaya (University of Western Australia) for his advice and encouragement while starting this book project.

I would like to thank Kennith Birman (Cornell University), Marcin Paprzycki (University of Southern Mississippi), Hamid R. Arabnia (The University of Georgia), and Gregory F. Pfister (IBM Server Group) for their critical comments and suggestions on improving the book.

I thank Toni Cortes (Universitat Politecnica de Catalunya) for his consistent support and invaluable LaTeX expertise.

I thank Mark Baker (University of Portsmouth), Erich Schikuta (Universitaet Wien), Dror G. Feitelson (Hebrew University of Jerusalem), Daniel F. Savarese and Thomas Sterling (California Institute of Technology), Ira Pramanick (Silicon Graphics Inc), and Daniel S. Katz (Jet Propulsion Laboratory, California Institute of Technology) for writing overviews for various parts of the book.

I thank my wife, Smrithi, and my daughter, Soumya, for their love and understanding (my long absences from home) during the preparation of the book.

I acknowledge the support of the Australian Government Overseas Postgraduate Research Scholarship, the Queensland University of Technology Postgraduate Research Award (Programming Languages and Systems Research Centre Scholarship), the Monash University Graduate Scholarship, the Monash Departmental Scholarship, and the Distributed Systems and Software Engineering Centre Scholarship.

I thank Clemens Szyperski (Queensland University of Technology) and David Abramson (Monash University) for advising my Ph.D research program.

Finally, I would like to thank the staff at Prentice Hall, particularly Greg Doench, Mary Treacy, Joan L. McNamara, Barbara Cotton, Mary Loudin, Lisa Iarkowski, and Bryan Gambrel. They were wonderful to work with!

Rajkumar Buyya
Monash University, Melbourne, Australia
(rajkumar@dgs.monash.edu.au / rajkumar@ieee.org)

February, 1999

Contents at a Glance

Contents