大学计算机教育国外著名教材、教参系列 (影印版)

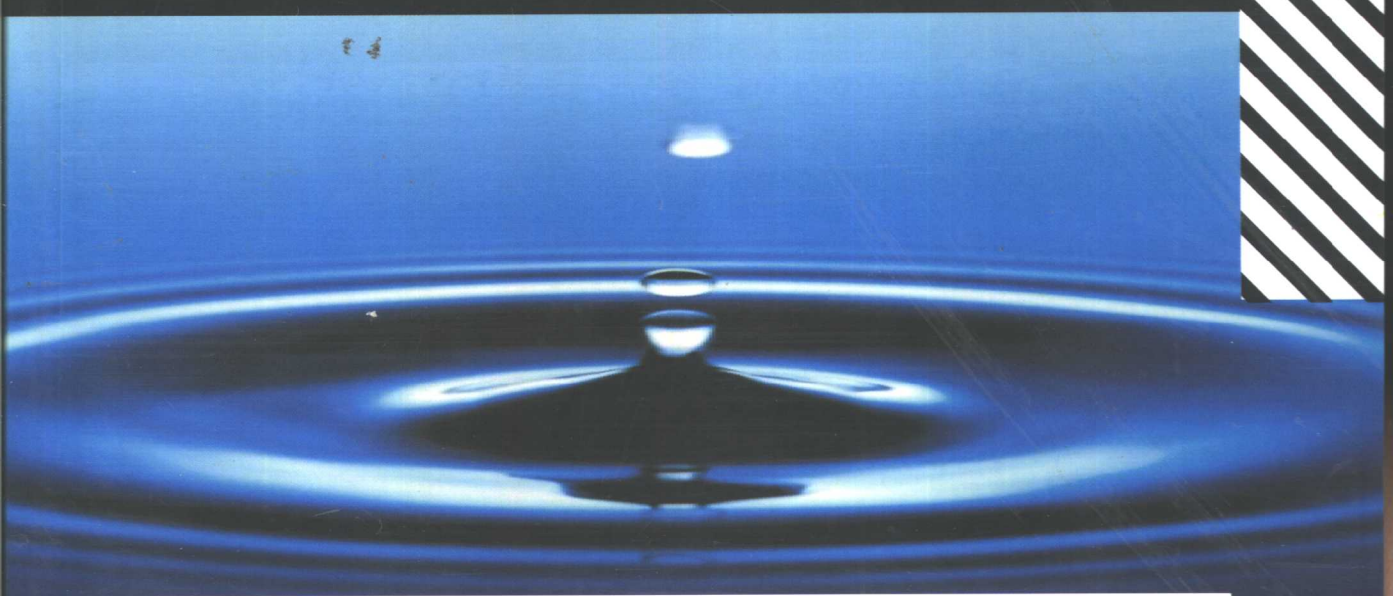
Principles of Distributed Database Systems

Second Edition

M. Tamer Özsu
Patrick Valduriez

分布式数据库系统原理

(第 2 版)



清华大学出版社
http://www.tup.tsinghua.edu.cn

Prentice Hall
PRENTICE-HALL,INC.
http://www.prenhall.com

Principles of Distributed Database Systems

Second Edition

分布式数据库系统原理

第 2 版

M. Tamer Özsu

University of Alberta
Edmonton, Canada

Patrick Valduriez

INRIA
Paris, France

清华大学出版社　　Prentice Hall, Inc.

（京）新登字 158 号

Principles of Distributed Database Systems 2nd ed.
M. Tamer Özsu, Patrick Valduriez

北京市版权局著作权合同登记号：图字：01-2002-0464

书　名：分布式数据库系统原理　第 2 版
作　者：M. Tamer Özsu　Patrick Valduriez
出版者：清华大学出版社（北京清华大学学研大厦，邮编 100084）
　　http:// www.tup.tsinghua.edu.cn
印刷者：清华大学印刷厂
发行者：新华书店总店北京发行所
开　本：787×960 1/16　印张：43.25
版　次：2002 年 6 月第 1 版　2002 年 6 月第 1 次印刷
书　号：ISBN 7-302-05493-2/TP・3230
印　数：0001～4000
定　价：57.00 元

出 版 说 明

进入 21 世纪，世界各国的经济、科技以及综合国力的竞争将更加激烈。竞争的中心无疑是对人才的争夺。谁拥有大量高素质的人才，谁就能在竞争中取得优势。高等教育，作为培养高素质人才的事业，必然受到高度重视。目前我国高等教育的教材更新较慢，为了加快教材的更新频率，教育部正在大力促进我国高校采用国外原版教材。

清华大学出版社从 1996 年开始，与国外著名出版公司合作，影印出版了“大学计算机教育丛书（影印版）”等一系列引进图书，受到了国内读者的欢迎和支持。跨入 21 世纪，我们本着为我国高等教育教材建设服务的初衷，在已有的基础上，进一步扩大选题内容，改变图书开本尺寸，一如既往地请有关专家挑选适用于我国高校本科及研究生计算机教育的国外经典教材或著名教材以及教学参考书，组成本套“大学计算机教育国外著名教材、教参系列（影印版）”，以飨读者。深切期盼读者及时将使用本系列教材、教参的效果和意见反馈给我们。更希望国内专家、教授积极向我们推荐国外计算机教育的优秀教材，以利我们把“大学计算机教育国外著名教材、教参系列（影印版）”做得更好，更适合高校师生的需要。

计算机引进版图书编辑室

2002.3

PREFACE TO THE SECOND EDITION

Many things have changed since the publication of the first edition of this book in 1991. At the time, we reported projections that, by 1998, centralized database managers (DBMSs) would be an "antique curiosity" and most organizations would move towards distributed database managers. Distribution was slowly starting and "client/server" had just started to enter our daily jargon. These systems were generally multiple client/single server systems in which the distribution was mostly in terms of functionality, not data. If multiple servers were used, clients were responsible for managing the connections to these servers. Thus, transparency of access was not widely supported, and each client had to "know" the location of the required data. The distribution of data among multiple servers was very primitive; systems did not support fragmentation or replication of data. Systems of the time were "homogeneous" in that each system could manage only data that were stored in its own database, with no linkage to other repositories.

Things have changed dramatically since then. Many vendors are much closer to achieving true distribution in their development cycle. Client/server systems remain the preferred solution in many cases, but they are much more sophisticated. For example, today's client/server systems provide significant transparency in accessing data from multiple servers, support distributed transactions to facilitate transparency, and execute queries over (horizontally) fragmented data. Further, new systems implement both synchronous and asynchronous replication protocols, and many vendors have introduced gateways to access other databases. In addition, significant achievements have taken place in the development and deployment of parallel database servers. Object database managers have entered the marketplace and have found a niche market in some classes of applications which are inherently distributed.

In parallel with these developments in the database system front, there have been phenomenal changes in the computer networking infrastructure that supports these systems. The relatively slow (10Mbit/sec) Ethernet has been replaced as the de facto local area network standard by much faster networks (FDDI or switched Ethernet) operating at around 100Mbit/sec, and broadband networks (particularly

the ATM technology) have been deployed for both local area and wide area networking. These networks, coupled with very low overhead networking protocols, such as SCI, reduce the differences between local area and wide area networks (other than latency considerations) and potentially eliminate the network as the major performance bottleneck. This, in turn, requires us to review our system development assumptions and performance tuning criteria. Use of the Internet—which is basically a heterogeneous network with links of varying capacities and capabilities—has exploded.

There is clearly a technology push/application pull in effect with respect to distributed DBMS development: new applications are requiring changes in DBMS capabilities, and new technological developments are making these changes possible. With these developments, it was time to prepare a revised second edition of the book. In the process, we have retained the fundamental characteristics and key features of the book as outlined in the Preface to the first edition. However, the material has been heavily edited. Every chapter has been revised—some in fundamental ways, others more superficially. The major changes are the following:

1. The query processing/optimization chapters (Chapters 7–9) have been revised to focus on the techniques employed in commercial systems. New algorithms, such as randomized search strategies, are now included.
2. The transaction management chapters (Chapters 10–12) now include material on advanced transaction models and workflows.
3. Chapter 13, which focused on the relationship of distributed DBMSs and distributed operating systems, has been dropped and some of the material is incorporated into the relevant chapters.
4. The first edition contained a chapter (Chapter 15) which discussed current issues at the time—parallel DBMSs, distributed knowledge-base systems (mainly deductive DBMSs), and distributed object DBMSs. In the intervening years, two of these topics have matured and become major forces in their own rights, while the third (deductive databases) has not achieved the same prominence. In this edition, we devote full chapters to parallel DBMSs (Chapter 13) and distributed object DBMSs (Chapter 14), and have dropped deductive DBMSs.
5. Following the same approach, we introduce a new chapter devoted to current issues (Chapter 16). This chapter now includes sections on data warehousing (from a distributed data management perspective), World Wide Web and databases, push-based technologies, and mobile DBMSs.
6. The chapter on multidatabase systems (Chapter 15 in the current edition) has been revised to include a discussion of general interoperability issues and distributed object platforms such as OMA/CORBA and DCOM/OLE.

We are quite satisfied with the result, which represents a compromise between our desire to address new and emerging issues, and maintain the main characteristic

of the book in addressing the principles of distributed data management. Certain chapters, in particular Chapters 15 and 16, require further depth, but those will be topics of future editions.

The guide to reading the book, introduced in the Preface to the first edition, is still valid in general terms. However, we now discuss, in Chapter 3, the relationship between distributed DBMSs and the new networking technologies. Thus, this chapter no longer serves simply as background and should be read (at least the relevant sections) following Chapter 1.

We have set up a Web site to communicate with our readers. The site is at http://www.cs.ualberta.ca/~database/distdb.html. This site contains presentation slides that accompany the book as well as other information regarding the book's use as a textbook.

Many colleagues have helped with the revisions. Maggie Dunham and Nandid Soparkar provided detailed and early comments on the overall structure and content of the book. Maggie also provided input for the mobile database management section (Section 16.4). Ioannis Nikolaidis helped immensely with the revisions to Chapter 3—he made us rewrite that chapter three times. Jari Veijaleinen provided many exercises which have been incorporated into this edition. Esther Pacitti provided input for replication protocols. Peter Triantafillou provided material on this topic as well. Alexander Thomasian's input for performance evaluation work was invaluable, as was Elliot Moss's critical review of the nested transaction discussion in the transaction processing chapters. Mukesh Singhal advised us of the new advances in distributed deadlock management. Luc Bouganim contributed significantly to Chapter 13 on parallel DBMSs. Ken Barker and Kamalakar Karlapalem provided the material that formed the basis of distributed object database design in Chapter 14. Kaladhar Voruganti wrote the first draft of the architectural and system issues sections of Chapter 14. Randal Peters read Chapter 14 and forced us to revise many parts of it. The distributed garbage collection section of that chapter is based on a draft provided by Laurent Amsaleg and Michael Franklin. Amit Sheth provided input on the revised outline for Chapter 15. Asuman Dogac read the complete chapter and provided feedback. Mokrane Bouzeghoub and Eric Simon helped on the data warehouse section of Chapter 16. Dana Florescu, Alon Levy, Ioana Manolescu and Anthony Tomasic provided input for research prototypes in the section on Web and databases in Chapter 16. The material in push-based technologies section was reviewed (a number of times) by Stan Zdonik and Mike Franklin. Both of them also provided significant feedback about the characterization of data delivery alternatives. We are indebted to all of them, as well as to those who helped with the original edition of the book and whom we cite in the Preface to the First Edition. Many other colleagues have asked questions and provided suggestions over the years; unfortunately, we have not kept their names. Our thanks to everyone who has provided input. We look forward to receiving more suggestions on the second edition.

We have had very good luck with our editors at Prentice Hall. Our current editor, Alan Apt, and our development editor, Sondra Chavez, have been tremendously helpful in both pushing us forward and providing the necessary institutional

support. Our production editors, Ed DeFelippis and Irwin Zucker, have managed the production process so that the production of earlier chapters could proceed in parallel with our writing of the later chapters. This allowed the revised edition to be ready within one year. Stephen Lee, as our copy editor, made the entire text significantly more readable. Anne Nield helped us in many ways—editing chapters, correcting the text and keeping us organized. Paul Iglinski wrote a number of scripts that helped immensely with cleaning up the bibliography. We thank them all.

M. Tamer Özsu (ozsu@cs.ualberta.ca)
Patrick Valduriez (Patrick.Valduriez@inria.fr)

PREFACE TO THE FIRST EDITION

Distributed database system technology is one of the major recent developments in the database systems area. There are claims that in the next ten years centralized database managers will be an "antique curiosity" and most organizations will move toward distributed database managers [Stonebraker, 1988, p. 189]. The intense interest in this subject in both the research community and the commercial marketplace certainly supports this claim. The extensive research activity in the last decade has generated results that now enable the introduction of commercial products into the marketplace. This book aims to introduce and explain the theory, algorithms, and methods that underly distributed database management systems (distributed DBMS). For the most part, our presentation emphasizes the principles that guide the design of such systems more than their use. However, the issues in designing a distributed database are also addressed.

With its emphasis on fundamentals, the book is meant to be used as a textbook for a one- or two-semester graduate-level course as well as a reference book. The material is currently being covered in a one-semester graduate course at the University of Alberta. If it is used in a two-semester course, the material can be complemented by current literature. The structure of the text also lends itself to be used as a companion text for undergraduate database courses. The key features of the text are as follows:

1. The book starts by placing the distributed database technology in its proper context vis-à-vis the distributed computing and database management technologies. The introductory chapters are also aimed at providing the necessary background in computer networks and in relational database systems that is necessary for following the subsequent material.

2. Coverage of each topic starts by an introductory overview that sets the framework and defines the problems that are addressed. The subsequent discussion elaborates these issues. In certain cases the introductory material is included within one chapter (e.g., Sections 5.1 and 5.2), whereas in others they are separated as independent chapters (e.g., Chapters 7 and 10). It is these parts

of the book, in addition to Chapter 1, that can be used to complement the undergraduate courses in database systems.

3. In addition to covering matured technology, the book also discusses current research areas such as distributed data servers, distributed object-oriented databases, and distributed knowledge bases. Thus, it serves not only to describe the technology that has been developed during the past decade, but it also provides an introduction to the technology that the researchers will be working on during the next one. Furthermore, there is coverage of issues related to the integration of distributed database systems and distributed operating systems. These issues have to be topics of intense research and experimentation if distributed database managers are to provide the performance, functionality, and extensibility expected of them.

4. A database design of an engineering organization is used consistently throughout as an example. This consistency enables the development of topics in a systematic fashion. The only section where a different example is used is in the transaction management chapters (10 through 12), where we opt for an airline reservation system, which is a favorite example of the database community as well as being a major application domain for transaction-based systems.

5. The book is structured so that two different uses, as a textbook and as a reference material, can be accommodated. On the one hand, the topics (e.g., distributed database design, semantic data control, distributed query processing and distributed transaction management) are developed systematically with almost no forward referencing. The backward references are few and clearly marked. This enables its use as a textbook where issues can be developed one at a time based on one another. On the other hand, each topic is covered as a self-contained module to the extent that this is possible. Thus, readers who have the background can simply refer to the topics they are interested in. In this mode of use as a reference material, the only important backward references are to previous examples.

6. Exercises are at the end of most chapters. However, chapters that serve as an introduction to topics (Chapters 1 through 3, 7, and 10) or which cover discussion of issues (Chapters 4 and 13) do not contain exercises. Where available, the questions are classified with respect to their difficulty. The number of asterisks (*) in front of a question indicates their level of difficulty.

Organization of the Text

The organization of the book and the dependencies of the chapters are shown in the following figure. The introductory chapter is followed by two chapters that provide an overview of relational database technology and computer networks. If the reader has the background, these chapters can be skipped without any effect on the rest

of the book. The only part of Chapter 2 that should be referenced is Example 2.1, which describes the engineering database example.

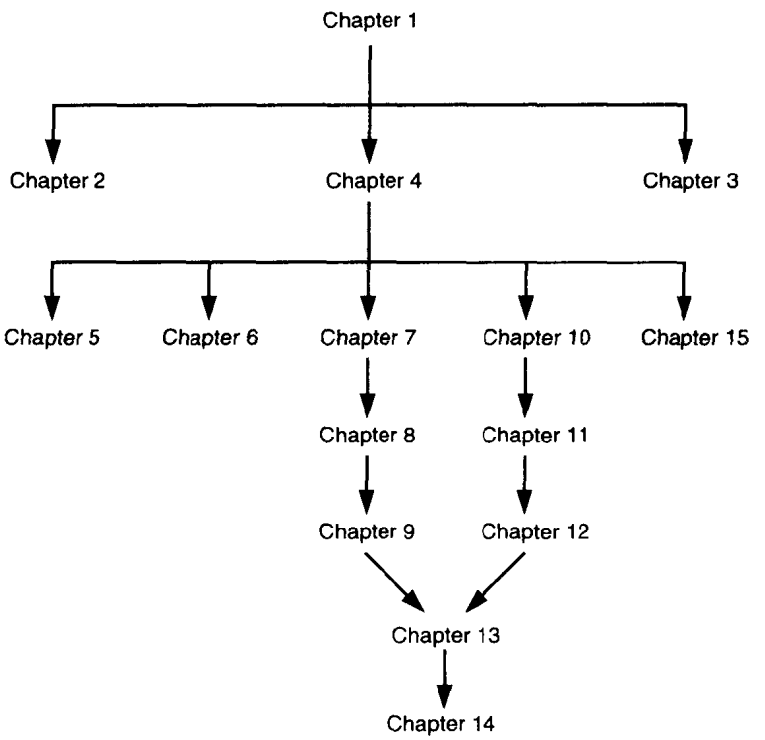


Figure 1. Organization of the Book

Chapter 4 covers the architectural issues, the types of transparencies that distributed DBMSs are supposed to provide and discusses the differences between what we consider distributed database systems and multidatabase systems. This separation is critical to the rest of the book; thus, this chapter should be covered. Most of the book addresses distributed database system issues; multidatabase issues are covered only in Chapter 14.

Chapter 5 describes the design of a distributed database. It is the only chapter of the book where we emphasize the use of a distributed DBMS rather than its development. A similar discussion relating to multidatabase systems is included in Chapter 14.

Chapter 6 covers a unique issue that is usually omitted from database textbooks—namely, semantic data control. Distributed semantic data control includes security aspects of distributed databases as well as integrity enforcement to ensure that the database is always consistent with respect to a set of semantic consistency rules.

Chapters 7 through 9 are devoted to a discussion of distributed query processing issues. The discussion starts with an introduction to the fundamental issues and the presentation of a methodology for carrying out this process. The following two chapters, on query decomposition and localization and distributed query optimization, discuss the steps of this methodology in more detail.

Chapters 10 through 12 are devoted to transaction management issues. The treatment is organized similar to query processing with an introductory chapter that defines the fundamental terms and presents the goals that transaction managers aim to achieve, and the subsequent chapters cover the two fundamental aspects of distributed transaction management: distributed concurrency control and the reliable execution of distributed transactions.

Chapter 13 is built on previous material, especially the distributed query processing and distributed transaction management issues, and the problems associated with implementing distributed DBMSs on top of distributed operating systems are discussed. This chapter also serves as a short introduction to operating systems issues for database researchers.

As we mentioned before, Chapter 14 is dedicated to a discussion of the issues related to multidatabase systems. They differ from what we call distributed database management environments in the high degree of autonomy that is associated with each data manager and their bottom-up design as opposed to the top-down approach utilized by distributed DBMSs. The treatment assumes knowledge of the related issued and solutions for distributed database systems.

Finally, Chapter 15 covers the current trends in distributed databases. Specifically, we address distributed data servers, distributed object-oriented databases, and distributed knowledge bases.

Acknowledgments

Sylvia Osborn read the entire manuscript and provided numerous suggestions. Her contributions to the text are invaluable. Janguk Kim reviewed the manuscript as well. A special thanks goes to both of them. Ahmed Kamal reviewed Chapter 3 and helped with the networking terminology. C. Mohan and Ahmed Elmagarmid provided critical comments on the transaction processing chapters. Ahmed, together with Amit Sheth, reviewed the multidatabase chapter and suggested many improvements. Ravi Krishnamurthy provided help on the query processing chapters, and Guy Lohman improved the precision of many aspects of the R* query optimizer. Eric Simon provided invaluable help on the semantic data control chapter.

The notes that form the basis of this book as well as the book's earlier versions were used in a graduate course on distributed database systems at the University of Alberta. The students who took this course in past years have tremendously helped its presentation. They debugged the text thoroughly and found subtle errors that could have otherwise gone unnoticed. We would like to extend to them our sincere appreciation for helping out as well as for putting up with the troubles of using a continuously changing set of notes as a textbook.

The graduate students in the Distributed Database Systems Group of the Uni-

versity of Alberta all made significant contributions to the text. The Ph.D. students, Ken Barker, Tse-Men Koon, Dave Straube and Randal Peters, all read parts of the manuscript and provided critical comments. The thesis of a former Ph.D. student, Abdel Farrag, provided important material for the transaction processing chapters. The works of M.Sc. students Christina Lau, Yan Li, David Meechan, and Mei-Fen Teo found their way into the book, especially in Chapter 13. Another M.Sc. student, Kok-Lung Wong, reviewed the distributed database design chapter and provided exercises for it. We thank them for all this effort over and above their own research.

The language of the text was edited by Suzanne Sauvé. If readers are not completely happy with some of the language in this edition, they should be grateful that they did not see the text before Suzanne went through it. The remaining errors are probably due to our stubbornness in not accepting some of her suggestions.

Throughout this effort, there was one person who maintained interest in the project perhaps even more than we did: our secretary Amanda Collins at the University of Alberta. She not only ably typed the text once, but then converted the full text to LaTeX. On top of all this, she kept pressing us to finish the writing so that she could start typing. We owe her a great deal for maintaining her enthusiasm and good nature even when things were not moving as smoothly as we all wanted.

M. Tamer Özsu would like to thank Lee White not only for creating an exciting environment within the Department of Computing Science at the University of Alberta during the period of writing this book, but also for the continuing friendship and many opportunities that he has provided over the years. Patrick Valduriez would like to thank Haran Boral and Georges Gardarin for their friendship and support as well as his colleagues of the System Architecture Group at MCC and the SABRE group at INRIA for the exceptional working environment.

We would like to thank our families. This project took valuable time away from them during the last four years. We appreciate their understanding and patience during this long period of time.

Another group who had to wait patiently during these years consists of our editors. We would like to thank Rick Williamson, for suggesting the project to us, and the editors at Prentice-Hall, Valerie Ashton, Marcia Horton and Thomas McElwee. They were all very patient and supportive. We would also like to acknowledge the professional help provided by our production editors, Christina Burghard and Jennifer Wenzel, during the production process.

To my family
and my parents
M.T.Ö.

To Esther, Sarah, Juliette,
and my parents
P.V.

CONTENTS

PREFACE TO THE SECOND EDITION xiii

PREFACE TO THE FIRST EDITION xvii

1 INTRODUCTION 1

1.1 DISTRIBUTED DATA PROCESSING 2

1.2 WHAT IS A DISTRIBUTED DATABASE SYSTEM? 4

1.3 PROMISES OF DDBSs 7

1.3.1 Transparent Management of Distributed and Replicated Data 8

1.3.2 Reliability Through Distributed Transactions 15

1.3.3 Improved Performance 16

1.3.4 Easier System Expansion 18

1.4 COMPLICATING FACTORS 19

1.5 PROBLEM AREAS 20

1.5.1 Distributed Database Design 20

1.5.2 Distributed Query Processing 20

1.5.3 Distributed Directory Management 21

1.5.4 Distributed Concurrency Control 21

1.5.5 Distributed Deadlock Management 21

1.5.6 Reliability of Distributed DBMS 21

1.5.7 Operating System Support 22

1.5.8 Heterogeneous Databases 22

1.5.9 Relationship among Problems 22

1.6 BIBLIOGRAPHIC NOTES 24

2 OVERVIEW OF RELATIONAL DBMS 25

2.1 RELATIONAL DATABASE CONCEPTS 26

2.2 NORMALIZATION 27
2.2.1 Dependency Structures 29
2.2.2 Normal Forms 32
2.3 INTEGRITY RULES 34
2.4 RELATIONAL DATA LANGUAGES 35
2.4.1 Relational Algebra 35
2.4.2 Relational Calculus 43
2.4.3 Interface with Programming Languages 46
2.5 RELATIONAL DBMS 49
2.6 BIBLIOGRAPHIC NOTES 51

3 REVIEW OF COMPUTER NETWORKS 52
3.1 DATA COMMUNICATION CONCEPTS 53
3.2 TYPES OF NETWORKS 55
3.2.1 Topology 56
3.2.2 Communication Schemes 59
3.2.3 Scale 61
3.3 PROTOCOL STANDARDS 63
3.4 BROADBAND NETWORKS 67
3.5 WIRELESS NETWORKS 69
3.6 INTERNET 70
3.7 CONCLUDING REMARKS 71
3.8 BIBLIOGRAPHIC NOTES 74

4 DISTRIBUTED DBMS ARCHITECTURE 75
4.1 DBMS STANDARDIZATION 76
4.2 ARCHITECTURAL MODELS FOR DISTRIBUTED DBMSs 82
4.2.1 Autonomy 82
4.2.2 Distribution 84
4.2.3 Heterogeneity 84
4.2.4 Architectural Alternatives 84
4.3 DISTRIBUTED DBMS ARCHITECTURE 87
4.3.1 Client/Server Systems 88
4.3.2 Peer-to-Peer Distributed Systems 90
4.3.3 MDBS Architecture 94
4.4 GLOBAL DIRECTORY ISSUES 97
4.5 CONCLUSION 100
4.6 BIBLIOGRAPHIC NOTES 100

5 DISTRIBUTED DATABASE DESIGN 102
5.1 ALTERNATIVE DESIGN STRATEGIES 104
5.1.1 Top-Down Design Process 104
5.1.2 Bottom-Up Design Process 106
5.2 DISTRIBUTION DESIGN ISSUES 107
5.2.1 Reasons for Fragmentation 107
5.2.2 Fragmentation Alternatives 108
5.2.3 Degree of Fragmentation 110
5.2.4 Correctness Rules of Fragmentation 110
5.2.5 Allocation Alternatives 111
5.2.6 Information Requirements 111
5.3 FRAGMENTATION 112
5.3.1 Horizontal Fragmentation 112
5.3.2 Vertical Fragmentation 131
5.3.3 Hybrid Fragmentation 146
5.4 ALLOCATION 147
5.4.1 Allocation Problem 147
5.4.2 Information Requirements 150
5.4.3 Allocation Model 151
5.4.4 Solution Methods 154
5.5 CONCLUSION 155
5.6 BIBLIOGRAPHIC NOTES 157
5.7 EXERCISES 158

6 SEMANTIC DATA CONTROL 161
6.1 VIEW MANAGEMENT 162
6.1.1 Views in Centralized DBMSs 162
6.1.2 Updates through Views 164
6.1.3 Views in Distributed DBMSs 165
6.2 DATA SECURITY 167
6.2.1 Centralized Authorization Control 167
6.2.2 Distributed Authorization Control 170
6.3 SEMANTIC INTEGRITY CONTROL 171
6.3.1 Centralized Semantic Integrity Control 173
6.3.2 Distributed Semantic Integrity Control 179
6.4 CONCLUSION 184
6.5 BIBLIOGRAPHIC NOTES 185
6.6 EXERCISES 186