



Table of Contents

Table of Contents	vii
Preface	xv
Acknowledgments	xix
1 Processor Design	1
1.1 The Evolution of Microprocessors	2
1.2 Instruction Set Processor Design	4
1.2.1 Digital Systems Design	4
1.2.2 Architecture, Implementation, and Realization	5
1.2.3 Instruction Set Architecture	6
1.2.4 Dynamic-Static Interface	8
1.3 Principles of Processor Performance	10
1.3.1 Processor Performance Equation	10
1.3.2 Processor Performance Optimizations	11
1.3.3 Performance Evaluation Method	13
1.4 Instruction-Level Parallel Processing	16
1.4.1 From Scalar to Superscalar	16
1.4.2 Limits of Instruction-Level Parallelism	24
1.4.3 Machines for Instruction-Level Parallelism	27
1.5 Summary	32
References	33
Homework Problems	35
2 Pipelined Processors	39
2.1 Pipelining Fundamentals	40
2.1.1 Pipelined Design	40
2.1.2 Arithmetic Pipeline Example	44
2.1.3 Pipelining Idealism	48
2.1.4 Instruction Pipelining	51
2.2 Pipelined Processor Design	54
2.2.1 Balancing Pipeline Stages	55
2.2.2 Unifying Instruction Types	61

2.2.3	Minimizing Pipeline Stalls	71
2.2.4	Commercial Pipelined Processors	87
2.3	Deeply Pipelined Processors	94
2.4	Summary	97
	References	98
	Homework Problems	98
3	Memory and I/O Systems	105
3.1	Introduction	105
3.2	Computer System Overview	106
3.3	Key Concepts: Latency and Bandwidth	107
3.4	Memory Hierarchy	110
3.4.1	Components of a Modern Memory Hierarchy	111
3.4.2	Temporal and Spatial Locality	113
3.4.3	Caching and Cache Memories	115
3.4.4	Main Memory	127
3.5	Virtual Memory Systems	136
3.5.1	Demand Paging	138
3.5.2	Memory Protection	141
3.5.3	Page Table Architectures	142
3.6	Memory Hierarchy Implementation	145
3.7	Input/Output Systems	153
3.7.1	Types of I/O Devices	154
3.7.2	Computer System Busses	161
3.7.3	Communication with I/O Devices	165
3.7.4	Interaction of I/O Devices and Memory Hierarchy	168
3.8	Summary	170
	References	171
	Homework Problems	172
4	Superscalar Organization	177
4.1	Limitations of Scalar Pipelines	178
4.1.1	Upper Bound on Scalar Pipeline Throughput	178
4.1.2	Inefficient Unification into a Single Pipeline	179
4.1.3	Performance Lost Due to a Rigid Pipeline	179
4.2	From Scalar to Superscalar Pipelines	181
4.2.1	Parallel Pipelines	181
4.2.2	Diversified Pipelines	184
4.2.3	Dynamic Pipelines	186
4.3	Superscalar Pipeline Overview	190
4.3.1	Instruction Fetching	191
4.3.2	Instruction Decoding	195

4.3.3	Instruction Dispatching	199
4.3.4	Instruction Execution	203
4.3.5	Instruction Completion and Retiring	206
4.4	Summary	209
	References	210
	Homework Problems	211
5	Superscalar Techniques	217
5.1	Instruction Flow Techniques	218
5.1.1	Program Control Flow and Control Dependences	218
5.1.2	Performance Degradation Due to Branches	219
5.1.3	Branch Prediction Techniques	223
5.1.4	Branch Misprediction Recovery	228
5.1.5	Advanced Branch Prediction Techniques	231
5.1.6	Other Instruction Flow Techniques	236
5.2	Register Data Flow Techniques	237
5.2.1	Register Reuse and False Data Dependences	237
5.2.2	Register Renaming Techniques	239
5.2.3	True Data Dependences and the Data Flow Limit	244
5.2.4	The Classic Tomasulo Algorithm	246
5.2.5	Dynamic Execution Core	254
5.2.6	Reservation Stations and Reorder Buffer	256
5.2.7	Dynamic Instruction Scheduler	260
5.2.8	Other Register Data Flow Techniques	261
5.3	Memory Data Flow Techniques	262
5.3.1	Memory Accessing Instructions	263
5.3.2	Ordering of Memory Accesses	266
5.3.3	Load Bypassing and Load Forwarding	267
5.3.4	Other Memory Data Flow Techniques	273
5.4	Summary	279
	References	279
	Homework Problems	281
6	The PowerPC 620	301
6.1	Introduction	302
6.2	Experimental Framework	305
6.3	Instruction Fetching	307
6.3.1	Branch Prediction	307
6.3.2	Fetching and Speculation	309

6.4	Instruction Dispatching	311
6.4.1	Instruction Buffer	311
6.4.2	Dispatch Stalls	311
6.4.3	Dispatch Effectiveness	313
6.5	Instruction Execution	316
6.5.1	Issue Stalls	316
6.5.2	Execution Parallelism	317
6.5.3	Execution Latency	317
6.6	Instruction Completion	318
6.6.1	Completion Parallelism	318
6.6.2	Cache Effects	318
6.7	Conclusions and Observations	320
6.8	Bridging to the IBM POWER3 and POWER4	322
6.9	Summary	324
	References	325
	Homework Problems	325
7	Intel's P6 Microarchitecture	329
7.1	Introduction	330
7.1.1	Basics of the P6 Microarchitecture	332
7.2	Pipelining	334
7.2.1	In-Order Front-End Pipeline	334
7.2.2	Out-of-Order Core Pipeline	336
7.2.3	Retirement Pipeline	337
7.3	The In-Order Front End	338
7.3.1	Instruction Cache and ITLB	338
7.3.2	Branch Prediction	341
7.3.3	Instruction Decoder	343
7.3.4	Register Alias Table	346
7.3.5	Allocator	353
7.4	The Out-of-Order Core	355
7.4.1	Reservation Station	355
7.5	Retirement	357
7.5.1	The Reorder Buffer	357
7.6	Memory Subsystem	361
7.6.1	Memory Access Ordering	362
7.6.2	Load Memory Operations	363
7.6.3	Basic Store Memory Operations	363
7.6.4	Deferring Memory Operations	363
7.6.5	Page Faults	364
7.7	Summary	364

7.8	Acknowledgments	365
	References	365
	Homework Problems	365
8	Survey of Superscalar Processors	369
8.1	Development of Superscalar Processors	369
8.1.1	Early Advances in Uniprocessor Parallelism: The IBM Stretch	369
8.1.2	First Superscalar Design: The IBM Advanced Computer System	372
8.1.3	Instruction-Level Parallelism Studies	377
8.1.4	By-Products of DAE: The First Multiple-Decoding Implementations	378
8.1.5	IBM Cheetah, Panther, and America	380
8.1.6	Decoupled Microarchitectures	382
8.1.7	Other Efforts in the 1980s	382
8.1.8	Wide Acceptance of Superscalar	382
8.2	A Classification of Recent Designs	384
8.2.1	RISC and CISC Retrofits	384
8.2.2	Speed Demons: Emphasis on Clock Cycle Time	386
8.2.3	Brainiacs: Emphasis on IPC	386
8.3	Processor Descriptions	387
8.3.1	Compaq / DEC Alpha	387
8.3.2	Hewlett-Packard PA-RISC Version 1.0	392
8.3.3	Hewlett-Packard PA-RISC Version 2.0	395
8.3.4	IBM POWER	397
8.3.5	Intel i960	402
8.3.6	Intel IA32—Native Approaches	405
8.3.7	Intel IA32—Decoupled Approaches	409
8.3.8	x86-64	417
8.3.9	MIPS	417
8.3.10	Motorola	422
8.3.11	PowerPC—32-bit Architecture	424
8.3.12	PowerPC—64-bit Architecture	429
8.3.13	PowerPC-AS	431
8.3.14	SPARC Version 8	432
8.3.15	SPARC Version 9	435
8.4	Verification of Superscalar Processors	439
8.5	Acknowledgments	440
	References	440
	Homework Problems	449

9	Advanced Instruction Flow Techniques	453
9.1	Introduction	453
9.2	Static Branch Prediction Techniques	454
9.2.1	Single-Direction Prediction	455
9.2.2	Backwards Taken/Forwards Not-Taken	456
9.2.3	Ball/Larus Heuristics	456
9.2.4	Profiling	457
9.3	Dynamic Branch Prediction Techniques	458
9.3.1	Basic Algorithms	459
9.3.2	Interference-Reducing Predictors	472
9.3.3	Predicting with Alternative Contexts	482
9.4	Hybrid Branch Predictors	491
9.4.1	The Tournament Predictor	491
9.4.2	Static Predictor Selection	493
9.4.3	Branch Classification	494
9.4.4	The Multihybrid Predictor	495
9.4.5	Prediction Fusion	496
9.5	Other Instruction Flow Issues and Techniques	497
9.5.1	Target Prediction	497
9.5.2	Branch Confidence Prediction	501
9.5.3	High-Bandwidth Fetch Mechanisms	504
9.5.4	High-Frequency Fetch Mechanisms	509
9.6	Summary	512
	References	513
	Homework Problems	516
10	Advanced Register Data Flow Techniques	519
10.1	Introduction	519
10.2	Value Locality and Redundant Execution	523
10.2.1	Causes of Value Locality	523
10.2.2	Quantifying Value Locality	525
10.3	Exploiting Value Locality without Speculation	527
10.3.1	Memoization	527
10.3.2	Instruction Reuse	529
10.3.3	Basic Block and Trace Reuse	533
10.3.4	Data Flow Region Reuse	534
10.3.5	Concluding Remarks	535
10.4	Exploiting Value Locality with Speculation	535
10.4.1	The Weak Dependence Model	535
10.4.2	Value Prediction	536
10.4.3	The Value Prediction Unit	537

10.4.4	Speculative Execution Using Predicted Values	542
10.4.5	Performance of Value Prediction	551
10.4.6	Concluding Remarks	553
10.5	Summary	554
	References	555
	Homework Problems	556
11	Executing Multiple Threads	559
11.1	Introduction	559
11.2	Synchronizing Shared-Memory Threads	562
11.3	Introduction to Multiprocessor Systems	565
11.3.1	Fully Shared Memory, Unit Latency, and Lack of Contention	566
11.3.2	Instantaneous Propagation of Writes	567
11.3.3	Coherent Shared Memory	567
11.3.4	Implementing Cache Coherence	571
11.3.5	Multilevel Caches, Inclusion, and Virtual Memory	574
11.3.6	Memory Consistency	576
11.3.7	The Coherent Memory Interface	581
11.3.8	Concluding Remarks	583
11.4	Explicitly Multithreaded Processors	584
11.4.1	Chip Multiprocessors	584
11.4.2	Fine-Grained Multithreading	588
11.4.3	Coarse-Grained Multithreading	589
11.4.4	Simultaneous Multithreading	592
11.5	Implicitly Multithreaded Processors	600
11.5.1	Resolving Control Dependences	601
11.5.2	Resolving Register Data Dependences	605
11.5.3	Resolving Memory Data Dependences	607
11.5.4	Concluding Remarks	610
11.6	Executing the Same Thread	610
11.6.1	Fault Detection	611
11.6.2	Prefetching	613
11.6.3	Branch Resolution	614
11.6.4	Concluding Remarks	615
11.7	Summary	616
	References	616
	Homework Problems	618
	Index	000