

Instant Recovery with Write-Ahead Logging

*Page Repair, System Restart, Media Restore, and
System Failover*

Second Edition

Synthesis Lectures on Data Management

Editor

Z. Meral Özsoyoğlu, *Case Western Reserve University*

Founding Editor

M. Tamer Özsu, *University of Waterloo*

Synthesis Lectures on Data Management is edited by Meral Özsoyoğlu of Case Western Reserve University. The series publishes 80- to 150-page publications on topics pertaining to data management. Topics include query languages, database system architectures, transaction management, data warehousing, XML and databases, data stream systems, wide scale data distribution, multimedia data management, data mining, and related subjects.

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, Media Restore, and System Failover

Second Edition

Goetz Graefe, Wey Guy, and Caetano Sauer

March 2016

Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics

Laure Berti-Équille, Javier Borge-Holthoefer

December 2015

Datalog and Logic Databases

Sergio Greco, Cristian Molinaro

November 2015

Big Data Integration

Xin Luna Dong, Divesh Srivastava

February 2015

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, and Media Restore

Goetz Graefe, Wey Guy, and Caetano Sauer

December 2014

Similarity Joins in Relational Database Systems

Nikolaus Augsten and Michael H. Böhlen

November 2013

Information and Influence Propagation in Social Networks

Wei Chen, Laks V.S. Lakshmanan, and Carlos Castillo

October 2013

Data Cleaning: A Practical Perspective

Venkatesh Ganti and Anish Das Sarma

September 2013

Data Processing on FPGAs

Jens Teubner and Louis Woods

June 2013

Perspectives on Business Intelligence

Raymond T. Ng, Patricia C. Arocena, Denilson Barbosa, Giuseppe Carenini, Luiz Gomes, Jr., Stephan Jou, Rock Anthony Leung, Evangelos Milios, Renée J. Miller, John Mylopoulos, Rachel A. Pottinger, Frank Tompa, and Eric Yu

April 2013

Semantics Empowered Web 3.0: Managing Enterprise, Social, Sensor, and Cloud-based Data and Services for Advanced Applications

Amit Sheth and Krishnaprasad Thirunarayan

December 2012

Data Management in the Cloud: Challenges and Opportunities

Divyakant Agrawal, Sudipto Das, and Amr El Abbadi

December 2012

Query Processing over Uncertain Databases

Lei Chen and Xiang Lian

December 2012

Foundations of Data Quality Management

Wenfei Fan and Floris Geerts

July 2012

Incomplete Data and Data Dependencies in Relational Databases

Sergio Greco, Cristian Molinaro, and Francesca Spezzano

July 2012

Business Processes: A Database Perspective

Daniel Deutch and Tova Milo

July 2012

Data Protection from Insider Threats

Elisa Bertino

June 2012

Deep Web Query Interface Understanding and Integration

Eduard C. Dragut, Weiyi Meng, and Clement T. Yu

June 2012

P2P Techniques for Decentralized Applications

Esther Pacitti, Reza Akbarinia, and Manal El-Dick

April 2012

Query Answer Authentication

HweeHwa Pang and Kian-Lee Tan

February 2012

Declarative Networking

Boon Thau Loo and Wenchao Zhou

January 2012

Full-Text (Substring) Indexes in External Memory

Marina Barsky, Ulrike Stege, and Alex Thomo

December 2011

Spatial Data Management

Nikos Mamoulis

November 2011

Database Repairing and Consistent Query Answering

Leopoldo Bertossi

August 2011

Managing Event Information: Modeling, Retrieval, and Applications

Amarnath Gupta and Ramesh Jain

July 2011

Fundamentals of Physical Design and Query Compilation

David Toman and Grant Weddell

July 2011

Methods for Mining and Summarizing Text Conversations

Giuseppe Carenini, Gabriel Murray, and Raymond Ng

June 2011

Probabilistic Databases

Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch

May 2011

Peer-to-Peer Data Management

Karl Aberer

May 2011

Probabilistic Ranking Techniques in Relational Databases

Ihab F. Ilyas and Mohamed A. Soliman

March 2011

Uncertain Schema Matching

Avigdor Gal

March 2011

Fundamentals of Object Databases: Object-Oriented and Object-Relational Design

Suzanne W. Dietrich and Susan D. Urban

2010

Advanced Metasearch Engine Technology

Weiyi Meng and Clement T. Yu

2010

Web Page Recommendation Models: Theory and Algorithms

Şule Gündüz-Ögüdücü

2010

Multidimensional Databases and Data Warehousing

Christian S. Jensen, Torben Bach Pedersen, and Christian Thomsen

2010

Database Replication

Bettina Kemme, Ricardo Jimenez-Peris, and Marta Patino-Martinez

2010

Relational and XML Data Exchange

Marcelo Arenas, Pablo Barcelo, Leonid Libkin, and Filip Murlak

2010

User-Centered Data Management

Tiziana Catarci, Alan Dix, Stephen Kimani, and Giuseppe Santucci
2010

Data Stream Management

Lukasz Golab and M. Tamer Özsu
2010

Access Control in Data Management Systems

Elena Ferrari
2010

An Introduction to Duplicate Detection

Felix Naumann and Melanie Herschel
2010

Privacy-Preserving Data Publishing: An Overview

Raymond Chi-Wing Wong and Ada Wai-Chee Fu
2010

Keyword Search in Databases

Jeffrey Xu Yu, Lu Qin, and Lijun Chang
2009

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2016

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Instant Recovery with Write-Ahead Logging: Page Repair, System Restart, Media Restore, and System Failover: Second Edition
Goetz Graefe, Wey Guy, Caetano Sauer

ISBN: 978-3-031-00729-3 print
ISBN: 978-3-031-01857-2 ebook

DOI 10.1007/978-3-031-01857-2

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MANAGEMENT #44
Series Editor: Z. Meral Özsoyoğlu, Case Western Reserve University
Founding Editor: M. Tamer Özsu, University of Waterloo

Series ISSN 2153-5418 Print 2153-5426 Electronic

Instant Recovery with Write-Ahead Logging

*Page Repair, System Restart, Media Restore, and
System Failover*

Second Edition

Goetz Graefe

Hewlett Packard Labs

Wey Guy

Caetano Sauer

University of Kaiserslautern

SYNTHESIS LECTURES ON DATA MANAGEMENT #44

ABSTRACT

Traditional theory and practice of write-ahead logging and of database recovery focus on three failure classes: transaction failures (typically due to deadlocks) resolved by transaction rollback; system failures (typically power or software faults) resolved by restart with log analysis, “redo,” and “undo” phases; and media failures (typically hardware faults) resolved by restore operations that combine multiple types of backups and log replay.

The recent addition of single-page failures and single-page recovery has opened new opportunities far beyond the original aim of immediate, lossless repair of single-page wear-out in novel or traditional storage hardware. In the contexts of system and media failures, efficient single-page recovery enables on-demand incremental “redo” and “undo” as part of system restart or media restore operations. This can give the illusion of practically instantaneous restart and restore: instant restart permits processing new queries and updates seconds after system reboot and instant restore permits resuming queries and updates on empty replacement media as if those were already fully recovered. In the context of node and network failures, instant restart and instant restore combine to enable practically instant failover from a failing database node to one holding merely an out-of-date backup and a log archive, yet without loss of data, updates, or transactional integrity.

In addition to these instant recovery techniques, the discussion introduces self-repairing indexes and much faster offline restore operations, which impose no slowdown in backup operations and hardly any slowdown in log archiving operations. The new restore techniques also render differential and incremental backups obsolete, complete backup commands on a database server practically instantly, and even permit taking full up-to-date backups without imposing any load on the database server.

Compared to the first version of this book, this second edition adds sections on applications of single-page repair, instant restart, single-pass restore, and instant restore. Moreover, it adds sections on instant failover among nodes in a cluster, applications of instant failover, recovery for file systems and data files, and the performance of instant restart and instant restore.

KEYWORDS

algorithms, databases, transactions, failures, recovery, availability, reliability, write-ahead logging, instant restart, log analysis, redo, undo, rollback, compensation, log replay, instant restore, single-pass restore, virtual backup, big data, file systems, key-value stores, clusters, log shipping, failover, elasticity, failover pool

Contents

	Preface	xv
	Acknowledgments	xvii
1	Introduction	1
2	Related Prior Work	5
	2.1 System Model	5
	2.2 ARIES	7
	2.3 Restart After a System Failure	8
	2.4 Database Backup and Log Archive	11
	2.5 Restore After a Media Failure	13
	2.6 Mirroring, Log Shipping, and Failover	14
	2.7 Allocation-Only Logging	16
	2.8 System Transactions	18
	2.9 Summary of Prior Work	20
3	Single-Page Recovery	21
	3.1 Detection of Single-Page Failures	21
	3.2 Recovery for Logged Updates	22
	3.3 Recovery for Non-Logged Updates	22
	3.4 Chains of Log Records	23
	3.5 Summary of Single-Page Recovery	26
4	Applications of Single-Page Recovery	27
	4.1 Self-Repairing Indexes	27
	4.2 Write Elision	29
	4.3 Read Elision	31
	4.4 Deferred “Undo”	32
	4.5 Summary of Single-Page Recovery Applications	33
5	Instant Restart after a System Failure	35
	5.1 Restart Techniques	37
	5.2 Restart Schedules	41
	5.3 Optimizing Log Scans	42

5.4	Summary of Instant Restart	44
6	Applications of Instant Restart	47
6.1	Parallel “Redo” and “Undo”	47
6.2	Distributed Transactions	47
6.3	Fast Reboot	48
6.4	Summary of Instant Restart Applications	48
7	Single-Pass Restore	49
7.1	Partially Sorted Log Archive	50
7.2	Archiving Logic	51
7.3	Restore Logic	53
7.4	Active Transactions	55
7.5	Summary of Single-Pass Restore	56
8	Applications of Single-Pass Restore	59
8.1	Pipeline Extensions	59
8.2	Instant Backup	60
7.3	Virtual Backups	62
8.4	Obsolete Incremental Backups	64
8.5	Summary of Single-Pass Restore Applications	65
9	Instant Restore after a Media Failure	67
9.1	Indexed Backup and Log Archive	67
9.2	Restore Techniques	68
9.3	Restore Schedules	69
9.4	Summary of Instant Restore	70
10	Applications of Instant Restore	73
10.1	Pipeline Extensions	73
10.2	Hot Restore	73
10.3	Restore Without Replacement Media	74
10.4	Online Database Migration	74
10.5	Summary of Instant Restore Applications	75
11	Multiple Page, System, and Media Failures	77
11.1	Single-Page Failure During Restore	77
11.2	Single-Page Failure During Restart	77
11.3	Multiple System Failures	78
11.4	Multiple Media Failures	79

11.5	System Failure During Media Restore	80
11.6	Media Failure During System Restart	81
11.7	Summary of Recovery After Multiple Failures.	81
12	Instant Failover	83
12.1	Log Shipping and Log Archiving	85
12.2	Recovery of Server State in a Failover	85
12.3	Recovery of Database Contents in a Failover.	86
12.4	Summary of Instant Failover	87
13	Applications of Instant Failover	91
13.1	Instant Failback	91
13.2	Failover Pools	92
13.3	Elasticity	94
13.4	Summary of Instant Failover Applications.	94
14	File Systems and Data Files	95
14.1	Fault Detection.	96
14.2	Fault Repair	96
14.3	Logging Small Updates	97
14.4	Logging Large Updates	97
14.5	Summary of Instant Recovery in File Systems.	98
15	Performance and Scalability	101
15.1	System Failure and Restart.	101
15.2	Media Failure and Restore	102
15.3	Summary of Performance and Scalability.	103
16	Conclusions	105
	References	109
	Author Biographies	113

Preface

It has been a pleasure developing and compiling this set of concepts and techniques in order to make them available to researchers and software developers around the world. While the foundation of the presented techniques is write-ahead logging as commonly found in database management systems, the techniques and their advantages apply similarly to key-value stores, file systems with journaling, etc.—in other words, to all storage management layers for important and big data. In all these systems, write-ahead logging can enable efficient single-page repair after a localized data loss, system restart after a software crash, and media restore after a failure in the storage hardware or firmware. Instead of copying each data page to multiple devices, as many file storage systems do today in order to achieve high availability, only a single copy is required, plus a log of changes.

Acknowledgments

Barb Peters and Arianna Lund encouraged combining all “instant recovery” techniques into a single book. Harumi Kuno participated in the research on single-page recovery and its applications.