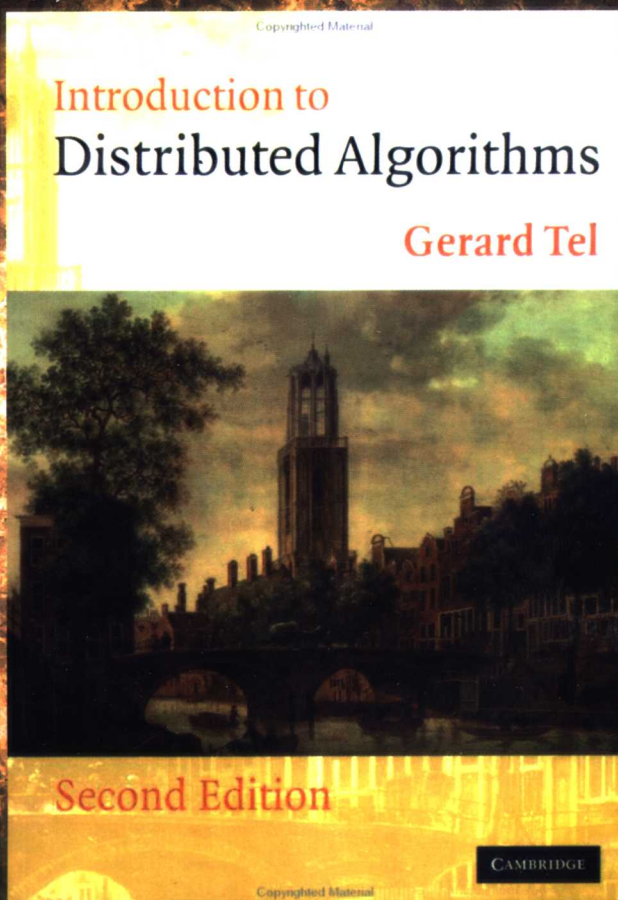


计 算 机 科 学 丛 书

原书第2版

分布式算法导论

(荷) Gerard Tel 著 霍红卫 译



Introduction to Distributed Algorithms
(Second Edition)



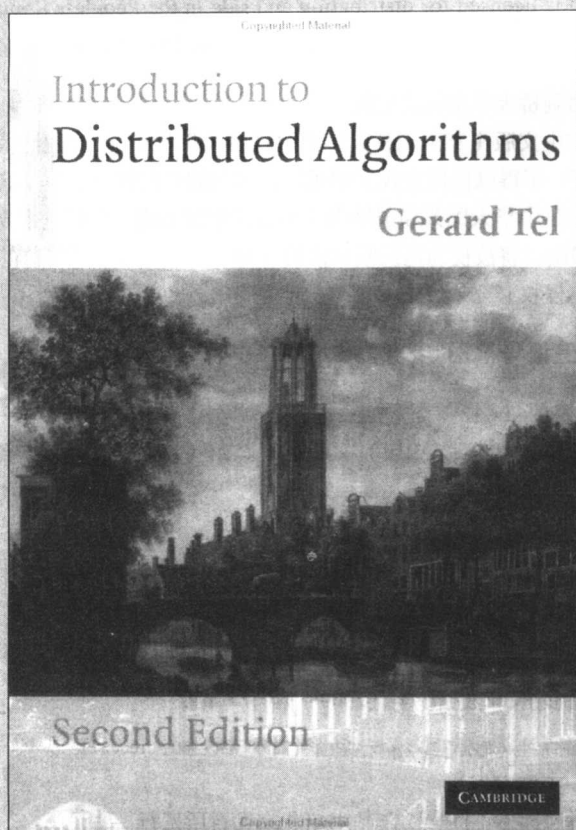
机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

原书第2版

分布式算法导论

(荷) Gerard Tel 著 霍红卫 译



**Introduction to Distributed Algorithms
(Second Edition)**



机械工业出版社
China Machine Press

本书详细介绍了分布式算法及其理论,结合大量定理、引理、命题等的证明,讨论了点对点消息传递模型上的算法、计算机通信网络中实现的算法,重点是分布式应用的控制算法(如波动算法、广播算法、选举算法、同步系统算法等),还涉及了利用分布式算法实现容错计算、方向侦听和故障检测器等方面的内容。本书条理清晰、深入浅出,适合作为大学本科高年级和研究生的分布式算法课程的教材和参考书,对于具有实践经验的专业人员也大有帮助。

Gerard Tel: Introduction to Distributed Algorithms, Second Edition(ISBN 0-521-79483-8).

Originally published by Cambridge University Press in 1994, 2000.

This Chinese edition is published with the permission of the Syndicate of the Press of the University of Cambridge, Cambridge, England.

Copyright © 2004 by Cambridge University Press.

This edition is licensed for distribution and sale in the People's Republic of China only, excluding Hong Kong, Taiwan and Macao and may not be distributed and sold elsewhere.

本书原版由剑桥大学出版社出版。

本书简体字中文版由英国剑桥大学出版社授权机械工业出版社独家出版。未经出版者预先书面许可,不得以任何方式复制或抄袭本书的任何部分。

此版本仅限在中华人民共和国境内(不包括中国香港、台湾、澳门地区)销售发行,未经授权的本书出口将被视为违反版权法的行为。

版权所有,侵权必究。

本书版权登记号: 图字: 01-2002-1180

图书在版编目(CIP)数据

分布式算法导论(原书第2版) / (荷)泰尔(Tel, G.)著; 霍红卫译. -北京: 机械工业出版社, 2004.9

(计算机科学丛书)

书名原文: Introduction to Distributed Algorithms, Second Edition

ISBN 7-111-14674-3

I. 分… II. ①泰… ②霍… III. 电子计算机-算法理论 IV. TP301.6

中国版本图书馆CIP数据核字(2004)第058300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: 王镇元

北京昌平奔腾印刷厂印刷·新华书店北京发行所发行

2004年9月第1版第1次印刷

787mm × 1092mm 1/16 · 25印张

印数: 0 001-3000 册

定价: 39.00元

凡购本书,如有倒页、脱页、缺页,由本社发行部调换
本社购书热线:(010) 68326294

前 言

在过去几年里，分布式系统和分布式信息处理得到广泛关注。几乎每一所大学都会开设至少一门关于分布式算法设计的课程。出现了许多关于分布式系统原理方面的书籍，例如Tanenbaum [Tan96] 和Sloman & Kramer [SK87]，但这些书籍主要是针对结构而不是算法。自从本书第1版问世以来，相继出版的分布式算法方面的著作有Barbosa [Bar96]、Lynch [Lyn96] 和Attiya & Welch [AW98]。

因为算法是计算机应用的基础，因此需要一本专门介绍分布式算法的书。本书的目的是展示过去20年来分布式算法方面的诸多理论。本书可作为分布式算法1~2学期的教学用书，一学期的课程可由教师从本书中选择若干专题来安排。

本书也可作为相关专业的工程师和从事分布式系统研究的科研人员的参考书。

练习。每章（除第1章和第13章）后面附有一些习题和项目。项目通常要求读者开发涉及该章内容的一些应用。大多数情况下，没有提供“答案”。

致谢。本书经以下人士的仔细校对。他们是：Erwin Bakker、Hans Bodlaender、Stefan Dobrev、Petra van Haaften、Ted Herman、Jan van Leeuwen、Patrick Lentfert、Friedemann Mattern、Pascale van der Put、Peter Ružička、Martin Rudalics、Anneke Schoone和Kaisa Sere。他们对手稿质量的改进提出了有益的意见。此外，在Utrecht大学选修秋季“分布式算法”课程的学生们也提供了有益的建议。计算机科学系为所需的文本处理和输出提供了技术支持。Susan Parkinson进行了文字编辑。

Gerard Tel

1994年4月/2000年2月

译者序

本书是关于分布式算法的最优秀的著作之一。它系统地阐述了分布式算法设计的理论、方法和应用实例。目前,国内尚缺少专门介绍分布式算法的著作。我们希望本书能对我国高等院校的计算机教育有所帮助。

在过去的二十年里,分布式算法一直是备受关注的研究课题。这本成功的教科书的第二版,介绍了分布式算法研究领域的最新进展。新增了两章关于方向侦听和故障检测器的内容,代表了当今该领域最新技术发展水平。

本书分四部分:协议(第2章~第5章)、基本算法(第6章~第12章)、容错(第13章~第17章)和附录(附录A、附录B)。书中内容全面阐述了过去20年来分布式算法方面的诸多理论。本书主要内容及特点如下:

- 第一部分介绍了分布式系统和通信网络的基本概念,讨论了平衡滑动窗口协议和基于计时器的协议,以严谨简明的形式对路由算法作了系统论述,最后讨论了缓冲区有限时无死锁的包交换问题。
- 第二部分讨论了基本算法。包括:波动算法、遍历算法、广播算法、选举算法、终止检测算法、匿名网络的随机算法、快照算法、方向侦听与定向算法、死锁检测算法和同步系统算法。
- 第三部分讨论了容错问题。引入了健壮算法和稳定算法的概念。证明了同步系统的健壮性要比异步系统更大。最后讨论了故障检测和稳定算法。
- 第四部分介绍了伪代码使用约定、图和网络中的一些基本概念和常用术语。

所有算法既给出严格的数学定义及类Pascal语言的形式描述,又以算法不变式作为手段给出算法正确性的形式证明,充分反映了作者在分布式算法方面的造诣。

本书适合作为高等院校分布式算法、分布式计算课程的本科生和研究生教材,同时可作为从事分布式系统设计与应用的专业人员的参考书。

由于时间较紧及译者水平有限,译文难免有错误及不妥之处,恳请读者批评指正。

霍红卫

西安电子科技大学计算机学院

2003年12月

译者简介



霍红卫,1963年8月出生,博士。现为西安电子科技大学计算机学院教授。主要研究方向:算法分析与设计、并行与分布式计算、遗传算法、生物信息学中的优化算法。著作有:《算法设计与分析》、《并行分类算法》和《Exercises & Solutions on Algorithms》。

作者序

With great pleasure I welcome this publication of the Chinese translation of my book——《Introduction to Distributed Algorithms》.

In recent years, the flourishing economy of China greatly promotes the development of science and technology. Cooperation between China and other countries in IT industry is increasingly strengthened. More and more Chinese researchers, along with their counterparts in other countries, participate in various distributed computing projects both at home and abroad.

All these distributed computing projects need algorithms for cooperation, coordination, information exchange, overcoming failures, etc.

The topics discussed in this text will promote a fundamental style of thinking about algorithms, mathematical proofs, specifications and models. They may not only be helpful in studying existing methods, but also lay a intellectual foundation for studying new problems. I hope that many Chinese readers will find the book useful.

Acknowledgements:

Thanks are due to the China Machine Press for planning and printing this edition of my book in the Chinese language.

Special thanks go to Prof. Hongwei Huo of Xidian University for she has undertaken the enormous work of translating this book into the Chinese language.

Gerard Tel, June 2004

欣闻我的著作——《分布式算法导论》中文版即将出版，非常高兴。

最近几年，繁荣的中国经济极大地促进了科学技术的发展。IT界的中外合作日益加强。越来越多的中国研究人员与国外同行一起参加到国内外各种分布式计算工程中。所有这些分布式计算工程都需要协调合作、信息交换和故障排除的算法。

本书讨论的课题将促进对算法、数学证明、说明以及模型的根本思考。它们不仅对研究已有的方法大有帮助，而且会为研究新的问题奠定知识基础。我希望广大中国读者会从本书中受益。

致谢：

非常感谢机械工业出版社策划和出版了本书的中文版。

还要特别感谢西安电子科技大学计算机学院的霍红卫教授承担了本书的翻译工作。

Gerard Tel
2004年6月

专家指导委员会

(按姓氏笔画顺序)

尤晋元	王 珊	冯博琴	史忠植	史美林
石教英	吕 建	孙玉芳	吴世忠	吴时霖
张立昂	李伟琴	李师贤	李建中	杨冬青
邵维忠	陆丽娜	陆鑫达	陈向群	周伯生
周立柱	周克定	周傲英	孟小峰	岳丽华
范 明	郑国梁	施伯乐	钟玉琢	唐世渭
袁崇义	高传善	梅 宏	程 旭	程时端
谢希仁	裘宗燕	戴 葵		

秘 书 组

武卫东

温莉芳

刘 江

杨海玲

出版者的话

文艺复兴以来，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭橥了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短、从业人员较少的现状下，美国等发达国家在其计算机科学发展的几十年间积淀的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章图文信息有限公司较早意识到“出版要为教育服务”。自1998年开始，华章公司就将工作重点放在了遴选、移译国外优秀教材上。经过几年的不懈努力，我们与Prentice Hall, Addison-Wesley, McGraw-Hill, Morgan Kaufmann等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出Tanenbaum, Stroustrup, Kernighan, Jim Gray等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及度藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专诚为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍，为进一步推广与发展打下了坚实的基础。

随着学科建设的初步完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都步入一个新的阶段。为此，华章公司将加大引进教材的力度，在“华章教育”的总规划之下出版三个系列的计算机教材：除“计算机科学丛书”之外，对影印版的教材，则单独开辟出“经典原版书库”；同时，引进全美通行的教学辅导书“Schaum's Outlines”系列组成“全美经典学习指导系列”。为了保证这三套丛书的权威性，同时也为了更好地为学校和老师服务，华章公司聘请了中国科学院、北京大学、清华大学、国防科技大学、复旦大学、上海交通大学、南京大学、浙江大学、中国科技大学、哈尔滨工业大学、西安交通大学、中国人民大学、北京航空航天大学、北京邮电大学、中山大学、解放军理工大学、郑州大学、湖北工学院、中国国家信息安全测评认证中心等国内重点大学和科研机构在计算机的各个领域的著名学者组成“专家指导委员会”，为我们提供选题意见和出版监督。

这三套丛书是响应教育部提出的使用外版教材的号召，为国内高校的计算机及相关专业

的教学度身订造的。其中许多教材均已为M. I. T., Stanford, U.C. Berkeley, C. M. U. 等世界名牌大学所采用。不仅涵盖了程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等国内大学计算机专业普遍开设的核心课程,而且各具特色——有的出自语言设计者之手、有的历经三十年而不衰、有的已被全世界的几百所高校采用。在这些圆熟通博的名师大作的指引之下,读者必将在计算机科学的宫殿中由登堂而入室。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑,这些因素使我们的图书有了质量的保证,但我们的目标是尽善尽美,而反馈的意见正是我们达到这一终极目标的重要帮助。教材的出版只是我们的后续服务的起点。华章公司欢迎老师和读者对我们的工作提出建议或给予指正,我们的联系方法如下:

电子邮件: hzedu@hzbook.com

联系电话: (010) 68995264

联系地址: 北京市西城区百万庄南街1号

邮政编码: 100037

4.5 分级路由	90
习题	92
第5章 无死锁的包交换	93
5.1 引言	93
5.2 有结构的方法	94
5.2.1 缓冲图	95
5.2.2 图G的定向	97
5.3 无结构的方法	100
5.3.1 前向计数控制器和后向计数 控制器	100
5.3.2 前向状态控制器和后向状态 控制器	101
5.4 需进一步研究的问题	102
5.4.1 拓扑变化	102
5.4.2 其他类型的死锁	103
5.4.3 活锁	104
习题	105

第二部分 基本算法

第6章 波动算法与遍历算法	107
6.1 波动算法的定义和使用	107
6.1.1 波动算法定义	107
6.1.2 波动算法的一些基本结果	109
6.1.3 具有反馈的信息传播	110
6.1.4 同步	111
6.1.5 计算下确界函数	111
6.2 波动算法集	112
6.2.1 环网算法	112
6.2.2 树算法	113
6.2.3 回波算法	115
6.2.4 轮询算法	116
6.2.5 相位算法	117
6.2.6 Finn算法	118
6.3 遍历算法	120
6.3.1 遍历团	121
6.3.2 遍历圆环	121
6.3.3 遍历超立方体	122
6.3.4 遍历连通网络	123
6.4 深度优先搜索的时间复杂度	124
6.4.1 分布式深度优先搜索	125

6.4.2 线性时间的深度优先搜索算法	126
6.4.3 具有近邻知识的深度优先搜索	130
6.5 遗留问题	130
6.5.1 波动算法综述	130
6.5.2 计算和	130
6.5.3 时间复杂度的另一种定义	132
习题	134
第7章 选举算法	137
7.1 引言	137
7.1.1 本章所做假设	138
7.1.2 选举和波动	138
7.2 环网	140
7.2.1 LeLann和Chang-Roberts算法	140
7.2.2 Peterson/Dolev-Klawe-Rodeh算法	144
7.2.3 一个下界	146
7.3 任意网	148
7.3.1 废止和快速算法	149
7.3.2 Gallager-Humblet-Spira算法	151
7.3.3 GHS算法的全局描述	152
7.3.4 GHS算法的详细描述	153
7.3.5 GHS算法的讨论和变化	157
7.4 Korach-Kutten-Moran算法	158
7.4.1 模块构造	158
7.4.2 KKM算法的应用	161
习题	162
第8章 终止检测	165
8.1 预备知识	165
8.1.1 定义	165
8.1.2 两个下界	167
8.1.3 终止进程	169
8.2 计算树和森林	169
8.2.1 Dijkstra-Scholten算法	169
8.2.2 Shavit-Francez算法	172
8.3 基于波动的方法	175
8.3.1 Dijkstra-Feijen-Van Gasteren算法	175
8.3.2 基本消息的计数: Safra算法	178
8.3.3 利用确认	181
8.3.4 带波动的终止检测	183
8.4 其他方法	184
8.4.1 信用-恢复算法	184

13.2.3 第14章到第16章综述	270	第16章 故障检测	315
13.2.4 本书中没有涉及的主题	271	16.1 模型和定义	315
13.3 稳定算法	271	16.1.1 四种基本检测器类型	316
第14章 异步系统中的容错	273	16.1.2 故障检测器的用途和缺陷	317
14.1 一致性的不可能性	273	16.2 用弱精确检测器解一致性问题	318
14.1.1 表示、定义及基本结果	273	16.3 最终弱精确检测器	319
14.1.2 不可能性证明	274	16.3.1 弹性上界	319
14.1.3 讨论	275	16.3.2 一致算法	320
14.2 初始死进程	276	16.4 故障检测器的实现	321
14.3 确定可实现实例	277	16.4.1 同步系统: 完美检测	321
14.3.1 可解问题: 重命名	278	16.4.2 部分同步系统: 最终完美检测	321
14.3.2 扩展的不可能性结果	280	16.4.3 小结	322
14.4 概率一致性算法	282	习题	323
14.4.1 损毁-健壮一致协议	282	第17章 稳定性	325
14.4.2 Byzantine-健壮一致性协议	285	17.1 引言	325
14.5 弱终止性	288	17.1.1 定义	325
习题	290	17.1.2 稳定系统中的通信	326
第15章 同步系统中的容错	293	17.1.3 例子: Dijkstra令牌环	327
15.1 同步判定协议	293	17.2 图论算法	329
15.1.1 弹性界限	294	17.2.1 环定向	329
15.1.2 Byzantine广播算法	295	17.2.2 最大匹配	331
15.1.3 多项式级的广播算法	297	17.2.3 选举和生成树构造	332
15.2 鉴别协议	300	17.3 稳定方法学	334
15.2.1 高度弹性的协议	301	17.3.1 协议组合	334
15.2.2 数字签名的实现	303	17.3.2 计算最小路径	338
15.2.3 ElGamal签名模式	303	17.3.3 结论和讨论	342
15.2.4 RSA签名模式	304	习题	342
15.2.5 Fiat-Shamir签名模式	305		
15.2.6 概述和讨论	306		
15.3 时钟同步	308		
15.3.1 读取远程时钟	308		
15.3.2 分布式时钟同步	310		
15.3.3 轮模型的实现	313		
习题	314		
		第四部分 附录	
		附录A 伪代码使用约定	345
		附录B 图和网络	349
		参考文献	359
		主题词索引	375

第1章 导论: 分布式系统

本章简要介绍了研制分布式算法所依据的硬件和软件系统。阐述了研究分布式算法的原因。通过一个分布式系统表明若干台计算机或处理器相互协作完成计算机应用。分布式系统的定义不仅包括广域计算机通信网, 还有局域网、多处理器计算机(每一台处理器都有自己的控制单元)及协同处理系统。

1.1节介绍了各种类型的分布式系统, 讨论了使用分布式系统的原因, 给出了现有分布式系统的例子。然而, 本书的主题既不是这些系统的构成, 也不是如何使用它们, 而是它们的工作原理。同时, 对分布式系统中所使用的算法做专门研究。

当然, 仅研究分布式系统的算法是不能完全理解它的整体结构和操作的。为了充分理解这一系统, 还必须研究硬件及软件的整个体系结构, 即把功能划分成模块。同时还需要研究与程序设计语言性质有关的重要问题, 而程序设计语言可用于构建分布式系统中的软件, 1.2节将会讨论这些主题。

现有的一些关于分布式系统方面的优秀著作, 比如, Tanenbaum [Tan96]、Sloman and Kramer [SK87]、Bal [Bal90]、Coulouris and Dollimore [CD88] or Goscinski [Gos91] 侧重于系统的体系结构和语言方面。正如已经提到的那样, 本书主要讨论分布式系统的算法。1.3节解释了分布式算法设计与集中式算法设计的差异。概略地描述了分布式算法的研究领域, 并概括了本书的其余部分。

1

1.1 分布式系统的定义

本章中的“分布式系统”表示自主计算机、进程或者处理器互连的集合。计算机、进程或者处理器称为分布式系统中的节点 (node) (在后续章节中, 将使用更具技术性的表示方法, 参见定义2.6)。“自主性”是指节点必须至少配备自己专用的控制单元。因此, 单指令多数据 (SIMD) 模型的并行计算机并不是分布式系统。“互连性”是指节点之间必须能够交换信息。

由于(软件)进程能够起着系统中节点的作用, 因此这个定义包括了作为通信进程集合而构造的软件系统, 即使是运行在单一硬件装置上。然而, 在大多数情况下, 一个分布式系统至少包含几个处理器, 这些处理器通过通信硬件互连。

在文献中可以看到对分布式系统更严格的定义。例如, Tanenbaum[Tan96]认为, 仅当系统中自主性节点的存在对用户是透明的时候, 一个系统才称为分布式系统。从这种意义上讲, 分布式系统就像是一个虚拟的、单一的计算机系统, 但要实现这种透明性, 需要开发复杂的分布式控制算法。

1.1.1 动机

分布式计算机系统比顺序系统更可取, 或者说, 它们的使用是必然的。我们将对其原因进

行分析, 下面的分析只是部分原因。分布式系统的选择受到下列诸多因素的影响, 有时可能源于其他原因, 但其优势随之显现。分布式系统的特征可能会随着其存在的形式不同而不同。

1.1.2节1.1.6节将会更详细地讨论这些问题。

2 (1) 信息交换 在60年代, 当多数大学和公司开始拥有自己的大型机时, 产生了在不同计算机间进行数据交换的需求。不同机构的人员通过这些机构的计算机交换数据, 使合作变得更为便利。从而引发了对所谓的广域网 (wide-area network, WAN) 的开发。当今因特网的前身ARPANET于1969年12月投入使用。广域网 (有时称远程网络, long-haul network) 中所连接的计算机通常都配有用户所需的各种设备, 如备份存储器、磁盘、各种应用程序及打印机等。

后来, 计算机体积变得越来越小, 价格越来越便宜, 很快, 每一个机构都有了许多计算机, 现在, 常常是人手一台计算机 (个人计算机或工作站)。因此, 机构人员间的信息 (电子) 交换要求自主性计算机互连。甚至有的个人或家庭将多台计算机连接成一小型个人家庭网络也是很常见的。

(2) 资源共享 尽管由于计算机的价格便宜, 机构可为每一位员工提供一台个人计算机, 但对于外围设备 (如打印机、存储设备和磁盘等) 情况却并非如此。在较小规模内, 每台计算机可以依靠专用服务器提供编译器和其他应用程序。此外, 在所有计算机上进行应用程序和相关文件资源的复制, 则效率低下; 除了浪费磁盘空间, 还会引起不必要的维护问题。因此, 用户计算机要依靠专用节点进行打印输出和磁盘服务。在一个组织内部连接而成的计算机网络称为局域网 (local-area network, LAN)。

对于机构而言, 建立小型计算机组成的网络, 而不是购买大型机, 不仅是为了降低系统成本, 而且是使系统具有可扩充性。首先, 较之大型计算机, 小型计算机系统有更好的性能价格比; 虽然典型的大型机要比个人计算机快50倍, 但其价格却是个人计算机的500倍之多。其次, 如果一个小型计算机系统的容量不足, 可根据机构需要在网络中增加机器 (文件服务器、打印机和工作站)。而如果单机系统的容量不够, 则只能更换。

3 (3) 通过复制提高可靠性 分布式系统较之单机系统更具可靠性。这是因为它们有局部故障 (partial-failure) 的性质。这表明, 如果系统中的某些节点发生故障, 其他节点仍可正常运行并且能够接管故障的部分。而单机系统的故障则会影响整个系统, 在这种情况下, 系统不可能继续运行。因此, 在设计高可靠性计算机系统时, 分布式体系结构常常更受到关注。

高可靠性系统一般有多个处理器, 是运行一个应用程序的单处理器的二至四倍用程序并利用表决机制决定机器的输出。因此, 当系统的某个部件发生故障时, 要使分布式系统正常运行, 需要相当复杂的算法支持。

(4) 通过并行化提高性能 由于分布式系统中存在多处理器, 人们可以把计算密集的作业进行划分, 并分布到若干台处理器上分别处理, 来减少作业的处理时间。

并行计算机的工作原理就是如此, 但通过将任务分配到其他工作站上进行并行处理, 局域网的用户也可受益。

(5) 通过规范简化设计 计算机系统设计相当复杂, 尤其是有相当多的功能要求时。将系统分成模块, 可简化系统设计, 每一模块实现部分功能, 并可与其他模块通信。

通过定义抽象数据类型和不同任务过程,可以得到独立的程序模块。也可将一个大的系统定义为协同进程的集合。在这两种情况下,模块都可以在单一计算机上执行。但在一个局域网中可能会有不同类型的计算机,如,有一台配备专门进行密集数值计算(number crunching)的硬件的计算机,有一台带有制图硬件的计算机,还有一台配备了磁盘等设备。

1.1.2 计算机网络

计算机网络是由通信设备把多个计算机互连而成的计算机集合,通过这些设备实现信息交换。交换是通过接收和发送信息实现的。计算机网络与分布式系统的定义不谋而合。根据计算机之间的距离,计算机网络可分为广域网和局域网。

广域网通常连接不同机构(各行业、大学之间等)的计算机。节点间的物理距离不少于10公里。网络中的每一节点是一台完整的计算机,包括所有的外围设备和相当数量的应用程序。广域网的主要目标是实现不同节点间用户的信息交换。

局域网通常是连接某一机构内部的计算机。节点间的物理距离不大于10公里。网络中的节点可以是工作站、文件服务器或者打印服务器,即机构内部的用于特定功能的相对较小的专用工作站。局域网的主要目标是实现信息交换和资源共享。

这两种类型的网络的界限并不总能清晰界定。从算法观点来看,这种区分不是很重要。因为所有计算机网络中的算法类似。以下列出与算法开发有关的因素。

(1) 可靠性参数 在广域网中,不能忽视消息传递过程中发生错误的可能性。通常,广域网的分布式算法就设计成要处理这种可能性。局域网就可靠得多。在算法设计时,通常假设通信是完全可靠的。然而,在这种情况下,出错的不可能事件可能会难以察觉地发生,并导致系统操作错误。

(2) 通信时间 广域网中的消息传输时间要比局域网中的消息传输时间大若干数量级。在广域网中,较之消息的传输时间,消息的处理时间几乎可以忽略。

(3) 同一性 即使在局域网中,所有的节点也未必相同。通常在一个机构内部,可能会赞同采用共同的软件和协议。在广域网中,有多种协议,这就使得在不同协议间转换和设计出能兼容不同标准的软件变得相当困难。

(4) 互信 在同一机构内部,所有用户都互相信任。但在广域网中,决不是这种情况。广域网需要开发安全算法,防范其他节点非法用户的入侵。

1.1.3节和1.1.4节分别简要讨论了广域网络和局域网络。

1.1.3 广域网络

1. 发展历史

在开发广域网络的过程中,大量早期开创性的工作是在美国国防部的高级研究计划局(Advanced Research Projects Agency, ARPA)的一个项目中完成的。1969年,ARPANET投入运行。当时连接了4个节点。现在这个网络已增长到数百个节点。利用类似的一些技术(MILNET、CYPRESS等),还建立了其他一些网络。ARPANET包括一些特殊节点(称为接口信息处理器,IMP),其惟一目的是处理大量信息。

当UNIX[⊖]系统广泛流行时,人们意识到需要在不同UNIX机器间进行信息交换。为了达

⊖ UNIX是AT&T贝尔实验室的注册商标。

到这一目的, 编写了uucp (UNIX-to-UNIX CoPy) 程序。使用此程序, 通过电话线就可以进行文件交换, 称为UUCP网络的UNIX用户网络迅速出现。由于ARPANET属于国防部, 仅有某些机构可以与之相连, 因此80年代, 开发出另一个主要的网络, 称为BITNET。

如今, 所有这些网络都是互连的; 有一些节点连接了两种网络 (称为网关), 允许在不同网络的节点之间进行信息交换。统一地址空间和通用协议的引入把网络变成一个单一的虚拟网络, 通常称为因特网。不同于“单一”网络, 因特网有许多用户, 并且没有一个权威机构。但其组织上的多样性对于用户是仔细隐藏的。作者的邮件地址 (gerard@cs.uu.nl) 并没有提供其部门所连接的网络的信息。

2. 组织和算法上的问题

6 广域网总是被组成点对点网络。这表明, 只有通过一种专门用于这两个节点的连接机制, 才会发生一对节点间的通信。这种装置可以是电话线、光纤或者卫星网等。点对点的互连结构可以用图直观地表示出来, 用圆圈或方框表示网络中的节点, 节点之间的边表示两个节点间的通信线路, 如图1-1所示。用更技术化的语言, 用图表示结构, 其中边代表网络的通信线路。关于图论中的一些术语在附录B中给出。

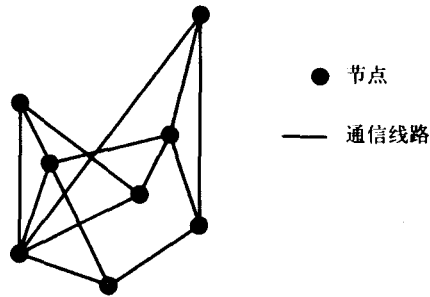


图1-1 点对点网络的一个例子

广域网的主要目的是进行信息交换, 如电子邮件、电子公告和远程文件与数据库访问。通过一个应用程序: 网页浏览器, 可以使用大多数服务。要完成一个能实现这些目标的适合的通信系统, 需要解决下列算法问题, 其中一些问题在本书第一部分讨论。

7 (1) 点对点数据交换的可靠性 (第3章) 通过连接两节点间的线路进行数据交换, 必须处理可能会发生的线路不可靠问题。由于大气噪声、动力受损以及其他的物质环境影响, 通过线路所发送的消息可能只接收到一部分, 甚至丢失。因此必须辨别这些传输故障并纠正。

这个问题不仅发生在用通信线路直接相连的两点间, 而且发生在借助中间节点通信的未直接相连的节点之间。在这种情况下, 问题更复杂, 因为消息到达的次序可能与发送次序不同, 可能在很长时间以后到达或者被复制。

(2) 通信路径的选择 (第4章) 在点对点的网络中, 由于费用昂贵, 不可能在每一对节点之间都提供通信线路。因此, 某些节点必须依靠其他节点进行通信。路由问题所关注的是如何在要通信的节点之间选择一条 (或多条) 路径。所用的路径选举算法与节点命名模式有关, 例如, 某节点用于向另一个节点发送消息的地址格式。中间节点的路径选择利用这一地址来进行, 而如果能将拓扑信息“编码”在地址中, 就能选择更有效的路径。

(3) 拥塞控制 如果同时传送大量信息, 通信网络的吞吐量就会急剧下降。必须控制各节点所产生的信息, 并使其适合网络的可用容量。文献 [Tan96, 5.3节] 讨论了几种避免拥