Lecture Notes in Computer Science 5959

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Alfred Kobsa
University of California, Irvine, CA, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
TU Dortmund University, Germany

Madhu Sudan
Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Bernadette Charron-Bost
Fernando Pedone
André Schiper (Eds.)

Replication

Theory and Practice

Springer

Volume Editors

Bernadette Charron-Bost
École Polytechnique, CNRS
Laboratoire d'Informatique (LIX)
91128 Palaiseau CEDEX, France
E-mail: charron@lix.polytechnique.fr

Fernando Pedone
Università della Svizzera italiana (USI)
Facoltà di Scienze informatiche
Via Giuseppe Buffi 6, 6900 Lugano, Switzerland
E-mail: fernando.pedone@usi.ch

André Schiper
École Polytechnique Fédérale de Lausanne (EPFL)
School of Computer and Communication Sciences
Station 14, 1015 Lausanne, Switzerland
E-mail: andre.schiper@epfl.ch

Cover illustration: M.C. Escher's "Symmetry Drawing E126"
© 2010 The M.C. Escher Company–Holland. All rights reserved.
www.mcescher.com

Library of Congress Control Number: 2009941316

CR Subject Classification (1998): H.2-4, H.2.4, C.4, C.2.4, E.3, E.1

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

ISSN 0302-9743
ISBN-10 3-642-11293-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-11293-5 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Markus Richter, Heidelberg
Printed on acid-free paper SPIN: 12823659 06/3180 5 4 3 2 1 0

Preface

This book is the result of the seminar "*A 30-Year Perspective on Replication*," which took place at Monte Verità, Ascona, Switzerland, in November 2007. As suggested by the title, the goal of the seminar was not to speculate about the future of replication, but rather to understand the present, by analyzing past successes and past failures, and to make an assessment of about 30 years of research on replication. Replication is a topic addressed by several communities: the distributed computing community, the distributed system community, and the database community. Each of these communities has looked at replication from different points of view and with different goals, e.g., performance vs. fault tolerance. Recently, these different goals have started to converge, and there has been work showing that efficiency and strong consistency can sometimes be reconciled.

During the seminar the observation was made that we had reached a point of understanding of the different issues of replication, and this knowledge should be materialized in a book covering the different aspects of replication. This book results from this observation. Its goal is to present a comprehensive view of the achievements of 30 years of research on replication. The book was written by most of the people who have contributed to developing the state-of-the-art replication techniques. It brings a comprehensive view of existing solutions, from a theoretical as well as from a practical point of view. It covers replication of processes/objects and of databases; replication for fault tolerance and replication for performance; benign faults and malicious (Byzantine) faults. By covering these different issues in an integrated way, we believe the book fills a gap, and as such it should find a place in the graduate teaching of distributed computing, distributed systems, and databases.

The book is organized in thirteen chapters. Chapter 1 introduces consistency models for replicated data, both in the context of process/object and database replication. Chapter 2 discusses replication techniques commonly used in process replication, focusing on primary back-up and related techniques; it considers both the fail-stop and the crash failure models. Chapter 3 considers modular approaches to process replication; it starts with state-machine replication based on atomic broadcast and shows how this can be built on top of consensus. Although the literature on consensus is vast, there are many misunderstandings, often involving dif-

ferent communities. Chapter 4 discusses these misunderstandings. Chapter 5 covers replication for performance; it contains different strategies and examples, and discusses trade-offs. Chapters 6 and 7 provide a historical account of the Virtual Synchrony Replication Model and Viewstamped Replication, two early replication systems and how they have evolved over the years. Chapters 8 and 9 are dedicated to state-machine replication with Byzantine faults; the first considers distributed trust systems, and the second introduces protocols for state-machine replication. Chapter 10 surveys Byzantine quorum systems, suitable for use when parts of the system cannot be trusted. Chapters 11 through 13 consider database replication. Chapter 11 bridges the gap between process/object replication and database replication, while Chap. 12 surveys database replication techniques; it discusses different replication approaches, consistency criteria for replicated databases and existing systems. Chapter 13 illustrates database replication with a case study: the details of an architecture for practical database replication.

Each one of the chapters in the book is self-contained, and can be read individually. Readers interested in certain specific aspects of replication, however, may prefer to focus on some of the chapters. Chapters 1 and 11 through 13 provide a detailed description of replication in the context of databases. Theoretical aspects of replication under benign failures are discussed in Chapters 1, 3 and 4. Chapters 5, 12 and 13 cover many issues involving practical replication issues. Chapters 8 through 10 address replication under malign failures (i.e., Byzantine failures). Readers mostly interested in historical aspects of replication should read Chaps. 6 and 7.

The Monte Verità seminar organizers are thankful to all the participants for accepting to take part in this unique seminar, and to all authors for taking their time to produce this book. We would also like to thank a number of institutions for the financial support to the seminar: the Monte Verità Foundation, the Hasler Foundation, Microsoft, Eidgenössische Technische Hochschule Zürich (ETHZ), École Polytechnique Fédérale de Lausanne (EPFL), Università della Svizzera italiana (USI), and the École polytechnique in Palaiseau.

October 2009

Bernadette Charron-Bost
Fernando Pedone
André Schiper

List of Authors

Marcos K. Aguilera	Microsoft Research Silicon Valley, USA
Gustavo Alonso	ETHZ, Switzerland
Ken Birman	Cornell University, USA
Christian Cachin	IBM Research - Zurich, Switzerland
Nuno Carvalho	Instituto Superior Técnico/INESC-ID, Portugal
Alfrânio Correia Jr.	Universidade do Minho, Portugal
Alan D. Fekete	University of Sydney, Australia
Rachid Guerraoui	EPFL, Switzerland
Ricardo Jiménez-Peris	Universidad Politécnica de Madrid, Spain
Bettina Kemme	McGill University, Canada
Barbara Liskov	MIT, USA
Michael G. Merideth	Carnegie Mellon University, USA
Rui Oliveira	Universidade do Minho, Portugal
Marta Patiño-Martínez	Universidad Politécnica de Madrid, Spain
Fernando Pedone	University of Lugano, Switzerland
José Pereira	Universidade do Minho, Portugal
Guillaume Pierre	VU University Amsterdam, The Netherlands
Krithi Ramamritham	I.I.T. Bombay, India
Michael K. Reiter	University of North Carolina, USA
Luís Rodrigues	Instituto Superior Técnico/INESC-ID, Portugal
André Schiper	EPFL, Switzerland
Fred B. Schneider	Cornell University, USA
Robbert van Renesse	Cornell University, USA
Maarten van Steen	VU University Amsterdam, The Netherlands
Lidong Zhou	Microsoft Research Asia, China

Contents

1 Consistency Models for Replicated Data 1
Alan D. Fekete and Krithi Ramamritham
1.1 Introduction 1
1.1.1 Contributions 2
1.2 Defining the Sequential Data Type 2
1.3 Strong Consistency 4
1.3.1 Relaxing Inter-Client Operation Ordering 7
1.4 Weak Consistency 10
1.5 Transactions 14
1.6 Discussion 15
1.7 Conclusion 16
References 17

2 Replication Techniques for Availability 19
Robbert van Renesse and Rachid Guerraoui
2.1 Introduction 19
2.2 Model 20
2.2.1 Environment 21
2.2.2 Specification 21
2.3 Fail-Stop Failure Model 22
2.3.1 Primary-Backup 23
2.3.2 Chain Replication 26
2.3.3 Queries 28
2.4 Crash Failure Model 30
2.4.1 Quorums 31
2.4.2 Stake Replication 31
2.4.3 Broker Replication 36
2.5 Recovery and Reconfiguration 38
2.6 Conclusion 39
References 39

3 Modular Approach to Replication for Availability 41
Fernando Pedone and André Schiper
3.1 Introduction 41
3.2 Atomic Broadcast for State Machine Replication 42
3.3 The Consensus Problem, or How to Implement Atomic Broadcast in a Modular Way 43
3.3.1 Consensus 43
3.3.2 Implementation of Atomic Broadcast 44
3.4 Solving Consensus 47
3.4.1 About System Models 47
3.4.2 Partially Synchronous Systems 47
3.4.3 Asynchronous System Augmented with Failure Detectors 50
3.4.4 Discussion 52
3.5 Generic Broadcast 53
3.6 Dynamic Groups 54
3.6.1 Group Membership Service 54
3.6.2 Group Communication in Dynamic Groups 55
3.7 Conclusion 56
References 57

4 Stumbling over Consensus Research: Misunderstandings and Issues . 59
Marcos K. Aguilera
4.1 Introduction 59
4.2 Misunderstandings 60
4.2.1 Asynchronous Systems 60
4.2.2 Eventually-Forever Assumptions 61
4.2.3 Eventual Guarantees 62
4.2.4 The Consensus Impossibility Result 63
4.2.5 Uses of Replication 65
4.2.6 Correlated Failures 66
4.3 Issues 67
4.3.1 The Application Interface 67
4.3.2 Violation of Abstraction Boundaries 69
4.3.3 Ambiguities and Errors 70
4.3.4 Unfriendly Formalisms 70
4.3.5 Lack of Feedback from Practitioners 71
4.3.6 Hidden Limitations in Algorithms 71
4.4 Conclusion 71
References 72

5 Replicating for Performance: Case Studies 73
Maarten van Steen and Guillaume Pierre
5.1 Introduction 73
5.2 Replication Strategies 75
5.2.1 Replica Placement 76
5.2.2 Content Distribution 77
5.2.3 Strategy Evaluation 78

5.3 Replication Granularity 78
5.3.1 Example 1: Content Delivery Networks 79
5.3.2 Example 2: Edge-Server Computing 81
5.3.3 Example 3: Decentralized Wikipedia 83
5.4 Replicating for Performance versus Consistency 84
5.5 Replication Management 86
5.6 Conclusions 87
References 87

6 A History of the Virtual Synchrony Replication Model 91
Ken Birman
6.1 Introduction 91
6.2 Distributed Consistency: Who Needs It? 94
6.3 Goals in This Chapter 95
6.4 Historical Context 95
6.4.1 Resilient Objects in Isis V1.0 98
6.4.2 Beyond Resilient Objects 101
6.4.3 The Isis Toolkit and the Virtual Synchrony Model 102
6.4.4 A Design Feature Motivated by Performance Considerations . 104
6.5 Dynamic Membership 106
6.5.1 Local Reads and Fast Updates 106
6.5.2 Partitionable Views 108
6.6 Causally Ordered Multicast: *cbcast* 108
6.7 Time-Critical Applications 110
6.8 A Series of Commercial Successes, but Ultimately, a Market Failure . 111
6.8.1 How Replication Was Used 112
6.8.2 Causal and Other Controversies 114
6.8.3 What Next? Live Objects and Quicksilver Scalable Multicast! 115
6.9 Closing Thoughts 116
References 117

7 From Viewstamped Replication to Byzantine Fault Tolerance 121
Barbara Liskov
7.1 Introduction 121
7.2 Prehistory 122
7.3 Viewstamped Replication 123
7.3.1 Replica Groups 124
7.3.2 Architecture 125
7.3.3 Approach 125
7.4 The VR Protocol 126
7.4.1 Normal Operation 127
7.4.2 View Changes 128
7.4.3 Recovery 131
7.5 Discussion of VR 134
7.5.1 Differences from the Original 134
7.5.2 Two-Phase Commit 134

7.5.3 Optimizations 135
7.5.4 Performance in the Normal Case 136
7.5.5 Performance of View Changes 137
7.5.6 State Management 137
7.5.7 Non-deterministic Operations 138
7.6 Byzantine Fault Tolerance 138
7.6.1 Approach 139
7.7 The PBFT Protocol 140
7.7.1 View Changes 142
7.8 Discussion of PBFT 144
7.8.1 Cryptography 144
7.8.2 Optimizations 144
7.8.3 Selecting the Primary 146
7.8.4 Recovery 146
7.8.5 Non-determinism 146
7.9 Conclusions 146
References 147

8 Implementing Trustworthy Services Using Replicated State Machines 151
Fred B. Schneider and Lidong Zhou
8.1 Introduction 151
8.2 The State-Machine Approach 152
8.3 Compromise and Proactive Recovery 153
8.4 Service Key Refresh and Scalability 154
8.4.1 Service Private Keys 155
8.4.2 Proactive Secret Sharing 156
8.5 Server Key Refresh 156
8.5.1 Trusted Hardware 156
8.5.2 Offline Keys 157
8.5.3 Attack Awareness 157
8.6 Processor Independence 158
8.7 Replica Coordination 160
8.8 Computing with Server Confidential Data 162
8.9 Discussion 163
References 165

9 State Machine Replication with Byzantine Faults 169
Christian Cachin
9.1 Introduction 169
9.2 Building Blocks 170
9.2.1 Broadcast Primitives 170
9.2.2 Distributed Cryptography 173
9.2.3 Byzantine Consensus 175
9.3 Atomic Broadcast Protocols 178
9.3.1 Consensus-Based Atomic Broadcast 179

9.3.2 Sequencer-Based Atomic Broadcast 179
9.3.3 Hybrid Atomic Broadcast 180
9.4 Service Replication 181
9.4.1 Replicating Cryptographic Services 181
9.4.2 Handling Responses Securely 181
9.4.3 Preserving Causality of Requests 182
9.5 Conclusion 182
References 183

10 Selected Results from the Latest Decade of Quorum Systems Research 185
Michael G. Merideth and Michael K. Reiter
10.1 Introduction 185
10.2 Quorum Systems for Byzantine Faults 186
10.2.1 Access Strategies and Load 187
10.2.2 Probabilistic Quorum Systems 188
10.3 Minimizing Delays of Quorum Accesses 189
10.4 Uses of Byzantine Quorums in Protocols 191
10.4.1 Read-Overwrite Protocols 192
10.4.2 State-Machine-Replication Protocols 195
10.5 Conclusion 205
References 205

11 From Object Replication to Database Replication 207
Fernando Pedone and André Schiper
11.1 Introduction 207
11.2 Replication Model and Consistency 208
11.2.1 Generic Functional Model 209
11.2.2 Object and Database Consistency 210
11.3 From Object Replication to Database Replication: Multi-primary Passive Replication 210
11.4 Deferred Update Database Replication 211
11.4.1 Additional Definitions 212
11.4.2 Atomic Commit-Based Termination 212
11.4.3 Atomic Broadcast-Based Termination 214
11.4.4 Reordering-Based Termination 215
11.4.5 Generic Broadcast-Based Termination 216
11.5 Final Remarks 217
References 218

12 Database Replication: A Tutorial 219
Bettina Kemme, Ricardo Jiménez-Peris, Marta Patiño-Martínez, and Gustavo Alonso
12.1 Introduction 219
12.1.1 Why Replication 219
12.1.2 Organization of the Chapter 220

12.2 Basic Taxonomy for Replica Control Approaches 221
12.2.1 Eager Primary Copy 223
12.2.2 Eager Update Anywhere 225
12.2.3 Lazy Primary Copy 228
12.2.4 Lazy Update Anywhere 230
12.2.5 Eager vs. Lazy 232
12.3 Correctness Criteria 233
12.3.1 Atomicity and Consistency 233
12.3.2 Isolation 235
12.3.3 Session Consistency 237
12.4 Other Parameters 237
12.4.1 Message Management 238
12.4.2 Executing Writes 239
12.4.3 Concurrency Control Mechanisms 240
12.4.4 Architectural Alternatives 243
12.4.5 Cluster vs. WAN Replication 245
12.4.6 Degree of Replication 246
12.4.7 Recovery 247
12.5 Existing Systems 247
12.5.1 Early Work 247
12.5.2 Commercial Systems 247
12.5.3 Lazy Replication Made Serializable 248
12.5.4 Cluster Replication 248
12.5.5 Other Issues 249
12.5.6 Related Areas of Research 249
12.6 Conclusions 250
References 250

13 Practical Database Replication 253
Alfrânio Correia Jr., José Pereira, Luís Rodrigues, Nuno Carvalho, and Rui Oliveira
13.1 Introduction 253
13.2 An Architecture for Practical Database Replication 254
13.3 Reflector: Replication-Friendly Database Support 257
13.3.1 Reflection for Replication 257
13.3.2 Processing Stages 258
13.3.3 Processing Contexts 259
13.3.4 Base-Level and Meta-level Calls 260
13.3.5 Exception Handling 261
13.3.6 Existing Reflector Bindings 261
13.4 GCS: Communication and Coordination Support 263
13.4.1 Architectural and Algorithmic Issues 264
13.4.2 Existing GCS Bindings 266
13.5 Replicator: Pluggable Replication Protocols 267

13.6 Consistent Database Replication 268
13.6.1 Replication with Conservative Execution 270
13.6.2 Replication with Optimistic Execution 272
13.6.3 Active Replication 275
13.6.4 Hybrid Replication 276
13.6.5 Evaluation 277
13.7 Conclusions 282
References 283

Index 287