

DISTRIBUTED OPERATING SYSTEMS

CONCEPTS & PRACTICE



DOREEN L. GALLI

Contents

1. Introduction to Distributed Systems	1
1.1 <i>What is an Operating System?</i>	1
1.2 <i>What is a Distributed System?</i>	2
1.2.1 Popular Network Topologies and Features	4
1.2.2 ISO/OSI Reference Model	8
1.2.3 Distributed Computing Models	12
1.2.4 Distributed versus Centralized Solutions	14
1.2.5 Network vs. Distributed Operating Systems	15
1.3 <i>What is a Real-Time System?</i>	16
1.3.1 Real-time Event Characteristics	17
1.3.2 Network Characteristics Affecting Distributed Real-Time Applications	17
1.4 <i>What is a Parallel System?</i>	19
1.4.1 Parallel Architectures	19
1.4.2 Parallel Software Paradigms	23
1.5 <i>Sample Distributed Application</i>	25
1.6 <i>Summary</i>	27
1.7 <i>References for Further Study</i>	27
<i>Exercises</i>	28
 2. The Kernel	 31
2.1 <i>Kernel Types</i>	31
2.2 <i>Processes and Threads</i>	33
2.2.1 Introduction to Multi-threaded Processes	33
2.2.2 Multi-threaded Process Paradigms	34
2.2.3 Multi-threaded support	37
2.3 <i>Process Management</i>	38
2.3.1 Types of Processes	40
2.3.2 Load Distribution & Process Migration	41
2.4 <i>Process Scheduling</i>	45
2.4.1 Identifying Processes for Scheduling	45
2.4.2 Scheduler Organization	45
2.5 <i>Summary</i>	46

<i>2.6 References for Further Study</i>	49
<i>Exercises</i>	49

3. Interprocess Communication 53

<i>3.1 Selection Factors</i>	54
<i>3.2 Message Passing</i>	54
3.2.1 Blocking Primitives	55
3.2.2 Non-blocking Primitives	58
3.2.3 Process Address	59
<i>3.3 Pipes</i>	61
3.3.1 Unnamed Pipes	61
3.3.2 Named Pipes	63
<i>3.4 Sockets</i>	63
3.4.1 UNIX Sockets	65
3.4.2 JAVA Support for Sockets	69
<i>3.5 Remote Procedure Calls</i>	72
3.5.1 Parameter Type	72
3.5.2 Data Type Support	72
3.5.3 Parameter Marshalling	73
3.5.4 RPC Binding	73
3.5.5 RPC Authentication	74
3.5.6 RPC Call Semantics	75
3.5.7 SUN's ONC RPC	76
<i>3.6 Summary</i>	76
<i>3.7 References for Further Study</i>	77
<i>Exercises</i>	77

4. Memory Management 81

<i>4.1 Review of Centralized Memory Management</i>	82
4.1.1 Virtual Memory	82
4.1.2 Pages and Segments	82
4.1.3 Page Replacement Algorithms	84
<i>4.2 Simple Memory Model</i>	85
<i>4.3. Shared Memory Model</i>	86
4.3.1 Shared Memory Performance	87
4.3.2 Cache Consistency	88
<i>4.4 Distributed Shared Memory</i>	89
4.3.1 Methods for Distributing Shared Data	90
4.3.2 DSM Performance Issues	93
<i>4.5 Memory Migration</i>	98
<i>4.6 Summary</i>	100

<i>4.7 References for Further Study</i>	<i>101</i>
<i>Exercises</i>	<i>102</i>

5. Concurrency Control **105**

<i>5.1 Mutual Exclusion & Critical Regions</i>	<i>105</i>
<i>5.2 Semaphores</i>	<i>107</i>
5.2.1 Semaphore Weakness	108
5.2.2 Semaphore Evaluation	108
<i>5.3 Monitors</i>	<i>108</i>
5.3.1 Condition Variables	110
5.3.2 Monitor Evaluation	111
<i>5.4 Locks</i>	<i>111</i>
5.4.1 Taking Turns	113
5.4.2 Atomic Operations and Hardware Support	113
<i>5.5 Software Lock Control</i>	<i>115</i>
5.5.1 Centralized Lock Manager	115
5.5.2 Distributed Lock Manager	116
<i>5.6 Token-Passing Mutual Exclusion</i>	<i>118</i>
<i>5.7 Deadlocks</i>	<i>119</i>
5.7.1 Prevent Deadlocks	121
5.7.2 Avoid Deadlocks	122
5.7.3 Ignore Deadlocks	122
5.7.4 Detect Deadlocks	122
<i>5.8 Summary</i>	<i>124</i>
<i>5.9 References for Further Study</i>	<i>125</i>
<i>Exercises</i>	<i>125</i>

6. Object-Based Operating Systems **127**

<i>6.1 Introduction to Objects</i>	<i>128</i>
6.1.1 Object Definitions	128
6.1.2 Evaluation of Objects	129
<i>6.2 The Clouds Object Approach</i>	<i>131</i>
6.2.1 Clouds' Objects	131
6.2.2 Clouds Threads	132
6.2.3 Clouds' Memory Storage	132
<i>6.3 Chorus V3 and COOL v2</i>	<i>132</i>
6.3.1 The Base Layer: COOL Memory Management	134
6.3.2 The Generic Runtime System Layer: COOL Objects	134
6.3.3 The Language-Specific Runtime System Layer	135
<i>6.4 Amoeba</i>	<i>135</i>
6.4.1 Identification and Protection of Amoeba Objects	137

6.4.2 Amoeba Object Communication	137
6.5 Distributed Component Object Model (DCOM)	138
6.5.1 Monikers	140
6.5.2 Remote Method Calls	140
6.5.3 Garbage Collection	141
6.5.4 Thread Models Supported in DCOM	141
6.5.5 DCOM Security Policies	142
6.6 CORBA™ Overview	143
6.6.1 CORBA's ORB	144
6.6.2 CORBA Object Adapters	145
6.6.3 CORBA Messaging	147
6.6.4 CORBA Compliance	147
6.6.5 CORBA to COM Mapping	148
6.7 Summary	148
6.7 References for Further Study	149
Exercises	149

7. Distributed Process Management	151
7.1 <i>Distributed Scheduling Algorithm Choices</i>	152
7.1.1 Level of Scheduling	152
7.1.2 Load Distribution Goals	153
7.1.3 Scheduling Efficiency Goals	153
7.1.4 Processor Binding Time	155
7.2 <i>Scheduling Algorithm Approaches</i>	157
7.2.1 Usage Points	158
7.2.2 Graph Theory	160
7.2.3 Probes	162
7.2.4 Scheduling Queues	163
7.2.5 Stochastic Learning	166
7.3 <i>Coordinator Elections</i>	167
7.4 <i>Orphan Processes</i>	168
7.4.1 Orphan Clean Up	170
7.4.2 Child Process Allowance	172
7.4.3 Process Version Numbers	173
7.5 <i>Summary</i>	173
7.6 <i>References for Further Study</i>	175
<i>Exercises</i>	176

8. Distributed File Systems	179
8.1 <i>Distributed Name Service</i>	179
8.1.1 File Types	180
8.1.2 Location Transparency	181

8.1.3 Global Naming & Name Transparency	181
8.2 <i>Distributed File Service</i>	187
8.2.1 File Variations	187
8.2.2 File Modification Notification	190
8.2.3 File Service Implementations	191
8.2.4 File Replication	192
8.3 <i>Distributed Directory Service</i>	195
8.3.1 Directory Structures	195
8.3.2 Directory Management	195
8.3.3 Directory Operations	196
8.4 <i>NFS™</i>	197
8.4.1 NFS File Service	197
8.4.2 NFS Directory Service	198
8.4.3 NFS Name Service	199
8.5 <i>X.500</i>	200
8.5.1 X.500 File and Name Service: The Information Model	201
8.5.2 X.500's Directory Service: The Directory Model	202
8.6 <i>Summary</i>	202
8.7 <i>References for Further Study</i>	203
<i>Exercises</i>	204

9. Transaction Management & Consistency Models	207
9.1 <i>Transaction Management Motivation</i>	208
9.1.1 The Lost Update	208
9.1.2 Retrieval Disparity	209
9.2 <i>ACID Properties of a Transaction</i>	213
9.3 <i>Consistency Models</i>	216
9.3.1 Strict Consistency Model	216
9.3.2 Sequential Consistency Model	216
9.3.3 Casual Consistency Model	217
9.3.4 PRAM Consistency Model	218
9.3.5 Processor Consistency Model	219
9.3.6 Weak Consistency Model	220
9.3.7 Release Consistency Model	222
9.3.8 Lazy Release Consistency	224
9.3.9 Entry Consistency Model	224
9.4 <i>Two-Phase Commit Protocol</i>	224
9.4.1 Prepare to Commit Phase	225
9.4.2 Commit Phase	227
9.5 <i>Nested Transactions</i>	228
9.6 <i>Implementation Issues for Transactions</i>	230

9.6.1 Premature Reads & Writes	231
9.6.2 The Domino Effect of aborts	231
9.6.3 Ensuring Recoverability	231
9.7 Summary	231
9.7 References for Further Study	233
Exercises	233
10. Distributed Synchronization	235
10.1 Introduction to Global Time	235
10.2 Physical Clocks	236
10.2.1 Obtaining an Accurate Physical Time	236
10.2.2 Synchronizing Physical Time	237
10.2.3 Centralized Physical Time Services	239
10.2.4 Distributed Physical Time Services	242
10.3 Network Time Protocol (NTP)	243
10.3.1 NTP Architecture	244
10.3.2 NTP Design Goals	244
10.3.3 NTP Synchronization Modes	247
10.3.4 Simple Network Time Protocol (SNTP)	247
10.4 Logical Clocks	250
10.4.1 Happen-Before Relationship	251
10.4.2 Logical Ordering	251
10.4.3 Total Ordering with Logical Clocks	254
10.5 Summary	254
10.6 References for Further Study	255
Exercises	255
11. Distributed Security	259
11.1 Cryptography & Digital Signatures	260
11.1.1 Symmetric Encryption	261
11.1.2 Asymmetric Encryption	265
11.2 Authentication	270
11.2.1 Certificate Lists	265
11.2.2 Centralized Certificate Distribution Center	271
11.3 Access Control (Firewalls)	279
11.3.1 Packet Filtering Gateways	280
11.3.2 Proxy Services	281
11.3.3 Firewall Architectures	282
11.4 Summary	284
11.5 References for Further Study	285
Exercises	285

12. CASE STUDY: Windows 2000™	289
12.1 <i>Overview: Windows 2000 Design</i>	291
12.2 <i>Kernel Mode Overview</i>	293
12.2.1 Kernel Objects	295
12.2.2 Hardware Abstraction Layer (HAL)	297
12.2.3 Device Drivers	297
12.2.4 The Executive	297
12.3 <i>Plug and Play</i>	299
12.4 <i>NT Files System in Windows 2000 (NTFS)</i>	303
12.4.1 Access Control Lists (ACLs)	304
12.4.2 Reparse Points	305
12.4.3 Storage Management	306
12.5 <i>Active Directory</i>	307
12.5.1 NameSpace	310
12.5.2 Change Journal for Replication and Scalability	312
12.5.3 Microsoft Index Server and HTTP Support	314
12.6 <i>Microsoft Management Console (MMC)</i>	317
12.7 <i>Cluster Service</i>	318
12.7.1 Cluster Service Overview	319
12.7.2 Cluster Abstractions	320
12.7.3 Cluster Service Architectures	320
12.7.4 Cluster Service Deployment to Applications	320
12.8 <i>Windows 2000 Security</i>	321
12.8.1 Security Configuration Editor	323
12.8.2 Encrypting File System	325
12.8.3 Microsoft Security Support Provider Interface	327
12.9 <i>HYDRA – A Thin Client</i>	327
12.10 <i>Summary</i>	327
12.11 <i>References for Further Study</i>	328
<i>Exercises</i>	328
 Appendix A. Surgical Scheduling Program	 331
 List of Acronyms	 413
 Glossary of Terms	 419
 Bibliography	 437
 Index	 455