

Table of Contents

About the Authors xv

Preface xvii

Acknowledgments xxi

Acronyms xxiii

1 Introduction 1

1.1 Advantages of Distributed Systems 1

1.2 Defining Distributed Systems 3

1.3 Challenges of a Distributed System 5

1.4 Goals of Distributed System 6

1.4.1 Single System View 7

1.4.2 Hiding Distributions 7

1.4.3 Degrees and Distribution of Hiding 9

1.4.4 Interoperability 10

1.4.5 Dynamic Reconfiguration 10

1.5 Architectural Organization 11

1.6 Organization of the Book 12

Bibliography 13

2 The Internet 15

2.1 Origin and Organization 15

2.1.1 ISPs and the Topology of the Internet 17

2.2 Addressing the Nodes 17

2.3 Network Connection Protocol 20

2.3.1 IP Protocol 22

2.3.2 Transmission Control Protocol 22

2.3.3 User Datagram Protocol 22

2.4 Dynamic Host Control Protocol 23

2.5 Domain Name Service 24

2.5.1 Reverse DNS Lookup 27

2.5.2 Client Server Architecture 30

2.6 Content Distribution Network 32

2.7 Conclusion 34

Exercises 34

Bibliography 35

3 Process to Process Communication 37

3.1 Communication Types and Interfaces 38

3.1.1 Sequential Type 38

3.1.2 Declarative Type 39

3.1.3 Shared States 40

3.1.4 Message Passing 41

3.1.5 Communication Interfaces 41

3.2 Socket Programming 42

3.2.1 Socket Data Structures 43

3.2.2 Socket Calls 44

3.3 Remote Procedure Call 48

3.3.1 Xml RPC 52

3.4 Remote Method Invocation 55

3.5 Conclusion 59

Exercises 59

Additional Web Resources 61

Bibliography 61

4 Microservices, Containerization, and MPI 63

4.1 Microservice Architecture 64

4.2 REST Requests and APIs 66

4.2.1 Weather Data Using REST API 67

4.3 Cross Platform Applications 68

4.4 Message Passing Interface 78

4.4.1 Process Communication Models 78

4.4.2 Programming with MPI 81

4.5 Conclusion 87

Exercises 88

Additional Internet Resources 89

Bibliography 89

5 Clock Synchronization and Event Ordering 91

5.1 The Notion of Clock Time 92

5.2 External Clock Based Mechanisms 93

5.2.1 Cristian's Algorithm 93

5.2.2 Berkeley Clock Protocol 94

5.2.3 Network Time Protocol 95

5.2.3.1 Symmetric Mode of Operation 96

5.3 Events and Temporal Ordering 97

5.3.1 Causal Dependency 99

5.4 Logical Clock 99

5.5 Causal Ordering of Messages 106

5.6 Multicast Message Ordering 107

5.6.1 Implementing FIFO Multicast	110
5.6.2 Implementing Causal Ordering	112
5.6.3 Implementing Total Ordering	113
5.6.4 Reliable Multicast	114
5.7 Interval Events	115
5.7.1 Conceptual Neighborhood	116
5.7.2 Spatial Events	118
5.8 Conclusion	120
Exercises	121
Bibliography	123
6 Global States and Termination Detection	127
6.1 Cuts and Global States	127
6.1.1 Global States	132
6.1.2 Recording of Global States	134
6.1.3 Problem in Recording Global State	138
6.2 Liveness and Safety	140
6.3 Termination Detection	143
6.3.1 Snapshot Based Termination Detection	144
6.3.2 Ring Method	145
6.3.3 Tree Method	148
6.3.4 Weight Throwing Method	151
6.4 Conclusion	153
Exercises	154
Bibliography	156
7 Leader Election	157

7.1	Impossibility Result	158
7.2	Bully Algorithm	159
7.3	Ring-Based Algorithms	160
7.3.1	Circulate IDs All the Way	161
7.3.2	As Far as an ID Can Go	162
7.4	Hirschberg and Sinclair Algorithm	163
7.5	Distributed Spanning Tree Algorithm	167
7.5.1	Single Initiator Spanning Tree	167
7.5.2	Multiple Initiators Spanning Tree	170
7.5.3	Minimum Spanning Tree	176
7.6	Leader Election in Trees	176
7.6.1	Overview of the Algorithm	176
7.6.2	Activation Stage	177
7.6.3	Saturation Stage	178
7.6.4	Resolution Stage	179
7.6.5	Two Nodes Enter SATURATED State	180
7.7	Leased Leader Election	182
7.8	Conclusion	184
	Exercises	185
	Bibliography	187
8	Mutual Exclusion	189
8.1	System Model	190
8.2	Coordinator-Based Solution	192
8.3	Assertion-Based Solutions	192
8.3.1	Lamport's Algorithm	192

8.3.2 Improvement to Lamport's Algorithm 195

8.3.3 Quorum-Based Algorithms 196

8.4 Token-Based Solutions 203

8.4.1 Suzuki and Kasami's Algorithm 203

8.4.2 Singhal's Heuristically Aided Algorithm 206

8.4.3 Raymond's Tree-Based Algorithm 212

8.5 Conclusion 214

Exercises 215

Bibliography 216

9 Agreements and Consensus 219

9.1 System Model 220

9.1.1 Failures in Distributed System 221

9.1.2 Problem Definition 222

9.1.3 Agreement Problem and Its Equivalence 223

9.2 Byzantine General Problem (BGP) 225

9.2.1 BGP Solution Using Oral Messages 228

9.2.2 Phase King Algorithm 232

9.3 Commit Protocols 233

9.3.1 Two-Phase Commit Protocol 234

9.3.2 Three-Phase Commit 238

9.4 Consensus 239

9.4.1 Consensus in Synchronous Systems 239

9.4.2 Consensus in Asynchronous Systems 241

9.4.3 Paxos Algorithm 242

9.4.4 Raft Algorithm 244

9.4.5 Leader Election 246

9.5 Conclusion 248

Exercises 249

Bibliography 250

10 Gossip Protocols 253

10.1 Direct Mail 254

10.2 Generic Gossip Protocol 255

10.3 Anti-entropy 256

10.3.1 Push-Based Anti-Entropy 257

10.3.2 Pull-Based Anti-Entropy 258

10.3.3 Hybrid Anti-Entropy 260

10.3.4 Control and Propagation in Anti-Entropy 260

10.4 Rumor-mongering Gossip 261

10.4.1 Analysis of Rumor Mongering 262

10.4.2 Fault-Tolerance 265

10.5 Implementation Issues 265

10.5.1 Network-Related Issues 266

10.6 Applications of Gossip 267

10.6.1 Peer Sampling 267

10.6.2 Failure Detectors 270

10.6.3 Distributed Social Networking 271

10.7 Gossip in IoT Communication 273

10.7.1 Context-Aware Gossip 273

10.7.2 Flow-Aware Gossip 274

10.7.2.1 Fire Fly Gossip 274

10.7.2.2 Trickle 275

10.8 Conclusion 278

Exercises 279

Bibliography 280

11 Message Diffusion Using Publish and Subscribe 283

11.1 Publish and Subscribe Paradigm 284

11.1.1 Broker Network 285

11.2 Filters and Notifications 287

11.2.1 Subscription and Advertisement 288

11.2.2 Covering Relation 288

11.2.3 Merging Filters 290

11.2.4 Algorithms 291

11.3 Notification Service 294

11.3.1 Siena 294

11.3.2 Rebeca 295

11.3.3 Routing of Notification 296

11.4 MQTT 297

11.5 Advanced Message Queuing Protocol 299

11.6 Effects of Technology on Performance 301

11.7 Conclusions 303

Exercises 304

Bibliography 305

12 Peer-to-Peer Systems 309

12.1 The Origin and the Definition of P2P 310

12.2 P2P Models 311

12.2.1 Routing in P2P Network 312

12.3 Chord Overlay 313

12.4 Pastry 321

12.5 Can 325

12.6 Kademlia 327

12.7 Conclusion 331

Exercises 332

Bibliography 333

13 Distributed Shared Memory 337

13.1 Multicore and S-DSM 338

13.1.1 Coherency by Delegation to a Central Server 339

13.2 Manycore Systems and S-DSM 340

13.3 Programming Abstractions 341

13.3.1 MapReduce 341

13.3.2 OpenMP 343

13.3.3 Merging Publish and Subscribe with DSM 345

13.4 Memory Consistency Models 347

13.4.1 Sequential Consistency 349

13.4.2 Linearizability or Atomic Consistency 351

13.4.3 Relaxed Consistency Models 352

13.4.3.1 Release Consistency 356

13.4.4 Comparison of Memory Models 357

13.5 DSM Access Algorithms 358

13.5.1 Central Sever Algorithm 359

13.5.2 Migration Algorithm 360

13.5.3 Read Replication Algorithm	361
13.5.4 Full Replication Algorithm	362
13.6 Conclusion	364
Exercises	364
Bibliography	367
14 Distributed Data Management	371
14.1 Distributed Storage Systems	372
14.1.1 Raid	372
14.1.2 Storage Area Networks	372
14.1.3 Cloud Storage	373
14.2 Distributed File Systems	375
14.3 Distributed Index	376
14.4 NoSQL Databases	377
14.4.1 Key-Value and Document Databases	378
14.4.1.1 MapReduce Algorithm	380
14.4.2 Wide Column Databases	381
14.4.3 Graph Databases	382
14.4.3.1 Pregel Algorithm	384
14.5 Distributed Data Analytics	386
14.5.1 Distributed Clustering Algorithms	388
14.5.1.1 Distributed K-Means Clustering Algorithm	388
14.5.2 Stream Clustering	391
14.5.2.1 BIRCH Algorithm	392
14.6 Conclusion	393
Exercises	394

Bibliography 395

15 Distributed Knowledge Management 399

15.1 Distributed Knowledge 400

15.2 Distributed Knowledge Representation 401

15.2.1 Resource Description Framework (RDF) 401

15.2.2 Web Ontology Language (OWL) 406

15.3 Linked Data 407

15.3.1 Friend of a Friend 407

15.3.2 DBpedia 408

15.4 Querying Distributed Knowledge 409

15.4.1 SPARQL Query Language 410

15.4.2 SPARQL Query Semantics 411

15.4.3 SPARQL Query Processing 413

15.4.4 Distributed SPARQL Query Processing 414

15.4.5 Federated and Peer-to-Peer SPARQL Query Processing 416

15.5 Data Integration in Distributed Sensor Networks 421

15.5.1 Semantic Data Integration 422

15.5.2 Data Integration in Constrained Systems 424

15.6 Conclusion 427

Exercises 428

Bibliography 429

16 Distributed Intelligence 433

16.1 Agents and Multi-Agent Systems 434

16.1.1 Agent Embodiment 436

16.1.2 Mobile Agents 436

16.1.3 Multi-Agent Systems	437
16.2 Communication in Agent-Based Systems	438
16.2.1 Agent Communication Protocols	439
16.2.2 Interaction Protocols	440
16.2.2.1 Request Interaction Protocol	441
16.3 Agent Middleware	441
16.3.1 FIPA Reference Model	442
16.3.2 FIPA Compliant Middleware	443
16.3.2.1 JADE: Java Agent Development Environment	443
16.3.2.2 MobileC	443
16.3.3 Agent Migration	444
16.4 Agent Coordination	445
16.4.1 Planning	447
16.4.1.1 Distributed Planning Paradigms	447
16.4.1.2 Distributed Plan Representation and Execution	448
16.4.2 Task Allocation	450
16.4.2.1 Contract-Net Protocol	450
16.4.2.2 Allocation of Multiple Tasks	452
16.4.3 Coordinating Through the Environment	453
16.4.3.1 Construct-Ant-Solution	455
16.4.3.2 Update-Pheromone	456
16.4.4 Coordination Without Communication	456
16.5 Conclusion	456
Exercises	457
Bibliography	459

17 Distributed Ledger	461
17.1 Cryptographic Techniques	462
17.2 Distributed Ledger Systems	464
17.2.1 Properties of Distributed Ledger Systems	465
17.2.2 A Framework for Distributed Ledger Systems	466
17.3 Blockchain	467
17.3.1 Distributed Consensus in Blockchain	468
17.3.2 Forking	470
17.3.3 Distributed Asset Tracking	471
17.3.4 Byzantine Fault Tolerance and Proof of Work	472
17.4 Other Techniques for Distributed Consensus	473
17.4.1 Alternative Proofs	473
17.4.2 Non-linear Data Structures	474
17.4.2.1 Tangle	474
17.4.2.2 Hashgraph	476
17.5 Scripts and Smart Contracts	480
17.6 Distributed Ledgers for Cyber-Physical Systems	483
17.6.1 Layered Architecture	484
17.6.2 Smart Contract in Cyber-Physical Systems	486
17.7 Conclusion	486
Exercises	487
Bibliography	488
18 Case Study	491
18.1 Collaborative E-Learning Systems	492
18.2 P2P E-Learning System	493

18.2.1 Web Conferencing Versus P2P-IPS	495
18.3 P2P Shared Whiteboard	497
18.3.1 Repainting Shared Whiteboard	497
18.3.2 Consistency of Board View at Peers	498
18.4 P2P Live Streaming	500
18.4.1 Peer Joining	500
18.4.2 Peer Leaving	503
18.4.3 Handling “Ask Doubt”	504
18.5 P2P-IPS for Stored Contents	504
18.5.1 De Bruijn Graphs for DHT Implementation	505
18.5.2 Node Information Structure	507
18.5.2.1 Join Example	510
18.5.3 Leaving of Peers	510
18.6 Searching, Sharing, and Indexing	511
18.6.1 Pre-processing of Files	511
18.6.2 File Indexing	512
18.6.3 File Lookup and Download	512
18.7 Annotations and Discussion Forum	513
18.7.1 Annotation Format	513
18.7.2 Storing Annotations	514
18.7.3 Audio and Video Annotation	514
18.7.4 PDF Annotation	514
18.7.5 Posts, Comments, and Announcements	514
18.7.6 Synchronization of Posts and Comments	515
18.7.6.1 Epidemic Dissemination	516

18.7.6.2 Reconciliation 516

18.8 Simulation Results 516

18.8.1 Live Streaming and Shared Whiteboard 517

18.8.2 De Bruijn Overlay 518

18.9 Conclusion 520

Bibliography 521

Index 525