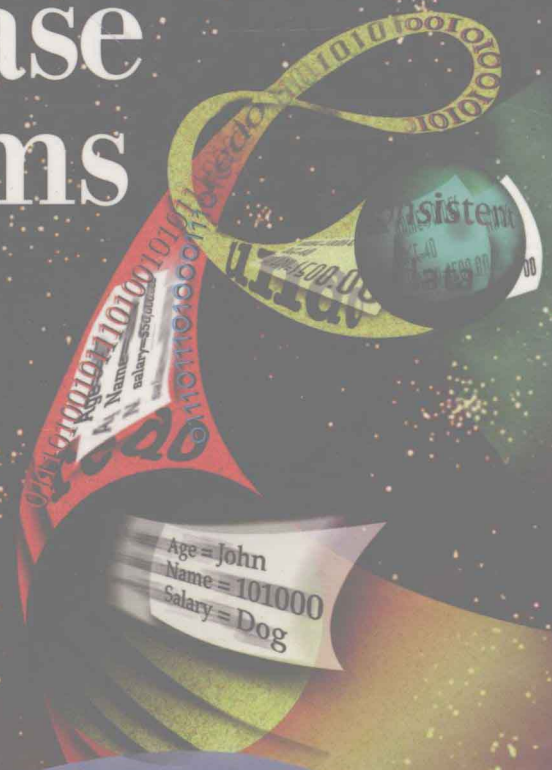
Recovery Mechanisms in Database Systems



Vijay Kumar Meichun Hsu

RECOVERY MECHANISMS IN DATABASE SYSTEMS

Vijay Kumar and Meichun Hsu

Editors

Prentice Hall PTR
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Recovery mechanisms in database systems / Vijay Kumar and Meichun Hsu, editors.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-614215-X
1. Database management. 2. Fault-tolerant computing. I. Kumar, Vijay. II. Hsu, Meichun.
QA76.9.D3R424 1998 97-44452
005.8'6--dc21 CIP

Cover design directors: *Jerry Votta and Amy Rosen*
Cover design: *Amy Rosen*
Cover illustration: *Wendy Grossman*
Manufacturing manager: *Alexis R. Heydt*
Acquisitions editor: *Mark L. Taub*
Editorial assistant: *Tara Ruggiero*
Marketing manager: *Dan Rush*

Published by Prentice Hall PTR
Prentice-Hall, Inc.
A Simon & Schuster Company
Upper Saddle River, New Jersey 07458

Prentice Hall books are widely used by corporations and government agencies for training, marketing, and resale.

The publisher offers discounts on this book when ordered in bulk quantities. For more information, contact Corporate Sales Department, Phone: 800-382-3419; FAX: 201- 236-7141;
E-mail: corpsales@prenhall.com
Or write: Prentice Hall PTR, Corporate Sales Dept., One Lake Street, Upper Saddle River, NJ 07458.

Product names mentioned herein are trademarks or registered trademarks of their respective owners.

Portions of this book have appeared in other publications. Permission acknowledgments are on p. xxx.

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN 0-13-614215-X

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

RECOVERY MECHANISMS IN DATABASE SYSTEMS

ISBN 0-13-614215-X
90000
9 780136 142157

This book is dedicated to its contributors.

Foreword

Robust recovery is an essential feature of any database system. Who would trust their data to a system that didn't reliably recover from failures? Without a recovery subsystem, a database system would be virtually useless.

Database recovery is one of the best success stories of software fault tolerance. It has been successful because it is useful and efficient. It is useful because it factors system reliability concerns from the application programmer. It allows application programmers to use the simple transaction bracketing operations–Start Transaction, Commit, and Abort–with the result that either all of the transaction's work is permanently installed or none of it is, even in the face of application, operating system, and disk (media) failures. These atomicity and durability properties of transactions allow application programmers to focus on errors of application logic and ignore those of the underlying system. Moreover, it is efficient, so customers really want the feature. Thus, virtually all commercial database system products apply database recovery techniques.

Initial work on database recovery was done in the context of commercial product development, primarily at IBM in the 1970s. Since then, researchers have had a major impact on database recovery technology, refining and extending those techniques, and developing new ones, too. Many of these techniques developed by researchers are now standard approaches in commercial products. More of them are finding their way into commercial products every year.

This book brings together many of the most important research papers and articles on database recovery in the last decade in one volume. Although some of them have already appeared in print, they were spread over many research journals and conference proceedings and were therefore accessible only to the few people with time and energy to ferret them out. Some of them are expanded versions of those original papers, now including many important details left out of earlier versions. Many of them are new works: descriptions of commercial implementations, analyses of existing methods, and presentations of new techniques.

The first work on database system recovery was done in the context of a sim-

ple transaction model, implemented on a centralized system with expensive main memory and unreliable disks. Many logging-based algorithms to solve this problem are now well known, the best of which are variations of the ARIES algorithm developed by C. Mohan and colleagues at IBM: a comprehensive description of ARIES is included in this volume.

Most of the new work on database recovery is driven by recent changes in the system environment where database recovery is applied. Hardware improvements change the cost-benefit trade-offs in recovery algorithm design, such as the availability of large main memory, uninterruptible power supplies to backup main memory, and redundant arrays of inexpensive disks (RAID). Distributing processing between clients and servers changes the design space of recovery algorithms, since updates are performed on the client but stored persistently in the server. An extreme example of this is laptop computers, whose stable storage and communications subsystems have special properties that affect the performance of standard recovery algorithms. There are also changes in the time dimension of recovery. Workflow systems to manage long-running multi-step activities are now part of many application products and tools. Workflow mechanisms are finding their way into the underlying system platform, where they are likely to be a standard function in a few years' time.

This is hardly the end of the line in improving the performance and breadth of applicability of database recovery solutions. The new wave of object-relational database systems that will cover a wider variety of data types will undoubtedly place new demands on the flexibility of recovery algorithms, perhaps making it more important for recovery algorithms to handle multilevel atomic actions and a broader range of concurrency control protocols. The changing performance ratios of computing, disk access, and communications will lead to new problems and opportunities. So will the greatly expanding volume of Internet commerce, in which recovery actions may have to span heterogeneous systems and scale up to millions of interconnected server systems. There is no shortage of new challenges. This book includes ideas to seed new work in these exciting areas, and no doubt others yet to be contemplated.

The editors and publisher have done the field a real service in collecting so many of the best works on database recovery here in one volume. The book will be of great interest to the database system expert, since database recovery is one of those "must know" areas that affects many other aspects of database system implementation. It will also help engineers and researchers in related areas–such as operating systems, communications systems, and fault tolerance–to understand how database recovery works its magic and how they might better support and use database recovery techniques in their own work. It is an excellent summary of the state of the art of database recovery and where the field is headed.

Philip A. Bernstein
Microsoft Corporation
Redmond, Washington

Preface

Recovery is a process of restitution. In the spiritual world the recovery process restores and reveals our true self. When we arrive in this world, we *are* in our true self. Our interaction with the materialistic world then makes us believe that the world as we perceive with our senses *is* real and it is the absolute truth, thus falling into the state of ignorance (called MAYA in Indian philosophy). The recovery process salvages us from this state of ignorance and establishes us in the state of bliss, where duality does not exist.

The theme of this book, however, is recovery as practiced in database systems. We describe here the concept and the process of recovery to database systems. To some of us, database recovery has been made more obscure than recovery in the spiritual world. The book, therefore, has tried to provide both the theoretical and the applied aspects of database recovery. It covers recovery in traditional database systems, as well as in emerging technologies such as main memory databases, mobile computing, and workflow systems. It compiles valuable past and present works. Some of the chapters have been exclusively written for the book, and some have been selected from previously published works. One of our main goals is to gather together in one place the different perspectives on the subject currently scattered over time in many places.

The book begins with a historical perspective on database recovery. Ron Obermarck is the narrator, and he has done an excellent job in capturing most of the interesting events in the early evolution of the subject. Ron was one of the members of this "recovery gang" whose motto was "to recover from failure without loss of the customer's work." He takes us back to the fall of 1968, at some corner of an IBM laboratory, when database recovery was "born." He describes the advent of a number of "magics" such as "Write-Ahead-Log Tape," "Undo," and "Redo." It is interesting to note that mother nature did play an active role in the birth of database recovery by stimulating events such as lightning and thunderstorms that led to the development of some of these techniques. Chapter 1, therefore, serves as an appetizer. Ron's history of recovery also complements nicely the history of database concurrency control, as given by Jim Gray in an earlier book edited by one of us.

Performance of Concurrency Control Mechanisms in Centralized Database Systems, published by Prentice Hall, in 1996.

Since then, database recovery has become an important area of research. However, it lags behind concurrency control in the level of conceptual abstraction.

Every chapter of this book exposes some aspect of database recovery. We will only mention a few here as examples. In Chapter 4, Weihl lucidly exposes the intricate relationship between recovery and concurrency control, and shows how some recovery methods place a set of constraints on concurrency control. In Chapter 18, Hsu and Kleissner introduce and describe their perspectives on recovery in workflow systems, a subject still being debated in the research community and evolving in commercial systems. In Chapters 23 and 24, Krishna and colleagues and Bertino and colleagues take readers to the area of mobile recovery. In Chapters 25 through 28, veteran researchers who have worked intensively with commercial database systems describe database recovery in practice. In Chapter 30, Thomasian presents performance issues of the RAID5 disk arrays.

It is our sincere hope that, with the help of the experts who contributed to this volume, we have compiled a book on database recovery that our readers will enjoy reading and consider a valuable source of reference.

Acknowledgments

This book would not have been completed without the generous contributions from the authors. Practically every one of them accepted our invitation to contribute an article to the book with little persuasion. Their cooperation also made our lives much easier during manuscript preparation. Some even offered to help in formatting the manuscript, a very time-consuming task.

We are grateful to so many people for helping us to complete this project that we would not attempt to provide a complete list. We will, however, give a special mention to Ron Obermarck, Dave Lomet, Dave DeWitt, Jim Gray, Elisa Bertino, and Krithi Ramamritham. We especially enjoyed communicating with Ron Obermarck and Dave Lomet, who not only provided us technical guidance, but also generously offered moral support, which we at times greatly needed. Phil Bernstein was very kind in accepting our invitation to write a foreword.

Preparing the camera-ready copy of the book proved to be a very time-consuming process. It would have been worse had we not been rescued by Panos Chrysanthis; Bala Jayabalan and Yutong Wong, Vijay's Master and Ph.D. students; and Professor Gian Paolo Rossi's Ph.D. student Elena Pagani. Panos was generous with his help. Whenever we had a LaTeX problem, we fired an e-mail to or called him and his response was immediate, and he never even threatened to change his e-mail address!

Bala laboriously incorporated copy editor's corrections to all chapters. Yutong helped us to redraw some of the graphs of our chapter, and Elena was very quick and precise in reformatting graphs of her chapter to fit in the book. We are thankful to these three wonderful students.

Our thanks also go to Mark Taub, Executive Editor, and Jane Bonnell at Prentice Hall PTR, and Cindy Kilborn, our representative at Prentice Hall. We communicated most of the time with Jane during the preparation of camera-ready chapters. Without her support and understanding we would not have succeeded in completing the camera-ready manuscript. Mark was very patient with us. Although he reminded us gently from time to time about the deadlines, he never exerted undue pressure, and was always kind enough to give us more time when we needed it. Cindy was instrumental in getting the project approved quickly.

Family members always play an important role in the completion of any project. Our children (Vijay's Krishna and Arjun, and Mei's baby Noelle) always greeted us with their wonderful smiles even when we stole some of their share of time in formatting the book chapters. The support and encouragement of Vijay's wife Elizabeth and Mei's husband Hoomin in completing this book have a special significance.

Finally, we wish to thank the Association for Computing Machinery, Institute of Electrical and Electronics Engineers, the VLDB Endowment, and Elsevier Science Publishing for granting permission to reprint the articles.

Vijay Kumar
University of Missouri
Kansas City, Missouri

Meichun Hsu
Hewlett-Packard Laboratories
Palo Alto, California

PERMISSION ACKNOWLEDGMENTS

Chapter 3. Haerder, T. and A. Reuter, "Principles of Transaction-Oriented Database Reccovery," *Computing Survey*, 15(4) (December 1983).

Chapter 5. Weihl, B. "The Impact of Recovery on Concurrency Control," *Journal of Computer and Systems Sciences*, 47(1) (August 1993).

Chapter 6. Lomet, D. and Mark R. Tuttle, "Redo Recovery after System Crashes," *Proceedings of Very Large Databases*, Zurich, September 1995.

Chapter 7. Lomet. D., "MLR: A Recovery Method for Multi-Level Systems," *ACM SIGMOD*, San Diego, 21(2) (June 1992).

Chapter 8. Mohan, C., Don Haderle, Bruce Lindsay, Hamid Pirahesh, and Peter Schwarz. "ARIES: A Transaction Recovery Method Supporting Fine-Granularity Locking and Partial Rollbacks Using Write-Ahead Logging," *ACM Transactions on Database Systems*, 17(1) (March 1992).

Chapter 9. Kumar, V. and Shown Moe. "Performance of Recovery Algorithms for Centralized Database Management Systems," *Information Sciences*, 86(1–3) (September 1995).

Chapter 10. Goes, P.B., and U. Sumita, "Stochastic Models for Performance Analysis of Database Recovery Control," *IEEE Transactions on Computers*, 44(4) (April 1995).

Chapter 12. Dan, A., Philip S. Yu, and A. Jhingran. "Recovery Analysis of Data Sharing Systems under Deferred Dirty Page Propagation Policies," *IEEE Transactions on Parallel and Distributed Systems*, 8(7) (July 1997).

Chapter 14. Franklin, M., Michael Zwilling, C.K. Tan, Michael Carey, and D. DeWitt, "Crash Recovery in Client-Server EXODUS," *ACM SIGMOD*, San Diego, 21(2) (June 1992).

Chapter 19. Copeland, G., T. Keller, R. Krishnamurthy, and M. Smith. "Case for Safe Ram," *Proc. of Very Large Databases*, Amsterdam, 1989.

Chapter 22. Kumar, V. and A. Bueger. "Performance Measurement of Main Memory Database Recovery Algorithms Based on Update-in-Place and Shadow Approaches," *IEEE Transactions on Knowledge and Data Engineering*, 4(6), (December 1992).

Chapter 23. Pradhan, D. K., P. Krishna, and N. H. Vaidya. "Recovery in Mobile Environment: Design and Trade-off Issues," *IEEE Proc. of Symposium for Fault-tolerant Computing*, June 1996.

Chapter 27. Gray, J., P. McJones, M. Blasgen, B. Lindsey, R. Lorie, T. Price, F. Putzolu, and I. Traiger. "The Recovery Manager of the System R Database Manager," *ACM Computing Surveys*, 13(2) (June 1981).

Chapter 28. White, Seth J., and David J. DeWitt. "Implementing Crash Recovery in QuickStore: A Performance Study," *ACM SIGMOD*, San Jose, 24(2)2 (May 1995).

Contents