

Contents

Preface	xix
Paths through Computer Architecture	xxix

PART I THE BEGINNING



1 Computer Systems Architecture	6
Where We're At	7
1.1 What is Computer Systems Architecture?	9
What is a Computer?	11
1.2 Architecture and Organization	14
1.2.1 Computer Systems and Technology	17
1.2.2 The Role of Computer Architecture in Computer Science	19
1.3 Development of Computers	21
1.3.1 Mechanical Computers	21
1.3.2 Electromechanical Computers	24
1.3.3 Early Electronic Computers	24
1.3.4 Minicomputers and the PC Revolution	25
1.3.5 Moore's Law and the March of Progress	25
1.3.6 The March of Memory Technology	26
1.3.7 Ubiquitous Computing	27
1.3.8 Multimedia Computers	28
1.4 The Stored Program Computer	28
1.4.1 The Problem	29
1.4.2 The Solution	29
1.4.3 Constructing an Algorithm	31
1.4.4 What Does a Computer Need to Solve a Problem?	32
1.4.5 The Memory	34
1.5 The Stored Program Concept	37
Two Address Instructions	39
One Address Instructions	40
Computer Categories	40
1.6 Overview of the Computer System	41
1.6.1 The Memory Hierarchy	41
1.6.2 The Bus	43
1.7 Modern Computing	45
Summary	46
Problems	46
2 Computer Arithmetic and Digital Logic	48
2.1 What is Data?	50
2.1.1 The Bit and Byte	50
2.1.2 Bit Patterns	51
Representing Information	51

2.2	Numbers	55
2.2.1	Positional Notation	56
2.3	Binary Arithmetic	57
2.4	Signed Integers	59
2.4.1	Sign and Magnitude Representation	60
2.4.2	Two's Complement Arithmetic	60
	Calculating Two's Complement Values	61
	Properties of Two's Complement Numbers	62
	Arithmetic Overflow	62
2.5	Introduction to Multiplication and Division	63
2.5.1	Shifting Operations	63
2.5.2	Unsigned Binary Multiplication	64
2.5.3	High-speed Multiplication	64
	Booth's Algorithm	66
2.5.4	Division	67
	Restoring Division	68
	Non-Restoring Division	70
2.6	Floating-Point Numbers	71
	Normalization of Floating-Point Numbers	72
	Biased Exponents	72
2.6.1	IEEE Floating-Point Numbers	72
	IEEE Floating-Point Format	73
	Characteristics of IEEE Floating-Point Numbers	75
2.7	Floating-Point Arithmetic	77
	Rounding and Truncation Errors	78
2.8	Floating-Point Arithmetic and the Programmer	79
2.8.1	Error Propagation in Floating-Point Arithmetic	81
2.8.2	Generating Mathematical Functions	81
	Using Functions to Generate New Functions	83
2.9	Computer Logic	84
2.9.1	Digital Systems and Gates	86
2.9.2	Gates	86
	Fundamental Gates	87
	The AND Gate	87
	The OR Gate	87
	The Inverter	88
	Derived Gates—the NOR (Not OR), NAND (Not AND), and Exclusive OR	89
2.9.3	Basic Circuits	91
	The Half Adder and Full Adder	93
	The Decoder	97
	The Multiplexer	97
	The Voting Circuit	98
	The Prioritizer	100
2.10	Sequential Circuits	101
2.10.1	Latches	102
	Clocked RS Flip-flops	104
	D Flip-flop	105
	The JK Flip-Flop	108
2.10.2	Registers	109
	Shift Register	110
	Left-Shift Register	111

2.10.3 Asynchronous Counters	113
Using a Counter to Create a Sequencer	114
2.10.4 Sequential Circuits	115
2.11 Buses and Tristate Gates	118
Registers, Buses, and Functional Units	120
Summary	122
Problems	123

PART II INSTRUCTION SET ARCHITECTURES



3 Architecture and Organization	130
3.1 Introduction to the Stored Program Machine	130
3.1.1 Extending the Processor: Dealing with Constants	136
3.1.2 Extending the Processor: Flow Control	139
Status Information	141
Example of a Branch Instruction	142
3.2 The Components of an ISA	146
3.2.1 Registers	146
General-Purpose Versus Special-Purpose Registers	147
3.2.2 Addressing Modes—an Overview	149
Memory and Register Addressing	151
3.2.3 Instruction Formats	151
3.2.4 Op-codes and Instructions	152
Two Address Machines	153
One Address Machines	153
Zero Address Machines	153
One-and-a-Half Address Machines	154
3.3 ARM Instruction Set Architecture	155
3.3.1 ARM's Register Set	156
3.3.2 ARM's Instruction Set	156
3.4 ARM Assembly Language	157
3.4.1 Structure of an ARM Program	158
3.4.2 The Assembler – Practical Considerations	161
3.4.3 Pseudoinstructions	164
3.5 ARM Data-processing Instructions	167
3.5.1 Arithmetic Instructions	167
Addition and Subtraction	167
Negation	168
Comparison	168
Multiplication	169
Division	170
3.5.2 Bitwise Logical Operations	170
3.5.3 Shift Operations	171
Arithmetic Shift	173
Rotate	173
Implementing a Shift Operation on the ARM	173
3.5.4 Instruction Encoding—An Insight Into the ARM's Architecture	175
3.6 ARM's Flow Control Instructions	176
3.6.1 Unconditional Branch	176
3.6.2 Conditional Branch	177

3.6.3	Compare and Test Instructions	178
3.6.4	Branching and Loop Constructs	178
	The FOR Loop	178
	The WHILE Loop	178
	The UNTIL loop	179
	Combination Loop	179
3.6.5	Conditional Execution	179
3.7	ARM Addressing Modes	181
3.7.1	Literal Addressing	182
	ARM's Way	183
3.7.2	Register Indirect Addressing	184
3.7.3	Register Indirect Addressing with an Offset	187
3.7.4	ARM's Autoindexing Pre-indexed Addressing Mode	190
3.7.5	ARM's Autoindexing Post-Indexing Mode	191
3.7.6	Program Counter Relative (PC-Relative) Addressing	192
3.7.7	ARM's Load and Store Encoding	193
3.8	Subroutine Call and Return	194
3.8.1	ARM Support for Subroutines	196
3.8.2	Conditional Subroutine Calls	197
3.9	Intermission: Examples of ARM Code	198
3.9.1	Extracting the Absolute Value	198
3.9.2	Byte Manipulation and Concatenation	198
3.9.3	Byte Reversal	199
3.9.4	Multiplication by $2^n - 1$ or $2^n + 1$	200
3.9.5	The Use of Multiple Conditions	200
3.9.6	With Just One Instruction...	200
3.9.7	Implementing Multiple Selection	201
3.9.8	Simple Bit-Level Logical Operations	201
3.9.9	Hexadecimal Character Conversion	201
3.9.10	Character Output in Hexadecimal	202
3.9.11	To Print a Banner	202
3.10	Subroutines and the Stack	203
3.10.1	Subroutine Call and Return	205
3.10.2	Nested Subroutines	206
3.10.3	Leaf Routines	207
3.11	Data Size and Arrangement	209
3.11.1	Data Organization and Endianism	209
3.11.2	Data Organization and the ARM	211
3.11.3	Block Move Instructions	216
	Block Moves and Stack Operations	217
	Applications of Block Move Instructions	219
3.12	Consolidation—Putting Things Together	220
	Four-Function Calculator Program	220
	Summary	223
	Problems	224
4	Instruction Set Architectures—Breadth and Depth	228
	Historical Background	230

4.1	The Stack and Data Storage	231
4.1.1	Storage and the Stack	232
	The Stack Frame and Local Variables	234
	Example of an ARM Processor Stack Frame	237
4.1.2	Passing Parameters via the Stack	239
	Pointers and C	242
	Functions and Parameters	243
	Pass-by-Reference	246
	Using Recursion	248
4.2	Privileged Modes and Exceptions	251
4.3	MIPS: Another RISC	254
	MIPS Instruction Format	255
	Conditional Branches	256
4.3.1	MIPS Data Processing Instructions	257
	Flow Control	258
	MIPS Example	259
	Other Loads and Stores	259
	MIPS and the ARM Processor	259
4.4	Data Processing and Data Movement	260
4.4.1	Indivisible Exchange Instructions	263
4.4.2	Double-Precision Shifting	264
4.4.3	Pack and Unpack Instructions	265
4.4.4	Bounds Testing	266
4.4.5	Bit Field Data	268
4.4.6	Mechanizing the Loop	272
4.5	Memory Indirect Addressing	273
	Using Memory Indirect Addressing to	
	Implement a switch Construct	277
	Using Memory Indirect Addressing to	
	Access Records	280
4.6	Compressed Code, RISC, Thumb, and MIPS16	282
4.6.1	Thumb ISA	282
	Design Decisions	283
4.6.2	MIPS16	287
4.7	Variable-Length Instructions	288
	Decoding Variable-Length Instructions	292
	Summary	294
	Problems	294
5	Computer Architecture and Multimedia	298
5.1	Applications of High-Performance Computing	299
	Computer Graphics	301
5.1.1	Operations On Images	303
	Noise Filtering	303
	Contrast Enhancement	303
	Edge Enhancement	304
	Lossy Compression	305
	JPEG	305
	MPEG	308
	MP3	308

	Digital Signal Processing	309
	DSP Architectures	312
	The SHARC Family of Digital Signal Processors	312
5.2	Multimedia Influences—Reinventing the CISC	314
	Architectural Progress	315
5.3	Introduction to SIMD Processing	318
	Packed Operations	319
	Saturating Arithmetic	321
	Packed Shifting	323
	Packed Multiplication	323
	Parallel Comparison	324
	Packing and Unpacking	325
	Coexisting with Floating-Point	326
5.3.1	Applications of SIMD Technology	328
	Chroma Keying	328
	Fade In and Out	330
	Clipping	332
5.4	Streaming Extensions and the Development of SIMD Technology	333
5.4.1	Floating-point Software Extensions	336
5.4.2	Intel's Third Layer of Multimedia Extensions	338
5.4.3	Intel's SSE3 and SSE4 Instructions	338
5.4.4	ARM Family Multimedia Instructions	340
	Summary	342
	Problems	343



PART III ORGANIZATION AND EFFICIENCY

6	Performance—Meaning and Metrics	348
6.1	Progress and Computer Technology	351
	Moore's Law	351
	Semiconductor Progress	352
	Memory Progress	354
6.2	The Performance of a Computer	356
6.3	Computer Metrics	358
6.3.1	Terminology	359
	Efficiency	359
	Throughput	360
	Latency	360
	Relative Performance	360
	Time and Rate	361
6.3.2	Clock Rate	361
	The Clock and the Consumer	365
6.3.3	MIPS	365
	Instruction Cycles and MIPS	367
6.3.4	MFLOPS	369
6.4	Amdahl's Law	371
	Examples of the Use of Amdahl's Law	372
6.5	Benchmarks	374
	LINPACK and LAPACK	374

	Oracle Applications Standard Benchmark	375
	PC Benchmarks	376
	Comparison of High-Performance Processors	376
	PCMARK7 A Commercial Benchmark for PCs	378
6.6	SPEC	382
	SPEC Methodology	384
	The SPEC CPU2006 Benchmarks	386
	SPEC and Power	389
6.7	Averaging Metrics	391
	Geometric Mean	392
	Harmonic Mean	393
	Weighted Means	394
	Summary	394
	Problems	395
7	Processor Control	398
7.1	The Generic Digital Processor	401
7.1.1	The Microprogram	404
	Modifying the Processor Organization	406
7.1.2	Generating the Microoperations	410
7.2	RISC Organization	414
7.2.1	The Register-to-register Data Path	416
	Load and Store Operations	417
	Jump and Branch Operations	418
7.2.2	Controlling the Single-cycle	
	Flow-through Computer	419
	Execution Time	422
7.3	Introduction to Pipelining	423
7.3.1	Speedup Ratio	427
7.3.2	Implementing Pipelining	427
	From PC to Operands	429
	Implementing Branch and Literal Operations	430
7.3.3	Hazards	434
	Delayed Branch	436
	Data Hazards	437
7.4	Branches and the Branch Penalty	442
7.4.1	Branch Direction	443
7.4.2	The Effect of a Branch on	
	the Pipeline	444
7.4.3	The Cost of Branches	445
7.4.4	The Delayed Branch	448
7.5	Branch Prediction	451
	Static and Dynamic Branch Prediction	453
7.6	Dynamic Branch Prediction	454
7.6.1	Branch Target Buffer	456
7.6.2	Two-Level Branch Prediction	459
	Combining Instruction Addresses	
	and Branch History	463
	Summary	464
	Problems	465

8	Beyond RISC: Superscalar, VLIW, and Itanium	472
	Overview of Chapter 8	473
8.1	Superscalar Architecture	473
	In-Order and Out-of-Order Execution	479
8.1.1	Instruction Level Parallelism (ILP)	482
	Data Dependencies and Register Renaming	484
8.1.2	Superscalar Instruction Issue	486
	Control Dependencies	488
	Examples of Superscalar Processors	490
	The Alpha	490
	The Pentium	492
8.1.3	VLIW Processors	499
	Interrupts and Superscalar Processing	502
8.2	Binary Translation	504
	The IA-32 code	505
8.2.1	The Transmeta Crusoe	506
8.3	EPIC Architecture	510
8.3.1	Itanium Overview	512
	IA64 Assembler Conventions	514
8.3.2	The Itanium Register Set	515
	The Not a Thing Bit	517
	Predicate and Branch Registers	517
	Other Itanium Registers	518
8.3.3	IA64 Instruction Format	518
8.3.4	IA64 Instructions and Addressing Modes	519
	Addressing Modes	523
8.3.5	Instructions, Bundles, and Breaks	524
	IA64 Bundles, STOPs, and Assembly Language Notation	527
8.3.6	Itanium Organization	529
	The McKinley—The Itanium 2	531
	The Itanium 9300 Tukwila Processor	532
	The Itanium Poulson Processor	532
	Is the IA64 a VLIW Processor?	532
8.3.7	Predication	532
	Compare Instructions in Detail	534
	Preventing False Data Dependency in Predicated Computing	537
	Branch Syntax	538
8.3.8	Memory Access and Speculation	539
	Control Speculation	540
	The Advanced Load	541
8.3.9	The IA64 and Software Pipelining	543
	Registers and Function Calls	548
	Summary	549
	Problems	549

PART IV THE SYSTEM

9 Cache Memory and Virtual Memory 554

Memory Hierarchy	554
------------------	-----

9.1	Introduction to Cache Memory	558
9.1.1	Structure of Cache Memory	560
	Principle of Locality of Reference	560
9.2	Performance of Cache Memory	561
9.3	Cache Organization	565
9.3.1	Fully Associative Mapped Cache	566
	Associative Memory	569
9.3.2	Direct-Mapped Cache	570
9.3.3	Set-Associative Cache	574
9.3.4	Pseudo-Associative, Victim, Annex, and Trace Caches	579
9.4	Considerations in Cache Design	581
9.4.1	Physical versus Logical Cache	581
9.4.2	Cache Electronics	582
9.4.3	Cache Coherency	582
9.4.4	Line Size	583
9.4.5	Fetch Policy	585
9.4.6	Multi-Level Cache Memory	586
9.4.7	Instruction and Data Caches	587
9.4.8	Writing to Cache	589
9.5	Virtual Memory and Memory Management	592
9.5.1	Memory Management	592
9.5.2	Virtual Memory	595
	Memory Management and Multitasking	595
	Address Translation	596
	Two-Level Tables	598
	Summary	601
	Problems	602
10	Main Memory	606
10.1	Introduction	606
10.1.1	Principles and Parameters of Memory Systems	608
	Random Access and Sequential Access Memory	608
	Volatile and Nonvolatile Memory	609
	Read/Write and Read-Only Memory	609
	Static and Dynamic Memory	609
	Memory Parameters	610
10.1.2	Memory Hierarchy	611
10.2	Primary Memory	612
10.2.1	Static RAM	612
	The Static RAM Memory System	615
	The Write Cycle	617
	Byte/Word Control	618
	Address Decoding	620
10.2.2	Interleaved Memory	622
10.3	DRAM	623
10.3.1	DRAM Timing	627
	Write-Cycle Timing	630

10.3.2	Developments in DRAM Technology	631
	SDRAM	632
	DDR DRAM	634
	DDR2 and DDR3 DRAM	634
	DDR4	636
10.4	The Read-Only Memory Family	637
10.4.1	The EPROM Family	638
	The EEPROM	639
	Flash Memory	639
	Multi-Level Flash Technology	640
	NAND and NOR Flash	641
	Wear Leveling in Flash Memories	643
10.5	New and Emerging Nonvolatile Technologies	646
10.5.1	Ferroelectric Hysteresis	648
10.5.2	MRAM—Magnetoresistive Random Access Memory	651
10.5.3	Ovonic Memory	652
	Summary	654
	Problems	654
11	Secondary Storage	658
11.1	Magnetic Disk Drives	659
11.2	Magnetism and Data Storage	660
11.2.1	The Read/Write Head	662
	The Recording Process	663
11.2.2	Limits to Magnetic Recording Density	664
11.2.3	Principles of Data Recording on Disk	666
	Platter Technology	670
	The GMR Head—A Giant Step in Read-Head Technology	671
	Pixie Dust	672
	The Optically Assisted Head	673
11.3	Data Organization on Disk	674
11.3.1	Tracks and Sectors	676
	Formatting a Disk	678
	Interleaving	679
11.3.2	Disk Parameters and Performance	679
	Accessing Sectors	682
	The Internal Disk Cache	684
	Transfer Rate	684
11.3.3	SMART Technology	684
	Effect of Temperature on Disk Reliability	686
11.4	Secure Memory and RAID Systems	688
	RAID Level 1	689
	RAID Level 2 and Level 3	690
	RAID Level 4 and Level 5	691
	Failure of RAID 5—An Example	692
	RAID Level 6	692
11.5	Solid-State Disk Drives	693
	Special Features of SSDs	695
11.6	Magnetic Tape	698

11.7 Optical Storage Technology	700
11.7.1 Digital Audio	701
11.7.2 Reading Data from a CD	702
Disk Speed	705
The Optical Read-Head	706
Focusing and Tracking	706
Buffer Underrun	707
11.7.3 Low-Level Data Encoding	708
11.7.4 Recordable Disks	711
Re-Writable CDs	711
Magneto-Optical Storage	713
11.7.5 The DVD	714
Recordable DVDs	715
11.7.6 Blu-ray	715
Summary	717
Problems	717
 12 Input/Output	 720
12.1 Fundamental Principles of I/O	721
Memory-Mapped Peripherals	723
12.1.1 Peripheral Register Addressing Mechanisms	725
12.1.2 Peripheral Access and Bus Width	727
Preserving Order in I/O Operations	729
Side Effects	730
12.2 Data Transfer	731
12.2.1 Open-Loop Data Transfers	731
12.2.2 Closed-Loop Data Transfers	732
12.2.3 Buffering Data	733
The FIFO	734
12.3 I/O Strategy	739
12.3.1 Programmed I/O	739
12.3.2 Interrupt-driven I/O	740
Interrupt Processing	741
Nonmaskable Interrupts	742
Prioritized Interrupts	742
Nested Interrupts	743
Vectored Interrupts	745
Interrupt Timing	746
12.3.3 Direct Memory Access	749
12.4 Performance of I/O Systems	751
12.5 The Bus	752
12.5.1 Bus Structures and Topologies	753
12.5.2 The Structure of a Bus	755
The Data Bus	756
Bus Speed	756
The Address Bus	759
The Control Bus	760
12.6 Arbitrating for the Bus	761
12.6.1 Localized Arbitration and the VMEbus	763
Releasing the Bus	766

The Arbitration Process	766
VMEbus Arbitration Algorithms	767
12.6.2 Distributed Arbitration	768
NuBus Arbitration	768
12.7 The PCI and PCIe Buses	772
12.7.1 The PCI Bus	772
Data Transactions on the PCI Bus	776
12.7.2 The PCI Express Bus	781
PCIe Data Link Layer	784
12.7.3 CardBus, the PC Card, and ExpressCard	785
CardBus Cards	787
ExpressCard Cards	788
12.8 The SCSI and SAS Interfaces	789
SCSI Signals	790
SCSI Bus Transactions	792
SCSI Messages and Commands	792
12.9 Serial Interface Buses	794
12.9.1 The Ethernet	795
12.9.2 FireWire 1394 Serial Bus	797
Serial Bus Addressing	800
The Physical Layer	800
Arbitration	803
Initialization	804
The Link Layer	804
12.9.3 USB	805
USB – The First Two Generations	805
Electrical Characteristics	806
Physical Layer Data Transmission	808
Logical Layer	809
USB 3.0	811
Summary	812
Problems	813



PART V PROCESSOR-LEVEL PARALLELISM

13	Processor-Level Parallelism	820
	Dimensions of Parallel Processing	822
	A Brief History of Parallel Computing	823
13.1	Why Parallel Processing?	825
13.1.1	Power—The Final Frontier	826
13.2	Performance Revisited	829
	Performance Measurement	831
13.3	Flynn's Taxonomy and Multiprocessor Topologies	833
13.4	Multiprocessor Topologies	835
13.5	Memory in Multiprocessor Systems	842
13.5.1	NUMA Architectures	842
13.5.2	Cache Coherency in Multiprocessor Systems	843
	The MESI Protocol	844
	False Sharing	847

13.6 Multithreading	847
13.7 Multi-core Processors	851
Homogeneous and Heterogeneous Processors	852
13.7.1 Homogeneous Multiprocessors	852
Intel Nehalem Multi-Core Processor	854
AMD Multi-Core Processors	854
ARM Cortex A9 Multi Core	856
IBM Power7	857
The GPU	858
13.7.2 Heterogeneous Multiprocessors	861
The Cell Architecture	861
13.7.3 Networks on a Chip	862
13.8 Parallel Programming	865
13.8.1 Parallel Processing and Programming	867
OpenMP	868
13.8.2 Message Passing Interface	870
13.8.3 Partitioned Global Address Space	871
13.8.4 Synchronization	872
The Spinlock	873
Summary	874
Problems	874
 Bibliography	 876
Index	888