Table of Contents

Chapter 1: Motivation, Problem Statement, and Thesis **1**

Chapter 2: Fault-Tolerant Disk Storage **7**

2.1 Terminology 8
2.1.1 Faults, Errors, and Failure 8
2.1.2 Fault Models and Semantics 9
2.2 Metrics 10
2.3 Improving System Dependability 11
2.3.1 Detection, Diagnosis, and Isolation 11
2.3.2 Recovery, Reconfiguration, and Repair 13
2.3.3 A Closer Look at Error Recovery 14
2.3.4 Transactions 15
2.3.5 Discussion 18
2.4 Disk Drives 18
2.4.1 Disk Technology 19
2.4.2 Fault Model 21
2.4.2.1 Fault Model Used in This Work 22
2.4.3 Discussion 23
2.5 Disk Arrays 24
2.5.1 Striping for Performance 25
2.5.2 Redundant Disk Arrays 26
2.5.2.1 Encoding 26
2.5.2.2 Algorithms for Accessing Information 27
2.5.2.3 The Berkeley RAID Taxonomy 32
2.5.3 Fault Model 33
2.5.4 Beyond the RAID Taxonomy 37
2.5.4.1 Improving Dependability 38
2.5.4.2 Improving Performance 39
2.5.5 Discussion 42
2.6 Conclusions 43

Chapter 3: Mechanizing the Execution of Array Operations **45**

3.1 Goals of an Ideal Approach 46

3.2 Isolating Action-Specific Recovery..48
3.2.1 Creating Pass/Fail Actions..48
3.2.2 Actions Commonly Used in Redundant Disk Array Algorithms 50
3.2.2.1 Symbol Access..51
3.2.2.2 Resource Manipulation..52
3.2.2.3 Computation..52
3.2.2.4 Predicates..53
3.3 Representing Array Operations as Flow Graphs 53
3.3.1 Flow Graphs..53
3.3.2 Predicate Nodes 56
3.3.3 Simplifying Constraints..56
3.3.4 Graph Optimization 57
3.3.5 Automating Correctness Verification 59
3.3.6 Discussion..59
3.4 Execution Based Upon Forward Error Recovery is Unreasonable........................60
3.4.1 Correct Design is Not Obvious..61
3.4.2 Exhaustive Testing is Required 61
3.4.3 Recovery Code is Architecture-Specific..62
3.5 Simplifying Execution Through Mechanization..63
3.5.1 Undoing Completed Actions..63
3.5.1.1 Symbol Access..64
3.5.1.2 Computation..65
3.5.1.3 Resource Manipulation..65
3.5.1.4 Predicates..66
3.5.2 Node States and Transitions..67
3.5.3 Sequencing a Graph..68
3.5.3.1 Sequencing Graphs with Predicate Nodes..69
3.5.4 Automating Error Recovery..70
3.5.4.1 Coping With Deadlock 71
3.5.5 Distributing Graph Execution..73
3.6 Fault Model..74
3.6.1 Node Failures..75
3.6.2 Crash Recovery and Restart..76
3.6.3 Controller Failover..77
3.7 Summary..77

Chapter 4: RAIDframe: Putting Theory Into Practice..............................79

4.1 Motivation..79
4.2 Architecture..81
4.2.1 Design Decisions 81
4.2.2 Libraries..84
4.2.3 Processing a User Request..84
4.2.3.1 Locking 86
4.2.3.2 Error Recovery..86
4.3 Evaluation 87

4.3.1 Setup 88
4.3.1.1 Workload Generation 89
4.3.2 Extensibility 89
4.3.3 Efficiency 92
4.3.4 Verification 95
4.4 Conclusions 99

Chapter 5: Roll-Away Error Recovery 101

5.1 Full Undo Logging is Expensive 101
5.2 Reducing Undo Logging Requirements 103
5.2.1 Limiting the Scope of Rollback 103
5.2.2 Reclassifying Actions From Undoable to Real 104
5.3 Roll-Away Error Recovery 104
5.3.1 Properties of Phase-I Subgraphs 105
5.3.2 Properties of Phase-II Subgraphs 106
5.3.3 Commit Node Determines Direction of Recovery 106
5.3.3.1 Inserting a Commit Node Into a Read Graph 107
5.3.3.2 Inserting a Commit Node Into a Write Graph 107
5.3.4 Adjusting Graph Execution Rules 109
5.3.5 Fault Model 114
5.3.5.1 Adjusting Node Properties 116
5.4 Performance Evaluation 116
5.5 Correctness Testing 117
5.6 Summary 119

Chapter 6: Conclusions and Recommendations 123

6.1 Validating the Problem 123
6.2 Eliminating Architecture-Specific Error Recovery Code 124
6.2.1 Reducing Logging Penalties 125
6.2.2 Enabling Correctness Validation 125
6.3 Practicality 126
6.4 Suggestions for Future Work 126

References 131

Appendix A: Flow Graphs for Popular Array Architectures 141

A.1 RAID Level 0 142
A.2 RAID Level 1, Interleaved Declustering, and Chained Declustering 143
A.3 RAID Level 3 144
A.4 RAID Levels 4 and 5 and Parity Declustering 147
A.5 Parity Logging 150
A.6 RAID Level 6 and EVENODD 154

A.7 Two-Dimensional Parity 161

Appendix B: Modifying Graphs for Roll-Away Recovery **169**

Appendix C: Data **189**

C.1 Algorithm for Inserting a Commit Point Into a Write Graph..............................190
C.2 Raw Data 196
C.3 Sample Configuration File 230

Index **231**