

Contents

	Acknowledgments	ix
Chapter 1	Fundamentals of Quantitative Design and Analysis	
	1.1 Introduction	2
	1.2 Classes of Computers	6
	1.3 Defining Computer Architecture	12
	1.4 Trends in Technology	20
	1.5 Trends in Power and Energy in Integrated Circuits	24
	1.6 Trends in Cost	32
	1.7 Dependability	42
	1.8 Security	46
	1.9 Measuring, Reporting, and Summarizing Performance	47
	1.10 Quantitative Principles of Computer Design	57
	1.11 Putting It All Together: Performance, Price, and Power	66
	1.12 Fallacies and Pitfalls	69
	1.13 Concluding Remarks	76
	1.14 Historical Perspectives and References	79
	Case Studies and Exercises by Gregory D. Peterson	79
	Reference	91
Chapter 2	Memory Hierarchy Design	
	2.1 Introduction	94
	2.2 Memory Technology and Optimizations	101
	2.3 Ten Advanced Performance Optimizations for Memory Hierarchies	111
	2.4 Virtual Memory and Protection	135
	2.5 Cross-Cutting Issues: The Design of Memory Hierarchies	139
	2.6 Putting It All Together: Memory Hierarchies in the ARM Cortex-A53 and Intel Core i9 12900	141
	2.7 Fallacies and Pitfalls	152
	2.8 Concluding Remarks: Looking Ahead	156
	2.9 Historical Perspectives and References	158
	Case Studies and Exercises by Rajeev Balasubramonian, Norman P. Jouppi, Naveen Muralimanohar, and Sheng Li	158

Chapter 3	Instruction-Level Parallelism and Its Exploitation	
	3.1 Introduction	176
	3.2 Basic Compiler Techniques for Exposing ILP	184
	3.3 Key ILP Concepts in Modern Superscalar Processors	190
	3.4 Reducing Branch Costs with Advanced Branch Prediction	198
	3.5 Overcoming Name Dependences with Register Renaming	215
	3.6 Overcoming Data Hazards with Dynamic Scheduling	221
	3.7 Overcoming Memory Dependences with Dynamic Disambiguation	230
	3.8 Advanced Issues in Modern Superscalar Processors	234
	3.9 Exploiting ILP Using Multiple Issue and Static Scheduling	240
	3.10 Cross-Cutting Issues	244
	3.11 Multithreading: Exploiting Thread-Level Parallelism to Improve Single Core Throughput	246
	3.12 Microarchitecture Side-Channel Attacks	253
	3.13 Putting It All Together: The Arm Cortex-A53 and the Intel Golden Cove Processors	256
	3.14 Fallacies and Pitfalls	266
	3.15 Concluding Remarks	273
	3.16 Historical Perspective and References	275
	Case Studies and Exercises by Jason D. Bakos	275
	Additional References	284
Chapter 4	Data-Level Parallelism in Vector, SIMD, and GPU Architectures	
	4.1 Introduction	288
	4.2 Vector Architecture	289
	4.3 SIMD Instruction Set Extensions for Multimedia	309
	4.4 Graphics Processing Units	318
	4.5 Detecting and Enhancing Loop-Level Parallelism	347
	4.6 Cross-Cutting Issues	356
	4.7 Putting It All Together: Embedded Versus Server GPUs and Tesla Versus Core i7	357
	4.8 Fallacies and Pitfalls	366
	4.9 Concluding Remarks	368
	4.10 Historical Perspective and References	369
	Case Study and Exercises by Jason D. Bakos	369
	New References	376
Chapter 5	Thread-Level Parallelism	
	5.1 Introduction	380
	5.2 Multiprocessor Cache Coherence	388

5.3	Maintaining Cache Coherence with Snooping	392
5.4	Maintaining Cache Coherence with Directories	410
5.5	Synchronization: The Basics	423
5.6	Models of Memory Consistency: An Introduction	428
5.7	Cross-Cutting Issues	433
5.8	Putting It All Together: Multicore Processors and Their Performance	436
5.9	Fallacies and Pitfalls	443
5.10	The Future of Multicore Scaling	449
5.11	Concluding Remarks	452
5.12	Historical Perspectives and References	453
	Case Studies and Exercises by Amr Zaky	453
Chapter 6	Warehouse-Scale Architectures for Utility Computing	
6.1	Introduction	470
6.2	Cloud Computing: The Return of Utility Computing	475
6.3	Hardware Support for Virtualization	484
6.4	Computer Architecture of Warehouse-Scale Computers	496
6.5	The Architecture of High-Performance I/O Devices	517
6.6	WSC Power Distribution and Cooling	524
6.7	The Cost and Efficiency of Warehouse-scale Computing	531
6.8	Putting It All Together: Custom Silicon in the AWS Cloud	540
6.9	Fallacies and Pitfalls	553
6.9	Concluding Remarks	557
6.10	Historical Perspectives and References	558
	Case Studies and Exercises by Parthasarathy Ranganathan	558
	References	581
Chapter 7	Domain-Specific Architectures	
7.1	Introduction	586
7.2	Guidelines for DSAs	588
7.3	Example Domain: Deep Neural Networks	590
7.4	Google's Tensor Processing Unit v4 and v4 lite, Data Center DNN Accelerators	602
7.5	The NVIDIA A100 and T4 GPUs, Graphics, and DNN Accelerators for the Data Center	609
7.6	Graphcore IPU Bow, a Data Center Accelerator for Training	616
7.7	The Samsung Neural Processing Unit (NPU), a Smartphone Inference Accelerator	618
7.8	Cross-Cutting Issues	621
7.9	Putting It All Together: Comparing DNN Accelerators	625
7.10	Fallacies and Pitfalls	629
7.11	Concluding Remarks	632

7.12	Historical Perspectives and References	633
	Case Studies and Exercises by Cliff Young	633
	References	648
Appendix A	Instruction Set Principles	
A.1	Introduction	A-2
A.2	Classifying Instruction Set Architectures	A-3
A.3	Memory Addressing	A-7
A.4	Type and Size of Operands	A-13
A.5	Operations in the Instruction Set	A-15
A.6	Instructions for Control Flow	A-16
A.7	Encoding an Instruction Set	A-21
A.8	Cross-Cutting Issues: The Role of Compilers	A-24
A.9	Putting It All Together: The RISC-V Architecture	A-33
A.10	Fallacies and Pitfalls	A-42
A.11	Concluding Remarks	A-46
A.12	Historical Perspective and References	A-47
	Exercises by Gregory D. Peterson	A-47
Appendix B	Review of Memory Hierarchy	
B.1	Introduction	B-2
B.2	Cache Performance	B-15
B.3	Six Basic Cache Optimizations	B-22
B.4	Virtual Memory	B-41
B.5	Protection and Examples of Virtual Memory	B-50
B.6	Fallacies and Pitfalls	B-58
B.7	Concluding Remarks	B-60
B.8	Historical Perspective and References	B-60
	Exercises by Amr Zaky	B-61
Appendix C	Pipelining: Basic and Intermediate Concepts	
C.1	Introduction	C-2
C.2	The Major Hurdle of Pipelining—Pipeline Hazards	C-10
C.3	How Is Pipelining Implemented?	C-27
C.4	What Makes Pipelining Hard to Implement?	C-39
C.5	Extending the RISC-V Integer Pipeline to Handle Multicycle Operations	C-46
C.6	Putting It All Together: The MIPS R4000 Pipeline	C-56
C.7	Cross-Cutting Issues	C-66
C.8	Fallacies and Pitfalls	C-71
C.9	Concluding Remarks	C-72
C.10	Historical Perspective and References	C-72
	Updated Exercises by Diana Franklin	C-72

Appendix D	Storage Systems	
Appendix E	Embedded Systems <i>By Thomas M. Conte</i>	
Appendix F	Interconnection Networks <i>by Timothy M. Pinkston and José Duato</i>	
Appendix G	Vector Processors in More Depth <i>by Krste Asanovic</i>	
Appendix H	Hardware and Software for VLIW and EPIC	
Appendix I	Large-Scale Multiprocessors and Scientific Applications	
Appendix J	Computer Arithmetic <i>by David Goldberg</i>	
Appendix K	Survey of Instruction Set Architectures	
Appendix L	Advanced Concepts on Address Translation <i>by Abhishek Bhattacharjee</i>	
Appendix M	Historical Perspectives and References	
	References	R-1
	Index	I-1

For additional information on the topics covered in the book, visit the companion site: <https://www.elsevier.com/books-and-journals/book-companion/9780443154065>

For additional materials for qualified instructors, please visit the instructor site: <https://educate.elsevier.com/9780443154065>