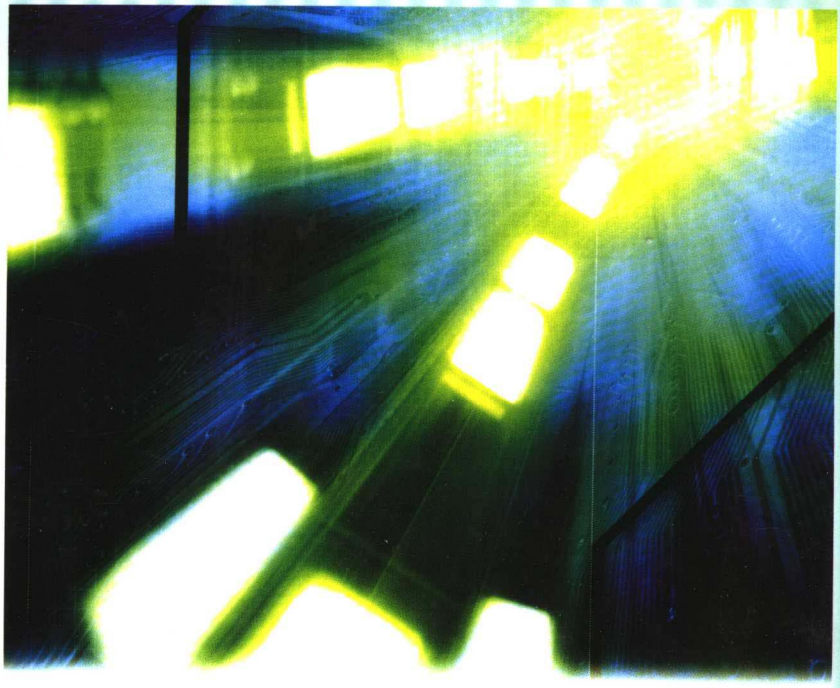
经 典 原 版 书 库

# 分布式系统
## 概念与设计

新版

（英文版 · 第4版）



fourth edition

# DISTRIBUTED SYSTEMS
## CONCEPTS AND DESIGN

George Coulouris
Jean Dollimore
Tim Kindberg

（英）George Coulouris Jean Dollimore Tim Kindberg 著

机械工业出版社
China Machine Press

经 典 原 版 书 库

# 分布式系统
## 概念与设计
（英文版 · 第4版）

# Distributed Systems
## Concepts and Design

(Fourth Edition)

(英) George Coulouris
Jean Dollimore 著
Tim Kindberg

机械工业出版社
China Machine Press

George Coulouris, Jean Dollimore, and Tim Kindberg: Distributed Systems: Concepts and Design, Fourth Edition (ISBN 0-321-26354-5).



**本书版权登记号：图字：01-2005-4693**

**图书在版编目（CIP）数据**

分布式系统：概念与设计（英文版·第4版）/（英）库劳里斯（Coulouris, G.）等著；　-北京：机械工业出版社，2006.1

（经典原版书库）

书名原文：Distributed Systems: Concepts and Design, Fourth Edition

ISBN 7-111-17366-X

Ⅰ.分…　Ⅱ.库…　Ⅲ.分布式操作系统-英文　Ⅳ.TP316.4

中国版本图书馆CIP数据核字（2005）第104938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：迟振春

北京瑞德印刷有限公司印刷·新华书店北京发行所发行

2006年1月第1版第1次印刷

718mm × 1020mm 1/16 · 59印张

印数：0 001- 3 000册

定价：89.00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

# 出版者的话

文艺复兴以降，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、研究及庋藏。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，有的还专程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们服务，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：hzjsj@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

# 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 | | | |

# PREFACE

This fourth edition of our textbook appears at a time when the Internet and the Web are mature systems, supporting a wide variety of distributed applications on a scale far greater than could have been anticipated when our third edition was published almost five years ago.

The book aims to provide an understanding of the principles on which the Internet and other distributed systems are based, their architecture, algorithms and design. We begin with two conceptual overview chapters that outline the characteristics of distributed systems and the challenges that must be addressed in their design: scalability, heterogeneity, security and failure handling being the most significant. These chapters also develop abstract models for understanding process interaction, failure and security. They are followed by foundational chapters devoted to the study of networking, interprocess communication, remote invocation and middleware, operating system support and naming.

We then cover the well-established topics of security, data replication, group communication, distributed file systems, distributed transactions, CORBA, distributed shared memory and multimedia systems together with several new ones: Web Services, XML, the Grid, peer-to-peer, mobile and ubiquitous systems. Algorithms associated with all these topics are covered as they arise and also in separate chapters devoted to timing, coordination and agreement.

## Purposes and readership

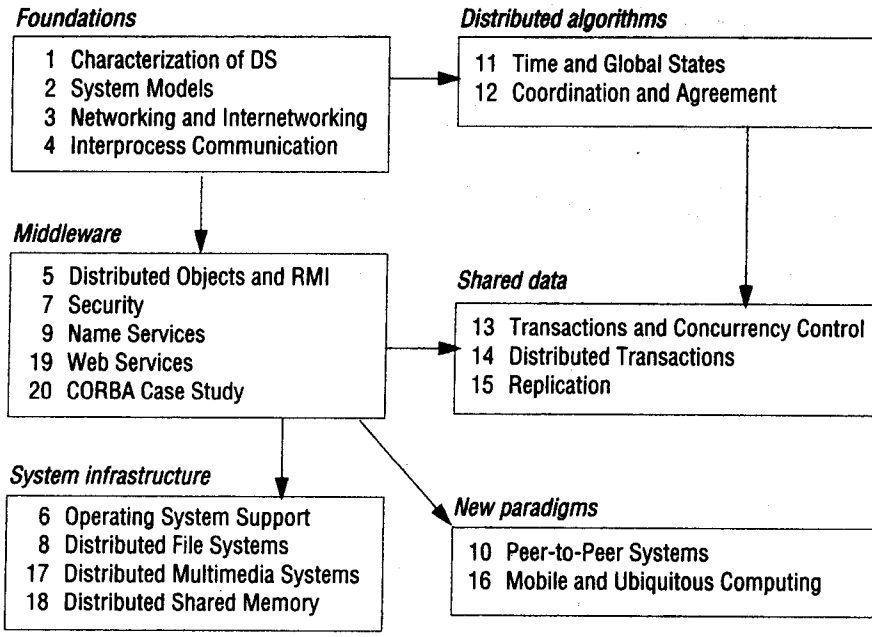
The book is intended for use in undergraduate and introductory postgraduate courses. It can equally be used for self-study. We take a top-down approach, addressing the issues to be resolved in the design of distributed systems and describing successful approaches in the form of abstract models, algorithms and detailed case studies of widely-used systems. We cover the field in sufficient depth and breadth to enable readers to go on to study most research papers in the literature on distributed systems.

We aim to make the subject accessible to students who have a basic knowledge of object oriented programming, operating systems and elementary computer architecture. The book includes coverage of those aspects of computer networks relevant to distributed systems, including the underlying technologies for the Internet, wide area, local area and wireless networks. Algorithms and interfaces are presented throughout

the book in Java or, in a few cases, ANSI C. For brevity and clarity of presentation, a form of pseudo-code derived from Java/C is also used.

## Organization of the book

The following diagram shows the book's chapters under six main topic areas. It is intended to provide a guide to the book's structure and to indicate recommended navigation routes for instructors wishing to provide, or readers wishing to achieve, understanding of the various subfields of distributed system design:




## References

The existence of the World Wide Web has changed the way in which a book such as this can be linked to source material, including research papers, technical specifications and standards. Many of the source documents are now available on the Web; some are available only there. For reasons of brevity and readability, we employ a special form of reference to web material which loosely resembles a URL: references such as [www.omg.org] and [www.rsasecurity.com I] refer to documentation that is available only on the Web. They can be looked up in the reference list at the end of the book, but the full URLs are given only in an online version of the reference list at the book's web site: www.cdk4.net/refs where they take the form of clickable links. Both versions of the reference list include a more detailed explanation of this scheme.

| *Entirely new chapters:* | |
|---|---|
| 10 Peer-to-Peer Systems | Chapters from 10 onwards have new numbering in this edition. |
| 16 Mobile and Ubiquitous Computing | |
| 19 Web Services | |
| ***Chapters to which new material has been added, but without structural changes:*** | |
| 1 Characterization of DS | Section 1.3.1: updated to introduce web services |
| 2 System Models | Section 2.2.2: updated to introduce peer-to-peer |
| 3 Networking and Internetworking | Many updates<br>New Section 3.5.3: Case Study: Bluetooth |
| 4 Interprocess Communication | New Section 4.3.3: XML |
| 7 Security | Several updates<br>New Section 7.6.4: weaknesses of WiFi |
| 9 Name Services | Section 9.1.1: section on URIs updated |
| 20 CORBA Case Study | Section 20.2.1: upgraded to Java 2 vn 1.4<br>Section 20.2.6: integration with web services |

*The remaining chapters have received only minor modifications.*

## Changes relative to the third edition

Before embarking on the writing of this new edition, we carried out a survey of teachers who used the third edition. From the results, we identified the new material required and the changes to be made. This led to our writing three entirely new chapters and making numerous insertions throughout the book. All the chapters have been changed to reflect new information that has become available about the systems described. However, to help teachers who used the third edition, we have left the structure of the existing chapters almost unchanged. The new chapters and those containing substantial changes are listed in the table above. The Mach case study chapter has been removed and is available from the book's web site, together with several smaller case studies that were removed from the second and third editions.

## Acknowledgements

We are very grateful to the following teachers who participated in our survey: Kay Robbins, Kohei Honda, Stefan Leue and Ian Wakeman.

We would like to thank the following people who reviewed the new chapters or provided other substantial help: John Barton, Arne Glenstrup, Roy Logie, Friedemann Mattern, Christian Mortensen, Anthony Rowstron, Bo Sanden, Dave Scott, Ben Smyth, Mirjana Spasojevic, Salman Taherian, Andrew Twigg, Jim Waldo, Eiko Yoneki, Kan Zhang and Ben Zhao.

The Department of Computer Science, Queen Mary College, University of London, has hosted the companion web site for the third edition and has agreed to host the site for the fourth edition. We thank the department for its support and Keith Clarke and the systems team for their help in setting up and maintaining these sites.

Finally, we thank Simon Plumtree, Bridget Allen, Mary Lince and Owen Knight of Pearson Education/Addison-Wesley for essential support throughout the arduous process of getting the book into print.

## Web site

As before, we shall maintain a web site with a wide range of material designed to assist teachers and readers. This web site can be accessed via either of the URLs:

www.cdk4.net www.pearsoned.co.uk/coulouris

The web site includes:

Instructor's Guide: Comprising:

- Complete artwork of the book available as PowerPoint files.
- Solutions to the exercises (protected by a password available only to teachers).
- Chapter-by-chapter teaching hints.
- Suggested laboratory projects.

Reference list: The list of references that can be found at the end of the book is replicated at the web site. The web version of the reference list includes active links for material that is available online.

Errata list: A list of known errors in the book, with corrections for each one. As with the third edition, the errors will be corrected in new impressions and a separate errata list will be provided for each impression.

Supplementary material: We maintain a set of supplementary material for each chapter. This consists of source code for the programs in the book and relevant reading material that was present in previous editions of the book but was removed for reasons of space. References to this supplementary material appear in the book with links such as www.cdk4.net/ipc.

Links to web sites for courses using the book: The web site for the third edition contains links to 15 courses using our book, which make available a wealth of useful lecture notes, slides, exercises and laboratory projects. We hope to get permission from the teachers of these courses to put these references on the new web site. Other teachers are asked to notify us of their courses with web sites for inclusion in the list.

*George Coulouris*
*Jean Dollimore*
*Tim Kindberg*
London and Bristol, March 2005
***<authors@cdk4.net>***

# CONTENTS


PREFACE vii

1 CHARACTERIZATION OF DISTRIBUTED SYSTEMS 1
- 1.1 Introduction 2
- 1.2 Examples of distributed systems 3
- 1.3 Resource sharing and the Web 7
- 1.4 Challenges 16
- 1.5 Summary 25

2 SYSTEM MODELS 29
- 2.1 Introduction 30
- 2.2 Architectural models 31
- 2.3 Fundamental models 47
- 2.4 Summary 61

3 NETWORKING AND INTERNETWORKING 65
- 3.1 Introduction 66
- 3.2 Types of network 69
- 3.3 Network principles 73
- 3.4 Internet protocols 89
- 3.5 Case studies: Ethernet, WiFi, Bluetooth and ATM 112
- 3.6 Summary 127

4 INTERPROCESS COMMUNICATION 131

4.1 Introduction 132
4.2 The API for the Internet protocols 133
4.3 External data representation and marshalling 144
4.4 Client-server communication 155
4.5 Group communication 164
4.6 Case study: interprocess communication in UNIX 168
4.7 Summary 172

5 DISTRIBUTED OBJECTS AND REMOTE INVOCATION 177

5.1 Introduction 178
5.2 Communication between distributed objects 181
5.3 Remote procedure call 197
5.4 Events and notifications 201
5.5 Case study: Java RMI 208
5.6 Summary 216

6 OPERATING SYSTEM SUPPORT 221

6.1 Introduction 222
6.2 The operating system layer 223
6.3 Protection 226
6.4 Processes and threads 228
6.5 Communication and invocation 245
6.6 Operating system architecture 256
6.7 Summary 260

7 SECURITY 265

7.1 Introduction 266
7.2 Overview of security techniques 274
7.3 Cryptographic algorithms 286
7.4 Digital signatures 295
7.5 Cryptography pragmatics 302
7.6 Case studies: Needham–Schroeder, Kerberos, TLS, 802.11 WiFi 305
7.7 Summary 319

8 DISTRIBUTED FILE SYSTEMS 323
8.1 Introduction 324
8.2 File service architecture 332
8.3 Case study: Sun Network File System 337
8.4 Case study: The Andrew File System 349
8.5 Enhancements and further developments 359
8.6 Summary 364

9 NAME SERVICES 367
9.1 Introduction 368
9.2 Name services and the Domain Name System 371
9.3 Directory services 386
9.4 Case study of the Global Name Service 387
9.5 Case study of the X.500 Directory Service 390
9.6 Summary 394

10 PEER-TO-PEER SYSTEMS 397
10.1 Introduction 398
10.2 Napster and its legacy 402
10.3 Peer-to-peer middleware 404
10.4 Routing overlays 406
10.5 Overlay case studies: Pastry, Tapestry 410
10.6 Application case studies: Squirrel, OceanStore, Ivy 419
10.7 Summary 429

11 TIME AND GLOBAL STATES 433
11.1 Introduction 434
11.2 Clocks, events and process states 435
11.3 Synchronizing physical clocks 437
11.4 Logical time and logical clocks 445
11.5 Global states 448
11.6 Distributed debugging 457
11.7 Summary 464

12 COORDINATION AND AGREEMENT 467
12.1 Introduction 468
12.2 Distributed mutual exclusion 471
12.3 Elections 479
12.4 Multicast communication 484
12.5 Consensus and related problems 499
12.6 Summary 510

13 TRANSACTIONS AND CONCURRENCY CONTROL 513
13.1 Introduction 514
13.2 Transactions 517
13.3 Nested transactions 528
13.4 Locks 530
13.5 Optimistic concurrency control 545
13.6 Timestamp ordering 549
13.7 Comparison of methods for concurrency control 556
13.8 Summary 557

14 DISTRIBUTED TRANSACTIONS 565
14.1 Introduction 566
14.2 Flat and nested distributed transactions 566
14.3 Atomic commit protocols 569
14.4 Concurrency control in distributed transactions 578
14.5 Distributed deadlocks 581
14.6 Transaction recovery 589
14.7 Summary 599

15 REPLICATION 603
15.1 Introduction 604
15.2 System model and group communication 606
15.3 Fault-tolerant services 615
15.4 Case studies of highly available services: the gossip architecture, Bayou and Coda 622
15.5 Transactions with replicated data 641
15.6 Summary 653

16 MOBILE AND UBIQUITOUS COMPUTING 657
16.1 Introduction 658
16.2 Association 666
16.3 *Interoperation* 675
16.4 Sensing and context-awareness 683
16.5 Security and privacy 696
16.6 Adaptation 705
16.7 Case study of Cooltown 710
16.8 Summary 717

17 DISTRIBUTED MULTIMEDIA SYSTEMS 721
17.1 Introduction 722
17.2 Characteristics of multimedia data 727
17.3 Quality of service management 728
17.4 Resource management 738
17.5 Stream adaptation 740
17.6 Case study: the Tiger video file server 742
17.7 Summary 746

18 DISTRIBUTED SHARED MEMORY 749
18.1 Introduction 750
18.2 Design and implementation issues 754
18.3 Sequential consistency and Ivy case study 763
18.4 Release consistency and Munin case study 771
18.5 *Other consistency models* 777
18.6 Summary 778

19 WEB SERVICES 783
19.1 *Introduction* 784
19.2 Web services 786
19.3 Service descriptions and IDL for web services 800
19.4 A directory service for use with web services 805
19.5 XML security 807
19.6 Coordination of web services 812
19.7 Case study: the Grid 814
19.8 Summary 824

20 CORBA CASE STUDY 827
20.1 Introduction 828
20.2 CORBA RMI 829
20.3 CORBA services 847
20.4 Summary 855

REFERENCES 859

INDEX 909

# 1

# CHARACTERIZATION OF DISTRIBUTED SYSTEMS

1.1 Introduction
1.2 Examples of distributed systems
1.3 Resource sharing and the Web
1.4 Challenges
1.5 Summary

A distributed system is one in which components located at networked computers communicate and coordinate their actions only by passing messages. This definition leads to the following characteristics of distributed systems: concurrency of components, lack of a global clock and independent failures of components.

We give three examples of distributed systems:

- the Internet;
- an intranet, which is a portion of the Internet managed by an organization;
- mobile and ubiquitous computing.

The sharing of resources is a main motivation for constructing distributed systems. Resources may be managed by servers and accessed by clients or they may be encapsulated as objects and accessed by other client objects. The Web is discussed as an example of resource sharing and its main features are introduced.

The challenges arising from the construction of distributed systems are the heterogeneity of its components, openness, which allows components to be added or replaced, security, scalability – the ability to work well when the number of users increases – failure handling, concurrency of components and transparency.