Computer Architecture
A Quantitative Approach

Third Edition

John L. Hennessy
Stanford University

David A. Patterson
University of California at Berkeley

With Contributions by

David Goldberg
Xerox Palo Alto Research Center

Krste Asanovic
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

MORGAN KAUFMANN PUBLISHERS

AN IMPRINT OF ELSEVIER SCIENCE

AMSTERDAM BOSTON LONDON NEW YORK
OXFORD PARIS SAN DIEGO SAN FRANCISCO
SINGAPORE SYDNEY TOKYO

Contents

Foreword vii

Preface xvii

Acknowledgments xxv

Chapter 1 **Fundamentals of Computer Design**

1.1 Introduction 2
1.2 The Changing Face of Computing and the Task of the Computer Designer 4
1.3 Technology Trends 11
1.4 Cost, Price, and Their Trends 14
1.5 Measuring and Reporting Performance 24
1.6 Quantitative Principles of Computer Design 39
1.7 Putting It All Together: Performance and Price-Performance 48
1.8 Another View: Power Consumption and Efficiency as the Metric 56
1.9 Fallacies and Pitfalls 57
1.10 Concluding Remarks 65
1.11 Historical Perspective and References 67
Exercises 74

Chapter 2 **Instruction Set Principles and Examples**

2.1 Introduction 90
2.2 Classifying Instruction Set Architectures 92
2.3 Memory Addressing 95
2.4 Addressing Modes for Signal Processing 101
2.5 Type and Size of Operands 104

2.6	Operands for Media and Signal Processing	105
2.7	Operations in the Instruction Set	108
2.8	Operations for Media and Signal Processing	109
2.9	Instructions for Control Flow	111
2.10	Encoding an Instruction Set	117
2.11	Crosscutting Issues: The Role of Compilers	120
2.12	Putting It All Together: The MIPS Architecture	129
2.13	Another View: The Trimedia TM32 CPU	136
2.14	Fallacies and Pitfalls	142
2.15	Concluding Remarks	147
2.16	Historical Perspective and References	148
	Exercises	161

Chapter 3 **Instruction-Level Parallelism and Its Dynamic Exploitation**

3.1	Instruction-Level Parallelism: Concepts and Challenges	172
3.2	Overcoming Data Hazards with Dynamic Scheduling	181
3.3	Dynamic Scheduling: Examples and the Algorithm	189
3.4	Reducing Branch Costs with Dynamic Hardware Prediction	196
3.5	High-Performance Instruction Delivery	209
3.6	Taking Advantage of More ILP with Multiple Issue	215
3.7	Hardware-Based Speculation	224
3.8	Studies of the Limitations of ILP	240
3.9	Limitations on ILP for Realizable Processors	253
3.10	Putting It All Together: The P6 Microarchitecture	259
3.11	Another View: Thread-Level Parallelism	272
3.12	Crosscutting Issues: Using an ILP Data Path to Exploit TLP	273
3.13	Fallacies and Pitfalls	273
3.14	Concluding Remarks	276
3.15	Historical Perspective and References	280
	Exercises	288

Chapter 4 **Exploiting Instruction-Level Parallelism with Software Approaches**

4.1	Basic Compiler Techniques for Exposing ILP	304
4.2	Static Branch Prediction	313
4.3	Static Multiple Issue: The VLIW Approach	315
4.4	Advanced Compiler Support for Exposing and Exploiting ILP	319
4.5	Hardware Support for Exposing More Parallelism at Compile Time	340

4.6	Crosscutting Issues: Hardware versus Software Speculation Mechanisms	350
4.7	Putting It All Together: The Intel IA-64 Architecture and Itanium Processor	351
4.8	Another View: ILP in the Embedded and Mobile Markets	363
4.9	Fallacies and Pitfalls	370
4.10	Concluding Remarks	372
4.11	Historical Perspective and References	373
	Exercises	378

Chapter 5 **Memory Hierarchy Design**

5.1	Introduction	390
5.2	Review of the ABCs of Caches	392
5.3	Cache Performance	406
5.4	Reducing Cache Miss Penalty	413
5.5	Reducing Miss Rate	423
5.6	Reducing Cache Miss Penalty or Miss Rate via Parallelism	435
5.7	Reducing Hit Time	443
5.8	Main Memory and Organizations for Improving Performance	448
5.9	Memory Technology	454
5.10	Virtual Memory	460
5.11	Protection and Examples of Virtual Memory	469
5.12	Crosscutting Issues: The Design of Memory Hierarchies	478
5.13	Putting It All Together: Alpha 21264 Memory Hierarchy	482
5.14	Another View: The Emotion Engine of the Sony Playstation 2	490
5.15	Another View: The Sun Fire 6800 Server	494
5.16	Fallacies and Pitfalls	498
5.17	Concluding Remarks	504
5.18	Historical Perspective and References	504
	Exercises	513

Chapter 6 **Multiprocessors and Thread-Level Parallelism**

6.1	Introduction	528
6.2	Characteristics of Application Domains	540
6.3	Symmetric Shared-Memory Architectures	549
6.4	Performance of Symmetric Shared-Memory Multiprocessors	560
6.5	Distributed Shared-Memory Architectures	576
6.6	Performance of Distributed Shared-Memory Multiprocessors	584

6.7	Synchronization	590
6.8	Models of Memory Consistency: An Introduction	605
6.9	Multithreading: Exploiting Thread-Level Parallelism within a Processor	608
6.10	Crosscutting Issues	615
6.11	Putting It All Together: Sun's Wildfire Prototype	622
6.12	Another View: Multithreading in a Commercial Server	635
6.13	Another View: Embedded Multiprocessors	636
6.14	Fallacies and Pitfalls	637
6.15	Concluding Remarks	643
6.16	Historical Perspective and References	649
	Exercises	665

Chapter 7 **Storage Systems**

7.1	Introduction	678
7.2	Types of Storage Devices	679
7.3	Buses—Connecting I/O Devices to CPU/Memory	692
7.4	Reliability, Availability, and Dependability	702
7.5	RAID: Redundant Arrays of Inexpensive Disks	705
7.6	Errors and Failures in Real Systems	710
7.7	I/O Performance Measures	716
7.8	A Little Queuing Theory	720
7.9	Benchmarks of Storage Performance and Availability	731
7.10	Crosscutting Issues	737
7.11	Designing an I/O System in Five Easy Pieces	741
7.12	Putting It All Together: EMC Symmetrix and Celerra	754
7.13	Another View: Sanyo VPC-SX500 Digital Camera	760
7.14	Fallacies and Pitfalls	763
7.15	Concluding Remarks	769
7.16	Historical Perspective and References	770
	Exercises	778

Chapter 8 **Interconnection Networks and Clusters**

8.1	Introduction	788
8.2	A Simple Network	793
8.3	Interconnection Network Media	802
8.4	Connecting More Than Two Computers	805
8.5	Network Topology	814
8.6	Practical Issues for Commercial Interconnection Networks	821

8.7	Examples of Interconnection Networks	825
8.8	Internetworking	830
8.9	Crosscutting Issues for Interconnection Networks	834
8.10	Clusters	838
8.11	Designing a Cluster	843
8.12	Putting It All Together: The Google Cluster of PCs	855
8.13	Another View: Inside a Cell Phone	862
8.14	Fallacies and Pitfalls	867
8.15	Concluding Remarks	870
8.16	Historical Perspective and References	871
	Exercises	877

Appendix A **Pipelining: Basic and Intermediate Concepts**

A.1	Introduction	A-2
A.2	The Major Hurdle of Pipelining—Pipeline Hazards	A-11
A.3	How Is Pipelining Implemented?	A-26
A.4	What Makes Pipelining Hard to Implement?	A-37
A.5	Extending the MIPS Pipeline to Handle Multicycle Operations	A-47
A.6	Putting It All Together: The MIPS R4000 Pipeline	A-57
A.7	Another View: The MIPS R4300 Pipeline	A-66
A.8	Crosscutting Issues	A-67
A.9	Fallacies and Pitfalls	A-77
A.10	Concluding Remarks	A-78
A.11	Historical Perspective and References	A-78
	Exercises	A-81

Appendix B **Solutions to Selected Exercises**

	Introduction	B-2
B.1	Chapter 1 Solutions	B-2
B.2	Chapter 2 Solutions	B-7
B.3	Chapter 3 Solutions	B-11
B.4	Chapter 4 Solutions	B-16
B.5	Chapter 5 Solutions	B-21
B.6	Chapter 6 Solutions	B-25
B.7	Chapter 7 Solutions	B-29
B.8	Chapter 8 Solutions	B-30
B.9	Appendix A Solutions	B-35

Online Appendices *(www.mkp.com/CA3/)*

Appendix C **A Survey of RISC Architectures for Desktop, Server, and Embedded Computers**

Appendix D **An Alternative to RISC: The Intel 80x86**

Appendix E **Another Alternative to RISC: The VAX Architecture**

Appendix F **The IBM 360/370 Architecture for Mainframe Computers**

Appendix G **Vector Processors**
Revised by Krste Asanovic

Appendix H **Computer Arithmetic**
by David Goldberg

Appendix I **Implementing Coherence Protocols**

References R-1

Index I-1