

国外计算机科学教材系列

高性能集群计算：编程与应用

(第二卷)

High Performance Cluster Computing
Programming and Applications
Volume 2

[美] Rajkumar Buyya 编

郑纬民 汪东升 石 威 等译

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

本书全面系统地介绍了高性能集群计算技术的概念、理论研究和工程实践成果。全书由两卷构成:系统与结构(第一卷)和编程与应用(第二卷)。在系统与结构卷中,主要论述了集群计算的需求和一般问题,网络、协议与 I/O 技术,进程调度、负载共享与平衡技术和有代表性的集群计算机系统等。在编程与应用卷中,主要介绍编程环境与开发工具,Java 在高性能计算中的应用、算法与应用等问题。本书内容丰富,既介绍了当今科学家在集群计算领域的最新成果,同时也介绍了集群计算在科学与工程计算及信息服务领域中应用的大量实践经验。本书既可以作为高等院校计算机专业高年级学生和研究生教科书或参考书,也可以供利用并行计算机从事科学与工程计算和信息服务的科技人员学习参考。

Authorized translation from the English language edition published by Prentice - Hall, Inc. Copyright © 1999.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from the Publisher.

Simplified Chinese language edition published by Publishing House of Electronics Industry, Copyright © 2001.

本书中文简体版专有翻译出版权由 Pearson 教育集团所属的 Prentice - Hall, Inc. 授予电子工业出版社。其原文版权及中文翻译出版权受法律保护。未经许可,不得以任何形式或手段复制或抄袭本书内容。

图书在版编目(CIP)数据

高性能集群计算. 第二卷, 编程与应用/(美)布亚(Buyya, R.)著;郑纬民等译.

—北京:电子工业出版社, 2001.7

(国外计算机科学教材系列)

ISBN 7-5053-6785-4

I. 高... II. ①布... ②郑... III. ①电子计算机-集群-计算方法 ②集群-计算-程序设计
IV. TP301.6

中国版本图书馆 CIP 数据核字(2001)第 040086 号

书 名: 高性能集群计算: 编程与应用(第二卷)

原 书 名: High Performance Cluster Computing: Programming and Applications, Volume 2

编 者: [美] Rajkumar Buyya

译 者: 郑纬民 汪东升 石 威 等

责任编辑: 周宏敏

排版制作: 电子工业出版社计算机排版室监制

印 刷 者: 北京东光印刷厂

出版发行: 电子工业出版社 URL: <http://www.phei.com.cn>

北京市海淀区万寿路 173 信箱 邮编 100036

经 销: 各地新华书店

开 本: 787×1092 1/16 印张: 29.5 字数: 755 千字

版 次: 2001 年 7 月第 1 版 2001 年 7 月第 1 次印刷

书 号: ISBN 7-5053-6785-4
TP·3814

定 价: 48.00 元

版权贸易合同登记号 图字: 01-2000-4288

凡购买电子工业出版社的图书, 如有缺页、倒页、脱页、所附磁盘或光盘有问题者, 请向购买书店调换。

若书店售缺, 请与本社发行部联系调换。联系电话: 88211980 68279077

出版说明

随着 21 世纪的到来, 计算机技术的发展更加迅猛, 在各行各业的应用更加广泛, 越来越多的高等院校增设了有关计算机科学的课程内容, 或对现有计算机课程设置进行了适当调整, 以紧跟前沿技术。在这个教学体系和学科结构变革的大环境下, 对适合不同院系、不同专业、不同层次的教材的需求量与日俱增。此时, 如果能够借鉴、学习国外一流大学的先进教学体系, 引进具有先进性、实用性和权威性的国外一流大学计算机教材, 汲取其精华, 必能更好地促进中国高等院校教学的全面改革。

美国 Prentice Hall 出版公司是享誉世界的高校教材出版商, 自 1913 年成立以来, 一直致力于教材的出版, 所出版的计算机教材为美国众多大学采用, 其中有不少是专业领域中的经典名著, 已翻译成多种文字在世界各地的大学中使用, 成为全人类的共同财富。许多蜚声世界的教授、学者都是该公司的资深作者, 如道格拉斯·科默 (Douglas E. Comer)、威廉·斯大林 (William Stallings) 等。早在 1997 年, 电子工业出版社就从 Prentice Hall 引进了一套计算机英文版专业教材, 并将其翻译出版, 同时定名为《国外计算机科学教材系列》(下称: 第一轮教材)。截至 2000 年 12 月, 该系列教材已出版 23 种, 深受读者欢迎, 被许多大学选为高年级学生和研究生教材或参考书。

4 年过去了, 已出版的教材中多数已经有了后续版本。因此, 我们开始设计新一轮教材 (第二轮教材) 的出版, 成立了由我国计算机界著名专家和教授组成的“教材出版委员会”, 并结合第一轮教材的使用情况和师生反馈意见, 组织了第二轮《国外计算机科学教材系列》出版工作。

第二轮教材的出版原则为:

1. 引进 Prentice Hall 出版公司 2000 年和 2001 年推出的新版教材, 作为替换版本。
2. 在著名高校教授的建议下, 除了从 Prentice Hall 新选了一些教材之外, 还从 McGraw-Hill 和 Addison Wesley Longman 等著名专业教材出版社、麻省理工学院出版社和剑桥大学出版社等著名大学出版社引进了一些经典教材, 作为增补版本。
3. 对于第一轮中无新版本的优秀教材, 我们将其作为沿用版本, 直接进入第二轮使用。
4. 对于第一轮中翻译质量较好且无新版本的教材, 我们将其进行了修订, 也作为沿用版本, 进入第二轮使用。

这次推出的教材覆盖学科范围广、领域宽、层次多, 既有本科专业课程教材, 也有研究生课程教材, 以适应不同院系、不同专业、不同层次的师生对教材的需求。广大师生可自由选择 and 自由组合使用。

按照计划, 本轮教材规划出版 37 种, 其中替换版本 8 种, 新增版本 14 种, 沿用版本 15 种。教材内容涉及的学科方向包括网络与通信、操作系统、计算机组织与结构、算法与数据结构、数据库与信息处理、编程语言、图形图像与多媒体、软件工程等。本轮教材计划于 2001 年 7 月前全部出版。教材的使用年限平均为 3 年。我们还将陆续推出一些教材的参考课件, 希望能为授课老师提供帮助。

为了保证本轮教材的选题质量和翻译质量, 我们约请了清华大学、北京大学、北京航空航天大学、复旦大学、上海交通大学、南京大学、浙江大学、哈尔滨工业大学、华中科技大学、西安交通

大学、国防科学技术大学、解放军理工大学等著名高校的教授和骨干教师参与了本轮教材的选题、翻译和审校工作。他们中既有讲授同类教材的骨干教师和博士,也有积累了几十年教学经验的教授和博士生导师。

在本轮教材的选题、翻译和编辑加工过程中,为提高教材质量,我们做了大量细致的工作,包括:

1. 对于新选题和新版本进行了全面论证。
2. 对于沿用版本,认真审查了前一版本教材,修改了其中的印刷错误。
3. 对于译者和编辑的选择,达到了专业对口。
4. 对于从英文原书中发现的错误,我们通过与作者联络、从网上下载勘误表等方式,一一做了修改。
5. 对于翻译、审校、编辑、排版、印刷质量进行了严格的审查把关。

通过这些工作,保证了本轮教材的质量较前一轮有明显的提高。相信读者一定能够从字里行间体会到我们的这些努力。

今后,我们将继续加强与各高校教师的密切联系,为广大师生引进更多的国外优秀教材和参考书,为我国计算机科学教学体系与国际教学体系的接轨做出努力。

由于我们对国际计算机科学、我国高校计算机教育的发展存在认识上的不足,在选题、翻译、出版等方面的工作中还有许多有待提高之处,恳请广大师生和读者提出批评和建议。

电子工业出版社
2001 年春

教材出版委员会

主 任	杨芙清	北京大学教授 中国科学院院士 北京大学信息与工程学部主任 北京大学软件工程研究所所长
委 员	王 珊	中国人民大学信息学院院长、教授
	胡道元	清华大学计算机科学与技术系教授 国际信息处理联合会通信系统中国代表
	钟玉琢	清华大学计算机科学与技术系教授 中国计算机学会多媒体专业委员会主任
	谢希仁	中国人民解放军理工大学教授 全军网络技术研究中心主任、博士生导师
	尤晋元	上海交通大学计算机科学与工程系教授 上海分布计算技术中心主任
	施伯乐	上海国际数据库研究中心主任、复旦大学教授 中国计算机学会常务理事、上海市计算机学会理事长
	邹 鹏	国防科学技术大学计算机学院教授、博士生导师 教育部计算机基础教学课程指导委员会副主任委员
	张昆藏	青岛大学信息工程学院教授

译者序

自 20 世纪 80 年代以来,国际上出现了所谓的第三类科学,即除了理论科学和实验科学以外的计算科学。该类科学在不少领域计算已取代了实验,而且随着计算机硬件、软件及算法的进步,这种趋势还会加强。作为三大科学研究手段之一的高性能计算机,其发展的根本动力来自于国防、科技、经济和社会对高性能计算技术的永无止境的需求。高性能计算机采用各种并行技术的并行机。集群计算机具有投资风险小、可扩展性好、可继承现有软硬件资源和开发周期短、可编程性好等特点,目前已成为并行处理的热点和主流。集群计算技术迅猛发展的主要原因来自于微处理器性能的不断提高和网络速度的快速提升,应用对计算机性能的无止境需求和并行编程环境的不断成熟也是促进集群计算技术发展的根本动力。

本书全面系统地介绍了高性能集群计算技术的概念、理论研究和工程实践成果。全书由两卷构成:结构与系统(第一卷)和编程与应用(第二卷)。在结构与系统卷中,主要论述了集群计算的需求和一般问题,网络、协议与 I/O 技术,进程调度、负载共享与平衡技术和有代表性的集群计算机系统等。在编程与应用卷中,主要介绍编程环境与开发工具,Java 在高性能计算中的应用,算法与应用等问题。本书内容丰富全面,既介绍了当今科学家在集群计算领域的最新成果,也介绍了集群计算在科学与工程计算及信息服务领域应用的大量实践经验。

本书以指南的形式把这一领域的最新技术和研究成果汇集在一起。该书不仅反映了集群计算的研究与应用现状,同时也指出了未来集群计算研究和开发的可能方向。

郑纬民、汪东升、石威组织翻译并审校了本书,感谢清华大学计算机系研究生杨博、陈永键、史树民、靳超、张武生、黄启峰、朱旋等人在翻译过程中所付出的辛勤努力。由于时间仓促、译者水平有限,文中定有不当或欠妥之处,还望读者批评指正。

目 录

第一部分 编程环境和开发工具	1
第 1 章 并行编程模型和范例	3
1.1 引言	3
1.2 集群计算机及其结构	4
1.3 并行程序及其开发	5
1.4 代码粒度和并行层次	7
1.5 并行编程模型和工具	8
1.6 并行算法的系统化设计	11
1.7 并行编程模式	12
1.8 编程框架和模板	17
1.9 小结	18
1.10 参考文献	19
第 2 章 并行编程语言与环境	21
2.1 引言	21
2.2 早期的机制	22
2.3 共享存储器环境	25
2.4 分布式存储环境	29
2.5 并行说明性环境	32
2.6 小结	34
2.7 参考文献	34
第 3 章 MPI 和 PVM 编程技术	36
3.1 引言	36
3.2 MPI 和 PVM 的对比	36
3.3 任意两点间的最短路径问题	37
3.4 MPI 编程环境	41
3.5 PVM 编程环境	49
3.6 移植	59
3.7 小结	61
3.8 参考文献	61
第 4 章 链接消息传递环境	63
4.1 消息传递接口之间的互操作	63
4.2 PLUS 库简介	64
4.3 系统结构	68
4.4 加入新的消息传递环境	70

4.5	性能测试结果	70
4.6	相关的工作	71
4.7	小结	71
4.8	参考文献	72
第 5 章	主动对象	74
5.1	基于集群的并行系统中的对象	74
5.2	主动和被动对象	74
5.3	对象和原子操作	77
5.4	BaLinda K 对象	78
5.5	投机处理	80
5.6	小结	83
5.7	参考文献	83
第 6 章	用作用域行为来优化数据共享模式	84
6.1	引言	84
6.2	动机:数据共享模式	85
6.3	Aurora:分布式共享数据系统	85
6.4	实现概述	89
6.5	并行程序模拟	90
6.6	讨论及相关工作	95
6.7	小结	95
6.8	参考文献	96
第 7 章	基于组件的开发方法	98
7.1	引言	98
7.2	基于组件的应用开发	99
7.3	高级特征	104
7.4	在分布式环境中复用仿真软件	110
7.5	各种方法的比较	111
7.6	小结	112
7.7	参考文献	112
第 8 章	使用 LiPS 的超计算	114
8.1	生成通信	114
8.2	使用 LiPS	119
8.3	LiPS 运行时系统	122
8.4	LiPS 开发系统	126
8.5	参考文献	129
第 9 章	高效元组空间编程环境	130
9.1	引言	130

9.2 元组空间编程	131
9.3 编译环境	134
9.4 运行时环境	139
9.5 扩展	143
9.6 小结	143
9.7 参考文献	144
第 10 章 调试并行程序	146
10.1 引言	146
10.2 自动并行	147
10.3 调试中面临的问题	147
10.4 使用 Code Liberation 进行调试	148
10.5 实验结果	163
10.6 小结	164
10.7 参考文献	164
第 11 章 WebOS——应用于广域网的操作系统	167
11.1 引言	167
11.2 WebOS 概述	168
11.3 命名	170
11.4 一致性共享状态	173
11.5 安全和认证	174
11.6 进程控制	175
11.7 Rent-A-Server	176
11.8 相关工作	179
11.9 小结	180
11.10 参考文献	180
第二部分 用于高性能计算的 Java	183
第 12 章 分布对象计算	184
12.1 引言	184
12.2 CORBA	185
12.3 Java RMI	190
12.4 DCOM	194
12.5 Voyager	199
12.6 简单的性能测试	202
12.7 参考文献	203
第 13 章 Java 以及不同的并行程序设计模型	204
13.1 引言	204
13.2 Java 线程对并行和并发的内部支持	204

13.3	并行程序设计模型	210
13.4	小结	215
13.5	参考文献	215
第 14 章	HPspmd 模型及其 Java 绑定	217
14.1	引言	217
14.2	Java 语言绑定	218
14.3	HPspmd 编程模型的 Java 程序包	223
14.4	编程实例	225
14.5	语言设计中的问题	226
14.6	项目进展	228
14.7	小结	229
14.8	参考文献	230
第 15 章	基于 Web 的 Java 并行计算	232
15.1	引言	232
15.2	基于 Web 的并行计算	232
15.3	集群与基于 Web 并行计算的比较	233
15.4	基于 Internet 并行计算的例子	234
15.5	Java 能用于基于 Web 的并行计算吗	235
15.6	基于 Web 的并行计算需要解决的问题	235
15.7	JET 平台	237
15.8	一些性能结果	240
15.9	小结	242
15.10	参考文献	242
第三部分	算法和应用	245
第 16 章	并行遗传算法的面向对象的实现	247
16.1	引言	247
16.2	GA 系统概览	248
16.3	PGA 的面向对象的实现	249
16.4	表示个体的类	249
16.5	局部遗传操作	251
16.6	岛模型	252
16.7	全局种群模型	253
16.8	负载平衡	254
16.9	文件和输入/输出操作	255
16.10	PGA 系统的框架	256
16.11	实验结果	256
16.12	小结	259

16.13 参考文献	259
第 17 章 异构系统上特定应用的负载平衡	261
17.1 引言	261
17.2 系统概述	262
17.3 复数 FDTD 方程的实现	263
17.4 负载平衡	267
17.5 分析	274
17.6 小结	277
17.7 参考文献	278
第 18 章 并行仿真中的时间管理	279
18.1 引言	279
18.2 并行仿真中的主要问题	280
18.3 并行仿真的原理	280
18.4 保守的同步协议	284
18.5 小结	291
18.6 参考文献	291
第 19 章 硬件系统仿真	293
19.1 引言	293
19.2 NEPSi	293
19.3 讨论	307
19.4 参考文献	308
第 20 章 实时资源管理中间件:开放系统及其应用	309
20.1 引言	309
20.2 动态 QoS 管理中间件的体系结构	311
20.3 自适应的资源分配	312
20.4 自适应资源管理服务的实际应用	318
20.5 小结	324
20.6 参考文献	324
第 21 章 无共享数据库中的数据分布	326
21.1 引言	326
21.2 数据安排	327
21.3 划分	328
21.4 安排	330
21.5 二次分配	332
21.6 动态重组	333
21.7 小结	334

21.8 参考文献	334
第 22 章 大型知识库的并行接口:实时推理的连接方法	336
22.1 引言	336
22.2 SHRUTI: 结构化连接推理系统	337
22.3 SHRUTI 到并行机器上的映射	340
22.4 SHRUTI 在 CM-5 上的设计和实现	341
22.5 SHRUTI-CM5 的数学分析	346
22.6 SHRUTI-CM5 上对大规模知识库进行的试验	350
22.7 小结	355
22.8 参考文献	356
第 23 章 MaRT:用于并行光线跟踪的惰性计算	358
23.1 引言	358
23.2 光线追踪并行化技术	360
23.3 MaRT:惰性光线追踪器	363
23.4 并行 MaRT	366
23.5 小结	371
23.6 参考文献	371
第 24 章 基于内容的快速图像检索	373
24.1 引言	373
24.2 图像特征提取	374
24.3 动态图像索引	379
24.4 图像相似性度量	380
24.5 图像搜索	380
24.6 并行实现	381
24.7 实验结果	382
24.8 并行图像特征提取	383
24.9 小结	386
24.10 参考文献	386
第 25 章 海洋气候建模	388
25.1 引言	388
25.2 模型描述	389
25.3 不规则几何体的并行分解	389
25.4 不同系统中的海洋模型	392
25.5 科学结论	395
25.6 小结	396
25.7 参考文献	397

第 26 章 计算电磁学	398
26.1 引言.....	398
26.2 物理光学方法.....	400
26.3 有限差分的时间域方法.....	403
26.4 有穷单元积分方程耦合方法.....	405
26.5 小结.....	408
26.6 参考文献.....	409
第 27 章 CFD 模拟——软件工程中的实例研究	411
27.1 引言.....	411
27.2 TtC——代表当前水平的工业 CFD 包	412
27.3 并行 CFD 模拟的条件	415
27.4 ParTtC 的设计与实现	417
27.5 科学软件的面向对象设计.....	422
27.6 并行科学计算软件的生产应用.....	423
27.7 参考文献.....	425
第 28 章 量子反应散射计算	427
28.1 引言.....	427
28.2 多体问题的描述、分解和解决	428
28.3 并行化策略.....	431
28.4 CRAY T3E 上的并行实现	437
28.5 SGI Origin 2000 上的并行实现	439
28.6 元计算机上的并行实现.....	440
28.7 小结.....	442
28.8 参考文献.....	442
第 29 章 生物医学应用模型	444
29.1 引言.....	444
29.2 染色体重构问题.....	444
29.3 染色体重构的 PVM 算法	445
29.4 心率可变性和 Kolmogorov 熵	450
29.5 使用 PVM 计算 K_2 熵的并行算法	453
29.6 最佳缩放区域决定算法.....	454
29.7 实验结果.....	454
29.8 小结.....	457
29.9 参考文献.....	457

第一部分 编程环境和开发工具

一般情况下为了有效地开发并行计算机尤其是集群的计算能力,好的编程环境和开发工具是必不可少的条件。如果能让并行处理过程和分布式计算不仅仅局限于某个专门的专家小组,而是具有一般程序员所习惯的编程模型特点,则这一点尤为重要。

在过去的二十年里,各种不同的编程环境得到了很大的发展,同时得到发展的还有很多相应的编程辅助工具,包括并行调试器和各种跟踪监测工具。许多这样的环境作为标准出现,这些标准反过来又促进了可移植并行环境模型的发展。所有这些引发了大量的普及并行计算的行动。本书在这一部分主要讲述在最近二十年里出现的一些主要并行环境和相关的编程工具。

第1章介绍了目前实现并行应用程序时常用的并行编程模型。主要讨论主/从模型、单控制流多数据流模型(single-program multiple-data, SPMD)、数据流水线、分治模型、投机并行和各种混合方法。这一章的最后一节将给出一个并行编程模板。

第2章描述了一些主流并行编程语言和环境。首先讨论的是在并行编程初期使用的并行构造方法,包括各种原语,比如信号量、监控程序、套接字和远地过程调用;然后描述通常使用的共享存储模型,包括纯粹和虚拟的共享存储环境;接下来重点讨论各种分布存储编程环境;最后以并行声明语言(parallel declarative languages)结尾。

虽然在第2章中我们试着为系统提供一个共享存储的接口,但是分布存储模型最适合集群计算。第3章将描述两种最流行的分布存储消息传递环境:MPI和PVM,它们已经成为现在消息传递编程的标准。本章分别利用MPI和PVM模型实现两种不同的用于计算任意两点间最短距离的并行算法,并藉此说明在这两种模型下的编程特点和方法。除了说明两种环境下基本函数的用法之外,我们还给出了不同环境下相关函数的区别。本章结尾部分给出了如何在MPI和PVM之间进行相互转换的一些方法。

第4章主要讲述用于集群计算的各种不同消息传递环境之间的协作、互连问题。讲述了三种用于连接MPI和PVM的软件包:PACX-MPI、PLUS和PVMPI;然后详细描述了PLUS库提供的资源管理支持;接下来讲述如何将一个新的消息系统加到某个已有环境中,同时还给出该章结尾处所给例子的性能测试结果。

第5章探讨了集群计算中的主动对象。它首先讲述了主动与被动对象的区别,然后讨论对象的原子性要求。本章有一小节专门讲述BaLinda K对象,还有一节讨论当该对象应用于主动对象时产生的投机处理现象。

第6章以Aurora系统为例描述了用于优化数据共享的域化行为(scoped behavior)。Aurora是一个分布式共享数据系统。本章首先通过一个矩阵相乘的例子简要叙述Aurora中相关的处理模型、共享数据对象和域化行为的概念;然后简要解释Aurora的实现,并且给出三个不同的应用程序作为例子:矩阵相乘、二维漫射和基于正则抽样的并行排序。

第7章演示了如何在集群上利用基于组件的方法开发并行应用程序。本章首先讲述现在

使用的几种主要方法,然后具体地描述 TRACS 方法,重点介绍该方法对设计组件的重用能力。本章还介绍了 TRACS 方法的以下几方面:API、组件重用能力、如何在该模型内编译并运行分布式应用程序以及另一个可选的设计风格。最后,列举了一个在分布式环境中重用串行软件

的例子。

第 8 章讲述一个具有容错功能的元组空间系统:LiPS(Library for Parallel System:并行系统库)。本章描述了 LiPS 中的协作运行时系统,重点放在容错、系统运行时间和应用程序运行时间几个方面。接着结合分布式数据结构、活动(living)数据结构、消息传递和文件使用的例子讨论了 LiPS 的开发系统。

第 9 章讨论集群环境下的元组空间编程。在对元组空间编程进行简单介绍之后,本章接着在基本翻译机制和优化编译器方面讨论这些系统的编译环境;然后,再从集中式数据和分布式数据两个方面讨论运行时环境;最后,详细论述元组空间的几种扩展形式:多元组空间、持久元组空间、具有容错能力的元组空间以及具有适应能力的并行性。

第 10 章讨论的是集群计算中的一个重要方面,即并程序调试。本章的重点是利用代码释放进行调试的概念,并将讨论全局重命名、名字回收和在此上下文中的调试接口。本章还有一节讲述集群环境下代码解放带来的其他好处。

第一部分的最后一章——第 11 章的重点是 Web 操作系统,即一个支持物理上分布、高可用性、增益可扩展性、动态重构能力的框架。Web 操作系统提供了对广域命名、一致性存储、安全性和执行进程的支持。本章介绍了如何将这些基本功能组合起来开发 Rent-A-Server, Rent-A-Server 是一种能够在广域范围内根据客户的请求类型动态地复制自己的 Web 服务器。性能测试表明,与现有的静态复制方法相比,这个系统能够改善延迟并减少所消耗的带宽。

第 1 章 并行编程模型和范例

Luis Moura e Silva⁺ And Rajkumar Buyya⁺

⁺ Departamento de Engenharia Informática
Universidade de Coimbra, Polo II
3030 Coimbra-Portugal

⁺ School of Computer Science and Software Engineering
Monash University
Melbourne, Australia

Email: luis@dei.uc.pt, rajkumar@ieee.org

1.1 引言

在 20 世纪 80 年代,我们认为计算机性能的极大提高得益于不断出现的更快、更有效的处理器。如今,这种看法遇到了并行处理的挑战,这里“并行”的本质含义是指将两个或更多的计算机连接起来解决同一个计算问题。从 20 世纪 90 年代初期开始,从昂贵的、面向特定应用的并行超级计算机(向量机和大规模并行处理机)向网络计算机(计算机可以是 PC 或者工作站或者 SMP 机器)转变的趋势越来越明显。这些技术使得计算机网络或者集群变成一个非常有吸引力的工具,可以用在性价比要求高的并行处理中,同时还为实现低价的商用超级计算机指明了方向。

从 PC 或工作站(同质或异质)集群到 SMP 集群的可扩展计算机集群正很快地成为高性能和大规模计算的标准平台。这种系统最主要的优点是,它们建立在容易购买且价格低廉的商品化硬件(比如奔腾 PC 机)、快速局域网(比如 Myrinet)和标准软件组件(比如 UNIX、MPI 和 PVM 并行编程环境)之上。这些系统是可扩展的,也就是说,它们可以根据预算和计算能力的要求做出调整,并且能够有效地运行高要求的并行和串行程序。这样的系统有 Berkeley NOW、HPVM、Beowulf 和 Solaris-MC,本书的第一卷已经讨论过这些系统了^[8]。

集群利用一定的智能机制实现动态的、网络范围内的资源共享,并随着资源供求的变化而变化。这些机制给集群带来了可扩展性,而且允许灵活地使用工作站,因为集群或网络范围内的可用资源应该比集群中某个节点/工作站上的资源充足。智能机制同样使得集群能支持多用户、分时共享和并行执行环境。在这些系统中,可以通过共享资源和动态分配负载从而有效利用全局资源的功能^[4]。

有效地开发工作站集群计算能力的想法已经在高性能计算领域内得到高度重视,当前的

趋势也更倾向于这种商品化的超级计算机。出现这种现象的原因是:绝大部分的科学团体希望减小经济上的风险,同时希望并行计算能建立在成熟的、商品化的技术之上。集群计算被认为是未来解决大型科学或商业计算的主流方案。

我们已经列举了一些为什么在高性能并行计算领域内集群能够得到越来越广泛应用的原因。在下面的小节中,将讨论集群计算机的一般结构、并行的层次、编程环境与模型、编写并行程序常见的方法和两种实现并行的主要手段(显式的和隐式的)。在这两种方法中,我们将简要地总结开发并行能力时所有可能的选择:使用并行编译器、并行编程语言、消息传递库、分布式共享内存、面向对象的程序设计和程序设计框架。但是,本章的重点还是描述常用的并行编程模型,并讲述如何在已有的应用中辨认出这些模型。在本章结束时,我们给出了一些并行库、并行工具和并行环境的例子,它们通过使用框架或者模板的方法为并行编程提供高层支持。这种方法显示了一些有趣的优点,比如具有代码重用功能、高灵活性,并且能不断提高并行编程开发人员的生产率。

1.2 集群计算机及其结构

集群是一种并行或者分布处理系统,它由一组互连的单机组成,这些单机协调工作提供一个单一的、完整的计算资源。

一个计算机节点可以是一个具有存储器、I/O 设备和操作系统的单机或者多处理器系统(PC 机、工作站或者 SMP 机器)。一个集群一般指的是两个或者两个以上互相连接起来的计算机(节点)。这些节点可以放在同一个机柜中,也可以分布在不同的地方并通过局域网连接。一个(通过局域网)互连的集群对使用者和应用程序来说表现为一个单一的系统。这样的系统能够廉价地获得原来只有昂贵的私有共享存储系统才具有的特征和优点(快捷而且可靠的服务)。集群的典型结构如图 1.1 所示。

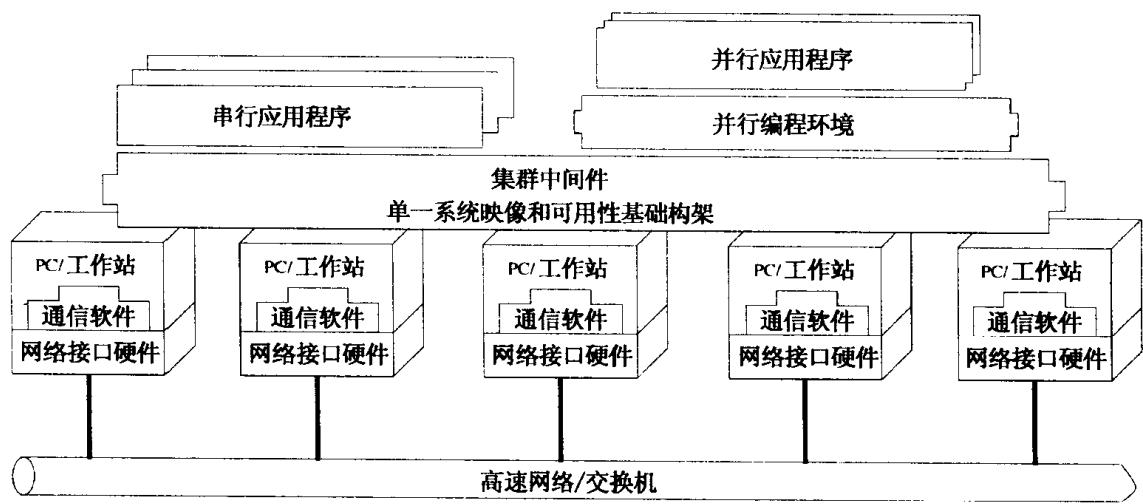


图 1.1 集群结构图

下面是集群的一些显著特征:

- 多种高性能计算机(PC 机、工作站或者 SMP 机器)
- 先进的操作系统(分层或者微内核结构)