

Alan Clements

COMPUTER ORGANIZATION AND ARCHITECTURE

Themes and Variations

Alan Clements

Teesside University



This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.



Computer Organization and Architecture: Themes and Variations, First Edition **Alan Clements**

Publisher, Global Engineering: Christopher M.

Shortt

Acquisitions Editor: Swati Meherishi

Developmental Editor: Amy Hill, Collaborative

Concepts, LLC

Assistant Development Editor: Farah Naseem

Editorial Assistant: Tanya Altieri Team Assistant: Carly Rizzo

Marketing Manager: Lauren Betsos

Media Editor: Chris Valentine

Content Project Manager: D. Jean Buttrom Production Service: RPK Editorial Services, Inc.

Copyeditor: Shelly Gerger-Knechtl

Proofreader: Pat Daly

Indexer: Shelly Gerger-Knechtl Compositor: MPS Limited

Senior Art Director: Michelle Kunkler

Cover and Internal Designer: Rokusek Design

Cover Image: © Dvpodt/Shutterstock

Rights Acquisition Director: Audrey Pettengill

Rights Acquisition Specialist, Text and Image:

Amber Hosea

Text and Image Permissions Researcher:

Kristiina Paul

Senior Manufacturing Planner: Doug Wilke

© 2014 Cengage Learning

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced, transmitted, stored, or used in any form or by any means graphic, electronic, or mechanical, including but not limited to photocopying, recording, scanning, digitizing, taping, web distribution, information networks, or information storage and retrieval systems, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the publisher.

> For product information and technology assistance, contact us at Cengage Learning Customer & Sales Support, 1-800-354-9706.

For permission to use material from this text or product, submit all requests online at www.cengage.com/permissions. Further permissions questions can be emailed to permissionrequest@cengage.com.

Library of Congress Control Number: 2012942868

ISBN-13: 978-1-111-98704-6

ISBN-10: 1-111-98704-1

Cengage Learning

200 First Stamford Place, Suite 400

Stamford, CT 06902

USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at:

international.cengage.com/region.

Cengage Learning products are represented in Canada by Nelson Education Ltd.

For your course and learning solutions, visit www.cengage.com/engineering.

Purchase any of our products at your local college store or at our preferred online store www.cengagebrain.com.

Dedication

Dedicated to the memory of my friend, colleague, and mentor Tomás López Jiménez.

Contents

Paths through Computer Architecture

Preface

PART I	THE	BEGINNING	"IISC M	\$ 1
1	Cor	mputer 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?

1.4.5 The Memory

1.6.2 The Bus

1.7 Modern Computing

2.1 What is Data?

Summary **Problems**

1.5 The Stored Program Concept

Two Address Instructions

One Address Instructions

Computer Categories 1.6 Overview of the Computer System

Computer Arithmetic and Digital Logic

Representing Information

2.1.1 The Bit and Byte

2.1.2 Bit Patterns

1.6.1 The Memory Hierarchy

xix

32

34

37

39

40 40

41

41

43

45 46

46

48

50

50

51 51 11500

xxix

2.2	Numb	ers	55
	2.2.1	Positional Notation	56
2.3	Binary	Arithmetic	57
2.4	Signed	d 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	Introd	uction 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	Floatir	ng-Point Numbers	71
		Normalization of Floating-Point Numbers	72
		Biased Exponents	72
	2.6.1	IEEE Floating-Point Numbers	72
	2.0.1	IEEE Floating-Point Format	73
		Characteristics of IEEE Floating-Point Numbers	75
2.7	Floatir	ng-Point Arithmetic	77
,	Houth	Rounding and Truncation Errors	78
2.8	Floatir	ng-Point Arithmetic and the Programmer	79
2.0	2.8.1	Error Propagation in Floating-Point Arithmetic	81
	2.8.2	Generating Mathematical Functions	81
	2.0.2	Using Functions to Generate New Functions	83
2.9	Comp	uter Logic	84
2.,	2.9.1	Digital Systems and Gates	86
	2.9.2	Gates	86
	2.7.2	Fundamental Gates	87
		The AND Gate	87
		The OR Gate	87
		The Inverter	88
			oc
		Derived Gates—the NOR (Not OR), NAND	90
	2.9.3	(Not AND), and Exclusive OR Basic Circuits	89 91
	2.9.3	The Half Adder and Full Adder	91
			93
		The Decoder	97
		The Multiplexer	
		The Point Circuit	98
2 10	Comus	The Prioritizer	100
2.10	-	ntial Circuits	101
	∠.1∪.1	Latches Clocked BS Flin floor	102
		Clocked RS Flip-flops	104
		D Flip-flop	105
	2 10 2	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	Arc	hitect	ure and Organization	130			
	3.1	Intro	duction 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 C	omponents 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 A	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			

		Compare and Test Instructions Branching and Loop Constructs	178 178
	3.0.4	-	178 178
		The FOR Loop	178 178
		The UNITH loop	178 179
		The UNTIL loop	179 179
	3.6.5	Combination Loop Conditional Execution	179
3.7		Addressing Modes	1/9
3./		Literal Addressing	182
	3.7.1	ARM's Way	183
	372	Register Indirect Addressing	184
			187
		Register Indirect Addressing with an Offset ARM's Autoindexing Pre-indexed	167
	3.7.4		190
	275	Addressing Mode ARM's Autoindexing Post-Indexing Mode	190
			191
	3.7.0	Program Counter Relative (PC-Relative) Addressing	192
	277	ARM's Load and Store Encoding	192
3.8		utine Call and Return	193
3.0		ARM Support for Subroutines	194
	3.8.2	**	190
3.9	···	nission: Examples of ARM Code	197
3.9		Extracting the Absolute Value	198
		Byte Manipulation and Concatenation	198
		Byte Reversal	198
		Multiplication by $2^n - 1$ or $2^n + 1$	200
		The Use of Multiple Conditions	200
		•	200
		Implementing Multiple Selection	200
		Simple Bit-Level Logical Operations	201
			201
		Character Output in Hexadecimal	201
		To Print a Banner	202
2 10		utines and the Stack	202
3.10		Subroutine Call and Return	205
		Nested Subroutines	206
		Leaf Routines	207
2 11		size and Arrangement	209
J.11		Data Organization and Endianism	209
		Data Organization and the ARM	211
		Block Move Instructions	216
	3.11.3	Block Moves and Stack Operations	217
		Applications of Block Move Instructions	219
3.12	Conso	lidation—Putting Things Together	220
J.12	COIISO	Four-Function Calculator Program	220
Sum	mary	1 out 1 direction Culvatator 1 logiani	223
	lems		224
Inst	ructio	n Set Architectures—Breadth and Depth	228

Historical Background

230

4

4.1	The S	tack 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	<u>*</u>	239
		Pointers and C	242
		Functions and Parameters	243
		Pass-by-Reference	246
		Using Recursion	248
4.2	Privil	eged Modes and Exceptions	251
4.3		: Another RISC	254
		MIPS Instruction Format	255
		Conditional Branches	256
	4.3.1	MIPS Data Processing Instructions	257
	1.5.1	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
•••		Indivisible Exchange Instructions	263
		Double-Precision Shifting	264
		Pack and Unpack Instructions	265
		Bounds Testing	266
		Bit Field Data	268
	4.4.6	Mechanizing the Loop	272
4.5		ory Indirect Addressing	273
7.5	Wicini	Using Memory Indirect Addressing to	273
		Implement a switch Construct	277
		Using Memory Indirect Addressing to	211
		Access Records	280
4.6	Comr	pressed Code, RISC, Thumb, and MIPS16	282
7.0	4.6.1	Thumb ISA	282
	7.0.1	Design Decisions	283
	4.6.2	MIPS16	287
4.7		ble-Length Instructions	288
7.7	varia	Decoding Variable-Length Instructions	292
Sum	mary	Decoding variable-Length instructions	292 294
	olems		294
FIUL	JICIIIS		294
Con	npute	er Architecture and Multimedia	298
5.1	Appli	cations 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

5

		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
		Sum	mary	342
			lems	343
1150 1	PART III	ORG	SANIZATION AND EFFICIENCY	
(%H ¥				- 40
	6	Per	formance—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	367 369
		6.4	6.3.4 MFLOPS Amdahl's Law	
		6.4	6.3.4 MFLOPS	369
		6.4 6.5	6.3.4 MFLOPS Amdahl's Law Examples of the Use of Amdahl's Law Benchmarks	369 371
			6.3.4 MFLOPS Amdahl's Law Examples of the Use of Amdahl's Law	369 371 372

Digital Signal Processing

Architectural Progress

5.2 Multimedia Influences—Reinventing the CISC

The SHARC Family of Digital Signal Processors

DSP Architectures

309

312

312

314

315

			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	Avera	aging Metrics	391
			Geometric Mean	392
			Harmonic Mean	393
	_		Weighted Means	394
		mary		394
	Prot	olems		395
7	Pro	cesso	r Control	398
	7.1	The G	ieneric Digital Processor	401
		7.1.1		404
			Modifying the Processor Organization	406
		7.1.2		410
	7.2		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	Introd	duction 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		hes 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	Branc	h Prediction	451
			Static and Dynamic Branch Prediction	453
	7.6	•	mic 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
		mary		464
	Drok	lame		165

8	Beyond RISC: Superscalar, VLIW, and Itanium			
			Overview of Chapter 8	473
	8.1	Supei	rscalar 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	Binar	y Translation	504
			The IA-32 code	505
		8.2.1	The Transmeta Crusoe	506
	8.3		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	548 549
		Summary		
	Prok	olems		549



PART IV THE SYSTEM

9 Cache Memory and Virtual Memory

554

Memory Hierarchy

	9.1	Introd	luction to Cache Memory	558
		9.1.1	Structure of Cache Memory	560
			Principle of Locality of Reference	560
	9.2	Perfor	mance 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
			Set-Associative Cache	574
		9.3.4	Pseudo-Associative, Victim,	
			Annex, and Trace Caches	579
	9.4	Consid	derations 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	Virtua	l 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
	Sum	mary		601
	Prok	olems		602
10	Mai	n Men	nory	606
			·	606
	10.1	Introd		606
		10.1.1	Principles and Parameters of	608
			Memory Systems Rendem Access and Sequential	000
			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		ry Memory	612
	10.2		Static RAM	612
		10.2.1	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
			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
	Sum	mary	654
	Prob	lems	654
11	Sec	ondary Storage	658
	11.1	Magnetic Disk Drives	659
		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	
		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
	Sum	mary	717
	Prob	lems	717
12	Inpu	ut/Output	720
	12.1	Fundamental Dringinles of I/O	721
	12.1	Fundamental Principles of I/O	
		Memory-Mapped Peripherals	723 725
		12.1.1 Peripheral Register Addressing Mechanisms	723 727
		12.1.2 Peripheral Access and Bus Width	727
		Preserving Order in I/O Operations	
	12.2	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
	12.2	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
		Performance of I/O Systems	751
	12.5	The Bus	752
		12.5.1 Bus Structures and Topologies	753 755
		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 PO	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 SC	789	
		SCSI Signals	790
		SCSI Bus Transactions	792
		SCSI Messages and Commands	792
12.9	Serial	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
Prob	lems		813

"115°C (T) -	PART V	PROCESSOR-LEVEL PARALLELISM	
(26) 5	TANT	THOCESSON LEVEL I ANALLELISM	
	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			
13.7	Multi-core Processors			
		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	Paralle	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	
Sum	Summary			
Problems				
Bibliography			876	
Index				

Preface

The twenty-first century is an age of scientific and technological wonders. Computers have proved to be everything people expected—and more. Bioengineering has unraveled the mysteries of the cell and enabled scientists to synthesize drugs that were inconceivable a decade ago. Nanotechnology provides a glimpse into a world where the computer revolution is combined with engineering at the atomic level to create microscopic autonomous machines that may, one day, be injected into the body to carry out internal repairs. Ubiquitous computing has given us cell phones, MP3 players, and digital cameras that keep us in touch with each other via the Internet. The computer is at the core of almost all modern technologies. This book explains how the computer works.

The discipline called *computing* has been taught in universities since the 1950s. In the beginning, computing was dominated by the large mainframe, and the subject consisted of a study of computers themselves, the operating systems that controlled the computers, languages and their compilers, databases, and business computing. Since then, computing has expanded exponentially and now embraces so many different areas that it's impossible for any university to cover computing in a comprehensive fashion. We have to concentrate on the essential elements of computing. At the heart of this discipline lies the machine itself: the computer. Of course, computing as a theoretical concept could exist quite happily without computers. Indeed, a considerable amount of work on the theoretical foundations of computer science was carried out in the 1930s and 1940s before the computer revolution took place. However, the way in which computing has progressed over the last 40 years is intimately tied up with the rise of the microprocessor. The Internet could not have taken off in the way it has if people didn't have access to very low-cost computers.

Since the computer itself has had such an effect on both the growth of computing and the path computing has taken, it's intuitively reasonable to expect that the computing curriculum should include a course on how computers actually work. University-level Computer Science and Computer Engineering CS programs invariably include a course on how computers work. Indeed, professional and course accreditation bodies specify computer architecture as a core requirement; for example, computer architecture is central to the joint IEEE Computer Society and ACM Computing Curriculum.

Courses dealing with the embodiment or realization of the computer are known by a variety of names. Some call them hardware courses, some call them computer architecture courses, and some call them computer organization courses (with all manner of combinations in between). Throughout this text, I will use the expression *computer architecture* to describe the discipline that studies the way in which computers are designed and how they operate. I will, of course, explain why this discipline has so many different names and point out that the computer can be viewed in different ways.

Like all areas of computer science, the field of computer architecture is advancing rapidly as developments take place in instruction set design, instruction level parallelism, cache memory technology, bus systems, speculative execution, multi-core computing, and so on. We examine all these topics in this book.

Computer architecture underpins computer science; for example, computer performance is of greater importance today than ever before, because even those who buy personal computers have to understand systems architecture in order to make the best choice.

Although most students will never design a new computer, today's students need a much broader overview of the computer than their predecessors. Students no longer have to be competent assembly language programmers, but they must understand how buses, interfaces, cache memories, and instruction set architectures determine the performance of a computer system.

Moreover, students with an understanding of computer architecture are better equipped to study other areas of computer science; for example, a knowledge of instruction set architectures gives students a valuable insight into the operation of compilers.

My motivation for writing this book springs from my experience in teaching an intermediate level course in computer architecture at the University of Teesside. I threw away the conventional curriculum that I'd inherited and taught what could be best described as *Great Ideas in Computer Architecture*. I used this course to teach topics that emphasized global concepts in computer science that helped my students with both their operating systems and C courses. This course was very successful, particularly in terms of student motivation.

Anyone writing a text on computer architecture must appreciate that this subject is taught in three different departments: electrical engineering (EE), electrical and computer engineering (ECE), and computer science (CS). These departments have their own cultures and each looks at the computer from their own viewpoint. EE and ECE departments focus on electronics and how the individual components of a computer operate. EE/ECE-oriented texts concentrate on gates, interfaces, signals, and computer organization. Many students in CS departments don't have the requisite background in electronics, so they can't follow texts that emphasize circuit design. Instead, computer science departments place more stress on the relationship between the low-level architecture of the processor and the higher-level abstractions in computer science.

Although it is near impossible to write a text optimized for use in both EE/ECE and CS departments, *Computer Organization and Architecture: Themes and Variations* is an effective compromise that provides sufficient detail at the logic and organizational levels for EE/ECE departments without including the degree of detail that would alienate CS readers.

Undergraduate computer architecture is taught at three levels: introductory, intermediate, and advanced. Some schools teach all three levels, some compress this sequence into two levels, and some provide only an introduction. This text is aimed at students taking first- and second-level courses in computer architecture and at professional engineers who would like an overview of current developments in microprocessor architecture. The only prerequisite is that the reader should be aware of the basic principles of a high-level language such as C and have a knowledge of basic algebra.

It is difficult to pitch a book at precisely the right level. Indeed, such an ideal level doesn't exist. Different students react in different ways to any specific text. If you make a book very focused and follow a narrow curriculum, you appeal only to students on a tiny handful of courses. Computer Organization and Architecture: Themes and Variations is well-suited to a wide range of courses, because it covers the basics and some of the more advanced topics in computer architecture.

Features of the Book

Why inflict yet another text on computer architecture on the world? Computer architecture is a fascinating topic. It's all about how you can take vast numbers of a single primitive element such as a NAND gate and make a computer. It's all about how common sense and technology meet. For example, the cache memory that makes processors so fast is conceptually no more complicated than the note on the back of an envelope. Equally, the way in which all processors operate uses a technique invented by Ford for car production: the *pipeline* or production line. I have tried to make the subject interesting and have covered a greater range of topics than absolutely necessary. For example, in this text we will look at memory devices that operate by moving an oxygen atom from one end of a crystal to the other.

The title of this text, Computer Organization and Architecture, emphasizes the structure of the complete computer system (CPU, memory, buses, and peripherals). The subtitle Themes and Variations indicates that there is a theme (i.e., the computer system) and also variations, for example, the different approaches to increasing the speed of a CPU or to organizing cache memory.

It is often easier to describe something in terms of what it isn't rather than what it is. This book is not concerned with the precise engineering details of microprocessor systems design, interfacing, and peripherals. It certainly isn't an assembly language primer. The central theme of this book is microcomputer *architecture* rather than microprocessor systems design. Computer architecture can be defined, for our present purposes, as the view of a computer seen by the machine language programmer. That is, a computer's architecture takes no account of its actual hardware or implementation and is concerned only with what it does. We will not consider some of the hardware and interfacing aspects of microprocessors, except where they impinge on its architecture (e.g., cache memory, memory management, and the bus).

The Target Architecture

Anyone writing an architecture text has to select a target architecture as a vehicle to teach the fundamentals of machine design and assembly language programming. Professors regularly debate with religious intensity the relative merits of illustrating a course with a real commercial processor or with a hypothetical generic processor. The generic machine is easy to understand and has a shallow learning curve. Students often find that absorbing the fine details of a real processor is time consuming and unrewarding. On the other hand, practical engineering is all about living with the limitations of the real world. Moreover, a real machine teaches students about the design decisions that engineers have to make in order to create a commercially viable product.

In the 1970s and 1980s DEC's PDP-11 minicomputer was widely adopted as a teaching vehicle. The PDP-11 gradually dropped out of the curriculum with the advent of 16-bit microprocessors such as the Motorola 68K. From the academic's point of view, the 68K (loosely based on the earlier PDP-11) was a dream machine, because its architecture is relatively regular and that made it easy for students to write programs in 68K assembly language. A casual observer might have expected the ubiquitous Intel IA32 family, which is found in most PCs, to have played a significant role in computer architecture education. After all, countless students get hands-on experience of Intel's processors. The 80x86 family has never really caught on in the academic world because its complex architecture grew in an ad-hoc fashion as each new member of the family was released and this presents students with an excessive burden. Some academics illustrate their course with a high-performance RISC processor, such as MIPS, which is both powerful and easy to understand. Such high-end RISC processors are found in workstations but are relatively unknown to many students (professors have observed that students often request PC-based technology due to their familiarity with the PC). However, RISC processors are used in both high-performance computers and most cell phones.

I have selected the ARM processor as a vehicle to introduce assembly language and computer organization. It is a processor that is powerful, elegant, yet easy to learn. Moreover, development tools for the ARM processor are widely available which means that students can write programs in ARM assembly language and run them in the lab or at home on their PCs.

A strong contender for the role of target architecture in a modern text is Intel's IA64 Itanium processor. This is a device of immense power and sophistication, yet its basic architecture is simpler than the 80x86 family. The rich and innovative features of the Itanium's architecture illustrate numerous concepts found in a computer architecture course–from the data stack to speculative execution, and from pipelining to instruction level parallelism (ILP). Consequently, I also introduce some features of this processor when we look at high-performance computing.

Computer Organization and Architecture: Themes and Variations isn't a conventional computer architecture text. I go beyond the conventional curriculum and cover material that is interesting, important, and relevant. One of my principal objectives is to provide students with an appropriate body of knowledge that they can absorb. All too often, students graduate from university with embarrassingly large gaps in their knowledge. I know of no other text that adopts my approach. For example, all computer architecture texts introduce floating-point arithmetic, yet very few discuss the codes for data compression required to store large volumes of textual and video information, and none describe the MP3 data compression that's at the heart of an entire industry. Similarly, computer architecture texts often lack

coverage of architectural features intended to support multimedia applications. Some of the highlights of this text are described below.

History

Books on computer architecture usually have a section on the history of the computer. Such history chapters are often inaccurate and have received criticism from experts in the field. However, I feel that a history chapter is important, because a knowledge of computer history helps students appreciate how and why developments took place. By knowing where computers came from, students are better able to understand how they are likely to develop in the future. In this text, I provide a short overview of the history of computing and include further historical background in the supplementary web-based material that accompanies this book.

OS Support

The operating system is intimately bound up with computer architecture. Computer Organization and Architecture: Themes and Variations covers topics in architecture of interest to those who also study operating systems (e.g., memory management, context switching, protection mechanisms).

Multimedia Support

The most important driving force behind modern computer architectures is the growth of multimedia systems with their insatiable demand for high performance and high bandwidths. This text demonstrates how modern architectures have been optimized for multimedia applications. We look at the effect of multimedia applications on both the architectures of computers and the design of buses and computer peripherals, such as hard disks for use in audiovisual applications.

Input/Output Systems

Today's computers are not only much faster than their predecessors, but they also provide more sophisticated means of getting information into and out of the computer. I/O was of relatively little importance when the typical computer was interfaced only to a keyboard, modem, and printer. Computers are now routinely interfaced to peripherals, such as digital video cameras that demand massive data transfer rates. We will look at some of the modern, high-performance I/O systems, such as the USB and FireWire interfaces. We will also delve more deeply into input/output-related topics such as handshaking and buffering.

Computer Memory Systems

Memory is the *Cinderella* of the computer world. Without high-density, high-performance memory systems, neither low-cost desktop systems nor digital cameras with 32GB of storage would be possible. I have divided memory systems into two chapters: the first dealing with semiconductor memory and the second dealing with magnetic and optical memory. We will also take a look at some of the interesting emerging memory technologies, such as Ovonic memory and ferroelectric memory.

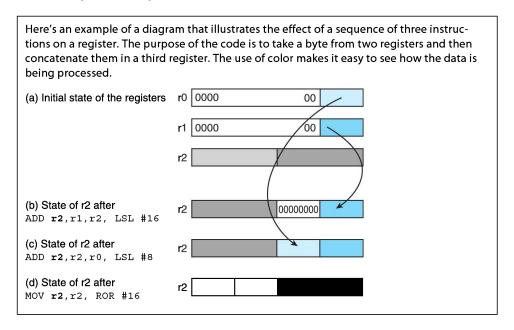
Approach

The books that I've most enjoyed are those where a little of the author's personality and view of the subject shines through. I hope that the same is true of this book. Computer architecture isn't something that can be expressed as a set of cold equations to be learned; it is a culture

that has developed over the years. At conferences, you will meet academics who passionately argue the relative merits of this computer over that one. I would be a poor educator if I did not at least hint to students that computer architecture can be as much fad as fact.

It's also true that different authors emphasize different topics. For example, most authors stress the design of the processor and have relatively little to say about memory, buses, and peripherals—even though all of these elements are necessary to create a computer system. Perhaps some feel that one aspect of a computer is more intrinsically interesting than others. I provide more coverage of memory, buses, and interfaces than many texts, because I feel that these topics are as important as the processor itself. Similarly, I've included details of memory elements such as Ovonic devices, which store data by melting a bead of glass and then storing a 1 or 0 by selecting how fast it cools. This is such a remarkable example of engineering ingenuity that I felt I had to include it. I'd like students to share the enthusiasm I have for this subject.

I find that a significant shortcoming of many texts is the quality of diagrams and illustrations. All too often a figure has far too little annotation, and its meaning is entirely lost. I have drawn virtually all of the included diagrams myself, and I hope that they do a good job in illustrating the meaning of the text.



Contents Overview

I have divided the book into five parts.

Part I: An Introduction to Computer Architecture covers the enabling material that allows us to discuss computer architecture. Chapter 1 takes an unusual approach to the introduction of a computer. I begin by presenting a problem, solving it, and then inventing a system to implement the solution—the computer. My aim is to demonstrate that the computer closely models the way in which we solve problems. Because so many topics in computer architecture are interrelated, I provide a brief overview of the computer system to allow me to mention topics, like cache memory, before I discuss them in detail.

We will look at the way in which information is encoded and represented in binary form; for example, I introduce the computer arithmetic of both signed and unsigned integers, demonstrate how floating-point numbers represent very large and very small quantities, and then

briefly introduce more complex mathematical functions. An introduction to gates and logic design is also included in Chapter 2. This topic always presents the author of a text on computer architecture with some difficulty. Should logic be omitted because some students take a separate course on logic design? Or should logic be relegated to an appendix so that only those who require it need delve into it? I've decided to put it in this section, because it allows a natural progression from numbers to computers and helps students understand some of the material that follows. Even students who have already taken a course in logic should go over this chapter lightly. I end the section on logic design with a very simple proto-computer.

In an ideal world, Part I would include a detailed history of the way in which computers have developed from simple mechanical calculators to today's powerful processors, because all students should have an understanding of how the profession developed. The history of the computer is a fascinating story that begins with a desire to make tedious arithmetical calculations easier, then moves to the telegraph and transatlantic cable project that did so much to "shrink" the world. In the last century, modern computing grew out of the technology used to create the telephone system and was driven by the need for high-speed calculation by scientists, the business community, and the military. No other human artifact has developed as fast as the computer in terms of its increase in performance and its decrease in cost. It is said that if the automobile had developed as rapidly as the computer we'd be driving cars at many times the speed of light using a single drop of gasoline a year.

Unfortunately, I just can't do justice to the history of the computer here. Students have to cover a certain body of knowledge in a computer architecture course, and they have to develop a range of skills in order to become practitioners in their chosen discipline. Consequently, I have located the material on computer history on the book's companion website, the details of which can be found in the Supplements and Resources section below... All students are strongly encouraged to read this material. Computer scientists have a tendency to reinvent the wheel, so concepts developed one year sometimes appear in a different context the following year.

Part II: Instruction Set Architectures is the heart of this course and includes three chapters. I first introduce the concept of an instruction set and then examine some of the important issues in computer architecture, such as the data structures that are accessed and manipulated by computers. Chapter 3 introduces the ARM family of microprocessors that have a truly elegant architecture and include facilities allowing us to cover several interesting topics, such as predicated execution where an instruction is executed only if certain conditions are met. It would be nice to cover all of the variations in the architecture of the computer, but unless you are going to dedicate the rest of your life to that end, there is not time. In Chapter 4, we look at some of the variations in computer architecture. For example, memory indirect addressing modes that let you access complex data structures with remarkable ease, and special compressed-code processors that squeeze conventional code to a fraction of its normal size.

The final chapter of Part II looks at the way in which processors have been adapted to suit modern multimedia applications such as video encoding and decoding. In this chapter, we also take a brief look at the background of multimedia processing and explain why high-performance computing is so necessary.

Part III: Organization and Efficiency is concerned with computer organization; that is, how computers work and are organized internally. Chapter 6 begins with an introduction to performance, which is the way the speed of computers is measured and how computers are compared. Some might argue that performance should be introduced earlier because of its importance. However, because so many topics influence performance, I thought it is better placed here where the reader is able to understand more of the background.

Chapter 7 looks at how computers actually work. We begin by looking at microprogramming, which is a technique that makes it possible to design computers of arbitrary complexity and is still used to implement some of the more complex instructions in Intel's IA32 processors. Then we introduce pipelining, which is fundamental to all modern computer design. Pipelining mirrors the industrial production line where instructions are executed bit by bit as they flow through the computer. However, pipelining encounters problems when

non-sequential instructions (branches) are encountered. The final part of Chapter 7 looks at how the problems caused by branch instructions can be limited.

Chapter 8 continues from where Chapter 7 left off. Here we look at how the performance of processors has been raised by executing instructions in parallel and even out-of-order. In Chapter 8, I introduce the very long instruction word computer (VLIW), which bundles several operations into a single instruction and executes them in parallel. As an example of modern high-performance computing, I introduce the remarkable IA64 Itanium computer that incorporates several powerful ways of accelerating the processor. You could even call the Itanium a computer architecture course on a chip.

Part IV: The Computer System turns our attention away from the central processing unit and devotes four chapters to the computer system, concentrating on memory, buses, and input/output. Chapter 9 focuses on two related topics: cache and virtual memory. Although the capacity of memory systems has expanded rapidly over the last few decades, their speed or access time has not improved at anything like the same rate as the CPUs. This situation has led to a bottleneck where memories can't provide data at the speed the CPUs can process it. Cache technology has been developed to make the best of a bad job by using a small amount of fast memory to do the work of a large amount of fast memory. I show how cache memories operate and describe their principal features. Chapter 9 also looks at virtual memory, which integrates main store with disk storage.

Chapters 10 and 11 cover memory technologies ranging from static semiconductor memory to disk and optical storage technologies.

Chapter 12 concludes Part IV by looking at the strategies used to get information into and out of the computer and then describing some of the modern high-speed interfaces used to support multimedia systems.

One of the key components of the modern computer is the bus, which distributes information between the various functional parts of the system. Indeed, the bus is one of the components most critical to the computer's performance. We look at the structure of computer buses, their functionality, and the way in which they permit competing devices to access the bus in multiprocessor systems. Finally, I describe one of the world's most popular high-performance buses: the PCI bus found in PCs.

Part V: The Near Future looks at the next direction in computer architecture, multiprocessing. Once upon a time, you could make a powerful computer by interconnecting an array of individual processors. Such computers had astronomical price tags. Today, it's possible to fabricate several processors on a single chip. Such multi-core processors are becoming increasingly popular, because they both increase processing power and reduce the amount of electrical energy needed. In Chapter 13, I provide a broad overview of multiprocessing covering the background, typical devices, interconnection topology, and the software of parallel processing.

Supplements and Resources

This book comes with a host of support material for both the instructor and the student. The supplements are housed on the book's companion website. To access the additional course materials, please visit www.cengagebrain.com. At the cengagebrain.com home page, search for the ISBN of your title (from the back cover of your book) using the search box at the top of the page. This will take you to the product page where these resources can be found.

Instructor Resources

An instructor's solutions manual (ISM) is available in both print and digital formats. The digital version is available to registered instructors at the publisher's website. This website also includes both a full set of PowerPoint slides containing all graphical images and tables in the text, and a set of LectureBuilder PowerPoint slides of all equations and example problems.

Student Resources

The student resources for this product include:

- a student workbook that provides a detailed introduction to writing programs in ARM processor assembly language and then running them on a simulator;
- additional study material, including a chapter on computer history and an overview of Karnaugh maps;
- useful links, including a link to download the student version of the Kiel simulator from ARM;
- code segments from the book; and
- handouts for all the lecture slides posted on the instructor website.

Acknowledgments

No one writes a textbook in a vacuum. All subjects have a history, background, and culture, and computer architecture is no exception. An author can go along with the flow or take a new direction. I could not have written this book without the contribution of all those texts that have preceded this, for example, the texts I used to study computer architecture myself when I was a student. Equally, I have to acknowledge the countless researchers that contributed to the body of knowledge making up computer architecture. The role of the writer is to take all of this information and create a path for students to follow. The writer has to decide what information is important and what can be omitted, what trends should be followed, and what trends can be relegated to the background. However, the writer is indebted to all those who have contributed to the body of knowledge.

Many people are involved with the production of a book as complex as this, but one stands out, the acquisitions editor. It was Swati Meherishi at Cengage Learning who began the long process of transforming a rough manuscript into the final polished article. An acquisitions editor has the ability to see beyond the manuscript and to understand how the book will fit into a complicated market. Swati had sufficient faith in me to help me endure hours of endless writing and rewriting. I'd like to thank Swati for putting up with me. The other key person in the creation of this book is the developmental editor, Amy Hill. Amy has spent a lot of time going through my manuscript to improve its structure and the way in which it all fits together. She has provided me with unfailingly good advice and gentle guidance. I appreciate the hard work and dedication of the entire team at Cengage Learning; without them this project would have remained a set of files on my hard disk. Also deserving of thanks are Rose Kernan and her team at RPK Editorial Services for their smooth management of all production related tasks, and Kristiina Paul for painstakingly researching and obtaining permissions for all third party material used in the book.

I must thank all the reviewers and editors that helped review and correct the manuscript. They made great suggestions about ways of improving it and helped point me in the right direction when I had misinterpreted source material.

I would like to acknowledge the technical reviewers and those who carefully read the text for errors and omissions. In particular, Loren Schwiebert did a great job of debugging Chapter 3. Sohum Sohoni of Oklahoma State University painstakingly provided a technical review of the entire manuscript. I must thank him for his hard work, especially for his excellent guidance on Chapter 8. My thanks also to Shuo Qin of the University of Southern California, who worked all the homework problems and checked the instructor's solution manual for errors.

The manuscript was reviewed at various stages of development by a number of instructors. I am thankful to all of them for their constructive comments. Below are the names of those who chose to be acknowledged:

Mokhtar Aboelaze York University

Manoj Franklin

University of Maryland-College Park

Israel Koren

University of Massachusetts-Amherst

Mikko Lipasti

University of Wisconsin-Madison

Rabi Mahapatra Texas A&M University

Xiannong Meng

Bucknell University

Prabhat Mishra University of Florida

William Mongan Drexel University

Vojin G. Oklobdzija Erik Jonsson School of Engineering,

University of Texas-Dallas

Soner Onder

Michigan Technological University

Füsun Özgüner Ohio State University

Richard J. Povinelli Marquette University Norman Ramsey Tufts University

Bill Reid

Clemson University

William H. Robinson Vanderbilt University

Carolyn J C Schauble Colorado State University

Aviral Shrivastava Arizona State

Sohum Sohoni

Oklahoma State University

Nozar Tabrizi Kettering University

Dean Tullsen

University of California-San Diego

Charles Weems Univ. of Mass.

Bilal Zafar

University of Southern California

Huiyang Zhou

North Carolina State University

Finally, I'd like to thank my wife Sue for her help in debugging the manuscript. Although Sue does not have a technical background, she has been very helpful in removing some of the ambiguities and clumsiness in my use of English.

Feedback from the readers, both critical and appreciative, is welcome. I would be particularly interested in suggestions for additional material that can be added to the companion website in order to help students with their studies. Please send your comments, concerns, and suggestions to globalengineering@cengage.com. You may also contact me directly at alanclements@ntlworld.com.

ALAN CLEMENTS