

Table of Contents

About the Authors viii

Foreword xix

Preface xx

Guide to Instructors/Readers xxiii

Part I Scalability and Clustering 1

Chapter 1 Scalable Computer Platforms and Models 3

1.1 Evolution of Computer Architecture 5

1.1.1 Computer Generations 5

1.1.2 Scalable Computer Architectures 6

1.1.3 Converging System Architectures 8

1.2 Dimensions of Scalability 9

1.2.1 Resource Scalability 9

1.2.2 Application Scalability 11

1.2.3 Technology Scalability 12

1.3 Parallel Computer Models 13

1.3.1 Semantic Attributes 14

1.3.2 Performance Attributes 17

1.3.3 Abstract Machine Models 18

1.3.4 Physical Machine Models 26

1.4 Basic Concepts of Clustering 30

1.4.1 Cluster Characteristics 30

1.4.2 Architectural Comparisons 31

1.4.3 Benefits and Difficulties of Clusters 32

1.5 Scalable Design Principles 37

1.5.1 Principle of Independence 37

1.5.2 Principle of Balanced Design 39

1.5.3 Design for Scalability 44

1.6 Bibliographic Notes and Problems 47

Chapter 2 Basics of Parallel Programming 51

2.1 Parallel Programming Overview 51

2.1.1 Why Is Parallel Programming Difficult? 52

2.1.2 Parallel Programming Environments 55

2.1.3 Parallel Programming Approaches 56

2.2 Processes, Tasks, and Threads 59

2.2.1	Definitions of an Abstract Process	59
2.2.2	Execution Mode	62
2.2.3	Address Space	63
2.2.4	Process Context	65
2.2.5	Process Descriptor	66
2.2.6	Process Control	67
2.2.7	Variations of Process	70
2.3	Parallelism Issues	71
2.3.1	Homogeneity in Processes	72
2.3.2	Static versus Dynamic Parallelism	74
2.3.3	Process Grouping	75
2.3.4	Allocation Issues	76
2.4	Interaction/Communication Issues	77
2.4.1	Interaction Operations	77
2.4.2	Interaction Modes	80
2.4.3	Interaction Patterns	82
2.4.4	Cooperative versus Competitive Interactions	84
2.5	Semantic Issues in Parallel Programs	85
2.5.1	Program Termination	85
2.5.2	Determinacy of Programs	86
2.6	Bibliographic Notes and Problems	87
Chapter 3	Performance Metrics and Benchmarks	91
3.1	System and Application Benchmarks	91
3.1.1	Micro Benchmarks	92
3.1.2	Parallel Computing Benchmarks	96
3.1.3	Business and TPC Benchmarks	98
3.1.4	SPEC Benchmark Family	100
3.2	Performance versus Cost	102
3.2.1	Execution Time and Throughput	103
3.2.2	Utilization and Cost-Effectiveness	104
3.3	Basic Performance Metrics	108
3.3.1	Workload and Speed Metrics	108
3.3.2	Caveats in Sequential Performance	111
3.4	Performance of Parallel Computers	113
3.4.1	Computational Characteristics	113
3.4.2	Parallelism and Interaction Overheads	115
3.4.3	Overhead Quantification	118
3.5	Performance of Parallel Programs	126
3.5.1	Performance Metrics	126
3.5.2	Available Parallelism in Benchmarks	131
3.6	Scalability and Speedup Analysis	134
3.6.1	Amdahl's Law: Fixed Problem Size	134
3.6.2	Gustafson's Law: Fixed Time	136
3.6.3	Sun and Ni's Law: Memory Bounding	139
3.6.4	Isoperformance Models	144
3.7	Bibliographic Notes and Problems	148

Part II Enabling Technologies 153

Chapter 4 Microprocessors as Building Blocks 155

4.1	System Development Trends	155
4.1.1	Advances in Hardware	156
4.1.2	Advances in Software	159
4.1.3	Advances in Applications	160
4.2	Principles of Processor Design	164
4.2.1	Basics of Instruction Pipeline	164
4.2.2	From CISC to RISC and Beyond	169
4.2.3	Architectural Enhancement Approaches	172
4.3	Microprocessor Architecture Families	174
4.3.1	Major Architecture Families	174
4.3.2	Superscalar versus Superpipelined Processors	175
4.3.3	Embedded Microprocessors	180
4.4	Case Studies of Microprocessors	182
4.4.1	Digital's Alpha 21164 Microprocessor	182
4.4.2	Intel Pentium Pro Processor	186
4.5	Post-RISC, Multimedia, and VLIW	191
4.5.1	Post-RISC Processor Features	191
4.5.2	Multimedia Extensions	195
4.5.3	The VLIW Architecture	199
4.6	The Future of Microprocessors	201
4.6.1	Hardware Trends and Physical Limits	201
4.6.2	Future Workloads and Challenges	203
4.6.3	Future Microprocessor Architectures	204
4.7	Bibliographic Notes and Problems	206

Chapter 5 Distributed Memory and Latency Tolerance 211

5.1	Hierarchical Memory Technology	211
5.1.1	Characteristics of Storage Devices	211
5.1.2	Memory Hierarchy Properties	214
5.1.3	Memory Capacity Planning	217
5.2	Cache Coherence Protocols	220
5.2.1	Cache Coherency Problem	220
5.2.2	Snoopy Coherency Protocols	222
5.2.3	The MESI Snoopy Protocol	224
5.3	Shared-Memory Consistency	228
5.3.1	Memory Event Ordering	228
5.3.2	Memory Consistency Models	231
5.3.3	Relaxed Memory Models	234
5.4	Distributed Cache/Memory Architecture	237
5.4.1	NORMA, NUMA, COMA, and DSM Models	237
5.4.2	Directory-Based Coherency Protocol	243
5.4.3	The Stanford Dash Multiprocessor	245
5.4.4	Directory-Based Protocol in Dash	248

5.5 Latency Tolerance Techniques	250
5.5.1 Latency Avoidance, Reduction, and Hiding	250
5.5.2 Distributed Coherent Caches	253
5.5.3 Data Prefetching Strategies	255
5.5.4 Effects of Relaxed Memory Consistency	257
5.6 Multithreaded Latency Hiding	257
5.6.1 Multithreaded Processor Model	258
5.6.2 Context-Switching Policies	260
5.6.3 Combining Latency Hiding Mechanisms	265
5.7 Bibliographic Notes and Problems	266
Chapter 6 System Interconnects and Gigabit Networks	273
6.1 Basics of Interconnection Network	273
6.1.1 Interconnection Environments	273
6.1.2 Network Components	276
6.1.3 Network Characteristics	277
6.1.4 Network Performance Metrics	280
6.2 Network Topologies and Properties	281
6.2.1 Topological and Functional Properties	281
6.2.2 Routing Schemes and Functions	283
6.2.3 Networking Topologies	286
6.3 Buses, Crossbar, and Multistage Switches	294
6.3.1 Multiprocessor Buses	294
6.3.2 Crossbar Switches	298
6.3.3 Multistage Interconnection Networks	301
6.3.4 Comparison of Switched Interconnects	305
6.4 Gigabit Network Technologies	307
6.4.1 Fiber Channel and FDDI Rings	307
6.4.2 Fast Ethernet and Gigabit Ethernet	310
6.4.3 Myrinet for SAN/LAN Construction	313
6.4.4 HiPPI and SuperHiPPI	314
6.5 ATM Switches and Networks	318
6.5.1 ATM Technology	318
6.5.2 ATM Network Interfaces	320
6.5.3 Four Layers of ATM Architecture	321
6.5.4 ATM Internetwork Connectivity	324
6.6 Scalable Coherence Interface	326
6.6.1 SCI Interconnects	327
6.6.2 Implementation Issues	329
6.6.3 SCI Coherence Protocol	332
6.7 Comparison of Network Technologies	334
6.7.1 Standard Networks and Perspectives	334
6.7.2 Network Performance and Applications	335
6.8 Bibliographic Notes and Problems	337
Chapter 7 Threading, Synchronization, and Communication	343
7.1 Software Multithreading	343

7.1.1 The Thread Concept	344
7.1.2 Threads Management	346
7.1.3 Thread Synchronization	348
7.2 Synchronization Mechanisms	349
7.2.1 Atomicity versus Mutual Exclusion	349
7.2.2 High-Level Synchronization Constructs	355
7.2.3 Low-Level Synchronization Primitives	360
7.2.4 Fast Locking Mechanisms	364
7.3 The TCP/IP Communication Protocol Suite	366
7.3.1 Features of The TCP/IP Suite	367
7.3.2 UDP, TCP, and IP	371
7.3.3 The Sockets Interface	375
7.4 Fast and Efficient Communication	376
7.4.1 Key Problems in Communication	377
7.4.2 The Log P Communication Model	384
7.4.3 Low-Level Communications Support	386
7.4.4 Communication Algorithms	396
7.5 Bibliographic Notes and Problems	398

Part III Systems Architecture 403

Chapter 8 Symmetric and CC-NUMA Multiprocessors 407

8.1 SMP and CC-NUMA Technology	407
8.1.1 Multiprocessor Architecture	407
8.1.2 Commercial SMP Servers	412
8.1.3 The Intel SHV Server Board	413
8.2 Sun Ultra Enterprise 10000 System	416
8.2.1 The Ultra E-10000 Architecture	416
8.2.2 System Board Architecture	418
8.2.3 Scalability and Availability Support	418
8.2.4 Dynamic Domains and Performance	420
8.3 HP/Convex Exemplar X-Class	421
8.3.1 The Exemplar X System Architecture	421
8.3.2 Exemplar Software Environment	424
8.4 The Sequent NUMA-Q 2000	425
8.4.1 The NUMA-Q 2000 Architecture	426
8.4.2 Software Environment of NUMA-Q	430
8.4.3 Performance of the NUMA-Q	431
8.5 The SGI/Cray Origin 2000 Superserver	434
8.5.1 Design Goals of Origin 2000 Series	434
8.5.2 The Origin 2000 Architecture	435
8.5.3 The Cellular IRIX Environment	443
8.5.4 Performance of the Origin 2000	447
8.6 Comparison of CC-NUMA Architectures	447
8.7 Bibliographic Notes and Problems	451

Chapter 9 Support of Clustering and Availability 453

- 9.1 Challenges in Clustering 453**
 - 9.1.1 Classification of Clusters 453
 - 9.1.2 Cluster Architectures 456
 - 9.1.3 Cluster Design Issues 457
- 9.2 Availability Support for Clustering 459**
 - 9.2.1 The Availability Concept 460
 - 9.2.2 Availability Techniques 463
 - 9.2.3 Checkpointing and Failure Recovery 468
- 9.3 Support for Single System Image 473**
 - 9.3.1 Single System Image Layers 473
 - 9.3.2 Single Entry and Single File Hierarchy 475
 - 9.3.3 Single I/O, Networking, and Memory Space 479
- 9.4 Single System Image in Solaris MC 482**
 - 9.4.1 Global File System 482
 - 9.4.2 Global Process Management 484
 - 9.4.3 Single I/O System Image 485
- 9.5 Job Management in Clusters 486**
 - 9.5.1 Job Management System 486
 - 9.5.2 Survey of Job Management Systems 492
 - 9.5.3 Load-Sharing Facility (LSF) 494
- 9.6 Bibliographic Notes and Problems 501**

Chapter 10 Clusters of Servers and Workstations 505

- 10.1 Cluster Products and Research Projects 505**
 - 10.1.1 Supporting Trend of Cluster Products 506
 - 10.1.2 Cluster of SMP Servers 508
 - 10.1.3 Cluster Research Projects 509
- 10.2 Microsoft Wolfpack for NT Clusters 511**
 - 10.2.1 Microsoft Wolfpack Configurations 512
 - 10.2.2 Hot Standby Multiserver Clusters 513
 - 10.2.3 Active Availability Clusters 514
 - 10.2.4 Fault-Tolerant Multiserver Cluster 516
- 10.3 The IBM SP System 518**
 - 10.3.1 Design Goals and Strategies 518
 - 10.3.2 The SP2 System Architecture 521
 - 10.3.3 I/O and Internetworking 523
 - 10.3.4 The SP System Software 526
 - 10.3.5 The SP2 and Beyond 530
- 10.4 The Digital TruCluster 531**
 - 10.4.1 The TruCluster Architecture 531
 - 10.4.2 The Memory Channel Interconnect 534
 - 10.4.3 Programming the TruCluster 537
 - 10.4.4 The TruCluster System Software 540
- 10.5 The Berkeley NOW Project 541**
 - 10.5.1 Active Messages for Fast Communication 541

- 10.5.2 GLUnix for Global Resource Management 547
- 10.5.3 The xFS Serverless Network File System 549
- 10.6 TreadMarks: A Software-Implemented DSM Cluster 556**
 - 10.6.1 Boundary Conditions 556
 - 10.6.2 User Interface for DSM 557
 - 10.6.3 Implementation Issues 559
- 10.7 Bibliographic Notes and Problems 561**

Chapter 11 MPP Architecture and Performance 565

- 11.1 An Overview of MPP Technology 565**
 - 11.1.1 MPP Characteristics and Issues 565
 - 11.1.2 MPP Systems – An Overview 569
- 11.2 The Cray T3E System 570**
 - 11.2.1 The System Architecture of T3E 571
 - 11.2.2 The System Software in T3E 573
- 11.3 New Generation of ASCI/MPPs 574**
 - 11.3.1 ASCI Scalable Design Strategy 574
 - 11.3.2 Hardware and Software Requirements 576
 - 11.3.3 Contracted ASCI/MPP Platforms 577
- 11.4 Intel/Sandia ASCI Option Red 579**
 - 11.4.1 The Option Red Architecture 579
 - 11.4.2 Option Red System Software 582
- 11.5 Parallel NAS Benchmark Results 584**
 - 11.5.1 The NAS Parallel Benchmarks 585
 - 11.5.2 Superstep Structure and Granularity 586
 - 11.5.3 Memory, I/O, and Communications 587
- 11.6 MPI and STAP Benchmark Results 590**
 - 11.6.1 MPI Performance Measurements 590
 - 11.6.2 MPI Latency and Aggregate Bandwidth 592
 - 11.6.3 STAP Benchmark Evaluation of MPPs 594
 - 11.6.4 MPP Architectural Implications 600
- 11.7 Bibliographic Notes and Problems 603**

Part IV Parallel Programming 607**Chapter 12 Parallel Paradigms and Programming Models 609**

- 12.1 Paradigms and Programmability 609**
 - 12.1.1 Algorithmic Paradigms 609
 - 12.1.2 Programmability Issues 612
 - 12.1.3 Parallel Programming Examples 614
- 12.2 Parallel Programming Models 617**
 - 12.2.1 Implicit Parallelism 617
 - 12.2.2 Explicit Parallel Models 621
 - 12.2.3 Comparison of Four Models 624
 - 12.2.4 Other Parallel Programming Models 627

12.3 Shared-Memory Programming	629
12.3.1 The ANSI X3H5 Shared-Memory Model	629
12.3.2 The POSIX Threads (Pthreads) Model	634
12.3.3 The OpenMP Standard	636
12.3.4 The SGI Power C Model	640
12.3.5 C//: A Structured Parallel C Language	643
12.4 Bibliographic Notes and Problems	649

Chapter 13 Message-Passing Programming 653

13.1 The Message-Passing Paradigm	653
13.1.1 Message-Passing Libraries	653
13.1.2 Message-Passing Modes	655
13.2 Message-Passing Interface (MPI)	658
13.2.1 MPI Messages	661
13.2.2 Message Envelope in MPI	668
13.2.3 Point-to-Point Communications	674
13.2.4 Collective MPI Communications	678
13.2.5 The MPI-2 Extensions	682
13.3 Parallel Virtual Machine (PVM)	686
13.3.1 Virtual Machine Construction	687
13.3.2 Process Management in PVM	689
13.3.3 Communication with PVM	693
13.4 Bibliographic Notes and Problems	699

Chapter 14 Data-Parallel Programming 705

14.1 The Data-Parallel Model	705
14.2 The Fortran 90 Approach	706
14.2.1 Parallel Array Operations	706
14.2.2 Intrinsic Functions in Fortran 90	708
14.3 High-Performance Fortran	711
14.3.1 Support for Data Parallelism	712
14.3.2 Data Mapping in HPF	715
14.3.3 Summary of Fortran 90 and HPF	721
14.4 Other Data-Parallel Approaches	725
14.4.1 Fortran 95 and Fortran 2001	725
14.4.2 The pC++ and Nesl Approaches	728
14.5 Bibliographic Notes and Problems	733

Bibliography 737

Web Resources List 765

Subject Index 787

Author Index 799