Contents

1 Introduction .................................................. 1

Kenneth P. Birman and Robbert van Renesse

1.1 Categorizing computing systems ......................... 2
1.2 A philosophical evolution ................................. 5
1.3 Goals of this book ......................................... 6
1.4 A brief history of distributed computing ................. 7
   1.4.1 Network computing systems ....................... 7
   1.4.2 Remote procedure calls ............................ 8
   1.4.3 Distributed computing systems .................. 9
   1.4.4 Controversy ....................................... 11
1.5 Isis background and development .......................... 13
1.6 Isis Applications .......................................... 15
1.7 Underlying Theory ........................................ 16
1.8 The Horus Architecture ................................... 20
1.9 A Security Architecture For Horus ...................... 21
1.10 Future Directions ......................................... 22
1.11 Outline of this Book .................................... 22
1.12 Acknowledgements ...................................... 22

iii
## CONTENTS

4.5 Group styles of computing ........................................ 74
4.6 Location transparency: Too much of a good thing? .............. 74
4.7 What is really needed is a new architecture .................... 76
  4.7.1 What should we adopt from the RPC model? .................. 76
  4.7.2 The new architecture .......................................... 76
4.8 Conclusions ...................................................... 78

5 Exploiting Virtual Synchrony in Distributed Systems ............. 79

*Kenneth P. Birman and Thomas Joseph*

5.1 A toolkit for distributed systems .................................. 79
5.2 Virtual synchrony .................................................. 81
  5.2.1 Assumptions ..................................................... 81
  5.2.2 Subproblems we wish to solve ................................. 81
  5.2.3 Existing methodologies ....................................... 82
  5.2.4 Virtually synchronous environments .......................... 83
  5.2.5 Other virtually synchronous tools ............................. 84
5.3 Virtually synchronous tools ........................................ 85
  5.3.1 Atomic multicast primitives .................................... 85
  5.3.2 Process groups and group RPC ................................ 86
  5.3.3 Cooperating to execute requests ............................... 87
  5.3.4 Concurrency .................................................... 88
  5.3.5 Semaphores ...................................................... 88
  5.3.6 Replicated data ................................................ 88
  5.3.7 Detecting and reacting to failures ............................. 89
  5.3.8 Recovery and reconfiguration .................................. 89
  5.3.9 News service ................................................... 90
  5.3.10 Protection ..................................................... 90
  5.3.11 Additional tools ............................................. 90
5.4 Miscellaneous system-level facilities ............................ 90
  5.4.1 Run time facilities ............................................ 91
  5.4.2 Machine Independence and scaling ............................ 92
5.5 A toolkit application .............................................. 92
5.6 Inside the coordinator-cohort tool ............................... 96
5.7 Performance ....................................................... 97
5.8 Status .......................................................... 100

6 Virtual Synchrony Model .......................................... 101

*Kenneth P. Birman*

6.1 Basic elements of the model ....................................... 101
6.2 Basic execution axioms ............................................ 102
6.3 Virtually synchronous execution .................................. 103
6.4 Failure model ..................................................... 104
## II Redesign

### 7 Design Alternatives for Process Group Membership and Multicast

*Kenneth P. Birman, Robert Cooper, and Barry Gleeson*

#### 7.1 Introduction

#### 7.2 Process groups

- **7.2.1 Group membership**
- **7.2.2 Why provide support for process groups?**
- **7.2.3 Which processes should be allowed to send to a group?**
- **7.2.4 Should group multicast provide “strong guarantees”?**
- **7.2.5 Why not layer group communication over RPC?**
- **7.2.6 Does multicast belong inside the operating system?**

#### 7.3 Detailed design choices for a single group

- **7.3.1 Group structure: Members and clients**
- **7.3.2 Atomicity**
- **7.3.3 Causal and total multicast orderings**

#### 7.4 Ordering properties that span group boundaries

- **7.4.1 Who uses overlapping groups?**
- **7.4.2 Should causality be preserved between groups?**
- **7.4.3 Should causality always be preserved between groups?**
- **7.4.4 How visible should causality information be?**
- **7.4.5 Should abcast be ordered between groups?**

#### 7.5 Extended example: Causal process pairs

#### 7.6 An implementation

#### 7.7 Conclusions

### 8 The Horus System

*Robbert van Renesse, Kenneth P. Birman, Robert Cooper, Brad Glade, and Patrick Stephenson*

#### 8.1 Introduction

#### 8.2 Design

#### 8.3 The Multicast Transport Service

#### 8.4 Multi-threading and Ordering

#### 8.5 The Vsync Subsystem

#### 8.6 Security layer and Real-time Layer

#### 8.7 Customizing Horus to a Port Interface

- **8.7.1 Ports**
- **8.7.2 Integrating Horus into a Port Interface**

#### 8.8 The Isis Toolkit

#### 8.9 Initial Performance Results and Conclusion

### 9 Integrating Security in a Group-Oriented Distributed System

*Michael Reiter, Kenneth P. Birman, and Li Gong*
CONTENTS

9.1 Introduction ................................................................ 148
9.2 The Isis abstractions .................................................. 150
9.3 The system model ......................................................... 151
9.4 Protecting the Isis abstractions ...................................... 152
  9.4.1 Multicast authentication ........................................... 154
  9.4.2 Joining groups ....................................................... 156
  9.4.3 Causal multicast .................................................... 158
9.5 Delegation and access control ......................................... 162
9.6 Conclusion and future work ........................................... 165

10 High Availability in a Real-Time System ......................... 167

Carlos Almeida, Brad Glade, Keith Marzullo, and Robbert van
Renesse

10.1 Introduction ................................................................ 167
10.2 System structure ......................................................... 168
  10.2.1 System interfaces .................................................. 169
  10.2.2 Example application .............................................. 170
10.3 Proposed architecture and implementation ...................... 171
10.4 Current Status .......................................................... 172

III Protocols .................................................................. 173

11 Reliable Communication in the Presence of Failures ............. 176

Kenneth P. Birman and Thomas Joseph

11.1 Introduction ................................................................ 176
11.2 System characteristics ............................................... 178
11.3 Fault tolerant process groups ........................................ 178
  11.3.1 Using the group broadcast primitive to maintain process group views 180
  11.3.2 The atomic broadcast primitive ................................ 180
  11.3.3 The causal broadcast primitive .............................. 181
  11.3.4 Additional broadcast primitives .............................. 183
  11.3.5 Synchronous and asynchronous uses of the primitives 183
  11.3.6 Group addressing ................................................ 184
  11.3.7 Checkpointing the group state and transferring state .... 185
  11.3.8 Example ............................................................. 186
11.4 Implementation ......................................................... 186
  11.4.1 The inter-site layer ............................................... 187
  11.4.2 Site view management ........................................... 188
  11.4.3 The protocols ....................................................... 190
  11.4.4 An associative store and distributed garbage collection facility 197
  11.4.5 CBCAST flush implementation ............................ 198
  11.4.6 Wide area networks .............................................. 199
  11.4.7 Liveness ............................................................. 199
11.5 Limitations ............................................. 199
11.6 Conclusions .............................................. 200

12 Lightweight Causal and Atomic Group Multicast .................. 201

Kenneth P. Birman, André Schiper, and Pat Stephenson

12.1 Introduction ............................................. 201
12.1.1 The Isis Toolkit ..................................... 201
12.2 Execution model ........................................ 202
12.2.1 Basic system model ................................... 203
12.2.2 Virtual synchrony requirements for multicast protocols ..... 204
12.2.3 Vector time ........................................... 205
12.3 The CBCAST and ABCAST protocol ........................ 206
12.3.1 CBCAST protocol ................................... 206
12.3.2 Causal ABCAST protocol ............................ 209
12.3.3 Multicast stability .................................... 210
12.3.4 VT compression ....................................... 211
12.3.5 Delivery atomicity and group membership changes ....... 211
12.4 Extensions to the basic protocol ............................................. 215
12.4.1 Extension of CBCAST to multiple groups ............... 216
12.4.2 Multiple-group ABCAST ..................................... 217
12.4.3 Extended VT compression .................................. 217
12.4.4 Garbage collection of vector timestamps ................. 219
12.4.5 Atomicity and group membership changes .................. 219
12.4.6 Use of communication structure ......................... 222
12.4.7 Extensions to dynamic communication structures .......... 225
12.5 Applying the protocols to Isis ............................................. 228
12.5.1 Optimizations for Client/Server Groups ................. 228
12.5.2 Point-to-point messages ................................... 229
12.5.3 System interface issues ................................... 230
12.5.4 Group view management ................................... 230
12.6 Performance .............................................. 231
12.7 Conclusions .............................................. 235

13 Consistent Process Membership in Asynchronous Environments ... 237

Aleta Ricciardi and Kenneth P. Birman

13.1 The Need for a Membership Service ............................ 237
13.2 System Model .............................................. 240
13.2.1 The Formal Language ................................... 240
13.2.2 Failure Suspicions ..................................... 241
13.2.3 Process Behavior and Local Observations .................. 241
13.3 Strong GMP Specification ................................... 243
13.4 Solving Strong GMP ....................................... 244
13.4.1 The Basic Algorithm .................................... 245
13.4.2 The Full Algorithm ..................................... 246
## CONTENTS

18.3.2 The Graphics Editor ........................................... 333
18.3.3 Specifying Program Properties ................................. 334
18.3.4 Compiling Paralex Programs .................................... 336
18.3.5 Executing Paralex Programs .................................. 336
18.4 Fault Tolerance ...................................................... 337
18.5 Dynamic Load Balancing ........................................... 339
18.6 Performance Results ............................................... 339
18.7 Related Work ........................................................ 341
18.8 Status and Conclusions ............................................ 341

19 IMIS: A Distributed Query and Report Formatting System 343
   *T. Anthony Allen, William Sheppard, and Steve Condon*

19.1 System Overview ..................................................... 343
19.2 Why Not Relational ................................................ 344
19.3 Transcending Relational Technology ............................. 346
19.4 The Report Formatter ............................................... 346
   19.4.1 Semantics and Data Structure ................................ 347
   19.4.2 The Tawk Calculation Language ............................... 348
   19.4.3 Data Updating ................................................ 349
19.5 IMIS User Interface ................................................ 350
   19.5.1 Custom Widgets .............................................. 350
   19.5.2 Modified Widgets ............................................ 350
   19.5.3 Building a Form ............................................. 351
   19.5.4 Inspectors .................................................. 352
   19.5.5 Print Quality .............................................. 353
19.6 IMIS and Distributed Processing .................................. 353
   19.6.1 The Isis Toolkit ............................................. 353
   19.6.2 Multiple Client/Multiple Server Model ...................... 354
   19.6.3 Server Groups .............................................. 355
   19.6.4 Client Groups .............................................. 355
   19.6.5 Monitoring Process Groups .................................. 356
   19.6.6 Performance ............................................... 356
19.7 Summary ............................................................ 358

20 Distributed Programming with Asynchronous Ordered Channels in Distributed ML
   *Robert Cooper and Clifford Krumvieda*

20.1 Introduction ........................................................ 359
20.2 Distributed ML ...................................................... 360
20.3 Multicasts ........................................................... 361
20.4 Multicast Ordering .................................................. 363
   20.4.1 Causal Ordering .............................................. 363
   20.4.2 Causal Completeness ......................................... 363
   20.4.3 Ordering Declarations ....................................... 364