

Contents

Foreword	vii
Foreword to the First Printing	ix
Preface	xix
Chapter 1 Introduction	1
1.1 Parallel Computing and Networks	2
1.2 Parallel Computer Architectures	3
1.3 Network Design Considerations	5
1.4 Classification of Interconnection Networks	7
1.5 Shared-Medium Networks	9
1.5.1 Shared-Medium Local Area Networks	9
1.5.2 Shared-Medium Backplane Bus	10
1.6 Direct Networks	12
1.6.1 Characterization of Direct Networks	14
1.6.2 Popular Network Topologies	15
1.6.3 Examples	19
1.7 Indirect Networks	20
1.7.1 Characterization of Indirect Networks	20
1.7.2 Crossbar Networks	22
1.7.3 Multistage Interconnection Networks	23
1.7.4 A Generalized MIN Model	24
1.7.5 Classification of Multistage Interconnection Networks	28
1.7.6 Examples	34
1.8 Hybrid Networks	34
1.8.1 Multiple Backplane Buses	36
1.8.2 Hierarchical Networks	37
1.8.3 Cluster-Based Networks	37
1.8.4 Other Hypergraph Topologies	38
1.9 A Unified View of Direct and Indirect Networks	39

Chapter 2	Message Switching Layer	43
2.1	Network and Router Model	44
2.2	Basic Concepts	45
2.3	Basic Switching Techniques	48
2.3.1	Circuit Switching	49
2.3.2	Packet Switching	52
2.3.3	Virtual Cut-Through (VCT) Switching	54
2.3.4	Wormhole Switching	55
2.3.5	Mad Postman Switching	58
2.4	Virtual Channels	61
2.5	Hybrid Switching Techniques	64
2.5.1	Buffered Wormhole Switching	64
2.5.2	Pipelined Circuit Switching	67
2.5.3	Scouting Switching	70
2.6	Optimizing Switching Techniques	72
2.7	A Comparison of Switching Techniques	75
2.8	Engineering Issues	76
2.9	Commented References	77
	Exercises	78
	Problems	80
Chapter 3	Deadlock, Livelock, and Starvation	83
3.1	A Theory of Deadlock Avoidance	86
3.1.1	Network and Router Models	86
3.1.2	Basic Definitions	87
3.1.3	Necessary and Sufficient Condition	90
3.1.4	Deadlock Avoidance in SAF and VCT Switching	95
3.1.5	Deadlock Avoidance in Wormhole Switching	98
3.2	Extensions	103
3.2.1	Channel Classes	103
3.2.2	Extending the Domain of the Routing Function	105
3.2.3	Central Queues	106
3.3	Alternative Approaches	107
3.3.1	Theoretical Approaches	107
3.3.2	Deflection Routing	109
3.3.3	Injection Limitation	110
3.4	Deadlock Avoidance in Switch-Based Networks	110
3.5	Deadlock Prevention in Circuit Switching and PCS	112

3.6	Deadlock Recovery	112
3.6.1	Deadlock Probability	113
3.6.2	Detection of Potential Deadlocks	116
3.6.3	Progressive and Regressive Recovery Techniques	117
3.7	Livelock Avoidance	122
3.8	Engineering Issues	123
3.9	Commented References	125
	Exercises	127
	Problems	136
Chapter 4	Routing Algorithms	139
4.1	Taxonomy of Routing Algorithms	140
4.2	Deterministic Routing Algorithms	145
4.3	Partially Adaptive Algorithms	150
4.3.1	Planar-Adaptive Routing	150
4.3.2	Turn Model	152
4.4	Fully Adaptive Algorithms	156
4.4.1	Algorithms Based on Structured Buffer Pools	157
4.4.2	Algorithms Derived from SAF Algorithms	158
4.4.3	Virtual Networks	159
4.4.4	Deterministic and Adaptive Subnetworks	165
4.5	Maximally Adaptive Routing Algorithms	169
4.5.1	Algorithms with Maximum Adaptivity	169
4.5.2	Algorithms with Minimum Buffer Requirements	173
4.5.3	True Fully Adaptive Routing Algorithms	177
4.6	Nonminimal Routing Algorithms	179
4.7	Backtracking Protocols	180
4.8	Routing in MINs	184
4.8.1	Blocking Condition in MINs	184
4.8.2	Self-Routing Algorithms for MINs	187
4.9	Routing in Switch-Based Networks with Irregular Topologies	190
4.10	Resource Allocation Policies	194
4.10.1	Selection Function	194
4.10.2	Policies for Arbitration and Resource Allocation	196
4.11	Engineering Issues	198
4.12	Commented References	200
	Exercises	201
	Problems	204

Chapter 5	Collective Communication Support	207
5.1	Collective Communication Services	208
5.1.1	Multiple One-to-One Communication	209
5.1.2	One-to-All Communication	209
5.1.3	All-to-One Communication	210
5.1.4	All-to-All Communication	211
5.1.5	Convenient Collective Communication Services	212
5.2	System Support for Collective Communication	213
5.3	Preliminary Considerations	214
5.3.1	The Need for Multicast Communication	214
5.3.2	Evaluation Criteria	215
5.4	Models for Multicast Communication	217
5.5	Hardware Implementations of Multicast	219
5.5.1	Multiaddress Encoding Schemes	220
5.5.2	Tree-Based Multicast Routing	221
5.5.3	Path-Based Multicast Communication	235
5.6	Hardware Support for Barrier Synchronization and Reduction	258
5.6.1	Barrier Synchronization on a Linear Array	258
5.6.2	Barrier Synchronization on Meshes	262
5.6.3	Reduction and Global Combining	264
5.7	Software Implementations of Multicast	267
5.7.1	Desirable Features in Multicast Trees	268
5.7.2	Dimension-Ordered Chains	271
5.7.3	Multicast in Hypercubes	274
5.7.4	Multicast in Meshes	274
5.8	Engineering Issues	277
5.9	Commented References	280
	Exercises	281
	Problems	284
Chapter 6	Fault-Tolerant Routing	287
6.1	Fault-Induced Deadlock and Livelock	288
6.2	Channel and Network Redundancy	290
6.3	Fault Models	293
6.4	Fault-Tolerant Routing in SAF and VCT Networks	297
6.4.1	Routing Algorithms Based on Local Information	298
6.4.2	Routing Algorithms Based on Nonlocal Information	301
6.4.3	Routing Algorithms Based on Graph Search	305
6.4.4	Deadlock and Livelock Freedom Issues	309

6.5	Fault-Tolerant Routing in Wormhole-Switched Networks	309
6.5.1	Rectangular Fault Regions	310
6.5.2	Software-Based Fault-Tolerant Routing	322
6.5.3	Unconstrained Fault Regions	325
6.6	Fault-Tolerant Routing in PCS and Scouting Networks	331
6.7	Dynamic Fault Recovery	337
6.7.1	Base Mechanisms	337
6.7.2	Flit-Level Recovery	338
6.7.3	Message-Level Recovery	341
6.7.4	Miscellaneous Issues	344
6.8	Engineering Issues	345
6.9	Commented References	351
	Exercises	353
	Problems	356
Chapter 7	Network Architectures	359
7.1	Network Topology and Physical Constraints	360
7.1.1	Bisection Bandwidth Constraints	360
7.1.2	Node Size Constraints	362
7.1.3	Wire Delay Constraints	363
7.1.4	A Latency Model	365
7.1.5	Analysis	367
7.1.6	Packaging Constraints	375
7.2	Router Architectures	378
7.2.1	Intrarouter Performance	379
7.2.2	Physical Channel Issues	384
7.2.3	Wormhole Routers	390
7.2.4	VCT Routers	409
7.2.5	Circuit Switching	423
7.2.6	Pipelined Circuit Switching	425
7.2.7	Buffered Wormhole Switching	431
7.2.8	Network of Workstations	434
7.3	Engineering Issues	439
7.4	Commented References	440
	Exercises	442
	Problems	443
Chapter 8	Messaging Layer Software	445
8.1	Functionality of the Messaging Layer	446
8.1.1	Packetization	447

8.1.2	Network Interface Control	447
8.1.3	User/Kernel Interface	448
8.1.4	Copying and Buffering	449
8.1.5	In-Order Message Delivery	450
8.1.6	Reliable Message Delivery	450
8.1.7	Protection	450
8.2	Impact of Message Processing Delays	451
8.3	Implementation of the Messaging Layer	453
8.3.1	Example: Active Messages	454
8.3.2	Example: Illinois Fast Messages	457
8.4	Application Programming Layer: The Message Passing Interface	459
8.4.1	General Concepts	460
8.4.2	Point-to-Point Communication	463
8.4.3	Collective Communication	465
8.5	Engineering Issues	469
8.6	Commented References	472
	Problems	473
Chapter 9	Performance Evaluation	475
9.1	Performance Metrics and Normalized Results	476
9.2	Workload Models	479
9.3	Comparison of Switching Techniques	482
9.4	Comparison of Routing Algorithms	485
9.4.1	Performance under Uniform Traffic	486
9.4.2	Performance under Local Traffic	492
9.4.3	Performance under Nonuniform Traffic	493
9.5	Effect of Message Length	495
9.6	Effect of Network Size	496
9.7	Impact of Design Parameters	496
9.7.1	Effect of the Number of Virtual Channels	497
9.7.2	Effect of the Number of Ports	499
9.7.3	Effect of Buffer Size	501
9.8	Comparison of Routing Algorithms for Irregular Topologies	502
9.9	Injection Limitation	506
9.10	Impact of Router Delays on Performance	510
9.10.1	A Speed Model	510
9.10.2	Performance under Uniform Traffic	514
9.10.3	Performance under Local Traffic	514
9.10.4	Performance under Nonuniform Traffic	516

9.10.5	Effect of the Number of Virtual Channels	517
9.11	Performance of Collective Communication	520
9.11.1	Comparison with Separate Addressing	520
9.11.2	Comparing Tree-Based and Path-Based Algorithms	520
9.11.3	Performance of Base Routing Conformed Path Multicast	524
9.11.4	Performance of Optimized Multicast Routing	524
9.11.5	Performance of Unicast- and Multidestination-Based Schemes	527
9.11.6	Performance of Path-Based Barrier Synchronization	528
9.12	Software Messaging Layer	530
9.12.1	Overhead of the Software Messaging Layer	530
9.12.2	Optimizations in the Software Messaging Layer	533
9.13	Performance of Fault-Tolerant Algorithms	535
9.13.1	Software-Based Fault-Tolerant Routing	535
9.13.2	Routing Algorithms for PCS	540
9.13.3	Routing Algorithms for Scouting	545
9.14	Conclusions	551
9.15	Commented References	557
Appendix A	Formal Definitions for Deadlock Avoidance	559
A.1	Assumptions	559
A.2	Definitions	560
A.3	Theorems	564
Appendix B	Acronyms	565
References		569
Index		593