

国外信息技术优秀图书选译

CAMBRIDGE

# 分布式计算

——原理、算法与系统

## Distributed Computing Principles, Algorithms, and Systems

Ajay D. Kshemkalyani Mukesh Singhal 著

余宏亮 张冬艳 译

 高等教育出版社  
HIGHER EDUCATION PRESS

国外信息技术优秀图书选译

FENBUSHI JISUAN

—YUANLI, SUANFA YU XITONG

# 分布式计算

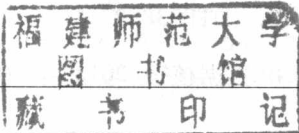
——原理、算法与系统

## Distributed Computing

### Principles, Algorithms, and Systems

Ajay D. Kshemkalyani Mukesh Singhal 著

余宏亮 张冬艳 译



T0991793

0991793



高等教育出版社·北京  
HIGHER EDUCATION PRESS BEIJING

图字:01 - 2009 - 7070 号

*Distributed Computing (Principles, Algorithms, and Systems)*, 1<sup>st</sup> edition, ISBN: 9780521876346, by Ajay D. Kshemkalyani, Mukesh Singhal, first published by Cambridge University Press 2008.

All rights reserved.

This simplified Chinese edition for the People's Republic of China is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Cambridge University Press & Higher Education Press, 2011

This book is in copyright. No reproduction of any part may take place without the written permission of Cambridge University Press or Higher Education Press.

This edition is for sale in the mainland of China only, excluding Hong Kong SAR, Macao SAR and Taiwan, and may not be bought for export therefrom.

此版本仅限于在中华人民共和国境内(但不允许在香港、澳门和中国台湾)销售。不得出口。

### 图书在版编目(CIP)数据

分布式计算:原理、算法与系统/(美)克谢姆卡亚尼(Kshemkalyani, A. D.), (美)辛哈(Singhal, M.)著;余宏亮,张冬艳译.--北京:高等教育出版社, 2012.6

书名原文:Distributed Computing Principles, Algorithms, and Systems

ISBN 978 - 7 - 04 - 032456 - 3

I. ①分… II. ①克…②辛…③余…④张… III.

①分布式计算机系统 IV. ①TP338.8

中国版本图书馆 CIP 数据核字(2012)第 009150 号

出版发行 高等教育出版社  
社 址 北京市西城区德外大街 4 号  
邮政编码 100120  
印 刷 涿州市京南印刷厂  
开 本 787mm × 1092mm 1/16  
印 张 41.25  
字 数 840 千字  
购书热线 010 - 58581118

咨询电话 400 - 810 - 0598  
网 址 <http://www.hep.edu.cn>  
<http://www.hep.com.cn>  
网上订购 <http://www.landaco.com>  
<http://www.landaco.com.cn>  
版 次 2012 年 6 月第 1 版  
印 次 2012 年 6 月第 1 次印刷  
定 价 79.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换  
版权所有 侵权必究  
物 料 号 32456 - 00

---

# 分布式计算——原理、算法与系统

分布式计算是指跨越由计算机网络连接的多处理平台的各种形式的计算、信息访问与信息交换。分布式计算系统的设计是一项复杂的工作,它要求对于设计的细节及其实现方案的理论和实践问题有深入细致的理解。本书全面阐述了支撑分布式计算理论、算法和系统的基本原理和模型。

本书涉及的理论广度和深度、兼顾与实际系统相关的问题,如互斥、死锁检测、认证以及故障恢复。算法经过认真选择,并清晰地呈现和描述,但没有给出复杂的证明,而是用简单的说明和图示来解释。有重大影响的新主题,例如 P2P 网络和网络安全,也包括在本书中。

本书同时拥有很多最新算法,大量的图示、实例、习题和作业对于电子与计算机工程、计算机科学专业的高年级本科生和研究生是十分有价值的,对于从事数据网络和传感器网络的实际工作者也是非常有价值的资源。

Ajay D. Kshemkalyani 是伊利诺伊大学芝加哥分校计算机科学系副教授。1991 年于俄亥俄州立大学获得计算机与信息科学博士学位。在转入学术领域之前,于 IBM Triangle Park 研究中心从事计算机网络工作数年。1999 年获得美国国家科学基金委 CAREER 奖。IEEE 高级会员。研究领域包括分布式计算、算法、计算机网络和并发系统。现为 Computer Networks 编委会成员。

Mukesh Singhal 是肯塔基大学计算机科学系网络工程教授并拥有 Gartner 集团荣誉教授。1986 年于马里兰大学获得计算机科学博士学位。2003 年获得 IEEE 技术成就奖。目前为 IEEE Transactions on Parallel and Distributed Systems 和 IEEE Transactions on Computers 的编委会成员。IEEE Fellow。研究领域包括分布式系统、计算机网络、无线和移动计算系统、性能评价和计算机安全。

献给我的父亲 *Shri Digambar* ,  
母亲 *Shrimati Vimala*

Ajay D. Kshemkalyani

献给我的母亲 *Chandra Prabha Singhal* ,  
父亲 *Brij Mohan Singhal* 以及  
女儿们 *Meenakshi* , *Malvika* 和 *Priyanka*

Mukesh Singhal

# 前 言

## 背景

分布式计算领域涵盖了通过任何形式,比如局域或广域的通信网络连接起来的、跨越多种处理元素之间的计算和信息获取的所有方面。自从 20 世纪 70 年代 Internet 的出现,需要分布式处理的新的应用持续增长。这得益于网络和硬件技术的进步、硬件成本的下降和终端用户意识的增强,并且这些因素也使得分布式计算高效益、高性能和容错属性成为现实。在千禧年之交,全球 Internet 的广度和效能有了爆炸性增长,并与日益增长的通过 WWW(World Wide Web)获取联网资源相匹配。再加上无线和移动网络领域同样不可思议的增长,以及带宽和存储设备价格的暴跌,我们正在见证分布式应用的急速增长以及大学、政府组织和私营机构对分布式计算领域的兴趣。

硬件技术的进步突然使得传感器网络成为现实,而嵌入式和传感器网络正迅速成为每个人生活中的组成部分——从互联的家庭网络到通过 GPS(global positioning system,全球定位系统)通信的汽车,再到使用 RFID 监控的完全网络化办公环境。在新兴的地球村中,分布式计算将成为计算机科学中所有计算和信息获取子学科的中心部分。很显然,这是一个非常重要的领域。而且,这个不断发展的领域其特点是将面临各种各样的挑战,而解决之道则需要有坚实的原理基础。

分布式计算领域如此重要,因此,一本优秀的综合介绍该领域的著作将有着极大的需求。本书全面深入地阐述了所有重要的主题,并给出了清晰易于理解的讲述。本书对于学术界和计算机工业界都具有特别的价值。编写这样一部综合性的著作是一项艰巨的似乎只有大力神才能承担起的任务,而知道我们能够完成它并且奉献给该领域,则有一种深深的满足感。

## 描述、方法与特点

本书重点介绍分布式计算所涉及的各个方面的基本原理和模型,包括分布式计算的理论、算法和系统方面的基本原理。本书算法的呈现清晰,并通过图示和简单的

说明来讲解算法的主要思想和直观知识,而不是纠缠于令人生畏的公式符号和冗长、难以理解的严格证明。每章主题的选择是广泛而综合的,全书以一定深度涵盖了所有重要的主题。每章的算法都经过仔细地选择,以阐明算法设计中一些新的和重要的技术。尽管本书集中于分布式计算中的算法和基础内容,但通过对背后的理论及其算法的介绍,它也深入彻底地讲解了所有实际的系统类型的问题(例如,互斥、死锁检测、终止检测、故障恢复、认证、全局状态和时间,等等)。本书写作时也时刻注意新兴主题对分布式计算的基础内容的影响,例如 P2P 计算和网络安全。

本书每章包含图表、实例、练习、小结和参考文献。

### 读者对象

本书目的是作为以下读者的教科书:

- 计算机科学和计算机工程专业的研究生和高年级本科生。
- 电子工程和数学专业的研究生。随着无线网络、P2P 网络和移动计算变得越来越重要,越来越多的电子工程系的学生会需要本书。
- 实际工作者、系统设计者/程序员以及工业界和研究实验室的咨询人员会发现这是一本非常有用的参考书,因为它包括了用来解决分布式系统中设计问题的原理和最新的算法,也包括最新的参考文献。

使用本书的软件和硬件前提条件包括:

- 要求学习了算法方面的本科课程。
- 操作系统和计算机网络方面的本科课程会很有用。
- 比较熟悉编程。

我们的目的是在一本书中非常综合全面地介绍分布式计算的模型和算法,使其成为读者可单独参考的著作。该书涵盖的讨论主题既有广度也有深度,并具有阐述清晰、易于理解的特点。现有的分布式计算方面的教材没有能具备上述所有的特点。

### 致谢

本书的编写来源于作者在一些大学开设分布式计算研究生课程的讲课笔记,这些大学包括俄亥俄州立大学、伊利诺伊大学芝加哥分校和肯塔基大学。我们感谢这些学校的研究生在许多方面对本书所做的贡献。

本书内容基于该领域许多研究学者所发表的研究成果。我们尽力用自己的文字介绍这些内容和材料,并对信息的原始来源给予指明。感谢所有在本书中介绍了他们工作的研究者。最后,我们感谢为本书出版提供了大力支持的剑桥大学出版社的同事们。

### 网络资源

下列网站将服务于本书。发现本书的任何错误以及对本书的任何评论都可以发送到 [ajayk@cs.uic.edu](mailto:ajayk@cs.uic.edu) 或者 [singhal@cs.uky.edu](mailto:singhal@cs.uky.edu)。有关本书更进一步的信息也可以

从作者主页上获得：

- [www.cs.uic.edu/~ajayk/DCS-Book](http://www.cs.uic.edu/~ajayk/DCS-Book)
- [www.cs.uky.edu/~singhal/DCS-Book](http://www.cs.uky.edu/~singhal/DCS-Book)



# 目 录

第一章 引言 .....	1
1.1 定义 .....	1
1.2 与计算机系统部件的关系 .....	2
1.3 动机 .....	3
1.4 与并行多处理器/多计算机系统的关系 .....	4
1.4.1 并行系统的特性 .....	4
1.4.2 Flynn 的分类法 .....	8
1.4.3 耦合、并行、并发及粒度 .....	9
1.5 消息传递系统与共享内存系统的对比 .....	11
1.5.1 在共享内存的系统上仿真消息传递 .....	11
1.5.2 在消息传递系统上仿真共享内存 .....	12
1.6 分布式通信的原语 .....	12
1.6.1 阻塞/非阻塞,同步/异步原语 .....	12
1.6.2 处理器同步性 .....	15
1.6.3 库与标准 .....	15
1.7 同步与异步执行 .....	16
1.7.1 通过同步系统仿真异步系统 .....	17
1.7.2 通过异步系统仿真同步系统 .....	17
1.7.3 仿真 .....	18
1.8 设计主题与挑战 .....	18
1.8.1 从系统角度看分布式系统的挑战 .....	19
1.8.2 分布式计算中的算法挑战 .....	20
1.8.3 分布式计算的应用以及更新的挑战 .....	25
1.9 关于主题的选择与覆盖 .....	27
1.10 本章小结 .....	27

1.11 习题 .....	28
1.12 参考文献说明 .....	29
参考文献 .....	30
<b>第二章 分布式计算模型</b> .....	<b>33</b>
2.1 分布式程序 .....	33
2.2 分布式运行模型 .....	33
2.3 通信网络模型 .....	36
2.4 分布式系统的全局状态 .....	36
2.4.1 全局状态 .....	37
2.5 分布式计算的运行分割 .....	38
2.6 事件的过去和未来锥面 .....	39
2.7 进程通信模型 .....	40
2.8 本章小结 .....	40
2.9 习题 .....	40
2.10 参考文献说明 .....	41
参考文献 .....	41
<b>第三章 逻辑时间</b> .....	<b>42</b>
3.1 引言 .....	42
3.2 逻辑时钟框架 .....	43
3.2.1 定义 .....	43
3.2.2 实现逻辑时钟 .....	43
3.3 标量时间 .....	44
3.3.1 定义 .....	44
3.3.2 基本性质 .....	45
3.4 向量时间 .....	46
3.4.1 定义 .....	46
3.4.2 基本性质 .....	47
3.4.3 有关向量时钟的大小 .....	48
3.5 向量时钟的有效实现 .....	49
3.5.1 Singhal - Kshemkalyani 的差量技术 .....	50
3.5.2 Fowler - Zwaenepoel 的直接依赖技术 .....	51
3.6 Jard - Jourdan 的自适应技术 .....	54
3.7 矩阵时间 .....	56
3.7.1 定义 .....	56

3.7.2 基本性质 .....	58
3.8 虚拟时间 .....	58
3.8.1 虚拟时间的定义 .....	58
3.8.2 与 Lamport 逻辑时钟比较 .....	59
3.8.3 时间变形机制 .....	60
3.8.4 本地控制机制 .....	60
3.8.5 全局控制机制 .....	62
3.9 物理时钟同步;NTP .....	64
3.9.1 动机 .....	64
3.9.2 定义及术语 .....	65
3.9.3 时钟不准确性 .....	65
3.10 本章小结 .....	67
3.11 习题 .....	68
3.12 参考文献说明 .....	68
参考文献 .....	69
<b>第四章 记录全局状态与快照算法 .....</b>	<b>71</b>
4.1 引言 .....	71
4.2 系统模型和定义 .....	73
4.2.1 系统模型 .....	73
4.2.2 一致性全局状态 .....	74
4.2.3 有关分割的解 .....	74
4.2.4 记录全局快照时遇到的问题 .....	75
4.3 FIFO 通道的快照算法 .....	75
4.3.1 Chandy - Lamport 算法 .....	75
4.3.2 被记录全局状态的性质 .....	77
4.4 Chandy - Lamport 算法的变种 .....	78
4.4.1 Spezialetti - Kearns 算法 .....	79
4.4.2 Venkatesan 快照增量算法 .....	80
4.4.3 Helary 波同步方法 .....	81
4.5 非 FIFO 通道的快照算法 .....	81
4.5.1 Lai - Yang 算法 .....	82
4.5.2 Li 等人的算法 .....	83
4.5.3 Mattern 算法 .....	84
4.6 因果传递系统快照 .....	85
4.6.1 进程状态记录 .....	85

4.6.2 Acharya - Badrinath 算法中的通道状态记录	86
4.6.3 Alagar - Venkatesan 算法中的通道状态记录	86
4.7 监控全局状态	88
4.8 一致性全局快照的必要和充分条件	88
4.8.1 Zigzag 路径和一致性全局快照	89
4.9 找出分布式计算中的一致性全局快照	92
4.9.1 找出一致性全局快照	92
4.9.2 枚举式一致性快照 Manivannan - Netzer - Singhal 算法	94
4.9.3 在分布式计算中找出 Z 路径	96
4.10 本章小结	97
4.11 习题	98
4.12 参考文献说明	99
参考文献	99
<b>第五章 术语和基本算法</b>	<b>102</b>
5.1 拓扑抽象和覆盖	102
5.2 分类和基本概念	104
5.2.1 应用执行和控制算法执行	104
5.2.2 集中式算法和分布式算法	104
5.2.3 对称算法和非对称算法	105
5.2.4 匿名算法	105
5.2.5 一致算法	105
5.2.6 自适应算法	105
5.2.7 确定性执行对非确定性执行	105
5.2.8 执行抑制	106
5.2.9 同步系统和异步系统	107
5.2.10 联机算法与脱机算法	107
5.2.11 故障模型	107
5.2.12 无需等待算法	108
5.2.13 通信通道	109
5.3 复杂度测量和度量	109
5.4 程序结构	110
5.5 图的基本算法	111
5.5.1 使用洪泛法的同步单一启动者生成树算法	111
5.5.2 使用洪泛法的异步单一启动者生成树算法	113
5.5.3 使用洪泛法的异步并发启动者生成树算法	115

5.5.4	异步并发启动者深度优先搜索生成树算法	118
5.5.5	在一棵树上广播和聚播	119
5.5.6	单一源最短路径算法:同步 Bellman - Ford	120
5.5.7	距离向量路径选择	121
5.5.8	单一源最短路径算法:异步 Bellman - Ford	122
5.5.9	全源最短路径:异步分布式 Floyd - Warshall	123
5.5.10	有约束的异步和同步洪泛法(W/O一棵生成树)	126
5.5.11	同步系统的最小权重生成树算法	128
5.5.12	异步系统的最小权重树	132
5.6	同步工具	133
5.7	最大独立集合	138
5.8	连通支配集	140
5.9	紧凑路由表	141
5.10	选领导者	142
5.11	设计分布式图算法的挑战	144
5.12	对象副本问题	144
5.12.1	问题定义	145
5.12.2	算法概要	145
5.12.3	读和写	146
5.12.4	收敛到一个副本模式	146
5.13	本章小结	149
5.14	习题	150
5.15	参考文献说明	152
	参考文献	153

<b>第六章</b>	<b>消息序与组通信</b>	<b>156</b>
6.1	消息序的模式	157
6.1.1	异步执行过程	157
6.1.2	先进先出执行过程	157
6.1.3	因果序执行过程	158
6.1.4	同步执行过程	160
6.2	使用同步通信的异步执行过程	161
6.2.1	用同步通信可实现的执行过程	162
6.2.2	序样式的层次体系	165
6.2.3	两种仿真	165
6.3	异步系统中同步程序的序	166

6.3.1	会合	167
6.3.2	双路会合算法	167
6.4	组通信	171
6.5	因果序	171
6.5.1	Raynal - Schiper - Toueg 算法 [22]	172
6.5.2	Kshemkalyani - Singhai 最优算法 [20,21]	173
6.6	全序	180
6.6.1	为全序设计的集中式算法	180
6.6.2	三阶段分布式算法	181
6.7	多播相关术语	185
6.8	多播传播树	186
6.9	应用层多播算法的分类	190
6.10	容错组通信的语义	192
6.11	网络层的分布式多播算法	194
6.11.1	泛洪约束的反向路径转发	194
6.11.2	Steiner 树	195
6.11.3	多播的成本函数	196
6.11.4	有界延迟的 Steiner 树	196
6.11.5	核基树	199
6.12	本章小结	199
6.13	习题	200
6.14	参考文献说明	201
	参考文献	202
<b>第七章</b>	<b>终止检测</b>	<b>204</b>
7.1	引言	204
7.2	分布式计算的系统模型	205
7.3	基于快照的终止检测算法	205
7.3.1	非形式化描述	206
7.3.2	形式化描述	206
7.3.3	讨论	207
7.4	信用 - 传递终止检测算法	207
7.4.1	形式化描述	208
7.4.2	算法的正确性	208
7.5	基于生成树的终止检测算法	209
7.5.1	定义	209

7.5.2	一个简单的算法	210
7.5.3	正确的算法	210
7.5.4	一个例子	211
7.5.5	算法性能分析	213
7.6	消息 - 优化终止检测算法	213
7.6.1	主要思想	213
7.6.2	算法描述	214
7.6.3	算法性能分析	216
7.7	通用分布式计算模型中的终止检测	216
7.7.1	模型定义和假设	217
7.7.2	符号	217
7.7.3	终止定义	217
7.7.4	一个静态终止检测算法	218
7.7.5	一个动态终止检测算法	219
7.8	原子计算模型中的终止检测	221
7.8.1	执行的原子模型	221
7.8.2	一个简单的计数方法	221
7.8.3	四计数器方法	222
7.8.4	怀疑论者算法	223
7.8.5	时间算法	223
7.8.6	向量计数算法	225
7.8.7	信道计数算法	226
7.9	带错分布式系统中的终止检测	229
7.9.1	流检测方案	229
7.9.2	捕获快照	230
7.9.3	算法描述	231
7.9.4	算法性能分析	234
7.10	本章小结	234
7.11	习题	235
7.12	参考文献说明	235
	参考文献	236
<b>第八章</b>	<b>知识推理</b>	<b>238</b>
8.1	Muddy children 难题	238
8.2	知识的逻辑	239
8.2.1	知识操作符	239

8.2.2	回到 muddy children 难题	240
8.2.3	克里普克结构	240
8.2.4	使用克里普克结构解决 muddy children 难题	241
8.2.5	知识的属性	243
8.3	同步系统中的知识	244
8.4	异步系统中的知识	244
8.4.1	逻辑和定义	244
8.4.2	异步系统中的达成一致	245
8.4.3	共同知识的变体	246
8.4.4	并发共同知识	246
8.5	知识转移	250
8.6	知识和时钟	251
8.7	本章小结	253
8.8	习题	253
8.9	参考文献说明	254
	参考文献	254
<b>第九章</b>	<b>分布式互斥算法</b>	<b>256</b>
9.1	引言	256
9.2	预备知识	257
9.2.1	系统模型	257
9.2.2	互斥算法的必备条件	257
9.2.3	性能指标	258
9.3	Lamport 算法	259
9.4	Ricart - Agrawala 算法	262
9.5	Singhal 动态信息结构算法	264
9.5.1	算法描述	266
9.5.2	正确性	268
9.5.3	性能分析	269
9.5.4	异构流量模式下的适应性	270
9.6	Lodha 和 Kshemkalyani 的公平互斥算法	270
9.6.1	系统模型	270
9.6.2	算法描述	270
9.6.3	安全性、公平性与活性	275
9.6.4	消息复杂性	275
9.7	基于仲裁团的互斥算法	275



9.8	Maekawa 算法	276
9.8.1	死锁问题	277
9.9	Agarwal - EI Abbadi 基于仲裁团的算法	278
9.9.1	构造树状结构仲裁团	279
9.9.2	构造树状结构仲裁团算法分析	280
9.9.3	有效性	280
9.9.4	树状结构仲裁团的例子	281
9.9.5	分布式互斥算法	282
9.9.6	正确性证明	283
9.10	基于令牌的算法	283
9.11	Suzuki - Kasami 广播算法	283
9.12	基于树的 Raymond 算法	285
9.12.1	HOLDER 变量	286
9.12.2	算法操作	287
9.12.3	算法描述	287
9.12.4	正确性	289
9.12.5	开销和性能分析	290
9.12.6	算法初始化	291
9.12.7	错误和恢复	291
9.13	本章小结	292
9.14	习题	292
9.15	参考文献说明	293
	参考文献	293

<b>第十章</b>	<b>死锁检测</b>	296
10.1	简介	296
10.2	系统模型	296
10.2.1	等待图	297
10.3	预备知识	297
10.3.1	死锁处理策略	297
10.3.2	死锁检测问题	298
10.4	死锁模型	298
10.4.1	单资源模型	299
10.4.2	AND 模型	299
10.4.3	OR 模型	299
10.4.4	AND - OR 模型	300