

- AMD Barcelona microprocessor, Google WSC server, **467**
- AMD Fusion, L-52
- AMD K-5, L-30
- AMD Opteron
 - address translation, B-38
 - Amazon Web Services, 457
 - architecture, 15
 - cache coherence, 361
 - data cache example, B-12 to B-15, **B-13**
 - Google WSC servers, 468–469
 - inclusion, 398
 - manufacturing cost, **62**
 - misses per instruction, **B-15**
 - MOESI protocol, 362
 - multicore processor performance, 400–401
 - multilevel exclusion, B-35
 - NetApp FAS6000 filer, D-42
 - paged virtual memory example, B-54 to B-57
 - vs. Pentium protection, B-57
 - real-world server considerations, 52–55
 - server energy savings, **25**
 - snooping limitations, 363–364
 - SPEC benchmarks, **43**
 - TLB during address translation, **B-47**
- AMD processors
 - architecture flaws vs. success, A-45
 - GPU computing history, L-52
 - power consumption, F-85
 - recent advances, L-33
 - RISC history, L-22
 - shared-memory multiprocessing workload, 378
 - terminology, 313–315
 - tournament predictors, 164
 - Virtual Machines, 110
 - VMMs, 129
- Amortization of overhead, sorting case study, D-64 to D-67
- AMPS, *see* Advanced mobile phone service (AMPS)
- Andreessen, Marc, F-98
- Android OS, **324**
- Annulling delayed branch, instructions, **K-25**
- Antenna, radio receiver, **E-23**
- Antialiasing, address translation, B-38
- Antidependences
 - compiler history, L-30 to L-31
 - definition, 152
 - finding, H-7 to H-8
 - loop-level parallelism calculations, 320
 - MIPS scoreboarding, C-72, C-79
- Apogee Software, **A-44**
- Apollo DN 10000, L-30
- Apple iPad
 - ARM Cortex-A8, 114
 - memory hierarchy basics, 78
- Application binary interface (ABI), control flow instructions, A-20
- Application layer, definition, **F-82**
- Applied Minds, L-74
- Arbitration algorithm
 - collision detection, F-23
 - commercial interconnection networks, **F-56**
 - examples, **F-49**
 - Intel SCCC, F-70
 - interconnection networks, F-21 to F-22, F-27, F-49 to F-50
 - network impact, F-52 to F-55
 - SAN characteristics, **F-76**
 - switched-media networks, F-24
 - switch microarchitecture, F-57 to F-58
 - switch microarchitecture pipelining, F-60
 - system area network history, F-100
- Architect-compiler writer relationship, A-29 to A-30
- Architecturally visible registers, register renaming vs. ROB, 208–209
- Architectural Support for Compilers and Operating Systems (ASPLOS), L-11
- Architecture, *see also* Computer architecture; CUDA (Compute Unified Device Architecture); Instruction set architecture (ISA); Vector architectures
 - compiler writer-architect relationship, A-29 to A-30
 - definition, 15
 - heterogeneous, 262
 - microarchitecture, 15–16, 247–254
 - stack, A-3, A-27, A-44 to A-45
- Areal density, disk storage, D-2
- Argument pointer, VAX, K-71
- Arithmetic intensity
 - as FP operation, **286**, 286–288
 - Roofline model, **326**, 326–327
- Arithmetic/logical instructions
 - desktop RISCs, **K-11**, **K-22**
 - embedded RISCs, **K-15**, **K-24**
 - Intel 80x86, **K-49**, **K-53**
 - SPARC, **K-31**
 - VAX, **B-73**
- Arithmetic-logical units (ALUs)
 - ARM Cortex-A8, **234**, 236
 - basic MIPS pipeline, C-36
 - branch condition evaluation, **A-19**
 - data forwarding, **C-40** to **C-41**
 - data hazards requiring stalls, C-19 to C-20
 - data hazard stall minimization, C-17 to C-19
 - DSP media extensions, E-10
 - effective address cycle, C-6
 - hardware-based execution, 185
 - hardware-based speculation, 200–201, **201**
 - IA-64 instructions, **H-35**
 - immediate operands, **A-12**
 - integer division, J-54
 - integer multiplication, J-48
 - integer shifting over zeros, J-45 to J-46
 - Intel Core i7, 238
 - ISA operands, A-4 to A-5
 - ISA performance and efficiency prediction, 241
 - load interlocks, **C-39**
 - microarchitectural techniques case study, 253
 - MIPS operations, A-35, **A-37**
 - MIPS pipeline control, C-38 to C-39
 - MIPS pipeline FP operations, C-52 to C-53
 - MIPS R4000, C-65
 - operand forwarding, **C-19**
 - operands per instruction example, **A-6**
 - parallelism, 45

- Arithmetic-logical units (*continued*)
 - pipeline branch issues, C-39 to C-41
 - pipeline execution rate, C-10 to C-11
 - power/DLP issues, 322
 - RISC architectures, K-5
 - RISC classic pipeline, C-7
 - RISC instruction set, C-4
 - simple MIPS implementation, C-31 to C-33
 - TX-2, L-49
- ARM (Advanced RISC Machine)
 - addressing modes, K-5, **K-6**
 - arithmetic/logical instructions, **K-15, K-24**
 - characteristics, **K-4**
 - condition codes, K-12 to K-13
 - constant extension, **K-9**
 - control flow instructions, 14
 - data transfer instructions, **K-23**
 - embedded instruction format, **K-8**
 - GPU computing history, L-52
 - ISA class, 11
 - memory addressing, 11
 - multiply-accumulate, **K-20**
 - operands, 12
 - RISC instruction set lineage, **K-43**
 - unique instructions, K-36 to K-37
- ARM AMBA, OCNs, F-3
- ARM Cortex-A8
 - dynamic scheduling, 170
 - ILP concepts, 148
 - instruction decode, **234**
 - ISA performance and efficiency prediction, 241–243
 - memory access penalty, **117**
 - memory hierarchy design, 78, 114–117, **115**
 - memory performance, 115–117
 - multibanked caches, 86
 - overview, 233
 - pipeline performance, 233–236, **235**
 - pipeline structure, **232**
 - processor comparison, **242**
 - way prediction, 81
- ARM Cortex-A9
 - vs. A8 performance, **236**
 - Tegra 2, mobile vs. server GPUs, 323–324, **324**
- ARM Thumb
 - addressing modes, **K-6**
 - arithmetic/logical instructions, **K-24**
 - characteristics, **K-4**
 - condition codes, K-14
 - constant extension, **K-9**
 - data transfer instructions, **K-23**
 - embedded instruction format, **K-8**
 - ISAs, 14
 - multiply-accumulate, **K-20**
 - RISC code size, A-23
 - unique instructions, K-37 to K-38
- ARPA (Advanced Research Project Agency)
 - LAN history, F-99 to F-100
 - WAN history, F-97
- ARPANET, WAN history, F-97 to F-98
- Array multiplier
 - example, **J-50**
 - integers, **J-50**
 - multipass system, **J-51**
- Arrays
 - access age, **91**
 - blocking, 89–90
 - bubble sort procedure, **K-76**
 - cluster server outage/anomaly statistics, **435**
 - examples, **90**
 - FFT kernel, I-7
 - Google WSC servers, 469
 - Layer 3 network linkage, **445**
 - loop interchange, 88–89
 - loop-level parallelism
 - dependences, 318–319
 - ocean application, I-9 to I-10
 - recurrences, H-12
 - WSC memory hierarchy, 445
 - WSCs, 443
- Array switch, WSCs, 443–444
- ASC, *see* Advanced Simulation and Computing (ASC) program
- ASCI, *see* Accelerated Strategic Computing Initiative (ASCI)
- ASCII character format, 12, A-14
- ASC Purple, F-67, F-100
- ASI, *see* Advanced Switching Interconnect (ASI)
- ASPLOS, *see* Architectural Support for Compilers and Operating Systems (ASPLOS)
- Assembly language, 2
- Association of Computing Machinery (ACM), L-3
- Associativity, *see also* Set associativity
 - cache block, B-9 to B-10, **B-10**
 - cache optimization, B-22 to B-24, B-26, B-28 to B-30
 - cloud computing, 460–461
 - loop-level parallelism, 322
 - multilevel inclusion, 398
 - Opteron data cache, B-14
 - shared-memory multiprocessors, **368**
- Astronautics ZS-1, L-29
- Asynchronous events, exception requirements, C-44 to C-45
- Asynchronous I/O, storage systems, D-35
- Asynchronous Transfer Mode (ATM)
 - interconnection networks, F-89
 - LAN history, F-99
 - packet format, **F-75**
 - total time statistics, **F-90**
 - VOQs, F-60
 - as WAN, F-79
 - WAN history, F-98
 - WANs, F-4
- ATA (Advanced Technology Attachment) disks
 - Berkeley's Tertiary Disk project, D-12
 - disk storage, D-4
 - historical background, L-81
 - power, D-5
 - RAID 6, D-9
 - server energy savings, **25**
- Atanasoff, John, L-5
- Atanasoff Berry Computer (ABC), L-5
- ATI Radeon 9700, L-51
- Atlas computer, L-9
- ATM, *see* Asynchronous Transfer Mode (ATM)
- ATM systems
 - server benchmarks, 41
 - TP benchmarks, D-18

- Atomic exchange
 - lock implementation, 389–390
 - synchronization, 387–388
 - Atomic instructions
 - barrier synchronization, **I-14**
 - Core i7, 329
 - Fermi GPU, 308
 - T1 multithreading uncore
 - performance, **229**
 - Atomicity-consistency-isolation-durability (ACID), vs. WSC storage, 439
 - Atomic operations
 - cache coherence, 360–361
 - snooping cache coherence
 - implementation, 365
 - “Atomic swap,” definition, K-20
 - Attributes field, IA-32 descriptor table, B-52
 - Autoincrement deferred addressing, VAX, K-67
 - Autonet, F-48
 - Availability
 - commercial interconnection networks, F-66
 - computer architecture, 11, 15
 - computer systems, D-43 to D-44, **D-44**
 - data on Internet, 344
 - fault detection, 57–58
 - I/O system design/evaluation, D-36
 - loop-level parallelism, 217–218
 - mainstream computing classes, 5
 - modules, 34
 - open-source software, 457
 - RAID systems, 60
 - as server characteristic, 7
 - servers, 16
 - source operands, C-74
 - WSCs, 8, 433–435, 438–439
 - Average instruction execution time, L-6
 - Average Memory Access Time (AMAT)
 - block size calculations, B-26 to B-28
 - cache optimizations, B-22, B-26 to B-32, B-36
 - cache performance, B-16 to B-21
 - calculation, B-16 to B-17
 - centralized shared-memory architectures, 351–352
 - definition, B-30 to B-31
 - memory hierarchy basics, 75–76
 - miss penalty reduction, B-32
 - via miss rates, **B-29**, B-29 to B-30
 - as processor performance predictor, B-17 to B-20
 - Average reception factor
 - centralized switched networks, F-32
 - multi-device interconnection networks, F-26
 - AVX, *see* Advanced Vector Extensions (AVX)
 - AWS, *see* Amazon Web Services (AWS)
- ## B
- Back-off time, shared-media networks, F-23
 - Backpressure, congestion management, F-65
 - Backside bus, centralized shared-memory multiprocessors, 351
 - Balanced systems, sorting case study, D-64 to D-67
 - Balanced tree, MINs with nonbliking, F-34
 - Bandwidth, *see also* Throughput arbitration, F-49
 - and cache miss, B-2 to B-3
 - centralized shared-memory multiprocessors, 351–352
 - communication mechanism, I-3
 - congestion management, F-64 to F-65
 - Cray Research T3D, **F-87**
 - DDR DRAMS and DIMMS, **101**
 - definition, F-13
 - DSM architecture, 379
 - Ethernet and bridges, **F-78**
 - FP arithmetic, J-62
 - GDRAM, 322–323
 - GPU computation, 327–328
 - GPU Memory, 327
 - ILP instruction fetch
 - basic considerations, 202–203
 - branch-target buffers, 203–206
 - integrated units, 207–208
 - return address predictors, 206–207
 - interconnection networks, **F-28**
 - multi-device networks, F-25 to F-29
 - performance considerations, F-89
 - two-device networks, F-12 to F-20
 - vs. latency, 18–19, **19**
 - memory, and vector performance, 332
 - memory hierarchy, 126
 - network performance and topology, F-41
 - OCN history, F-103
 - performance milestones, **20**
 - point-to-point links and switches, D-34
 - routing, F-50 to F-52
 - routing/arbitration/switching impact, F-52
 - shared- vs. switched-media networks, **F-22**
 - SMP limitations, 363
 - switched-media networks, F-24
 - system area network history, F-101
 - vs. TCP/IP reliance, F-95
 - and topology, F-39
 - vector load/store units, 276–277
 - WSC memory hierarchy, 443–444, **444**
- Bandwidth gap, disk storage, D-3
- Banerjee, Uptal, L-30 to L-31
- Bank busy time, vector memory systems, G-9
- Banked memory, *see also* Memory banks
 - and graphics memory, 322–323
 - vector architectures, **G-10**
- Banks, Fermi GPUs, 297
- Barcelona Supercomputer Center, **F-76**
- Barnes
 - characteristics, I-8 to I-9
 - distributed-memory multiprocessor, **I-32**
 - symmetric shared-memory multiprocessors, I-22, **I-23**, I-25

- Barnes-Hut n -body algorithm, basic concept, I-8 to I-9
- Barriers
 - commercial workloads, **370**
 - Cray X1, G-23
 - fetch-and-increment, I-20 to I-21
 - hardware primitives, 387
 - large-scale multiprocessor
 - synchronization, I-13 to I-16, **I-14**, **I-16**, **I-19**, I-20
 - synchronization, 298, **313**, 329
- BARRNet, *see* Bay Area Research Network (BARRNet)
- Based indexed addressing mode, Intel 80x86, K-49, **K-58**
- Base field, IA-32 descriptor table, B-52 to B-53
- Base station
 - cell phones, E-23
 - wireless networks, E-22
- Basic block, ILP, 149
- Batch processing workloads
 - WSC goals/requirements, 433
 - WSC MapReduce and Hadoop, 437–438
- Bay Area Research Network (BARRNet), **F-80**
- BBN Butterfly, L-60
- BBN Monarch, L-60
- Before rounding rule, J-36
- Benchmarking, *see also specific benchmark suites*
 - desktop, 38–40
 - EEMBC, **E-12**
 - embedded applications
 - basic considerations, E-12
 - power consumption and efficiency, E-13
 - fallacies, 56
 - instruction set operations, **A-15**
 - as performance measurement, 37–41
 - real-world server considerations, 52–55
 - response time restrictions, **D-18**
 - server performance, 40–41
 - sorting case study, D-64 to D-67
- Beneš topology
 - centralized switched networks, F-33
 - example, **F-33**
- BER, *see* Bit error rate (BER)
- Berkeley's Tertiary Disk project
 - failure statistics, **D-13**
 - overview, D-12
 - system log, **D-43**
- Berners-Lee, Tim, F-98
- Bertram, Jack, L-28
- Best-case lower bounds, multi-device
 - interconnection networks, F-25
- Best-case upper bounds
 - multi-device interconnection networks, F-26
 - network performance and topology, F-41
- Between instruction exceptions, definition, C-45
- Biased exponent, J-15
- Bidirectional multistage
 - interconnection networks
 - Beneš topology, **F-33**
 - characteristics, F-33 to F-34
 - SAN characteristics, **F-76**
- Bidirectional rings, topology, F-35 to F-36
- Big Endian
 - interconnection networks, F-12
 - memory address interpretation, A-7
 - MIPS core extensions, K-20 to K-21
 - MIPS data transfers, A-34
- Bigtable (Google), 438, 441
- BINAC, L-5
- Binary code compatibility
 - embedded systems, E-15
 - VLIW processors, 196
- Binary-coded decimal, definition, A-14
- Binary-to-decimal conversion, FP
 - precisions, J-34
- Bing search
 - delays and user behavior, **451**
 - latency effects, 450–452
 - WSC processor cost-performance, 473
- Bisection bandwidth
 - as network cost constraint, F-89
 - network performance and topology, F-41
 - NEWS communication, F-42
 - topology, F-39
- Bisection bandwidth, WSC array switch, 443
- Bisection traffic fraction, network
 - performance and topology, F-41
- Bit error rate (BER), wireless networks, E-21
- Bit rot, case study, D-61 to D-64
- Bit selection, block placement, B-7
- Black box network
 - basic concept, F-5 to F-6
 - effective bandwidth, F-17
 - performance, F-12
 - switched-media networks, F-24
 - switched network topologies, F-40
- Block addressing
 - block identification, B-7 to B-8
 - interleaved cache banks, **86**
 - memory hierarchy basics, 74
- Blocked floating point arithmetic, DSP, E-6
- Block identification
 - memory hierarchy considerations, B-7 to B-9
 - virtual memory, B-44 to B-45
- Blocking
 - benchmark fallacies, 56
 - centralized switched networks, F-32
 - direct networks, F-38
 - HOL, *see* Head-of-line (HOL) blocking
 - network performance and topology, F-41
- Blocking calls, shared-memory multiprocessor workload, 369
- Blocking factor, definition, 90
- Block multithreading, definition, L-34
- Block offset
 - block identification, B-7 to B-8
 - cache optimization, B-38
 - definition, B-7 to B-8
 - direct-mapped cache, **B-9**
 - example, B-9
 - main memory, B-44
 - Opteron data cache, **B-13**, B-13 to B-14

- Block placement
 - memory hierarchy considerations, B-7
 - virtual memory, B-44
- Block replacement
 - memory hierarchy considerations, B-9 to B-10
 - virtual memory, B-45
- Blocks, *see also* Cache block; Thread Block
 - ARM Cortex-A8, **115**
 - vs. bytes per reference, **378**
 - compiler optimizations, 89–90
 - definition, B-2
 - disk array deconstruction, D-51, **D-55**
 - disk deconstruction case study, D-48 to D-51
 - global code scheduling, H-15 to H-16
 - L3 cache size, misses per instruction, **371**
 - LU kernel, I-8
 - memory hierarchy basics, 74
 - memory in cache, **B-61**
 - placement in main memory, B-44
 - RAID performance prediction, D-57 to D-58
 - TI TMS320C55 DSP, E-8
 - uncached state, 384
- Block servers, vs. filers, D-34 to D-35
- Block size
 - vs. access time, **B-28**
 - memory hierarchy basics, 76
 - vs. miss rate, **B-27**
- Block transfer engine (BLT)
 - Cray Research T3D, **F-87**
 - interconnection network protection, F-87
- BLT, *see* Block transfer engine (BLT)
- Body of Vectorized Loop
 - definition, **292, 313**
 - GPU hardware, 295–296, 311
 - GPU Memory structure, **304**
 - NVIDIA GPU, 296
 - SIMD Lane Registers, **314**
 - Thread Block Scheduler, **314**
- Boggs, David, F-99
- BOMB, L-4
- Booth recoding, J-8 to J-9, **J-9**, J-10 to J-11
- chip comparison, J-60 to J-61
- integer multiplication, **J-49**
- Bose-Einstein formula, definition, 30
- Bounds checking, segmented virtual memory, B-52
- Branch byte, VAX, K-71
- Branch delay slot
 - characteristics, C-23 to C-25
 - control hazards, C-41
 - MIPS R4000, C-64
 - scheduling, **C-24**
- Branches
 - canceling, C-24 to C-25
 - conditional branches, 300–303, **A-17, A-19 to A-20, A-21**
 - control flow instructions, A-16, A-18
 - delayed, **C-23**
 - delay slot, **C-65**
 - IBM 360, K-86 to K-87
 - instructions, **K-25**
 - MIPS control flow instructions, A-38
 - MIPS operations, A-35
 - nullifying, C-24 to C-25
 - RISC instruction set, C-5
 - VAX, K-71 to K-72
 - WCET, E-4
- Branch folding, definition, 206
- Branch hazards
 - basic considerations, C-21
 - penalty reduction, C-22 to C-25
 - pipeline issues, C-39 to C-42
 - scheme performance, C-25 to C-26
 - stall reduction, **C-42**
- Branch history table, basic scheme, C-27 to C-30
- Branch offsets, control flow instructions, A-18
- Branch penalty
 - examples, **205**
 - instruction fetch bandwidth, 203–206
 - reduction, C-22 to C-25
 - simple scheme examples, **C-25**
- Branch prediction
 - accuracy, **C-30**
 - branch cost reduction, 162–167
 - correlation, 162–164
 - cost reduction, C-26
 - dynamic, C-27 to C-30
 - early schemes, L-27 to L-28
 - ideal processor, 214
 - ILP exploitation, **201**
 - instruction fetch bandwidth, 205
 - integrated instruction fetch units, 207
 - Intel Core i7, 166–167, 239–241
 - misprediction rates on SPEC89, **166**
 - static, C-26 to C-27
 - trace scheduling, H-19
 - two-bit predictor comparison, **165**
- Branch-prediction buffers, basic considerations, C-27 to C-30, **C-29**
- Branch registers
 - IA-64, H-34
 - PowerPC instructions, K-32 to K-33
- Branch stalls, MIPS R4000 pipeline, **C-67**
- Branch-target address
 - branch hazards, **C-42**
 - MIPS control flow instructions, A-38
 - MIPS pipeline, C-36, **C-37**
 - MIPS R4000, C-25
 - pipeline branches, C-39
 - RISC instruction set, C-5
- Branch-target buffers
 - ARM Cortex-A8, 233
 - branch hazard stalls, **C-42**
 - example, **203**
 - instruction fetch bandwidth, 203–206
 - instruction handling, **204**
 - MIPS control flow instructions, A-38
- Branch-target cache, *see* Branch-target buffers
- Brewer, Eric, L-73
- Bridges
 - and bandwidth, **F-78**
 - definition, F-78
- Bubbles
 - and deadlock, F-47
 - routing comparison, **F-54**
 - stall as, C-13
- Bubble sort, code example, **K-76**
- Buckets, D-26
- Buffered crossbar switch, switch microarchitecture, F-62
- Buffered wormhole switching, F-51

- definition, B-2
- example calculation, B-5
- Cache latency, nonblocking cache, 83–84
- Cache miss
 - and average memory access time, B-17 to B-20
 - block replacement, **B-10**
 - definition, B-2
 - distributed-memory multiprocessors, **I-32**
 - example calculations, 83–84
 - Intel Core i7, 122
 - interconnection network, F-87
 - large-scale multiprocessors, I-34 to I-35
 - nonblocking cache, **84**
 - single vs. multiple thread executions, **228**
 - WCET, E-4
- Cache-only memory architecture (COMA), L-61
- Cache optimizations
 - basic categories, B-22
 - basic optimizations, **B-40**
 - case studies, 131–133
 - compiler-controlled prefetching, 92–95
 - compiler optimizations, 87–90
 - critical word first, 86–87
 - energy consumption, **81**
 - hardware instruction prefetching, 91–92, **92**
 - hit time reduction, B-36 to B-40
 - miss categories, B-23 to B-26
 - miss penalty reduction
 - via multilevel caches, B-30 to B-35
 - read misses vs. writes, B-35 to B-36
 - miss rate reduction
 - via associativity, B-28 to B-30
 - via block size, B-26 to B-28
 - via cache size, B-28
 - multibanked caches, 85–86, **86**
 - nonblocking caches, 83–85, **84**
 - overview, 78–79
 - pipelined cache access, 82
 - simple first-level caches, 79–80
 - techniques overview, **96**
 - way prediction, 81–82
 - write buffer merging, 87, **88**
- Cache organization
 - blocks, B-7, **B-8**
 - Opteron data cache, B-12 to B-13, **B-13**
 - optimization, B-19
 - performance impact, B-19
- Cache performance
 - average memory access time, B-16 to B-20
 - basic considerations, B-3 to B-6, B-16
 - basic equations, **B-22**
 - basic optimizations, **B-40**
 - cache optimization, **96**
 - case study, 131–133
 - example calculation, B-16 to B-17
 - out-of-order processors, B-20 to B-22
 - prediction, 125–126
- Cache prefetch, cache optimization, 92
- Caches, *see also* Memory hierarchy
 - access time vs. block size, **B-28**
 - AMD Opteron example, B-12 to B-15, **B-13**, **B-15**
 - basic considerations, B-48 to B-49
 - coining of term, L-11
 - definition, B-2
 - early work, L-10
 - embedded systems, E-4 to E-5
 - Fermi GPU architecture, 306
 - ideal processor, 214
 - ILP for realizable processors, 216–218
 - Itanium 2, H-42
 - multichip multicore multiprocessor, **419**
 - parameter ranges, **B-42**
 - Sony PlayStation 2 Emotion Engine, E-18
 - vector processors, G-25
 - vs. virtual memory, B-42 to B-43
- Cache size
 - and access time, **77**
 - AMD Opteron example, B-13 to B-14
 - energy consumption, **81**
 - highly parallel memory systems, 133
 - memory hierarchy basics, 76
 - misses per instruction, **126**, **371**
 - miss rate, **B-24 to B-25**
 - vs. miss rate, **B-27**
- miss rate reduction, B-28
- multilevel caches, **B-33**
- and relative execution time, B-34
- scientific workloads
 - distributed-memory multiprocessors, **I-29 to I-31**
 - symmetric shared-memory multiprocessors, I-22 to I-23, **I-24**
 - shared-memory multiprocessing workload, 376
 - virtually addressed, **B-37**
- CACTI
 - cache optimization, 79–80, **81**
 - memory access times, **77**
- Caller saving, control flow instructions, A-19 to A-20
- Call gate
 - IA-32 segment descriptors, B-53
 - segmented virtual memory, B-54
- Calls
 - compiler structure, A-25 to A-26
 - control flow instructions, **A-17**, A-19 to A-21
 - CUDA Thread, 297
 - dependence analysis, 321
 - high-level instruction set, A-42 to A-43
 - Intel 80x86 integer operations, K-51
 - invocation options, A-19
 - ISAs, 14
 - MIPS control flow instructions, A-38
 - MIPS registers, 12
 - multiprogrammed workload, 378
 - NVIDIA GPU Memory structures, 304–305
 - return address predictors, 206
 - shared-memory multiprocessor workload, 369
 - user-to-OS gates, B-54
 - VAX, K-71 to K-72
- Canceling branch, branch delay slots, C-24 to C-25
- Canonical form, AMD64 paged virtual memory, B-55
- Capabilities, protection schemes, L-9 to L-10

- Capacity misses
 - blocking, 89–90
 - and cache size, **B-24**
 - definition, B-23
 - memory hierarchy basics, 75
 - scientific workloads on symmetric
 - shared-memory multiprocessors, I-22, **I-23**, I-24
 - shared-memory workload, 373
- CAPEX, *see* Capital expenditures (CAPEX)
- Capital expenditures (CAPEX)
 - WSC costs, 452–455, **453**
 - WSC Flash memory, 475
 - WSC TCO case study, 476–478
- Carrier sensing, shared-media networks, F-23
- Carrier signal, wireless networks, E-21
- Carry condition code, MIPS core, K-9 to K-16
- Carry-in, carry-skip adder, J-42
- Carry-lookahead adder (CLA)
 - chip comparison, J-60
 - early computer arithmetic, J-63
 - example, **J-38**
 - integer addition speedup, J-37 to J-41
 - with ripple-carry adder, **J-42**
 - tree, **J-40 to J-41**
- Carry-out
 - carry-lookahead circuit, **J-38**
 - floating-point addition speedup, J-25
- Carry-propagate adder (CPA)
 - integer multiplication, J-48, J-51
 - multipass array multiplier, **J-51**
- Carry-save adder (CSA)
 - integer division, J-54 to J-55
 - integer multiplication, J-47 to J-48, **J-48**
- Carry-select adder
 - characteristics, J-43 to J-44
 - chip comparison, J-60
 - example, **J-43**
- Carry-skip adder (CSA)
 - characteristics, J-41 to J-43
 - example, **J-42**, **J-44**
- CAS, *see* Column access strobe (CAS)
- Case statements
 - control flow instruction addressing modes, A-18
 - return address predictors, 206
- Case studies
 - advanced directory protocol, 420–426
 - cache optimization, 131–133
 - cell phones
 - block diagram, **E-23**
 - Nokia circuit board, **E-24**
 - overview, E-20
 - radio receiver, **E-23**
 - standards and evolution, E-25
 - wireless communication challenges, **E-21**
 - wireless networks, E-21 to E-22
 - chip fabrication cost, 61–62
 - computer system power consumption, 63–64
 - directory-based coherence, 418–420
 - dirty bits, D-61 to D-64
 - disk array deconstruction, D-51 to D-55, **D-52 to D-55**
 - disk deconstruction, D-48 to D-51, **D-50**
 - highly parallel memory systems, 133–136
 - instruction set principles, A-47 to A-54
 - I/O subsystem design, D-59 to D-61
 - memory hierarchy, B-60 to B-67
 - microarchitectural techniques, 247–254
 - pipelining example, C-82 to C-88
 - RAID performance prediction, D-57 to D-59
 - RAID reconstruction, D-55 to D-57
 - Sanyo VPC-SX500 digital camera, E-19
 - single-chip multicore processor, 412–418
 - Sony PlayStation 2 Emotion Engine, E-15 to E-18
 - sorting, D-64 to D-67
 - vector kernel on vector processor and GPU, 334–336
 - WSC resource allocation, 478–479
 - WSC TCO, 476–478
- CCD, *see* Charge-coupled device (CCD)
- C/C++ language
 - dependence analysis, H-6
 - GPU computing history, L-52
 - hardware impact on software development, 4
 - integer division/remainder, **J-12**
 - loop-level parallelism
 - dependences, 318, 320–321
 - NVIDIA GPU programming, 289
 - return address predictors, 206
- CDB, *see* Common data bus (CDB)
- CDC, *see* Control Data Corporation (CDC)
- CDF, datacenter, **487**
- CDMA, *see* Code division multiple access (CDMA)
- Cedar project, L-60
- Cell, Barnes-Hut *n*-body algorithm, I-9
- Cell phones
 - block diagram, **E-23**
 - embedded system case study
 - characteristics, E-22 to E-24
 - overview, E-20
 - radio receiver, **E-23**
 - standards and evolution, E-25
 - wireless network overview, E-21 to E-22
 - Flash memory, D-3
 - GPU features, **324**
 - Nokia circuit board, **E-24**
 - wireless communication
 - challenges, **E-21**
 - wireless networks, E-22
- Centralized shared-memory multiprocessors
 - basic considerations, 351–352
 - basic structure, 346–347, **347**
 - cache coherence, 352–353
 - cache coherence enforcement, 354–355
 - cache coherence example, 357–362
 - cache coherence extensions, 362–363
 - invalidate protocol
 - implementation, 356–357

- SMP and snooping limitations, 363–364
- snooping coherence
 - implementation, 365–366
 - snooping coherence protocols, 355–356
- Centralized switched networks
 - example, **F-31**
 - routing algorithms, F-48
 - topology, F-30 to F-34, **F-31**
- Centrally buffered switch,
 - microarchitecture, F-57
- Central processing unit (CPU)
 - Amdahl's law, 48
 - average memory access time, B-17
 - cache performance, B-4
 - coarse-grained multithreading, 224
 - early pipelined versions, L-26 to L-27
 - exception stopping/restarting, C-47
 - extensive pipelining, C-81
 - Google server usage, **440**
 - GPU computing history, L-52
 - vs. GPUs, 288
 - instruction set complications, C-50
 - MIPS implementation, C-33 to C-34
 - MIPS precise exceptions, C-59 to C-60
 - MIPS scoreboarding, C-77
 - performance measurement history, L-6
 - pipeline branch issues, C-41
 - pipelining exceptions, C-43 to C-46
 - pipelining performance, C-10
 - Sony PlayStation 2 Emotion Engine, E-17
 - SPEC server benchmarks, 40
 - TI TMS320C55 DSP, E-8
 - vector memory systems, **G-10**
- Central processing unit (CPU) time
 - execution time, 36
 - modeling, B-18
 - processor performance
 - calculations, B-19 to B-21
 - processor performance equation, 49–51
 - processor performance time, 49
- Cerf, Vint, F-97
- CERN, *see* European Center for Particle Research (CERN)
- CFM, *see* Current frame pointer (CFM)
- Chaining
 - convoys, DAXPY code, **G-16**
 - vector processor performance, G-11 to G-12, **G-12**
 - VMIPS, 268–269
- Channel adapter, *see* Network interface
- Channels, cell phones, E-24
- Character
 - floating-point performance, A-2
 - as operand type, A-13 to A-14
 - operand types/sizes, 12
- Charge-coupled device (CCD), Sanyo VPC-SX500 digital camera, E-19
- Checksum
 - dirty bits, D-61 to D-64
 - packet format, **F-7**
- Chillers
 - Google WSC, 466, 468
 - WSC containers, 464
 - WSC cooling systems, 448–449
- Chime
 - definition, **309**
 - GPUs vs. vector architectures, 308
 - multiple lanes, 272
 - NVIDIA GPU computational structures, 296
 - vector chaining, G-12
 - vector execution time, 269, G-4
 - vector performance, G-2
 - vector sequence calculations, 270
- Chip-crossing wire delay, F-70
- OCN history, F-103
- Chipkill
 - memory dependability, 104–105
 - WSCs, 473
- Choke packets, congestion
 - management, F-65
- Chunk
 - disk array deconstruction, D-51
 - Shear algorithm, **D-53**
- CIFS, *see* Common Internet File System (CIFS)
- Circuit switching
 - congestion management, F-64 to F-65
 - interconnected networks, F-50
- Circulating water system (CWS)
 - cooling system design, **448**
 - WSCs, 448
- CISC, *see* Complex Instruction Set Computer (CISC)
- CLA, *see* Carry-lookahead adder (CLA)
- Clean block, definition, B-11
- Climate Savers Computing Initiative,
 - power supply efficiencies, **462**
- Clock cycles
 - basic MIPS pipeline, C-34 to C-35
 - and branch penalties, **205**
 - cache performance, B-4
 - FP pipeline, **C-66**
 - and full associativity, B-23
 - GPU conditional branching, 303
 - ILP exploitation, 197, 200
 - ILP exposure, 157
 - instruction fetch bandwidth, 202–203
 - instruction steps, 173–175
 - Intel Core i7 branch predictor, 166
 - MIPS exceptions, C-48
 - MIPS pipeline, **C-52**
 - MIPS pipeline FP operations, C-52 to C-53
 - MIPS scoreboarding, C-77
 - miss rate calculations, B-31 to B-32
 - multithreading approaches, 225–226
 - pipelining performance, C-10
 - processor performance equation, 49
 - RISC classic pipeline, C-7
 - Sun T1 multithreading, 226–227
 - switch microarchitecture
 - pipelining, F-61
 - vector architectures, G-4
 - vector execution time, 269
 - vector multiple lanes, 271–273
 - VLIW processors, **195**
- Clock cycles per instruction (CPI)
 - addressing modes, A-10
 - ARM Cortex-A8, **235**
 - branch schemes, C-25 to C-26, **C-26**
 - cache behavior impact, B-18 to B-19
 - cache hit calculation, B-5
 - data hazards requiring stalls, C-20

- Clock cycles per instruction (*continued*)
 - extensive pipelining, C-81
 - floating-point calculations, 50–52
 - ILP concepts, 148–149, **149**
 - ILP exploitation, 192
 - Intel Core i7, 124, **240**, 240–241
 - microprocessor advances, L-33
 - MIPS R4000 performance, **C-69**
 - miss penalty reduction, B-32
 - multiprocessing/
 - multithreading-based performance, 398–400
 - multiprocessor communication
 - calculations, 350
 - pipeline branch issues, C-41
 - pipeline with stalls, C-12 to C-13
 - pipeline structural hazards, C-15 to C-16
 - pipelining concept, C-3
 - processor performance
 - calculations, 218–219
 - processor performance time, 49–51
 - and processor speed, 244
 - RISC history, L-21
 - shared-memory workloads, 369–370
 - simple MIPS implementation,
 - C-33 to C-34
 - structural hazards, C-13
 - Sun T1 multithreading uncore
 - performance, **229**
 - Sun T1 processor, **399**
 - Tomasulo's algorithm, 181
 - VAX 8700 vs. MIPS M2000, **K-82**
- Clock cycle time
 - and associativity, B-29
 - average memory access time, B-21 to B-22
 - cache optimization, B-19 to B-20, B-30
 - cache performance, B-4
 - CPU time equation, 49–50, B-18
 - MIPS implementation, **C-34**
 - miss penalties, 219
 - pipeline performance, C-12, C-14 to C-15
 - pipelining, C-3
 - shared- vs. switched-media networks, F-25
- Clock periods, processor performance
 - equation, 48–49
- Clock rate
 - DDR DRAMS and DIMMS, 101
 - ILP for realizable processors, 218
 - Intel Core i7, 236–237
 - microprocessor advances, L-33
 - microprocessors, **24**
 - MIPS pipeline FP operations, C-53
 - multicore processor performance, 400
 - and processor speed, **244**
- Clocks, processor performance
 - equation, 48–49
- Clock skew, pipelining performance, C-10
- Clock ticks
 - cache coherence, **391**
 - processor performance equation, 48–49
- Clos network
 - Beneš topology, **F-33**
 - as nonblocking, F-33
- Cloud computing
 - basic considerations, 455–461
 - clusters, 345
 - provider issues, 471–472
 - utility computing history, L-73 to L-74
- Clusters
 - characteristics, 8, **I-45**
 - cloud computing, 345
 - as computer class, **5**
 - containers, L-74 to L-75
 - Cray X1, G-22
 - Google WSC servers, 469
 - historical background, L-62 to L-64
 - IBM Blue Gene/L, I-41 to I-44, I-43 to I-44
 - interconnection network domains, F-3 to F-4
 - Internet Archive Cluster, *see* Internet Archive Cluster
 - large-scale multiprocessors, I-6
 - large-scale multiprocessor trends, L-62 to L-63
 - outage/anomaly statistics, **435**
 - power consumption, F-85
 - utility computing, L-73 to L-74
 - as WSC forerunners, 435–436, L-72 to L-73
 - WSC storage, 442–443
- Cm*, L-56
- C.mmp, L-56
- CMOS
 - DRAM, **99**
 - first vector computers, L-46, L-48
 - ripple-carry adder, J-3
 - vector processors, G-25 to G-27
- Coarse-grained multithreading,
 - definition, 224–226
- Cocke, John, L-19, L-28
- Code division multiple access (CDMA),
 - cell phones, E-25
- Code generation
 - compiler structure, A-25 to A-26, A-30
 - dependences, 220
 - general-purpose register computers, **A-6**
 - ILP limitation studies, 220
 - loop unrolling/scheduling, 162
- Code scheduling
 - example, **H-16**
 - parallelism, H-15 to H-23
 - superblock scheduling, H-21 to H-23, **H-22**
 - trace scheduling, H-19 to H-21, **H-20**
- Code size
 - architect-compiler considerations, A-30
 - benchmark information, A-2
 - comparisons, **A-44**
 - flawless architecture design, A-45
 - instruction set encoding, A-22 to A-23
 - ISA and compiler technology, A-43 to A-44
 - loop unrolling, 160–161
 - multiprogramming, 375–376
 - PMDs, 6
 - RISCs, A-23 to A-24
 - VAX design, A-45
 - VLIW model, 195–196
- Coefficient of variance, D-27
- Coerced exceptions
 - definition, C-45
 - exception types, **C-46**
- Coherence, *see* Cache coherence
- Coherence misses
 - definition, 366
 - multiprogramming, 376–377
 - role, 367
 - scientific workloads on symmetric
 - shared-memory multiprocessors, I-22
 - snooping protocols, 355–356

- Cold-start misses, definition, B-23
- Collision, shared-media networks, F-23
- Collision detection, shared-media networks, F-23
- Collision misses, definition, B-23
- Collocation sites, interconnection networks, F-85
- COLOSSUS, L-4
- Column access strobe (CAS), DRAM, 98–99
- Column major order
 - blocking, 89
 - stride, 278
- COMA, *see* Cache-only memory architecture (COMA)
- Combining tree, large-scale multiprocessor synchronization, I-18
- Command queue depth, *vs.* disk throughput, **D-4**
- Commercial interconnection networks
 - congestion management, F-64 to F-66
 - connectivity, F-62 to F-63
 - cross-company interoperability, F-63 to F-64
 - DECstation 5000 reboots, **F-69**
 - fault tolerance, F-66 to F-69
- Commercial workloads
 - execution time distribution, **369**
 - symmetric shared-memory multiprocessors, 367–374
- Commit stage, ROB instruction, 186–187, **188**
- Commodities
 - Amazon Web Services, 456–457
 - array switch, 443
 - cloud computing, 455
 - cost *vs.* price, 32–33
 - cost trends, 27–28, 32
 - Ethernet rack switch, 442
 - HPC hardware, 436
 - shared-memory multiprocessor, 441
 - WSCs, 441
- Commodity cluster, characteristics, I-45
- Common data bus (CDB)
 - dynamic scheduling with Tomasulo's algorithm, 172, 175
- FP unit with Tomasulo's algorithm, **185**
 - reservation stations/register tags, **177**
 - Tomasulo's algorithm, **180**, 182
- Common Internet File System (CIFS), D-35
- NetApp FAS6000 filer, D-41 to D-42
- Communication bandwidth, basic considerations, I-3
- Communication latency, basic considerations, I-3 to I-4
- Communication latency hiding, basic considerations, I-4
- Communication mechanism
 - adaptive routing, F-93 to F-94
 - internetworking, F-81 to F-82
 - large-scale multiprocessors
 - advantages, I-4 to I-6
 - metrics, I-3 to I-4
 - multiprocessor communication calculations, 350
 - network interfaces, F-7 to F-8
 - NEWS communication, F-42 to F-43
 - SMP limitations, 363
- Communication protocol, definition, F-8
- Communication subnets, *see* Interconnection networks
- Communication subsystems, *see* Interconnection networks
- Compare instruction, VAX, K-71
- Compares, MIPS core, K-9 to K-16
- Compare-select-store unit (CSSU), TI TMS320C55 DSP, E-8
- Compiler-controlled prefetching, miss penalty/rate reduction, 92–95
- Compiler optimizations
 - blocking, 89–90
 - cache optimization, 131–133
 - compiler assumptions, A-25 to A-26
 - and consistency model, 396
 - loop interchange, 88–89
 - miss rate reduction, 87–90
 - passes, **A-25**
 - performance impact, A-27
 - types and classes, **A-28**
- Compiler scheduling
 - data dependencies, 151
 - definition, C-71
 - hardware support, L-30 to L-31
 - IBM 360 architecture, 171
- Compiler speculation, hardware support
 - memory references, H-32
 - overview, H-27
 - preserving exception behavior, H-28 to H-32
- Compiler techniques
 - dependence analysis, H-7
 - global code scheduling, H-17 to H-18
 - ILP exposure, 156–162
 - vectorization, G-14
 - vector sparse matrices, G-12
- Compiler technology
 - and architecture decisions, A-27 to A-29
 - Cray X1, G-21 to G-22
 - ISA and code size, A-43 to A-44
 - multimedia instruction support, A-31 to A-32
 - register allocation, A-26 to A-27
 - structure, A-24 to A-26, **A-25**
- Compiler writer-architect relationship, A-29 to A-30
- Complex Instruction Set Computer (CISC)
 - RISC history, L-22
 - VAX as, K-65
- Compulsory misses
 - and cache size, **B-24**
 - definition, B-23
 - memory hierarchy basics, 75
 - shared-memory workload, 373
- Computation-to-communication ratios
 - parallel programs, I-10 to I-12
 - scaling, **I-11**
- Compute-optimized processors,
 - interconnection networks, F-88
- Computer aided design (CAD) tools,
 - cache optimization, 79–80
- Computer architecture, *see also* Architecture
 - coining of term, K-83 to K-84
 - computer design innovations, 4
 - defining, 11

- Computer architecture (*continued*)
 - definition, L-17 to L-18
 - exceptions, **C-44**
 - factors in improvement, 2
 - flawless design, K-81
 - flaws and success, K-81
 - floating-point addition, rules, **J-24**
 - goals/functions requirements, **15**, 15–16, **16**
 - high-level language, L-18 to L-19
 - instruction execution issues, K-81
 - ISA, 11–15
 - multiprocessor software
 - development, 407–409
 - parallel, 9–10
 - WSC basics, 432, 441–442
 - array switch, 443
 - memory hierarchy, 443–446
 - storage, 442–443
- Computer arithmetic
 - chip comparison, **J-58**, J-58 to J-61, **J-59** to **J-60**
 - floating point
 - exceptions, J-34 to J-35
 - fused multiply-add, J-32 to J-33
 - IEEE 754, **J-16**
 - iterative division, J-27 to J-31
 - and memory bandwidth, J-62
 - overview, J-13 to J-14
 - precisions, J-33 to J-34
 - remainder, J-31 to J-32
 - special values, J-16
 - special values and denormals, J-14 to J-15
 - underflow, J-36 to J-37, J-62
 - floating-point addition
 - denormals, J-26 to J-27
 - overview, J-21 to J-25
 - speedup, J-25 to J-26
 - floating-point multiplication
 - denormals, J-20 to J-21
 - examples, **J-19**
 - overview, J-17 to J-20
 - rounding, **J-18**
 - integer addition speedup
 - carry-lookahead, J-37 to J-41
 - carry-lookahead circuit, **J-38**
 - carry-lookahead tree, **J-40**
 - carry-lookahead tree adder, **J-41**
 - carry-select adder, **J-43**, J-43 to J-44, **J-44**
 - carry-skip adder, J-41 to J-43, **J-42**
 - overview, J-37
 - integer arithmetic
 - language comparison, **J-12**
 - overflow, **J-11**
 - Radix-2 multiplication/
 - division, **J-4**, J-4 to J-7
 - restoring/nonrestoring division, **J-6**
 - ripplly-carry addition, J-2 to J-3, **J-3**
 - signed numbers, J-7 to J-10
 - systems issues, J-10 to J-13
 - integer division
 - radix-2 division, **J-55**
 - radix-4 division, **J-56**
 - radix-4 SRT division, **J-57**
 - with single adder, J-54 to J-58
 - SRT division, J-45 to J-47, J-46
 - integer-FP conversions, J-62
 - integer multiplication
 - array multiplier, **J-50**
 - Booth recoding, **J-49**
 - even/odd array, **J-52**
 - with many adders, J-50 to J-54
 - multipass array multiplier, **J-51**
 - signed-digit addition table, **J-54**
 - with single adder, J-47 to J-49, **J-48**
 - Wallace tree, **J-53**
 - integer multiplication/division,
 - shifting over zeros, J-45 to J-47
 - overview, J-2
 - rounding modes, **J-20**
- Computer chip fabrication
 - cost case study, 61–62
 - Cray X1E, G-24
- Computer classes
 - desktops, 6
 - embedded computers, 8–9
 - example, **5**
 - overview, **5**
 - parallelism and parallel
 - architectures, 9–10
 - PMDs, 6
 - servers, 7
 - and system characteristics, **E-4**
 - warehouse-scale computers, 8
- Computer design principles
 - Amdahl's law, 46–48
 - common case, 45–46
 - parallelism, 44–45
 - principle of locality, 45
 - processor performance equation, 48–52
- Computer history, technology and architecture, 2–5
- Computer room air-conditioning (CRAC), WSC infrastructure, 448–449
- Compute tiles, OCNs, F-3
- Compute Unified Device Architecture, *see* CUDA (Compute Unified Device Architecture)
- Conditional branches
 - branch folding, 206
 - compare frequencies, A-20
 - compiler performance, C-24 to C-25
 - control flow instructions, 14, A-16, **A-17**, A-19, A-21
 - desktop RISCs, **K-17**
 - embedded RISCs, **K-17**
 - evaluation, **A-19**
 - global code scheduling, H-16, **H-16**
 - GPUs, 300–303
 - ideal processor, 214
 - ISAs, A-46
 - MIPS control flow instructions, A-38, **A-40**
 - MIPS core, K-9 to K-16
 - PA-RISC instructions, K-34, **K-34**
 - predictor misprediction rates, 166
 - PTX instruction set, 298–299
 - static branch prediction, C-26
 - types, **A-20**
 - vector-GPU comparison, 311
- Conditional instructions
 - exposing parallelism, H-23 to H-27
 - limitations, H-26 to H-27
- Condition codes
 - branch conditions, A-19
 - control flow instructions, 14
 - definition, C-5
 - high-level instruction set, A-43
 - instruction set complications, C-50
 - MIPS core, K-9 to K-16
 - pipeline branch penalties, C-23
 - VAX, K-71

- Conflict misses
 - and block size, B-28
 - cache coherence mechanism, 358
 - and cache size, **B-24**, B-26
 - definition, B-23
 - as kernel miss, 376
 - L3 caches, 371
 - memory hierarchy basics, 75
 - OLTP workload, 370
 - PIDs, B-37
 - shared-memory workload, 373
- Congestion control
 - commercial interconnection networks, F-64
 - system area network history, F-101
- Congestion management, commercial interconnection networks, F-64 to F-66
- Connectedness
 - dimension-order routing, F-47 to F-48
 - interconnection network topology, F-29
- Connection delay, multi-device interconnection networks, F-25
- Connection Machine CM-5, F-91, F-100
- Connection Multiprocessor 2, L-44, L-57
- Consistency, *see* Memory consistency
- Constant extension
 - desktop RISCs, **K-9**
 - embedded RISCs, **K-9**
- Constellation, characteristics, **I-45**
- Containers
 - airflow, **466**
 - cluster history, L-74 to L-75
 - Google WSCs, 464–465, **465**
- Context Switching
 - definition, 106, B-49
 - Fermi GPU, 307
- Control bits, messages, F-6
- Control Data Corporation (CDC), first vector computers, L-44 to L-45
- Control Data Corporation (CDC) 6600
 - computer architecture definition, L-18
 - dynamically scheduling with scoreboard, C-71 to C-72
 - early computer arithmetic, J-64
 - first dynamic scheduling, L-27
 - MIPS scoreboarding, C-75, C-77
 - multiple-issue processor
 - development, L-28
 - multithreading history, L-34
 - RISC history, L-19
- Control Data Corporation (CDC) STAR-100
 - first vector computers, L-44
 - peak performance *vs.* start-up overhead, 331
- Control Data Corporation (CDC) STAR processor, G-26
- Control dependences
 - conditional instructions, H-24
 - as data dependence, 150
 - global code scheduling, H-16
 - hardware-based speculation, 183
 - ILP, 154–156
 - ILP hardware model, 214
 - and Tomasulo's algorithm, 170
 - vector mask registers, 275–276
- Control flow instructions
 - addressing modes, A-17 to A-18
 - basic considerations, A-16 to A-17, A-20 to A-21
 - classes, **A-17**
 - conditional branch options, A-19
 - conditional instructions, H-27
 - hardware *vs.* software speculation, 221
 - Intel 80x86 integer operations, K-51
 - ISAs, 14
 - MIPS, A-37 to A-38, **A-38**
 - procedure invocation options, A-19 to A-20
- Control hazards
 - ARM Cortex-A8, 235
 - definition, C-11
- Control instructions
 - Intel 80x86, **K-53**
 - RISCs
 - desktop systems, **K-12**, **K-22**
 - embedded systems, **K-16**
 - VAX, **B-73**
- Controllers, historical background, L-80 to L-81
- Controller transitions
 - directory-based, **422**
 - snooping cache, **421**
- Control Processor
 - definition, **309**
 - GPUs, 333
 - SIMD, 10
 - Thread Block Scheduler, 294
 - vector processor, **310**, 310–311
 - vector unit structure, **273**
- Conventional datacenters, *vs.* WSCs, 436
- Convex Exemplar, L-61
- Convex processors, vector processor history, G-26
- Convolution, DSP, E-5
- Convoy
 - chained, DAXPY code, **G-16**
 - DAXPY on VMIPS, G-20
 - strip-mined loop, G-5
 - vector execution time, 269–270
 - vector starting times, **G-4**
- Conway, Lynn, L-28
- Cooling systems
 - Google WSC, 465–468
 - mechanical design, **448**
 - WSC infrastructure, 448–449
- Copper wiring
 - Ethernet, F-78
 - interconnection networks, F-9
- “Coprocessor operations,” MIPS core extensions, K-21
- Copy propagation, definition, H-10 to H-11
- Core definition, 15
- Core plus ASIC, embedded systems, E-3
- Correlating branch predictors, branch costs, 162–163
- Cosmic Cube, F-100, L-60
- Cost
 - Amazon EC2, **458**
 - Amazon Web Services, 457
 - bisection bandwidth, F-89
 - branch predictors, 162–167, C-26
 - chip fabrication case study, 61–62
 - cloud computing providers, 471–472
 - disk storage, D-2
 - DRAM/magnetic disk, **D-3**
 - interconnecting node calculations, F-31 to F-32, F-35
 - Internet Archive Cluster, D-38 to D-40
 - internetworking, F-80

- Cost (*continued*)
 - I/O system design/evaluation, D-36
 - magnetic storage history, L-78
 - MapReduce calculations, 458–459, **459**
 - memory hierarchy design, 72
 - MINs vs. direct networks, F-92
 - multiprocessor cost relationship, 409
 - multiprocessor linear speedup, 407
 - network topology, **F-40**
 - PMDs, 6
 - server calculations, **454**, 454–455
 - server usage, 7
 - SIMD supercomputer
 - development, L-43
 - speculation, 210
 - torus topology interconnections, F-36 to F-38
 - tournament predictors, 164–166
 - WSC array switch, 443
 - WSC vs. datacenters, 455–456
 - WSC efficiency, 450–452
 - WSC facilities, 472
 - WSC network bottleneck, 461
 - WSCs, 446–450, 452–455, **453**
 - WSCs vs. servers, 434
 - WSC TCO case study, 476–478
- Cost associativity, cloud computing, 460–461
- Cost-performance
 - commercial interconnection
 - networks, F-63
 - computer trends, 3
 - extensive pipelining, C-80 to C-81
 - IBM eServer p5 processor, **409**
 - sorting case study, D-64 to D-67
 - WSC Flash memory, 474–475
 - WSC goals/requirements, 433
 - WSC hardware inactivity, 474
 - WSC processors, 472–473
- Cost trends
 - integrated circuits, 28–32
 - manufacturing vs. operation, 33
 - overview, 27
 - vs. price, 32–33
 - time, volume, commoditization, 27–28
- Count register, PowerPC instructions, K-32 to K-33
- CP-67 program, L-10
- CPA, *see* Carry-propagate adder (CPA)
- CPI, *see* Clock cycles per instruction (CPI)
- CPU, *see* Central processing unit (CPU)
- CRAC, *see* Computer room
 - air-conditioning (CRAC)
- Cray, Seymour, G-25, G-27, L-44, L-47
- Cray-1
 - first vector computers, L-44 to L-45
 - peak performance vs. start-up overhead, 331
 - pipeline depths, G-4
 - RISC history, L-19
 - vector performance, 332
 - vector performance measures, G-16
 - as VMIPS basis, 264, 270–271, 276–277
- Cray-2
 - DRAM, G-25
 - first vector computers, L-47
 - tailgating, G-20
- Cray-3, G-27
- Cray-4, G-27
- Cray C90
 - first vector computers, L-46, L-48
 - vector performance calculations, G-8
- Cray J90, L-48
- Cray Research T3D, F-86 to F-87, **F-87**
- Cray supercomputers, early computer arithmetic, J-63 to J-64
- Cray T3D, F-100, L-60
- Cray T3E, F-67, F-94, F-100, L-48, L-60
- Cray T90, memory bank calculations, 276
- Cray X1
 - cluster history, L-63
 - first vector computers, L-46, L-48
 - MSP module, **G-22**, G-23 to G-24
 - overview, G-21 to G-23
 - peak performance, **58**
- Cray X1E, F-86, F-91
 - characteristics, G-24
- Cray X2, L-46 to L-47
 - first vector computers, L-48 to L-49
- Cray X-MP, L-45
 - first vector computers, L-47
- Cray XT3, L-58, L-63
- Cray XT3 SeaStar, F-63
- Cray Y-MP
 - first vector computers, L-45 to L-47
 - parallel processing debates, L-57
 - vector architecture programming, **281**, 281–282
- CRC, *see* Cyclic redundancy check (CRC)
- Create vector index instruction (CVI), sparse matrices, G-13
- Credit-based control flow
 - InfiniBand, F-74
 - interconnection networks, F-10, F-17
- CRISP, L-27
- Critical path
 - global code scheduling, H-16
 - trace scheduling, H-19 to H-21, **H-20**
- Critical word first, cache optimization, 86–87
- Crossbars
 - centralized switched networks, F-30, **F-31**
 - characteristics, **F-73**
 - Convex Exemplar, L-61
 - HOL blocking, **F-59**
 - OCN history, F-104
 - switch microarchitecture, F-62
 - switch microarchitecture
 - pipelining, F-60 to F-61, **F-61**
 - VMIPS, **265**
- Crossbar switch
 - centralized switched networks, F-30
 - interconnecting node calculations, F-31 to F-32
- Cross-company interoperability, commercial
 - interconnection networks, F-63 to F-64
- Crusoe, L-31
- Cryptanalysis, L-4
- CSA, *see* Carry-save adder (CSA); Carry-skip adder (CSA)
- C# language, hardware impact on software development, 4
- CSSU, *see* Compare-select-store unit (CSSU)

- CUDA (Compute Unified Device Architecture)
 - GPU computing history, L-52
 - GPU conditional branching, 303
 - GPUs *vs.* vector architectures, 310
 - NVIDIA GPU programming, 289
 - PTX, 298, 300
 - sample program, 289–290
 - SIMD instructions, 297
 - terminology, 313–315
- CUDA Thread
 - CUDA programming model, 300, 315
 - definition, **292, 313**
 - definitions and terms, **314**
 - GPU data addresses, 310
 - GPU Memory structures, **304**
 - NVIDIA parallelism, 289–290
 - vs.* POSIX Threads, 297
 - PTX Instructions, 298
 - SIMD Instructions, 303
 - Thread Block, **313**
- Current frame pointer (CFM), IA-64
 - register model, H-33 to H-34
- Custom cluster
 - characteristics, **I-45**
 - IBM Blue Gene/L, I-41 to I-44, I-43 to I-44
- Cut-through packet switching, F-51
 - routing comparison, **F-54**
- CVI, *see* Create vector index instruction (CVI)
- CWS, *see* Circulating water system (CWS)
- CYBER 180/990, precise exceptions, C-59
- CYBER 205
 - peak performance *vs.* start-up overhead, 331
 - vector processor history, G-26 to G-27
- CYBER 250, L-45
- Cycles, processor performance
 - equation, 49
- Cycle time, *see also* Clock cycle time
 - CPI calculations, 350
 - pipelining, C-81
 - scoreboarding, C-79
 - vector processors, 277
- Cyclic redundancy check (CRC)
 - IBM Blue Gene/L 3D torus network, F-73
 - network interface, F-8
- Cydrome Cydra 6, L-30, L-32
- D**
- DaCapo benchmarks
 - ISA, 242
 - SMT, 230–231, **231**
- DAMQs, *see* Dynamically allocatable multi-queues (DAMQs)
- DASH multiprocessor, L-61
- Database program speculation, *via* multiple branches, 211
- Data cache
 - ARM Cortex-A8, **236**
 - cache optimization, B-33, B-38
 - cache performance, B-16
 - GPU Memory, 306
 - ISA, 241
 - locality principle, B-60
 - MIPS R4000 pipeline, C-62 to C-63
 - multiprogramming, 374
 - page level write-through, B-56
 - RISC processor, C-7
 - structural hazards, C-15
 - TLB, B-46
- Data cache miss
 - applications *vs.* OS, **B-59**
 - cache optimization, B-25
 - Intel Core i7, 240
 - Opteron, B-12 to B-15
 - sizes and associativities, B-10
 - writes, B-10
- Data cache size, multiprogramming, 376–377
- Datacenters
 - CDF, **487**
 - containers, L-74
 - cooling systems, 449
 - layer 3 network example, 445
 - PUE statistics, **451**
 - tier classifications, **491**
 - vs.* WSC costs, 455–456
 - WSC efficiency measurement, 450–452
 - vs.* WSCs, 436
- Data dependences
 - conditional instructions, H-24
 - data hazards, 167–168
 - dynamically scheduling with scoreboard, C-71
 - example calculations, H-3 to H-4
 - hazards, 153–154
 - ILP, 150–152
 - ILP hardware model, 214–215
 - ILP limitation studies, 220
 - vector execution time, 269
- Data fetching
 - ARM Cortex-A8, **234**
 - directory-based cache coherence protocol example, 382–383
 - dynamically scheduled pipelines, C-70 to C-71
 - ILP, instruction bandwidth
 - basic considerations, 202–203
 - branch-target buffers, 203–206
 - return address predictors, 206–207
 - MIPS R4000, C-63
 - snooping coherence protocols, 355–356
- Data flow
 - control dependence, 154–156
 - dynamic scheduling, 168
 - global code scheduling, H-17
 - ILP limitation studies, 220
 - limit, L-33
- Data flow execution, hardware-based speculation, 184
- Datagrams, *see* Packets
- Data hazards
 - ARM Cortex-A8, 235
 - basic considerations, C-16
 - definition, C-11
 - dependences, 152–154
 - dynamic scheduling, 167–176
 - basic concept, 168–170
 - examples, 176–178
 - Tomasulo's algorithm, 170–176, 178–179
 - Tomasulo's algorithm loop-based example, 179–181
 - ILP limitation studies, 220
 - instruction set complications, C-50 to C-51
 - microarchitectural techniques case study, 247–254
 - MIPS pipeline, C-71
 - RAW, C-57 to C-58

Data hazards

- stall minimization by forwarding,
C-16 to C-19, **C-18**
- stall requirements, C-19 to C-21
- VMIPS, 264

Data-level parallelism (DLP)

definition, 9

GPUs

- basic considerations, 288
- basic PTX thread instructions,
299
- conditional branching, 300–303
- coprocessor relationship,
330–331
- Fermi GPU architecture
innovations, 305–308
- Fermi GTX 480 floorplan, **295**
- mapping examples, **293**
- Multimedia SIMD comparison,
312
- multithreaded SIMD Processor
block diagram, **294**
- NVIDIA computational
structures, 291–297
- NVIDIA/CUDA and AMD
terminology, 313–315
- NVIDIA GPU ISA, 298–300
- NVIDIA GPU Memory
structures, **304**, 304–305
- programming, 288–291
- SIMD thread scheduling, **297**
- terminology, **292**
- vs. vector architectures,
308–312, **310**

from ILP, 4–5

Multimedia SIMD Extensions

- basic considerations, 282–285
- programming, 285
- roofline visual performance
model, 285–288, **287**

and power, 322

vector architecture

- basic considerations, 264
- gather/scatter operations,
279–280
- multidimensional arrays,
278–279
- multiple lanes, 271–273
- peak performance vs. start-up
overhead, 331
- programming, 280–282

vector execution time, 268–271

vector-length registers,
274–275

vector load-store unit
bandwidth, 276–277

vector-mask registers, 275–276

vector processor example,
267–268

VMIPS, 264–267

vector kernel implementation,
334–336

vector performance and memory
bandwidth, 332

vector vs. scalar performance,
331–332

WSCs vs. servers, 433–434

Data link layer

definition, **F-82**

interconnection networks, F-10

Data parallelism, SIMD computer
history, L-55

Data-race-free, synchronized
programs, 394

Data races, synchronized programs, 394

Data transfers

- cache miss rate calculations, B-16
- computer architecture, 15
- desktop RISC instructions, **K-10**,
K-21

embedded RISCs, **K-14**, **K-23**

gather-scatter, 281, 291

instruction operators, **A-15**

Intel 80x86, **K-49**, **K-53 to K-54**

ISA, 12–13

MIPS, addressing modes, A-34

MIPS64, K-24 to K-26

MIPS64 instruction subset, **A-40**

MIPS64 ISA formats, 14

MIPS core extensions, K-20

MIPS operations, A-36 to A-37

MMX, 283

multimedia instruction compiler
support, A-31

operands, **A-12**

PTX, 305

SIMD extensions, 284

“typical” programs, A-43

VAX, **B-73**

vector vs. GPU, 300

Data trunks, MIPS scoreboarding,

C-75

Data types

architect-compiler writer
relationship, A-30

dependence analysis, H-10

desktop computing, A-2

Intel 80x86, K-50

MIPS, A-34, A-36

MIPS64 architecture, A-34

multimedia compiler support, A-31

operand types/sizes, A-14 to A-15

SIMD Multimedia Extensions,
282–283

SPARC, **K-31**

VAX, **K-66**, K-70

Dauber, Phil, L-28

DAXPY loop

chained convoys, **G-16**

on enhanced VMIPS, G-19 to G-21

memory bandwidth, 332

MIPS/VMIPS calculations,
267–268

peak performance vs. start-up
overhead, 331

vector performance measures,
G-16

VLRs, 274–275

on VMIPS, G-19 to G-20

VMIPS calculations, G-18

VMIPS on Linpack, G-18

VMIPS peak performance, G-17

D-caches

- case study examples, B-63
- way prediction, 81–82

DDR, *see* Double data rate (DDR)

Deadlock

- cache coherence, 361
- dimension-order routing, F-47 to
F-48
- directory protocols, 386
- Intel SCCC, F-70
- large-scale multiprocessor cache
coherence, I-34 to I-35,
I-38 to I-40
- mesh network routing, **F-46**
- network routing, F-44
- routing comparison, **F-54**
- synchronization, 388
- system area network history, F-101

Deadlock avoidance

- meshes and hypercubes, F-47
- routing, F-44 to F-45

- Deadlock recovery, routing, F-45
- Dead time
 - vector pipeline, **G-8**
 - vector processor, G-8
- Decimal operands, formats, A-14
- Decimal operations, PA-RISC
 - instructions, K-35
- Decision support system (DSS),
 - shared-memory workloads, 368–369, **369**, 369–370
- Decoder, radio receiver, **E-23**
- Decode stage, TI 320C55 DSP, E-7
- DEC PDP-11, address space, B-57 to B-58
- DECstation 5000, reboot
 - measurements, **F-69**
- DEC VAX
 - addressing modes, A-10 to A-11, **A-11**, K-66 to K-68
 - address space, B-58
 - architect-compiler writer
 - relationship, A-30
 - branch conditions, A-19
 - branches, **A-18**
 - jumps, procedure calls, K-71 to K-72
 - bubble sort, K-76
 - characteristics, K-42
 - cluster history, L-62, L-72
 - compiler writing-architecture
 - relationship, A-30
 - control flow instruction branches, A-18
 - data types, **K-66**
 - early computer arithmetic, J-63 to J-64
 - early pipelined CPUs, L-26
 - exceptions, **C-44**
 - extensive pipelining, C-81
 - failures, D-15
 - flawless architecture design, A-45, K-81
 - high-level instruction set, A-41 to A-43
 - high-level language computer
 - architecture, L-18 to L-19
 - history, 2–3
 - immediate value distribution, **A-13**
 - instruction classes, **B-73**
 - instruction encoding, K-68 to K-70, **K-69**
 - instruction execution issues, K-81
 - instruction operator categories, **A-15**
 - instruction set complications, C-49 to C-50
 - integer overflow, **J-11**
 - vs. MIPS, **K-82**
 - vs. MIPS32 sort, **K-80**
 - vs. MIPS code, K-75
 - miss rate vs. virtual addressing, **B-37**
 - operands, K-66 to K-68
 - operand specifiers, **K-68**
 - operands per ALU, **A-6**, A-8
 - operand types/sizes, A-14
 - operation count, K-70 to K-71
 - operations, K-70 to K-72
 - operators, **A-15**
 - overview, K-65 to K-66
 - precise exceptions, C-59
 - replacement by RISC, 2
 - RISC history, L-20 to L-21
 - RISC instruction set lineage, **K-43**
 - sort, K-76 to K-79
 - sort code, K-77 to K-79
 - sort register allocation, K-76
 - swap, K-72 to K-76
 - swap code, **B-74**, K-72, K-74
 - swap full procedure, K-75 to K-76
 - swap and register preservation, B-74 to B-75
 - unique instructions, K-28
- DEC VAX-11/780, L-6 to L-7, L-11, L-18
- DEC VAX 8700
 - vs. MIPS M2000, **K-82**, **L-21**
 - RISC history, L-21
- Dedicated link network
 - black box network, F-5 to F-6
 - effective bandwidth, F-17
 - example, **F-6**
- Defect tolerance, chip fabrication cost
 - case study, 61–62
- Deferred addressing, VAX, K-67
- Delayed branch
 - basic scheme, **C-23**
 - compiler history, L-31
 - instructions, **K-25**
 - stalls, **C-65**
- Dell Poweredge servers, prices, **53**
- Dell Poweredge Thunderbird, SAN
 - characteristics, **F-76**
- Dell servers
 - economies of scale, 456
 - real-world considerations, 52–55
 - WSC services, 441
- Demodulator, radio receiver, **E-23**
- Denormals, J-14 to J-16, J-20 to J-21
 - floating-point additions, J-26 to J-27
 - floating-point underflow, J-36
- Dense matrix multiplication, LU
 - kernel, I-8
- Density-optimized processors, vs. SPEC-optimized, F-85
- Dependability
 - benchmark examples, D-21 to D-23, **D-22**
 - definition, D-10 to D-11
 - disk operators, D-13 to D-15
 - integrated circuits, 33–36
 - Internet Archive Cluster, D-38 to D-40
 - memory systems, 104–105
 - WSC goals/requirements, 433
 - WSC memory, 473–474
 - WSC storage, 442–443
- Dependence analysis
 - basic approach, H-5
 - example calculations, H-7
 - limitations, H-8 to H-9
- Dependence distance, loop-carried
 - dependences, H-6
- Dependences
 - antidependences, 152, 320, C-72, C-79
 - CUDA, 290
 - as data dependence, 150
 - data hazards, 167–168
 - definition, 152–153, 315–316
 - dynamically scheduled pipelines, C-70 to C-71
 - dynamically scheduling with scoreboard, C-71
 - dynamic scheduling with Tomasulo's algorithm, 172
 - hardware-based speculation, 183
 - hazards, 153–154
 - ILP, 150–156
 - ILP hardware model, 214–215
 - ILP limitation studies, 220

Dependences (*continued*)

- loop-level parallelism, 318–322, H-3
 - dependence analysis, H-6 to H-10
- MIPS scoreboarding, C-79
- as program properties, 152
- sparse matrices, G-13
- and Tomasulo's algorithm, 170
- types, 150
- vector execution time, 269
- vector mask registers, 275–276
- VMIPS, 268
- Dependent computations, elimination, H-10 to H-12
- Descriptor privilege level (DPL), segmented virtual memory, B-53
- Descriptor table, IA-32, B-52
- Design faults, storage systems, D-11
- Desktop computers
 - characteristics, 6
 - compiler structure, A-24
 - as computer class, 5
 - interconnection networks, F-85
 - memory hierarchy basics, 78
 - multimedia support, **E-11**
 - multiprocessor importance, 344
 - performance benchmarks, 38–40
 - processor comparison, **242**
 - RAID history, L-80
 - RISC systems
 - addressing modes, **K-5**
 - addressing modes and instruction formats, K-5 to K-6
 - arithmetic/logical instructions, **K-22**
 - conditional branches, **K-17**
 - constant extension, **K-9**
 - control instructions, **K-12**
 - conventions, **K-13**
 - data transfer instructions, **K-10, K-21**
 - examples, K-3, **K-4**
 - features, **K-44**
 - FP instructions, **K-13, K-23**
 - instruction formats, **K-7**
 - multimedia extensions, K-16 to K-19, **K-18**
 - system characteristics, **E-4**
- Destination offset, IA-32 segment, **B-53**

Deterministic routing algorithm

- vs. adaptive routing, F-52 to F-55, **F-54**

DOR, F-46

Dies

- embedded systems, E-15
- integrated circuits, 28–30, 29
- Nehalem floorplan, **30**
- wafer example, **31**, 31–32

Die yield, basic equation, 30–31

Digital Alpha

- branches, **A-18**
- conditional instructions, H-27
- early pipelined CPUs, L-27
- RISC history, L-21
- RISC instruction set lineage, **K-43**
- synchronization history, L-64

Digital Alpha 21064, L-48

Digital Alpha 21264

- cache hierarchy, **368**
- floorplan, **143**

Digital Alpha MAX

- characteristics, **K-18**
- multimedia support, K-18

Digital Alpha processors

- addressing modes, **K-5**
- arithmetic/logical instructions, **K-11**
- branches, K-21
- conditional branches, K-12, **K-17**
- constant extension, **K-9**
- control flow instruction branches, A-18
- conventions, **K-13**
- data transfer instructions, **K-10**
- displacement addressing mode, **A-12**

- exception stopping/restarting, C-47
- FP instructions, **K-23**

immediate value distribution, A-13

MAX, multimedia support, **E-11**

MIPS precise exceptions, C-59

multimedia support, K-19

recent advances, L-33

as RISC systems, **K-4**

shared-memory workload, 367–369

unique instructions, K-27 to K-29

Digital Linear Tape, L-77

Digital signal processor (DSP)

- cell phones, E-23, **E-23**, E-23 to E-24
- definition, E-3

desktop multimedia support, **E-11**

embedded RISC extensions, K-19

examples and characteristics, **E-6**

media extensions, E-10 to E-11

overview, E-5 to E-7

saturating operations, K-18 to K-19

TI TMS320C6x, E-8 to E-10

TI TMS320C6x instruction packet, **E-10**TI TMS320C55, E-6 to E-7, **E-7 to E-8**TI TMS320C64x, **E-9**

Dimension-order routing (DOR), definition, F-46

DIMMs, *see* Dual inline memory modules (DIMMs)

Direct attached disks, definition, D-35

Direct-mapped cache

- address parts, **B-9**
- address translation, B-38
- block placement, B-7
- early work, L-10
- memory hierarchy basics, 74
- memory hierarchy, B-48
- optimization, 79–80

Direct memory access (DMA)

historical background, L-81

InfiniBand, F-76

network interface functions, F-7

Sanyo VPC-SX500 digital camera, E-19

Sony PlayStation 2 Emotion Engine, E-18

TI TMS320C55 DSP, E-8

zero-copy protocols, F-91

Direct networks

commercial system topologies, **F-37**

vs. high-dimensional networks, F-92

vs. MIN costs, F-92

topology, F-34 to F-40

Directory-based cache coherence

advanced directory protocol case study, 420–426

basic considerations, 378–380

case study, 418–420

definition, 354

distributed-memory multiprocessor, **380**

- large-scale multiprocessor history, L-61
- latencies, **425**
- protocol basics, 380–382
- protocol example, 382–386
- state transition diagram, **383**
- Directory-based multiprocessor
 - characteristics, **I-31**
 - performance, I-26
 - scientific workloads, I-29
 - synchronization, I-16, I-19 to I-20
- Directory controller, cache coherence, I-40 to I-41
- Dirty bit
 - case study, D-61 to D-64
 - definition, B-11
 - virtual memory fast address translation, B-46
- Dirty block
 - definition, B-11
 - read misses, B-36
- Discrete cosine transform, DSP, E-5
- Disk arrays
 - deconstruction case study, D-51 to D-55, **D-52 to D-55**
 - RAID 6, D-8 to D-9
 - RAID 10, D-8
 - RAID levels, D-6 to D-8, **D-7**
- Disk layout, RAID performance
 - prediction, D-57 to D-59
- Disk power, basic considerations, D-5
- Disk storage
 - access time gap, D-3
 - areal density, D-2 to D-5
 - cylinders, D-5
 - deconstruction case study, D-48 to D-51, **D-50**
 - DRAM/magnetic disk cost vs. access time, **D-3**
 - intelligent interfaces, D-4
 - internal microprocessors, D-4
 - real faults and failures, D-10 to D-11
 - throughput vs. command queue depth, **D-4**
- Disk technology
 - failure rate calculation, 48
 - Google WSC servers, 469
 - performance trends, 19–20, 20
 - WSC Flash memory, 474–475
- Dispatch stage
 - instruction steps, 174
- microarchitectural techniques case study, 247–254
- Displacement addressing mode
 - basic considerations, A-10
- MIPS, 12
- MIPS data transfers, A-34
- MIPS instruction format, A-35
- value distributions, **A-12**
- VAX, K-67
- Display lists, Sony PlayStation 2
 - Emotion Engine, E-17
- Distributed routing, basic concept, F-48
- Distributed shared memory (DSM)
 - basic considerations, 378–380
 - basic structure, 347–348, **348**
 - characteristics, **I-45**
 - directory-based cache coherence, 354, **380**, 418–420
 - multichip multicore
 - multiprocessor, **419**
 - snooping coherence protocols, 355
- Distributed shared-memory
 - multiprocessors
 - cache coherence implementation, I-36 to I-37
 - scientific application performance, I-26 to I-32, **I-28 to I-32**
- Distributed switched networks,
 - topology, F-34 to F-40
- Divide operations
 - chip comparison, J-60 to J-61
 - floating-point, stall, **C-68**
 - floating-point iterative, J-27 to J-31
 - integers, speedup
 - radix-2 division, **J-55**
 - radix-4 division, **J-56**
 - radix-4 SRT division, **J-57**
 - with single adder, J-54 to J-58
 - integer shifting over zeros, J-45 to J-47
 - language comparison, **J-12**
 - n*-bit unsigned integers, **J-4**
 - PA-RISC instructions, K-34 to K-35
 - Radix-2, J-4 to J-7
 - restoring/nonrestoring, **J-6**
 - SRT division, J-45 to J-47, **J-46**
 - unfinished instructions, **179**
- DLP, *see* Data-level parallelism (DLP)
- DLX
 - integer arithmetic, J-12
 - vs. Intel 80x86 operations, K-62, **K-63 to K-64**
- DMA, *see* Direct memory access (DMA)
- DOR, *see* Dimension-order routing (DOR)
- Double data rate (DDR)
 - ARM Cortex-A8, 117
 - DRAM performance, 100
 - DRAMs and DIMMs, **101**
 - Google WSC servers, 468–469
 - IBM Blue Gene/L, I-43
 - InfiniBand, F-77
 - Intel Core i7, 121
 - SDRAMs, 101
- Double data rate 2 (DDR2), SDRAM
 - timing diagram, **139**
- Double data rate 3 (DDR3)
 - DRAM internal organization, **98**
 - GDRAM, 102
 - Intel Core i7, 118
 - SDRAM power consumption, 102, **103**
- Double data rate 4 (DDR4), DRAM, **99**
- Double data rate 5 (DDR5), GDRAM, 102
- Double-extended floating-point
 - arithmetic, J-33 to J-34
- Double failures, RAID reconstruction, D-55 to D-57
- Double-precision floating point
 - add-divide, **C-68**
 - AVX for x86, **284**
 - chip comparison, **J-58**
 - data access benchmarks, **A-15**
 - DSP media extensions, E-10 to E-11
 - Fermi GPU architecture, 306
 - floating-point pipeline, C-65
 - GTX 280, 325, 328–330
 - IBM 360, 171
 - MIPS, 285, A-38 to A-39
 - MIPS data transfers, A-34
 - MIPS registers, 12, A-34
 - Multimedia SIMD vs. GPUs, 312
 - operand sizes/types, 12
 - as operand type, A-13 to A-14
 - operand usage, 297
 - pipeline timing, **C-54**

- Double-precision (*continued*)
 - Roofline model, **287**, **326**
 - SIMD Extensions, 283
 - VMIPS, **266**, 266–267
 - Double rounding
 - FP precisions, J-34
 - FP underflow, J-37
 - Double words
 - aligned/misaligned addresses, **A-8**
 - data access benchmarks, **A-15**
 - Intel 80x86, K-50
 - memory address interpretation, A-7 to A-8
 - MIPS data types, A-34
 - operand types/sizes, 12, A-14
 - stride, 278
 - DPL, *see* Descriptor privilege level (DPL)
 - DRAM, *see* Dynamic random-access memory (DRAM)
 - DDRDRAM, Sony PlayStation 2, E-16 to E-17
 - Driver domains, Xen VM, 111
 - DSM, *see* Distributed shared memory (DSM)
 - DSP, *see* Digital signal processor (DSP)
 - DSS, *see* Decision support system (DSS)
 - Dual inline memory modules (DIMMs)
 - clock rates, bandwidth, names, **101**
 - DRAM basics, 99
 - Google WSC server, **467**
 - Google WSC servers, 468–469
 - graphics memory, 322–323
 - Intel Core i7, 118, 121
 - Intel SCCC, F-70
 - SDRAMs, 101
 - WSC memory, 473–474
 - Dual SIMD Thread Scheduler,
 - example, 305–306
 - DVFS, *see* Dynamic voltage-frequency scaling (DVFS)
 - Dynamically allocatable multi-queues (DAMQs), switch
 - microarchitecture, F-56 to F-57
 - Dynamically scheduled pipelines
 - basic considerations, C-70 to C-71
 - with scoreboard, C-71 to C-80
 - Dynamically shared libraries, control
 - flow instruction addressing modes, A-18
 - Dynamic energy, definition, 23
 - Dynamic network reconfiguration,
 - fault tolerance, F-67 to F-68
 - Dynamic power
 - energy efficiency, 211
 - microprocessors, 23
 - vs. static power, 26
 - Dynamic random-access memory (DRAM)
 - bandwidth issues, 322–323
 - characteristics, 98–100
 - clock rates, bandwidth, names, **101**
 - cost vs. access time, **D-3**
 - cost trends, 27
 - Cray X1, G-22
 - CUDA, 290
 - dependability, 104
 - disk storage, D-3 to D-4
 - embedded benchmarks, E-13
 - errors and faults, D-11
 - first vector computers, L-45, L-47
 - Flash memory, 103–104
 - Google WSC servers, 468–469
 - GPU SIMD instructions, 296
 - IBM Blue Gene/L, I-43 to I-44
 - improvement over time, **17**
 - integrated circuit costs, 28
 - Intel Core i7, 121
 - internal organization, **98**
 - magnetic storage history, L-78
 - memory hierarchy design, 73, **73**
 - memory performance, 100–102
 - multibanked caches, 86
 - NVIDIA GPU Memory structures, 305
 - performance milestones, **20**
 - power consumption, **63**
 - real-world server considerations, 52–55
 - Roofline model, 286
 - server energy savings, **25**
 - Sony PlayStation 2, **E-16**, E-17
 - speed trends, **99**
 - technology trends, 17
 - vector memory systems, G-9
 - vector processor, G-25
 - WSC efficiency measurement, 450
 - WSC memory costs, 473–474
 - WSC memory hierarchy, 444–445
 - WSC power modes, 472
 - yield, 32
 - Dynamic scheduling
 - first use, L-27
 - ILP
 - basic concept, 168–169
 - definition, 168
 - example and algorithms, 176–178
 - with multiple issue and speculation, 197–202
 - overcoming data hazards, 167–176
 - Tomasulo's algorithm, 170–176, 178–179, 181–183
 - MIPS scoreboarding, C-79
 - SMT on superscalar processors, 230
 - and unoptimized code, C-81
 - Dynamic voltage-frequency scaling (DVFS)
 - energy efficiency, 25
 - Google WSC, 467
 - processor performance equation, 52
 - Dynamo (Amazon), 438, 452
- ## E
- Early restart, miss penalty reduction, 86
 - Earth Simulator, L-46, L-48, L-63
 - EBS, *see* Elastic Block Storage (EBS)
 - EC2, *see* Amazon Elastic Computer Cloud (EC2)
 - ECC, *see* Error-Correcting Code (ECC)
 - Eckert, J. Presper, L-2 to L-3, L-5, L-19
 - Eckert-Mauchly Computer Corporation, L-4 to L-5, L-56
 - ECL minicomputer, L-19
 - Economies of scale
 - WSC vs. datacenter costs, 455–456
 - WSCs, 434
 - EDSAC (Electronic Delay Storage Automatic Calculator), L-3
 - EDVAC (Electronic Discrete Variable Automatic Computer), L-2 to L-3

- EEMBC, *see* Electronic Design News Embedded Microprocessor Benchmark Consortium (EEMBC)
- EEPROM (Electrically Erasable Programmable Read-Only Memory)
 - compiler-code size considerations, **A-44**
 - Flash Memory, 102–104
 - memory hierarchy design, **72**
- Effective address
 - ALU, C-7, C-33
 - data dependences, 152
 - definition, A-9
 - execution/effective address cycle, C-6, C-31 to C-32, C-63
 - hardware-based speculation, 186, 190, 192
 - load interlocks, C-39
 - load-store, 174, 176, C-4
 - RISC instruction set, C-4 to C-5
 - simple MIPS implementation, C-31 to C-32
 - simple RISC implementation, C-6
 - TLB, B-49
 - Tomasulo's algorithm, **173**, 178, 182
- Effective bandwidth
 - definition, F-13
 - example calculations, F-18
 - vs. interconnected nodes, **F-28**
- interconnection networks
 - multi-device networks, F-25 to F-29
 - two-device networks, F-12 to F-20
 - vs. packet size, **F-19**
- Efficiency factor, F-52
- Eight-way set associativity
 - ARM Cortex-A8, 114
 - cache optimization, B-29
 - conflict misses, B-23
 - data cache misses, B-10
- Elapsed time, execution time, 36
- Elastic Block Storage (EBS), MapReduce cost calculations, 458–460, **459**
- Electrically Erasable Programmable Read-Only Memory, *see* EEPROM (Electrically Erasable Programmable Read-Only Memory)
- Electronic Delay Storage Automatic Calculator (EDSAC), L-3
- Electronic Design News Embedded Microprocessor Benchmark Consortium (EEMBC)
 - benchmark classes, E-12
 - ISA code size, **A-44**
 - kernel suites, **E-12**
 - performance benchmarks, 38
 - power consumption and efficiency metrics, E-13
- Electronic Discrete Variable Automatic Computer (EDVAC), L-2 to L-3
- Electronic Numerical Integrator and Calculator (ENIAC), L-2 to L-3, L-5 to L-6, L-77
- Element group, definition, **272**
- Embedded multiprocessors, characteristics, E-14 to E-15
- Embedded systems
 - benchmarks
 - basic considerations, E-12
 - power consumption and efficiency, E-13
 - cell phone case study
 - Nokia circuit board, **E-24**
 - overview, E-20
 - phone block diagram, **E-23**
 - phone characteristics, E-22 to E-24
 - radio receiver, **E-23**
 - standards and evolution, E-25
 - wireless networks, E-21 to E-22
 - characteristics, 8–9, **E-4**
 - as computer class, **5**
 - digital signal processors
 - definition, E-3
 - desktop multimedia support, **E-11**
 - examples and characteristics, **E-6**
 - media extensions, E-10 to E-11
 - overview, E-5 to E-7
 - TI TMS320C6x, E-8 to E-10
 - TI TMS320C6x instruction packet, **E-10**
 - TI TMS320C55, **E-6 to E-7**, E-7 to E-8
 - TI TMS320C64x, **E-9**
- EEMBC benchmark suite, **E-12**
- overview, E-2
- performance, **E-13 to E-14**
- real-time processing, E-3 to E-5
- RISC systems
 - addressing modes, **K-6**
 - addressing modes and instruction formats, K-5 to K-6
 - arithmetic/logical instructions, **K-24**
 - conditional branches, **K-17**
 - constant extension, **K-9**
 - control instructions, **K-16**
 - conventions, **K-16**
 - data transfer instructions, **K-14**, **K-23**
 - DSP extensions, K-19
 - examples, K-3, **K-4**
 - instruction formats, **K-8**
 - multiply-accumulate, **K-20**
- Sanyo digital camera SOC, **E-20**
- Sanyo VPC-SX500 digital camera
 - case study, E-19
- Sony PlayStation 2 block diagram, **E-16**
- Sony PlayStation 2 Emotion Engine case study, E-15 to E-18
- Sony PlayStation 2 Emotion Engine organization, **E-18**
- EMC, L-80
- Emotion Engine
 - organization modes, **E-18**
 - Sony PlayStation 2 case study, E-15 to E-18
- empowerTel Networks, MXP processor, E-14
- Encoding
 - control flow instructions, A-18
 - erasure encoding, 439
 - instruction set, A-21 to A-24, **A-22**
 - Intel 80x86 instructions, K-55, **K-58**

- Encoding (*continued*)
 - ISAs, 14, A-5 to A-6
 - MIPS ISA, A-33
 - MIPS pipeline, C-36
 - opcode, A-13
 - VAX instructions, K-68 to K-70, **K-69**
 - VLIW model, 195–196
- Encore Multimax, L-59
- End-to-end flow control
 - congestion management, F-65
 - vs. network-only features, F-94 to F-95
- Energy efficiency, *see also* Power consumption
 - Climate Savers Computing Initiative, **462**
 - embedded benchmarks, E-13
 - hardware fallacies, 56
 - ILP exploitation, 201
 - Intel Core i7, 401–405
 - ISA, 241–243
 - microprocessor, 23–26
 - PMDs, 6
 - processor performance equation, 52
 - servers, **25**
 - and speculation, 211–212
 - system trends, 21–23
 - WSC, measurement, 450–452
 - WSC goals/requirements, 433
 - WSC infrastructure, 447–449
 - WSC servers, 462–464
- Energy proportionality, WSC servers, **462**
- Engineering Research Associates (ERA), L-4 to L-5
- ENIAC (Electronic Numerical Integrator and Calculator), L-2 to L-3, L-5 to L-6, L-77
- Enigma coding machine, L-4
- Entry time, transactions, D-16, **D-17**
- Environmental faults, storage systems, D-11
- EPIC approach
 - historical background, L-32
 - IA-64, H-33
 - VLIW processors, **194**, 196
- Equal condition code, PowerPC, K-10 to K-11
- ERA, *see* Engineering Research Associates (ERA)
- Erasure encoding, WSCs, 439
- Error-Correcting Code (ECC)
 - disk storage, D-11
 - fault detection pitfalls, 58
 - Fermi GPU architecture, 307
 - hardware dependability, D-15
 - memory dependability, 104
 - RAID 2, D-6
 - and WSCs, 473–474
- Error handling, interconnection networks, F-12
- Errors, definition, D-10 to D-11
- Escape resource set, F-47
- ETA processor, vector processor
 - history, G-26 to G-27
- Ethernet
 - and bandwidth, **F-78**
 - commercial interconnection networks, F-63
 - cross-company interoperability, F-64
 - interconnection networks, F-89
 - as LAN, F-77 to F-79
 - LAN history, F-99
 - LANs, F-4
 - packet format, **F-75**
 - shared-media networks, F-23
 - shared- vs. switched-media networks, **F-22**
 - storage area network history, F-102
 - switch vs. NIC, **F-86**
 - system area networks, F-100
 - total time statistics, **F-90**
 - WAN history, F-98
- Ethernet switches
 - architecture considerations, 16
 - Dell servers, 53
 - Google WSC, 464–465, 469
 - historical performance milestones, **20**
 - WSCs, 441–444
- European Center for Particle Research (CERN), F-98
- Even/odd array
 - example, **J-52**
 - integer multiplication, J-52
- EVEN-ODD scheme, development, D-10
- EX, *see* Execution address cycle (EX)
- Example calculations
 - average memory access time, B-16 to B-17
- barrier synchronization, I-15
- block size and average memory access time, B-26 to B-28
- branch predictors, 164
- branch schemes, C-25 to C-26
- branch-target buffer branch penalty, 205–206
- bundles, H-35 to H-36
- cache behavior impact, B-18, B-21
- cache hits, B-5
- cache misses, 83–84, 93–95
- cache organization impact, B-19 to B-20
- carry-lookahead adder, J-39
- chime approximation, G-2
- compiler-based speculation, H-29 to H-31
- conditional instructions, H-23 to H-24
- CPI and FP, 50–51
- credit-based control flow, F-10 to F-11
- crossbar switch interconnections, F-31 to F-32
- data dependences, H-3 to H-4
- DAXPY on VMIPS, G-18 to G-20
- dependence analysis, H-7 to H-8
- deterministic vs. adaptive routing, F-52 to F-55
- dies, 29
- die yield, 31
- dimension-order routing, F-47 to F-48
- disk subsystem failure rates, 48
- fault tolerance, F-68
- fetch-and-increment barrier, I-20 to I-21
- FFT, I-27 to I-29
- fixed-point arithmetic, E-5 to E-6
- floating-point addition, J-24 to J-25
- floating-point square root, 47–48
- GCD test, 319, H-7
- geometric means, 43–44
- hardware-based speculation, 200–201
- inclusion, 397
- information tables, 176–177
- integer multiplication, J-9
- interconnecting node costs, F-35
- interconnection network latency and effective bandwidth, F-26 to F-28

- I/O system utilization, D-26
- L1 cache speed, 80
- large-scale multiprocessor locks, I-20
- large-scale multiprocessor
 - synchronization, I-12 to I-13
- loop-carried dependences, 316, H-4 to H-5
- loop-level parallelism, 317
- loop-level parallelism
 - dependences, 320
- loop unrolling, 158–160
- MapReduce cost on EC2, 458–460
- memory banks, 276
- microprocessor dynamic energy/
 - power, 23
- MIPS/VMIPS for DAXPY loop, 267–268
- miss penalty, B-33 to B-34
- miss rates, B-6, B-31 to B-32
- miss rates and cache sizes, B-29 to B-30
- miss support, 85
- M/M/1 model, D-33
- MTTF, 34–35
- multimedia instruction compiler
 - support, A-31 to A-32
- multiplication algorithm, J-19
- network effective bandwidth, F-18
- network topologies, F-41 to F-43
- Ocean application, I-11 to I-12
- packet latency, F-14 to F-15
- parallel processing, 349–350, I-33 to I-34
- pipeline execution rate, C-10 to C-11
- pipeline structural hazards, C-14 to C-15
- power-performance benchmarks, 439–440
- predicated instructions, H-25
- processor performance
 - comparison, 218–219
- queue I/O requests, D-29
- queue waiting time, D-28 to D-29
- queuing, D-31
- radix-4 SRT division, J-56
- redundant power supply reliability, 35
- ROB commit, 187
- ROB instructions, 189
- scoreboarding, C-77
- sequential consistency, 393
- server costs, 454–455
- server power, 463
- signed-digit numbers, J-53
- signed numbers, J-7
- SIMD multimedia instructions, 284–285
- single-precision numbers, J-15, J-17
- software pipelining, H-13 to H-14
- speedup, 47
- status tables, 178
- strides, 279
- TB-80 cluster MTTF, D-41
- TB-80 IOPS, D-39 to D-40
- torus topology interconnections, F-36 to F-38
- true sharing misses and false sharing, 366–367
- VAX instructions, K-67
- vector memory systems, G-9
- vector performance, G-8
- vector vs. scalar operation, G-19
- vector sequence chimes, 270
- VLIW processors, 195
- VMIPS vector operation, G-6 to G-7
- way selection, 82
- write buffer and read misses, B-35 to B-36
- write vs. no-write allocate, B-12
- WSC memory latency, 445
- WSC running service availability, 434–435
- WSC server data transfer, 446
- Exceptions
 - ALU instructions, C-4
 - architecture-specific examples, **C-44**
 - categories, **C-46**
 - control dependence, 154–155
 - floating-point arithmetic, J-34 to J-35
 - hardware-based speculation, 190
 - imprecise, 169–170, 188
 - long latency pipelines, C-55
 - MIPS, **C-48**, C-48 to C-49
 - out-of-order completion, 169–170
 - precise, C-47, C-58 to C-60
 - preservation via hardware support, H-28 to H-32
- return address buffer, 207
- ROB instructions, 190
- speculative execution, 222
- stopping/restarting, C-46 to C-47
- types and requirements, C-43 to C-46
- Execute step
 - instruction steps, 174
 - Itanium 2, H-42
 - ROB instruction, 186
 - TI 320C55 DSP, E-7
- Execution address cycle (EX)
 - basic MIPS pipeline, C-36
 - data hazards requiring stalls, **C-21**
 - data hazard stall minimization, C-17
 - exception stopping/restarting, C-46 to C-47
 - hazards and forwarding, C-56 to C-57
 - MIPS FP operations, basic considerations, C-51 to C-53
 - MIPS pipeline, **C-52**
 - MIPS pipeline control, C-36 to C-39
 - MIPS R4000, C-63 to C-64, **C-64**
 - MIPS scoreboarding, C-72, C-74, C-77
 - out-of-order execution, C-71
 - pipeline branch issues, **C-40**, **C-42**
 - RISC classic pipeline, C-10
 - simple MIPS implementation, C-31 to C-32
 - simple RISC implementation, C-6
- Execution time
 - Amdahl's law, 46–47, 406
 - application/OS misses, **B-59**
 - cache performance, B-3 to B-4, B-16
 - calculation, 36
 - commercial workloads, 369–370, **370**
 - energy efficiency, 211
 - integrated circuits, 22
 - loop unrolling, 160
 - multilevel caches, B-32 to B-34
 - multiprocessor performance, 405–406
 - multiprogrammed parallel “make” workload, **375**
 - multithreading, **232**

Execution time (*continued*)

- performance equations, **B-22**
- pipelining performance, C-3, C-10 to C-11
- PMDs, 6
- principle of locality, 45
- processor comparisons, **243**
- processor performance equation, 49, 51
- reduction, B-19
- second-level cache size, **B-34**
- SPEC benchmarks, 42–44, **43**, 56
- and stall time, B-21
- vector length, **G-7**
- vector mask registers, 276
- vector operations, 268–271

Expand-down field, **B-53**

Explicit operands, ISA classifications, A-3 to A-4

Explicit parallelism, IA-64, H-34 to H-35

Explicit unit stride, GPUs vs. vector architectures, 310

Exponential back-off

- large-scale multiprocessor synchronization, I-17
- spin lock, **I-17**

Exponential distribution, definition, D-27

Extended accumulator

- flawed architectures, A-44
- ISA classification, A-3

F

Facebook, 460

Failures, *see also* Mean time between failures (MTBF); Mean time to failure (MTTF)

- Amdahl's law, 56
- Berkeley's Tertiary Disk project, D-12
- cloud computing, 455
- definition, D-10
- dependability, 33–35
- dirty bits, D-61 to D-64
- DRAM, 473
- example calculation, 48
- Google WSC networking, 469–470
- power failure, C-43 to C-44, C-46
- power utilities, 435
- RAID reconstruction, D-55 to D-57

RAID row-diagonal parity, **D-9**

- rate calculations, 48
- servers, 7, 434
- SLA states, 34
- storage system components, D-43
- storage systems, D-6 to D-10
- TDP, 22
- Tertiary Disk, **D-13**
- WSC running service, 434–435
- WSCs, 8, 438–439
- WSC storage, 442–443

False sharing

- definition, 366–367
- shared-memory workload, 373

FarmVille, 460

Fast Fourier transformation (FFT)

- characteristics, I-7
- distributed-memory multiprocessor, **I-32**
- example calculations, I-27 to I-29
- symmetric shared-memory multiprocessors, I-22, **I-23**, I-25

Fast traps, SPARC instructions, K-30

Fat trees

- definition, F-34
- NEWS communication, F-43
- routing algorithms, F-48
- SAN characteristics, **F-76**
- topology, F-38 to F-39
- torus topology interconnections, F-36 to F-38

Fault detection, pitfalls, 57–58

Fault-induced deadlock, routing, F-44

Faulting prefetches, cache

- optimization, 92

Faults, *see also* Exceptions; Page

- faults

- address fault, B-42
- definition, D-10
- and dependability, 33
- dependability benchmarks, D-21
- programming mistakes, D-11
- storage systems, D-6 to D-10
- Tandem Computers, D-12 to D-13
- VAX systems, **C-44**

Fault tolerance

- and adaptive routing, F-94
- commercial interconnection networks, F-66 to F-69
- DECstation 5000 reboots, **F-69**
- dependability benchmarks, D-21

RAID, **D-7**

- SAN example, F-74
- WSC memory, 473–474
- WSC network, 461

Fault-tolerant routing, commercial interconnection networks, F-66 to F-67

FC, *see* Fibre Channel (FC)FC-AL, *see* Fibre Channel Arbitrated Loop (FC-AL)FC-SW, *see* Fibre Channel Switched (FC-SW)

Feature size

- dependability, 33
- integrated circuits, 19–21

FEC, *see* Forward error correction (FEC)

Federal Communications Commission (FCC), telephone company outages, D-15

Fermi GPU

- architectural innovations, 305–308
- future features, 333
- Grid mapping, **293**
- multithreaded SIMD Processor, **307**
- NVIDIA, 291, 305
- SIMD, 296–297
- SIMD Thread Scheduler, **306**

Fermi Tesla, GPU computing history, L-52

Fermi Tesla GTX 280

- GPU comparison, 324–325, **325**
- memory bandwidth, 328
- raw/relative GPU performance, **328**
- synchronization, 329
- weaknesses, 330

Fermi Tesla GTX 480

- floorplan, **295**
- GPU comparisons, 323–330, **325**

Fetch-and-increment

- large-scale multiprocessor synchronization, I-20 to I-21
- sense-reversing barrier, **I-21**
- synchronization, 388

Fetching, *see* Data fetching

Fetch stage, TI 320C55 DSP, E-7

FFT, *see* Fast Fourier transformation (FFT)

Fibre Channel (FC), F-64, F-67, F-102

- file system benchmarking, **D-20**
- NetApp FAS6000 filer, D-42
- Fibre Channel Arbitrated Loop (FC-AL), F-102
 - block servers vs. filers, D-35
 - SCSI history, L-81
- Fibre Channel Switched (FC-SW), F-102
- Field-programmable gate arrays (FPGAs), WSC array switch, 443
- FIFO, *see* First-in first-out (FIFO)
- Filers
 - vs. block servers, D-34 to D-35
 - NetApp FAS6000 filer, D-41 to D-42
- Filer servers, SPEC benchmarking, D-20 to D-21
- Filters, radio receiver, **E-23**
- Fine-grained multithreading
 - definition, 224–226
 - Sun T1 effectiveness, 226–229
- Fingerprint, storage system, D-49
- Finite-state machine, routing
 - implementation, F-57
- Firmware, network interfaces, F-7
- First-in first-out (FIFO)
 - block replacement, B-9
 - cache misses, **B-10**
 - definition, D-26
 - Tomasulo's algorithm, **173**
- First-level caches, *see also* L1 caches
 - ARM Cortex-A8, 114
 - cache optimization, B-30 to B-32
 - hit time/power reduction, 79–80
 - inclusion, B-35
 - interconnection network, F-87
 - Itanium 2, **H-41**
 - memory hierarchy, B-48 to B-49
 - miss rate calculations, B-31 to B-35
 - parameter ranges, **B-42**
 - technology trends, 18
 - virtual memory, B-42
- First-reference misses, definition, B-23
- FIT rates, WSC memory, 473–474
- Fixed-field decoding, simple RISC implementation, C-6
- Fixed-length encoding
 - general-purpose registers, **A-6**
 - instruction sets, **A-22**
- ISAs, 14
- Fixed-length vector
 - SIMD, 284
 - vector registers, 264
- Fixed-point arithmetic, DSP, E-5 to E-6
- Flags
 - performance benchmarks, 37
 - performance reporting, 41
 - scoreboarding, C-75
- Flash memory
 - characteristics, 102–104
 - dependability, 104
 - disk storage, D-3 to D-4
 - embedded benchmarks, E-13
 - memory hierarchy design, **72**
 - technology trends, 18
 - WSC cost-performance, 474–475
- FLASH multiprocessor, L-61
- Flexible chaining
 - vector execution time, 269
 - vector processor, G-11
- Floating-point (FP) operations
 - addition
 - denormals, J-26 to J-27
 - overview, J-21 to J-25
 - rules, **J-24**
 - speedup, J-25 to J-26
 - arithmetic intensity, 285–288, **286**
 - branch condition evaluation, A-19
 - branches, **A-20**
 - cache misses, 83–84
 - chip comparison, **J-58**
 - control flow instructions, A-21
 - CPI calculations, 50–51
 - data access benchmarks, **A-15**
 - data dependences, 151
 - data hazards, 169
 - denormal multiplication, J-20 to J-21
 - denormals, J-14 to J-15
 - desktop RISCs, **K-13**, **K-17**, **K-23**
 - DSP media extensions, E-10 to E-11
 - dynamic scheduling with
 - Tomasulo's algorithm, 171–172, **173**
 - early computer arithmetic, J-64 to J-65
 - exceptions, J-34 to J-35
 - exception stopping/restarting, C-47
 - fused multiply-add, J-32 to J-33
 - IBM 360, K-85
- IEEE 754 FP standard, **J-16**
- ILP exploitation, 197–199
- ILP exposure, 157–158
- ILP in perfect processor, **215**
- ILP for realizable processors, 216–218
- independent, **C-54**
- instruction operator categories, **A-15**
- integer conversions, J-62
- Intel Core i7, **240**, 241
- Intel 80x86, K-52 to K-55, **K-54**, **K-61**
- Intel 80x86 registers, **K-48**
- ISA performance and efficiency
 - prediction, 241
- Itanium 2, **H-41**
- iterative division, J-27 to J-31
- latencies, **157**
- and memory bandwidth, J-62
- MIPS, A-38 to A-39
 - Tomasulo's algorithm, **173**
- MIPS exceptions, C-49
- MIPS operations, A-35
- MIPS pipeline, **C-52**
 - basic considerations, C-51 to C-54
 - execution, C-71
 - performance, C-60 to C-61, **C-61**
 - scoreboarding, C-72
 - stalls, **C-62**
- MIPS precise exceptions, C-58 to C-60
- MIPS R4000, C-65 to C-67, **C-66 to C-67**
- MIPS scoreboarding, C-77
- MIPS with scoreboard, C-73
- misspeculation instructions, **212**
- Multimedia SIMD Extensions, 285
- multimedia support, K-19
- multiple lane vector unit, **273**
- multiple outstanding, **C-54**
- multiplication
 - examples, **J-19**
 - overview, J-17 to J-20
- multiplication precision, J-21
- number representation, J-15 to J-16
- operand sizes/types, 12
- overflow, J-11
- overview, J-13 to J-14
- parallelism vs. window size, **217**

- Floating-point operations (*continued*)
 - pipeline hazards and forwarding, C-55 to C-57
 - pipeline structural hazards, C-16
 - precisions, J-33 to J-34
 - remainder, J-31 to J-32
 - ROB commit, 187
 - SMT, 398–400
 - SPARC, K-31
 - SPEC benchmarks, **39**
 - special values, J-14 to J-15
 - stalls from RAW hazards, **C-55**
 - static branch prediction, C-26 to C-27
 - Tomasulo's algorithm, **185**
 - underflow, J-36 to J-37, J-62
 - VAX, **B-73**
 - vector chaining, G-11
 - vector sequence chimes, 270
 - VLIW processors, **195**
 - VMIPS, 264
 - Floating-point registers (FPRs)
 - IA-64, H-34
 - IBM Blue Gene/L, I-42
 - MIPS data transfers, A-34
 - MIPS operations, A-36
 - MIPS64 architecture, A-34
 - write-back, **C-56**
 - Floating-point square root (FPSQR)
 - calculation, 47–48
 - CPI calculations, 50–51
 - Floating Point Systems AP-120B, L-28
 - Floppy disks, L-78
 - Flow-balanced state, **D-23**
 - Flow control
 - and arbitration, F-21
 - congestion management, F-65
 - direct networks, F-38 to F-39
 - format, F-58
 - interconnection networks, F-10 to F-11
 - system area network history, F-100 to F-101
 - Fluent, F-76, **F-77**
 - Flush, branch penalty reduction, C-22
 - FM, *see* Frequency modulation (FM)
 - Form factor, interconnection networks, F-9 to F-12
 - FORTRAN
 - compiler types and classes, A-28
 - compiler vectorization, G-14, **G-15**
 - dependence analysis, H-6
 - integer division/remainder, **J-12**
 - loop-level parallelism
 - dependences, 320–321
 - MIPS scoreboarding, C-77
 - performance measurement history, L-6
 - return address predictors, 206
 - Forward error correction (FEC), DSP, E-5 to E-7
 - Forwarding, *see also* Bypassing
 - ALUs, **C-40 to C-41**
 - data hazard stall minimization, C-16 to C-19, **C-18**
 - dynamically scheduled pipelines, C-70 to C-71
 - load instruction, **C-20**
 - longer latency pipelines, C-54 to C-58
 - operand, **C-19**
 - Forwarding table
 - routing implementation, F-57
 - switch microarchitecture
 - pipelining, F-60
 - Forward path, cell phones, E-24
 - Fourier-Motzkin algorithm, L-31
 - Fourier transform, DSP, E-5
 - Four-way conflict misses, definition, B-23
 - FP, *see* Floating-point (FP) operations
 - FPGAs, *see* Field-programmable gate arrays (FPGAs)
 - FPRs, *see* Floating-point registers (FPRs)
 - FPSQR, *see* Floating-point square root (FPSQR)
 - Frame pointer, VAX, K-71
 - Freeze, branch penalty reduction, C-22
 - Frequency modulation (FM), wireless networks, E-21
 - Front-end stage, Itanium 2, H-42
 - FU, *see* Functional unit (FU)
 - Fujitsu Primergy BX3000 blade server, F-85
 - Fujitsu VP100, L-45, L-47
 - Fujitsu VP200, L-45, L-47
 - Full access
 - dimension-order routing, F-47 to F-48
 - interconnection network topology, F-29
 - Full adders, J-2, **J-3**
 - Fully associative cache
 - block placement, B-7
 - conflict misses, B-23
 - direct-mapped cache, **B-9**
 - memory hierarchy basics, 74
 - Fully connected topology
 - distributed switched networks, F-34
 - NEWS communication, F-43
 - Functional hazards
 - ARM Cortex-A8, 233
 - microarchitectural techniques case study, 247–254
 - Functional unit (FU)
 - FP operations, **C-66**
 - instruction execution example, **C-80**
 - Intel Core i7, **237**
 - Itanium 2, H-41 to H-43
 - latencies, **C-53**
 - MIPS pipeline, **C-52**
 - MIPS scoreboarding, C-75 to C-80
 - OCNs, F-3
 - vector add instruction, **272**, 272–273
 - VMIPS, 264
 - Function calls
 - GPU programming, 289
 - NVIDIA GPU Memory structures, 304–305
 - PTX assembler, 301
 - Function pointers, control flow
 - instruction addressing modes, A-18
 - Fused multiply-add, floating point, J-32 to J-33
 - Future file, precise exceptions, C-59
- ## G
- Gateways, Ethernet, F-79
 - Gather-Scatter
 - definition, **309**
 - GPU comparisons, 329
 - multimedia instruction compiler support, A-31
 - sparse matrices, G-13 to G-14
 - vector architectures, 279–280
 - GCD, *see* Greatest common divisor (GCD) test
 - GDDR, *see* Graphics double data rate (GDDR)

- GDram, *see* Graphics dynamic random-access memory (GDram)
- GE 645, L-9
- General-Purpose Computing on GPUs (GPGPU), L-51 to L-52
- General-purpose electronic computers, historical background, L-2 to L-4
- General-purpose registers (GPRs)
 - advantages/disadvantages, **A-6**
 - IA-64, H-38
 - Intel 80x86, **K-48**
 - ISA classification, A-3 to A-5
 - MIPS data transfers, A-34
 - MIPS operations, A-36
 - MIPS64, A-34
 - VMIPS, 265
- GENI, *see* Global Environment for Network Innovation (GENI)
- Geometric means, example
 - calculations, 43–44
- GFS, *see* Google File System (GFS)
- Gibson mix, L-6
- Giga Thread Engine, definition, **292, 314**
- Global address space, segmented
 - virtual memory, B-52
- Global code scheduling
 - example, **H-16**
 - parallelism, H-15 to H-23
 - superblock scheduling, H-21 to H-23, **H-22**
 - trace scheduling, H-19 to H-21, **H-20**
- Global common subexpression
 - elimination, compiler structure, A-26
- Global data area, and compiler technology, A-27
- Global Environment for Network Innovation (GENI), F-98
- Global load/store, definition, **309**
- Global Memory
 - definition, **292, 314**
 - GPU programming, 290
 - locks via coherence, 390
- Global miss rate
 - definition, B-31
 - multilevel caches, B-33
- Global optimizations
 - compilers, A-26, A-29
 - optimization types, **A-28**
- Global Positioning System, CDMA, E-25
- Global predictors
 - Intel Core i7, 166
 - tournament predictors, 164–166
- Global scheduling, ILP, VLIW
 - processor, 194
- Global system for mobile
 - communication (GSM), cell phones, E-25
- Goldschmidt's division algorithm, J-29, J-61
- Goldstine, Herman, L-2 to L-3
- Google
 - Bigtable, 438, 441
 - cloud computing, 455
 - cluster history, L-62
 - containers, L-74
 - MapReduce, **437, 458–459, 459**
 - server CPUs, **440**
 - server power-performance
 - benchmarks, 439–441
 - WSCs, 432, 449
 - containers, 464–465, **465**
 - cooling and power, 465–468
 - monitoring and repairing, 469–470
 - PUE, **468**
 - servers, **467, 468–469**
- Google App Engine, L-74
- Google Clusters
 - memory dependability, 104
 - power consumption, F-85
- Google File System (GFS)
 - MapReduce, 438
 - WSC storage, 442–443
- Google Goggles
 - PMDs, 6
 - user experience, 4
- Google search
 - shared-memory workloads, 369
 - workload demands, 439
- Gordon Bell Prize, L-57
- GPGPU (General-Purpose Computing on GPUs), L-51 to L-52
- GPRs, *see* General-purpose registers (GPRs)
- GPU (Graphics Processing Unit)
 - banked and graphics memory, 322–323
- computing history, L-52
- definition, 9
- DLP
 - basic considerations, 288
 - basic PTX thread instructions, **299**
 - conditional branching, 300–303
 - coprocessor relationship, 330–331
 - definitions, **309**
 - Fermi GPU architecture
 - innovations, 305–308
 - Fermi GTX 480 floorplan, **295**
 - GPUs vs. vector architectures, 308–312, **310**
 - mapping examples, **293**
 - Multimedia SIMD comparison, **312**
 - multithreaded SIMD Processor
 - block diagram, **294**
 - NVIDIA computational structures, 291–297
 - NVIDIA/CUDA and AMD terminology, 313–315
 - NVIDIA GPU ISA, 298–300
 - NVIDIA GPU Memory structures, **304, 304–305**
 - programming, 288–291
 - SIMD thread scheduling, **297**
 - terminology, **292**
- fine-grained multithreading, 224
- future features, 332
- gather/scatter operations, 280
- historical background, L-50
- loop-level parallelism, 150
- vs. MIMD with Multimedia SIMD, 324–330
- mobile client/server features, 324, **324**
- power/DLP issues, 322
- raw/relative performance, **328**
- Roofline model, **326**
- scalable, L-50 to L-51
- strided access-TLB interactions, 323
- thread count and memory
 - performance, 332
- TLP, 346
- vector kernel implementation, 334–336
- vs. vector processor operation, 276

- GPU Memory
 - caches, 306
 - CUDA program, 289
 - definition, **292, 309, 314**
 - future architectures, 333
 - GPU programming, 288
 - NVIDIA, **304**, 304–305
 - splitting from main memory, 330
 - Gradual underflow, J-15, J-36
 - Grain size
 - MIMD, 10
 - TLP, 346
 - Grant phase, arbitration, F-49
 - Graph coloring, register allocation,
 - A-26 to A-27
 - Graphics double data rate (GDDR)
 - characteristics, 102
 - Fermi GTX 480 GPU, 295, 324
 - Graphics dynamic random-access memory (GDRAM)
 - bandwidth issues, 322–323
 - characteristics, 102
 - Graphics-intensive benchmarks,
 - desktop performance, 38
 - Graphics pipelines, historical
 - background, L-51
 - Graphics Processing Unit, *see* GPU (Graphics Processing Unit)
 - Graphics synchronous dynamic
 - random-access memory (GSDRAM),
 - characteristics, 102
 - Graphics Synthesizer, Sony
 - PlayStation 2, **E-16**, E-16 to E-17
 - Greater than condition code,
 - PowerPC, K-10 to K-11
 - Greatest common divisor (GCD) test,
 - loop-level parallelism
 - dependences, 319, H-7
 - Grid
 - arithmetic intensity, 286
 - CUDA parallelism, 290
 - definition, **292, 309, 313**
 - and GPU, 291
 - GPU Memory structures, **304**
 - GPU terms, 308
 - mapping example, **293**
 - NVIDIA GPU computational
 - structures, 291
 - SIMD Processors, 295
 - Thread Blocks, 295
 - Grid computing, L-73 to L-74
 - Grid topology
 - characteristics, F-36
 - direct networks, **F-37**
 - GSDRAM, *see* Graphics synchronous
 - dynamic random-access
 - memory (GSDRAM)
 - GSM, *see* Global system for mobile
 - communication (GSM)
 - Guest definition, 108
 - Guest domains, Xen VM, 111
- ## H
- Hadoop, WSC batch processing, 437
 - Half adders, J-2
 - Half words
 - aligned/misaligned addresses, **A-8**
 - memory address interpretation,
 - A-7 to A-8
 - MIPS data types, A-34
 - operand sizes/types, 12
 - as operand type, A-13 to A-14
 - Handshaking, interconnection
 - networks, F-10
 - Hard drive, power consumption, **63**
 - Hard real-time systems, definition, E-3
 - to E-4
 - Hardware
 - as architecture component, 15
 - cache optimization, **96**
 - compiler scheduling support, L-30
 - to L-31
 - compiler speculation support
 - memory references, H-32
 - overview, H-27
 - preserving exception behavior,
 - H-28 to H-32
 - description notation, **K-25**
 - energy/performance fallacies, 56
 - for exposing parallelism, H-23 to
 - H-27
 - ILP approaches, 148, 214–215
 - interconnection networks, F-9
 - pipeline hazard detection, **C-38**
 - Virtual Machines protection, 108
 - WSC cost-performance, 474
 - WSC running service, 434–435
 - Hardware-based speculation
 - basic algorithm, **191**
 - data flow execution, 184
 - FP unit using Tomasulo's
 - algorithm, 185
 - ILP
 - data flow execution, 184
 - with dynamic scheduling and
 - multiple issue, 197–202
 - FP unit using Tomasulo's
 - algorithm, 185
 - key ideas, 183–184
 - multiple-issue processors, **198**
 - reorder buffer, 184–192
 - vs. software speculation,
 - 221–222
 - key ideas, 183–184
 - Hardware faults, storage systems,
 - D-11
 - Hardware prefetching
 - cache optimization, 131–133
 - miss penalty/rate reduction, 91–92
 - NVIDIA GPU Memory structures,
 - 305
 - SPEC benchmarks, **92**
 - Hardware primitives
 - basic types, 387–389
 - large-scale multiprocessor
 - synchronization, I-18 to
 - I-21
 - synchronization mechanisms,
 - 387–389
 - Harvard architecture, L-4
 - Hazards, *see also* Data hazards
 - branch hazards, C-21 to C-26,
 - C-39 to C-42, **C-42**
 - control hazards, 235, C-11
 - detection, hardware, **C-38**
 - dynamically scheduled pipelines,
 - C-70 to C-71
 - execution sequences, C-80
 - functional hazards, 233, 247–254
 - instruction set complications, C-50
 - longer latency pipelines, C-54 to
 - C-58
 - structural hazards, 268–269, C-11,
 - C-13 to C-16, C-71,
 - C-78 to C-79
 - HCAs, *see* Host channel adapters
 - (HCAs)
 - Header
 - messages, F-6
 - packet format, **F-7**

- switch microarchitecture
 - pipelining, F-60
 - TCP/IP, **F-84**
 - Head-of-line (HOL) blocking
 - congestion management, F-64
 - switch microarchitecture, F-58 to F-59, **F-59**, F-60, F-62
 - system area network history, F-101
 - virtual channels and throughput, F-93
 - Heap, and compiler technology, A-27 to A-28
 - HEP processor, L-34
 - Heterogeneous architecture,
 - definition, 262
 - Hewlett-Packard AlphaServer,
 - F-100
 - Hewlett-Packard PA-RISC
 - addressing modes, **K-5**
 - arithmetic/logical instructions, **K-11**
 - characteristics, **K-4**
 - conditional branches, K-12, **K-17**, **K-34**
 - constant extension, **K-9**
 - conventions, **K-13**
 - data transfer instructions, **K-10**
 - EPIC, L-32
 - features, **K-44**
 - floating-point precisions, J-33
 - FP instructions, **K-23**
 - MIPS core extensions, K-23
 - multimedia support, K-18, **K-18**, K-19
 - unique instructions, K-33 to K-36
 - Hewlett-Packard PA-RISC MAX2,
 - multimedia support, **E-11**
 - Hewlett-Packard Precision
 - Architecture, integer arithmetic, J-12
 - Hewlett-Packard ProLiant BL10e G2
 - Blade server, F-85
 - Hewlett-Packard ProLiant SL2x170z
 - G6, SPECPower benchmarks, **463**
 - Hewlett-Packard RISC
 - microprocessors, vector processor history, G-26
 - Higher-radix division, J-54 to J-55
 - Higher-radix multiplication, integer, J-48
 - High-level language computer
 - architecture (HLLCA), L-18 to L-19
 - High-level optimizations, compilers, A-26
 - Highly parallel memory systems, case studies, 133–136
 - High-order functions, control flow
 - instruction addressing modes, A-18
 - High-performance computing (HPC)
 - InfiniBand, F-74
 - interconnection network
 - characteristics, **F-20**
 - interconnection network topology, **F-44**
 - storage area network history, F-102
 - switch microarchitecture, F-56
 - vector processor history, G-27
 - write strategy, B-10
 - vs. WSCs, 432, 435–436
 - Hillis, Danny, L-58, L-74
 - Histogram, D-26 to D-27
 - History file, precise exceptions, C-59
 - Hitachi S810, L-45, L-47
 - Hitachi SuperH
 - addressing modes, K-5, **K-6**
 - arithmetic/logical instructions, **K-24**
 - branches, K-21
 - characteristics, **K-4**
 - condition codes, K-14
 - data transfer instructions, **K-23**
 - embedded instruction format, **K-8**
 - multiply-accumulate, **K-20**
 - unique instructions, K-38 to K-39
 - Hit time
 - average memory access time, B-16 to B-17
 - first-level caches, 79–80
 - memory hierarchy basics, 77–78
 - reduction, 78, B-36 to B-40
 - way prediction, 81–82
 - HLLCA, *see* High-level language computer architecture (HLLCA)
 - HOL, *see* Head-of-line blocking (HOL)
 - Home node, directory-based cache
 - coherence protocol basics, 382
 - Hop count, definition, F-30
 - Hops
 - direct network topologies, F-38
 - routing, F-44
 - switched network topologies, F-40
 - switching, F-50
 - Host channel adapters (HCAs)
 - historical background, L-81
 - switch vs. NIC, F-86
 - Host definition, 108, 305
 - Hot swapping, fault tolerance, F-67
 - HPC, *see* High-performance computing (HPC)
 - HPC Challenge, vector processor
 - history, G-28
 - HP-Compaq servers
 - price-performance differences, 441
 - SMT, 230
 - HPSm, L-29
 - Hypercube networks
 - characteristics, F-36
 - deadlock, F-47
 - direct networks, **F-37**
 - vs. direct networks, F-92
 - NEWS communication, F-43
 - HyperTransport, F-63
 - NetApp FAS6000 filer, D-42
 - Hypertransport, AMD Opteron cache
 - coherence, 361
 - Hypervisor, characteristics, 108
- I**
- IAS machine, L-3, L-5 to L-6
 - IBM
 - Chipkill, 104
 - cluster history, L-62, L-72
 - computer history, L-5 to L-6
 - early VM work, L-10
 - magnetic storage, L-77 to L-78
 - multiple-issue processor
 - development, L-28
 - RAID history, L-79 to L-80
 - IBM 360
 - address space, B-58
 - architecture, K-83 to K-84
 - architecture flaws and success, K-81
 - branch instructions, K-86
 - characteristics, **K-42**
 - computer architecture definition, L-17 to L-18
 - instruction execution frequencies, **K-89**

- IBM 360 (*continued*)
 - instruction operator categories, **A-15**
 - instruction set, K-85 to K-88
 - instruction set complications, C-49 to C-50
 - integer/FP R-R operations, K-85
 - I/O bus history, L-81
 - memory hierarchy development, L-9 to L-10
 - parallel processing debates, L-57
 - protection and ISA, 112
 - R-R instructions, K-86
 - RS and SI format instructions, K-87
 - RX format instructions, K-86 to K-87
 - SS format instructions, K-85 to K-88
- IBM 360/85, L-10 to L-11, L-27
- IBM 360/91
 - dynamic scheduling with Tomasulo's algorithm, 170–171
 - early computer arithmetic, J-63
 - history, L-27
 - speculation concept origins, L-29
- IBM 370
 - architecture, K-83 to K-84
 - characteristics, **K-42**
 - early computer arithmetic, J-63
 - integer overflow, **J-11**
 - protection and ISA, 112
 - vector processor history, G-27
 - Virtual Machines, 110
- IBM 370/158, L-7
- IBM 650, L-6
- IBM 701, L-5 to L-6
- IBM 702, L-5 to L-6
- IBM 704, L-6, L-26
- IBM 705, L-6
- IBM 801, L-19
- IBM 3081, L-61
- IBM 3090 Vector Facility, vector processor history, G-27
- IBM 3840 cartridge, L-77
- IBM 7030, L-26
- IBM 9840 cartridge, L-77
- IBM AS/400, L-79
- IBM Blue Gene/L, F-4
 - adaptive routing, F-93
 - cluster history, L-63
 - commercial interconnection networks, F-63
 - computing node, I-42 to I-44, **I-43**
 - as custom cluster, I-41 to I-42
 - deterministic vs. adaptive routing, F-52 to F-55
 - fault tolerance, F-66 to F-67
 - link bandwidth, F-89
 - low-dimensional topologies, F-100
 - parallel processing debates, L-58
 - software overhead, F-91
 - switch microarchitecture, F-62
 - system, **I-44**
 - system area network history, F-101 to F-102
 - 3D torus network, F-72 to F-74
 - topology, F-30, F-39
- IBM CodePack, RISC code size, A-23
- IBM CoreConnect
 - cross-company interoperability, F-64
 - OCNs, F-3
- IBM eServer p5 processor
 - performance/cost benchmarks, **409**
 - SMT and ST performance, **399**
 - speedup benchmarks, **408**, 408–409
- IBM Federation network interfaces, F-17 to F-18
- IBM J9 JVM
 - real-world server considerations, 52–55
 - WSC performance, **463**
- IBM PCs, architecture flaws vs. success, A-45
- IBM Power processors
 - branch-prediction buffers, **C-29**
 - characteristics, **247**
 - exception stopping/restarting, C-47
 - MIPS precise exceptions, C-59
 - shared-memory multiprogramming workload, 378
- IBM Power 1, L-29
- IBM Power 2, L-29
- IBM Power 4
 - multithreading history, L-35
 - peak performance, **58**
 - recent advances, L-33 to L-34
- IBM Power 5
 - characteristics, **F-73**
 - Itanium 2 comparison, **H-43**
 - manufacturing cost, **62**
- multithreading/
 - multithreading-based performance, 398–400
 - multithreading history, L-35
- IBM Power 7
 - vs. Google WSC, 436
- ideal processors, 214–215
- multicore processor performance, 400–401
- multithreading, **225**
- IBM Pulsar processor, L-34
- IBM RP3, L-60
- IBM RS/6000, L-57
- IBM RT-PC, L-20
- IBM SAGE, L-81
- IBM servers, economies of scale, 456
- IBM Stretch, L-6
- IBM zSeries, vector processor history, G-27
- IC, *see* Instruction count (IC)
- I-caches
 - case study examples, B-63
 - way prediction, 81–82
- ICR, *see* Idle Control Register (ICR)
- ID, *see* Instruction decode (ID)
- Ideal pipeline cycles per instruction, ILP concepts, 149
- Ideal processors, ILP hardware model, 214–215, 219–220
- IDE disks, Berkeley's Tertiary Disk project, D-12
- Idle Control Register (ICR), TI TMS320C55 DSP, E-8
- Idle domains, TI TMS320C55 DSP, E-8
- IEEE 754 floating-point standard, **J-16**
- IEEE 1394, Sony PlayStation 2 Emotion Engine case study, E-15
- IEEE arithmetic
 - floating point, J-13 to J-14
 - addition, J-21 to J-25
 - exceptions, J-34 to J-35
 - remainder, J-31 to J-32
 - underflow, J-36
 - historical background, J-63 to J-64
 - iterative division, J-30
 - x vs. 0 –x, J-62
 - NaN, J-14
 - rounding modes, **J-20**
 - single-precision numbers, J-15 to J-16

- IEEE standard 802.3 (Ethernet), F-77
 - to F-79
 - LAN history, F-99
- IF, *see* Instruction fetch (IF) cycle
- IF statement handling
 - control dependences, 154
 - GPU conditional branching, 300, 302–303
 - memory consistency, 392
 - vectorization in code, 271
 - vector-mask registers, 267, 275–276
- Illiac IV, F-100, L-43, L-55
- ILP, *see* Instruction-level parallelism (ILP)
- Immediate addressing mode
 - ALU operations, **A-12**
 - basic considerations, A-10 to A-11
 - MIPS, 12
 - MIPS instruction format, A-35
 - MIPS operations, A-37
 - value distribution, **A-13**
- IMPACT, L-31
- Implicit operands, ISA classifications, A-3
- Implicit unit stride, GPUs vs. vector architectures, 310
- Imprecise exceptions
 - data hazards, 169–170
 - floating-point, 188
- IMT-2000, *see* International Mobile Telephony 2000 (IMT-2000)
- Inactive power modes, WSCs, 472
- Inclusion
 - cache hierarchy, 397–398
 - implementation, 397–398
 - invalidate protocols, 357
 - memory hierarchy history, L-11
- Indexed addressing
 - Intel 80x86, K-49, **K-58**
 - VAX, K-67
- Indexes
 - address translation during, B-36 to B-40
 - AMD Opteron data cache, B-13 to B-14
 - ARM Cortex-A8, **115**
 - recurrences, H-12
 - size equations, B-22
- Index field, block identification, B-8
- Index vector, gather/scatter operations, 279–280
- Indirect addressing, VAX, K-67
- Indirect networks, definition, F-31
- Inexact exception
 - floating-point arithmetic, J-35
 - floating-point underflow, J-36
- InfiniBand, F-64, F-67, F-74 to F-77
 - cluster history, L-63
 - packet format, **F-75**
 - storage area network history, F-102
 - switch vs. NIC, **F-86**
 - system area network history, F-101
- Infinite population model, queuing model, D-30
- In flight instructions, ILP hardware model, 214
- Information tables, examples, 176–177
- Infrastructure costs
 - WSC, 446–450, 452–455, **453**
 - WSC efficiency, 450–452
- Initiation interval, MIPS pipeline FP operations, C-52 to C-53
- Initiation rate
 - floating-point pipeline, C-65 to C-66
 - memory banks, 276–277
 - vector execution time, 269
- Inktomi, L-62, L-73
- In-order commit
 - hardware-based speculation, 188–189
 - speculation concept origins, L-29
- In-order execution
 - average memory access time, B-17 to B-18
 - cache behavior calculations, B-18
 - cache miss, B-2 to B-3
 - dynamic scheduling, 168–169
 - IBM Power processors, **247**
 - ILP exploitation, 193–194
 - multiple-issue processors, **194**
 - superscalar processors, 193
- In-order floating-point pipeline, dynamic scheduling, 169
- In-order issue
 - ARM Cortex-A8, 233
 - dynamic scheduling, 168–170, C-71
 - ISA, 241
- In-order scalar processors, VMIPS, 267
- Input buffered switch
 - HOL blocking, **F-59**, F-60
 - microarchitecture, F-57, **F-57**
 - pipelined version, **F-61**
- Input-output buffered switch, microarchitecture, F-57
- Instruction cache
 - AMD Opteron example, **B-15**
 - antialiasing, B-38
 - application/OS misses, **B-59**
 - branch prediction, C-28
 - commercial workload, 373
 - GPU Memory, 306
 - instruction fetch, 202–203, 237
 - ISA, 241
 - MIPS R4000 pipeline, C-63
 - miss rates, 161
 - multiprogramming workload, 374–375
 - prefetch, 236
 - RISCs, A-23
 - TI TMS320C55 DSP, E-8
- Instruction commit
 - hardware-based speculation, 184–185, 187–188, **188**, 190
 - instruction set complications, C-49
 - Intel Core i7, 237
 - speculation support, 208–209
- Instruction count (IC)
 - addressing modes, A-10
 - cache performance, B-4, B-16
 - compiler optimization, **A-29**, A-29 to A-30
 - processor performance time, 49–51
 - RISC history, L-22
- Instruction decode (ID)
 - basic MIPS pipeline, C-36
 - branch hazards, C-21
 - data hazards, 169
 - hazards and forwarding, C-55 to C-57
 - MIPS pipeline, C-71
 - MIPS pipeline control, C-36 to C-39
 - MIPS pipeline FP operations, C-53
 - MIPS scoreboarding, C-72 to C-74
 - out-of-order execution, 170
 - pipeline branch issues, C-39 to C-41, **C-42**
 - RISC classic pipeline, C-7 to C-8, C-10

- Instruction decode (*continued*)
 - simple MIPS implementation, C-31
 - simple RISC implementation, C-5 to C-6
- Instruction delivery stage, Itanium 2, H-42
- Instruction fetch (IF) cycle
 - basic MIPS pipeline, C-35 to C-36
 - branch hazards, C-21
 - branch-prediction buffers, C-28
 - exception stopping/restarting, C-46 to C-47
 - MIPS exceptions, C-48
 - MIPS R4000, C-63
 - pipeline branch issues, **C-42**
 - RISC classic pipeline, C-7, C-10
 - simple MIPS implementation, C-31
 - simple RISC implementation, C-5
- Instruction fetch units
 - integrated, 207–208
 - Intel Core i7, 237
- Instruction formats
 - ARM-unique, K-36 to K-37
 - high-level language computer architecture, L-18
 - IA-64 ISA, H-34 to H-35, H-38, **H-39**
 - IBM 360, K-85 to K-88
 - Intel 80x86, **K-49, K-52, K-56 to K-57**
 - M32R-unique, K-39 to K-40
 - MIPS16-unique, K-40 to K-42
 - PA-RISC unique, K-33 to K-36
 - PowerPC-unique, K-32 to K-33
 - RISCs, **K-43**
 - Alpha-unique, K-27 to K-29
 - arithmetic/logical, **K-11, K-15**
 - branches, **K-25**
 - control instructions, **K-12, K-16**
 - data transfers, **K-10, K-14, K-21**
 - desktop/server, **K-7**
 - desktop/server systems, **K-7**
 - embedded DSP extensions, K-19
 - embedded systems, **K-8**
 - FP instructions, **K-13**
 - hardware description notation, **K-25**
 - MIPS64-unique, K-24 to K-27
 - MIPS core, K-6 to K-9
 - MIPS core extensions, K-19 to K-24
 - MIPS unaligned word reads, **K-26**
 - multimedia extensions, K-16 to K-19
 - overview, K-5 to K-6
 - SPARC-unique, K-29 to K-32
 - SuperH-unique, K-38 to K-39
 - Thumb-unique, K-37 to K-38
- Instruction groups, IA-64, H-34
- Instruction issue
 - definition, C-36
 - DLP, 322
 - dynamic scheduling, 168–169, C-71 to C-72
 - ILP, 197, 216–217
 - instruction-level parallelism, 2
 - Intel Core i7, 238
 - Itanium 2, H-41 to H-43
 - MIPS pipeline, C-52
 - multiple issue processor, **198**
 - multithreading, 223, 226
 - parallelism measurement, 215
 - precise exceptions, C-58, C-60
 - processor comparison, 323
 - ROB, 186
 - speculation support, 208, 210
 - Tomasulo's scheme, 175, 182
- Instruction-level parallelism (ILP)
 - ARM Cortex-A8, 233–236, **235–236**
 - basic concepts/challenges, 148–149, **149**
 - “big and dumb” processors, 245
 - branch-prediction buffers, **C-29, C-29 to C-30**
 - compiler scheduling, L-31
 - compiler techniques for exposure, 156–162
 - control dependence, 154–156
 - data dependences, 150–152
 - data flow limit, L-33
 - definition, 9, 149–150
 - dynamic scheduling
 - basic concept, 168–169
 - definition, 168
 - example and algorithms, 176–178
 - multiple issue, speculation, 197–202
 - overcoming data hazards, 167–176
 - Tomasulo's algorithm, 170–176, 178–179, 181–183
- early studies, L-32 to L-33
- exploitation methods, H-22 to H-23
- exploitation statically, H-2
- exposing with hardware support, H-23
- GPU programming, 289
- hardware-based speculation, 183–192
- hardware vs. software speculation, 221–222
- IA-64, H-32
- instruction fetch bandwidth
 - basic considerations, 202–203
 - branch-target buffers, 203–206, **204**
 - integrated units, 207–208
 - return address predictors, 206–207
- Intel Core i7, 236–241
- limitation studies, 213–221
- microarchitectural techniques case study, 247–254
- MIPS scoreboarding, C-77 to C-79
- multicore performance/energy efficiency, 404
- multicore processor performance, 400
- multiple-issue processors, L-30
- multiple issue/static scheduling, 192–196
- multiprocessor importance, 344
- multithreading, basic
 - considerations, 223–226
- multithreading history, L-34 to L-35
- name dependences, 152–153
- perfect processor, **215**
- pipeline scheduling/loop unrolling, 157–162
- processor clock rates, **244**
- realizable processor limitations, 216–218
- RISC development, 2
- SMT on superscalar processors, 230–232
- speculation advantages/disadvantages, 210–211

- speculation and energy efficiency, 211–212
- speculation support, 208–210
- speculation through multiple branches, 211
- speculative execution, 222–223
- Sun T1 fine-grained multithreading effectiveness, 226–229
- switch to DLP/TLP/RLP, 4–5
- TI 320C6x DSP, E-8
- value prediction, 212–213
- Instruction path length, processor performance time, 49
- Instruction prefetch
 - integrated instruction fetch units, 208
 - miss penalty/rate reduction, 91–92
 - SPEC benchmarks, **92**
- Instruction register (IR)
 - basic MIPS pipeline, C-35
 - dynamic scheduling, 170
 - MIPS implementation, C-31
- Instruction set architecture (ISA), *see also* Intel 80x86 processors; Reduced Instruction Set Computer (RISC)
 - addressing modes, A-9 to A-10
 - architect-compiler writer relationship, A-29 to A-30
 - ARM Cortex-A8, 114
 - case studies, A-47 to A-54
 - class code sequence example, **A-4**
 - classification, A-3 to A-7
 - code size-compiler considerations, A-43 to A-44
 - compiler optimization and performance, A-27
 - compiler register allocation, A-26 to A-27
 - compiler structure, A-24 to A-26
 - compiler technology and architecture decisions, A-27 to A-29
 - compiler types and classes, **A-28**
 - complications, C-49 to C-51
 - computer architecture definition, L-17 to L-18
 - control flow instructions
 - addressing modes, A-17 to A-18
 - basic considerations, A-16 to A-17, A-20 to A-21
 - conditional branch options, A-19
 - procedure invocation options, A-19 to A-20
 - Cray X1, G-21 to G-22
 - data access distribution example, **A-15**
 - definition and types, 11–15
 - displacement addressing mode, A-10
 - encoding considerations, A-21 to A-24, **A-22**, A-24
 - first vector computers, L-48
 - flawless design, A-45
 - flaws vs. success, A-44 to A-45
 - GPR advantages/disadvantages, **A-6**
 - high-level considerations, A-39, A-41 to A-43
 - high-level language computer architecture, L-18 to L-19
- IA-64
 - instruction formats, **H-39**
 - instructions, **H-35 to H-37**
 - instruction set basics, H-38
 - overview, H-32 to H-33
 - predication and speculation, H-38 to H-40
- IBM 360, K-85 to K-88
- immediate addressing mode, A-10 to A-11
- literal addressing mode, A-10 to A-11
- memory addressing, A-11 to A-13
- memory address interpretation, A-7 to A-8
- MIPS
 - addressing modes for data transfer, A-34
 - basic considerations, A-32 to A-33
 - control flow instructions, A-37 to A-38
 - data types, A-34
 - dynamic instruction mix, A-41 to A-42, **A-42**
 - FP operations, A-38 to A-39
 - instruction format, **A-35**
 - MIPS operations, A-35 to A-37
 - registers, A-34
 - usage, A-39
- MIPS64, **14**, **A-40**
- multimedia instruction compiler support, A-31 to A-32
- NVIDIA GPU, 298–300
- operand locations, **A-4**
- operands per ALU instruction, **A-6**
- operand type and size, A-13 to A-14
- operations, A-14 to A-16
- operator categories, **A-15**
- overview, K-2
- performance and efficiency
 - prediction, 241–243
- and protection, 112
- RISC code size, A-23 to A-24
- RISC history, L-19 to L-22, **L-21**
- stack architectures, L-16 to L-17
- top 80x86 instructions, **A-16**
- “typical” program fallacy, A-43
- Virtual Machines protection, 107–108
- Virtual Machines support, 109–110
- VMIPS, 264–265
- VMM implementation, 128–129
- Instructions per clock (IPC)
 - ARM Cortex-A8, 236
 - flawless architecture design, A-45
 - ILP for realizable processors, 216–218
- MIPS scoreboarding, C-72
- multiprocessing/
 - multithreading-based performance, 398–400
- processor performance time, 49
- Sun T1 multithreading uncore performance, **229**
- Sun T1 processor, **399**
- Instruction status
 - dynamic scheduling, 177
 - MIPS scoreboarding, C-75
- Integer arithmetic
 - addition speedup
 - carry-lookahead, J-37 to J-41
 - carry-lookahead circuit, **J-38**
 - carry-lookahead tree, **J-40**
 - carry-lookahead tree adder, **J-41**
 - carry-select adder, **J-43**, J-43 to J-44, **J-44**

Integer arithmetic (*continued*)

- carry-skip adder, J-41 to J-43, **J-42**
 - overview, J-37
 - division
 - radix-2 division, **J-55**
 - radix-4 division, **J-56**
 - radix-4 SRT division, **J-57**
 - with single adder, J-54 to J-58
 - FP conversions, J-62
 - language comparison, **J-12**
 - multiplication
 - array multiplier, **J-50**
 - Booth recoding, **J-49**
 - even/odd array, **J-52**
 - with many adders, J-50 to J-54
 - multipass array multiplier, **J-51**
 - signed-digit addition table, **J-54**
 - with single adder, J-47 to J-49, **J-48**
 - Wallace tree, **J-53**
 - multiplication/division, shifting
 - over zeros, J-45 to J-47
 - overflow, **J-11**
 - Radix-2 multiplication/division, **J-4**, J-4 to J-7
 - restoring/nonrestoring division, **J-6**
 - ripple-carry addition, J-2 to J-3, **J-3**
 - signed numbers, J-7 to J-10
 - SRT division, J-45 to J-47, **J-46**
 - systems issues, J-10 to J-13
- Integer operand
- flawed architecture, A-44
 - GCD, 319
 - graph coloring, A-27
 - instruction set encoding, A-23
 - MIPS data types, A-34
 - as operand type, 12, A-13 to A-14
- Integer operations
- addressing modes, A-11
 - ALUs, **A-12**, C-54
 - ARM Cortex-A8, **116**, **232**, **235**, 236
 - benchmarks, **167**, C-69
 - branches, A-18 to A-20, **A-20**
 - cache misses, 83–84
 - data access distribution, **A-15**
 - data dependences, 151
 - dependences, 322

- desktop benchmarks, 38–39
 - displacement values, **A-12**
 - exceptions, C-43, C-45
 - hardware ILP model, **215**
 - hardware vs. software speculation, 221
 - hazards, C-57
 - IBM 360, K-85
 - ILP, 197–200
 - instruction set operations, A-16
 - Intel Core i7, 238, **240**
 - Intel 80x86, K-50 to K-51
 - ISA, 242, A-2
 - Itanium 2, **H-41**
 - longer latency pipelines, C-55
 - MIPS, C-31 to C-32, C-36, C-49, C-51 to C-53
 - MIPS64 ISA, 14
 - MIPS FP pipeline, C-60
 - MIPS R4000 pipeline, C-61, C-63, C-70
 - misspeculation, **212**
 - MVL, 274
 - pipeline scheduling, 157
 - precise exceptions, C-47, C-58, C-60
 - processor clock rate, 244
 - R4000 pipeline, **C-63**
 - realizable processor ILP, 216–218
 - RISC, C-5, C-11
 - scoreboarding, C-72 to C-73, **C-76**
 - SIMD processor, 307
 - SPARC, K-31
 - SPEC benchmarks, **39**
 - speculation through multiple branches, 211
 - static branch prediction, C-26 to C-27
 - T1 multithreading uncore performance, 227–229
 - Tomasulo's algorithm, 181
 - tournament predictors, 164
 - VMIPS, 265
- Integer registers
- hardware-based speculation, 192
 - IA-64, H-33 to H-34
 - MIPS dynamic instructions, A-41 to A-42
 - MIPS floating-point operations, A-39
 - MIPS64 architecture, A-34
 - VLIW, 194

Integrated circuit basics

- cell phones, E-24, **E-24**
 - cost trends, 28–32
 - dependability, 33–36
 - logic technology, 17
 - microprocessor developments, 2
 - power and energy, 21–23
 - scaling, 19–21
- Intel 80286, L-9
- Intel Atom 230
- processor comparison, **242**
 - single-threaded benchmarks, **243**
- Intel Atom processors
- ISA performance and efficiency prediction, 241–243
 - performance measurement, 405–406
 - SMT, 231
 - WSC memory, 474
 - WSC processor cost-performance, 473
- Intel Core i7
- vs. Alpha processors, **368**
 - architecture, 15
 - basic function, 236–238
 - “big and dumb” processors, 245
 - branch predictor, 166–167
 - clock rate, **244**
 - dynamic scheduling, 170
 - GPU comparisons, 324–330, **325**
 - hardware prefetching, 91
 - ISA performance and efficiency prediction, 241–243
 - L2/L3 miss rates, **125**
 - memory hierarchy basics, 78, 117–124, **119**
 - memory hierarchy design, 73
 - memory performance, 122–124
 - MESIF protocol, 362
 - microprocessor die example, **29**
 - miss rate benchmarks, **123**
 - multibanked caches, 86
 - multithreading, **225**
 - nonblocking cache, 83
 - performance, **239**, 239–241, **240**
 - performance/energy efficiency, 401–405
 - pipelined cache access, 82
 - pipeline structure, **237**
 - processor comparison, **242**
 - raw/relative GPU performance, **328**
 - Roofline model, 286–288, **287**

- Intel Core i7 (*continued*)
 - single-threaded benchmarks, **243**
 - SMP limitations, 363
 - SMT, 230–231
 - snooping cache coherence
 - implementation, 365
 - three-level cache hierarchy, **118**
 - TLB structure, **118**
 - write invalid protocol, 356
- Intel 80x86 processors
 - address encoding, **K-58**
 - addressing modes, **K-58**
 - address space, B-58
 - architecture flaws and success, K-81
 - architecture flaws *vs.* success,
 - A-44 to A-45
 - Atom, 231
 - cache performance, B-6
 - characteristics, **K-42**
 - common exceptions, **C-44**
 - comparative operation
 - measurements, K-62 to K-64
 - floating-point operations, K-52 to K-55, **K-54**, **K-61**
 - instruction formats, **K-56 to K-57**
 - instruction lengths, **K-60**
 - instruction mix, **K-61 to K-62**
 - instructions *vs.* DLX, **K-63 to K-64**
 - instruction set encoding, A-23, K-55
 - instruction set usage
 - measurements, K-56 to K-64
 - instructions and functions, **K-52**
 - instruction types, **K-49**
 - integer operations, K-50 to K-51
 - integer overflow, **J-11**
 - Intel Core i7, 117
 - ISA, 11–12, 14–15, A-2
 - memory accesses, B-6
 - memory addressing, A-8
 - memory hierarchy development, L-9
 - multimedia support, K-17
 - operand addressing mode, **K-59**, K-59 to K-60
 - operand type distribution, **K-59**
 - overview, K-45 to K-47
 - process protection, B-50
 - vs.* RISC, 2, A-3
 - segmented scheme, **K-50**
 - system evolution, **K-48**
 - top instructions, **A-16**
 - typical operations, **K-53**
 - variable encoding, A-22 to A-23
 - virtualization issues, **128**
 - Virtual Machines ISA support, 109
 - Virtual Machines and virtual
 - memory and I/O, 110
- Intel 8087, floating point remainder, J-31
- Intel i860, K-16 to K-17, L-49, L-60
- Intel IA-32 architecture
 - call gate, B-54
 - descriptor table, B-52
 - instruction set complications, C-49 to C-51
 - OCNs, F-3, F-70
 - segment descriptors, **B-53**
 - segmented virtual memory, B-51 to B-54
- Intel IA-64 architecture
 - compiler scheduling history, L-31
 - conditional instructions, H-27
 - explicit parallelism, H-34 to H-35
 - historical background, L-32
 - ISA
 - instruction formats, **H-39**
 - instructions, **H-35 to H-37**
 - instruction set basics, H-38
 - overview, H-32 to H-33
 - predication and speculation, H-38 to H-40
 - Itanium 2 processor
 - instruction latency, **H-41**
 - overview, H-40 to H-41
 - performance, H-43, **H-43**
 - multiple issue processor
 - approaches, **194**
 - parallelism exploitation statically, H-2
 - register model, H-33 to H-34
 - RISC history, L-22
 - software pipelining, H-15
 - synchronization history, L-64
- Intel iPSC 860, L-60
- Intel Itanium 2
 - “big and dumb” processors, 245
 - clock rate, **244**
- IA-64
 - functional units and instruction
 - issue, H-41 to H-43
 - instruction latency, **H-41**
 - overview, H-40 to H-41
 - performance, H-43
 - peak performance, **58**
 - SPEC benchmarks, **43**
- Intelligent devices, historical
 - background, L-80
- Intel MMX, multimedia instruction
 - compiler support, A-31 to A-32
- Intel Nehalem
 - characteristics, **411**
 - floorplan, **30**
 - WSC processor cost-performance, 473
- Intel Paragon, F-100, L-60
- Intel Pentium 4
 - hardware prefetching, **92**
 - Itanium 2 comparison, **H-43**
 - multithreading history, L-35
- Intel Pentium 4 Extreme, L-33 to L-34
- Intel Pentium II, L-33
- Intel Pentium III
 - pipelined cache access, 82
 - power consumption, F-85
- Intel Pentium M, power consumption, F-85
- Intel Pentium MMX, multimedia
 - support, **E-11**
- Intel Pentium Pro, 82, L-33
- Intel Pentium processors
 - “big and dumb” processors, 245
 - clock rate, **244**
 - early computer arithmetic, J-64 to J-65
 - vs.* Opteron memory protection, B-57
 - pipelining performance, C-10
 - segmented virtual memory
 - example, B-51 to B-54
 - SMT, 230
- Intel processors
 - early RISC designs, 2
 - power consumption, F-85
- Intel Single-Chip Cloud Computing (SCCC)
 - as interconnection example, F-70 to F-72
- OCNs, F-3

- Intel Streaming SIMD Extension (SSE)
 - basic function, 283
 - Multimedia SIMD Extensions, A-31
 - vs. vector architectures, 282
- Intel Teraflops processors, OCNs, F-3
- Intel Thunder Tiger 4 QsNet^{II}, F-63, **F-76**
- Intel VT-x, 129
- Intel x86
 - Amazon Web Services, 456
 - AVX instructions, **284**
 - clock rates, 244
 - computer architecture, 15
 - conditional instructions, H-27
 - GPUs as coprocessors, 330–331
 - Intel Core i7, 237–238
 - Multimedia SIMD Extensions, 282–283
 - NVIDIA GPU ISA, 298
 - parallelism, 262–263
 - performance and energy efficiency, 241
 - vs. PTX, 298
 - RISC, 2
 - speedup via parallelism, **263**
- Intel Xeon
 - Amazon Web Services, 457
 - cache coherence, 361
 - file system benchmarking, **D-20**
 - InfiniBand, F-76
 - multicore processor performance, 400–401
 - performance, 400
 - performance measurement, 405–406
 - SMP limitations, 363
 - SPECPower benchmarks, **463**
 - WSC processor cost-performance, 473
- Interactive workloads, WSC goals/requirements, 433
- Interarrival times, queuing model, D-30
- Interconnection networks
 - adaptive routing, F-93 to F-94
 - adaptive routing and fault tolerance, F-94
 - arbitration, **F-49**, F-49 to F-50
 - basic characteristics, F-2, **F-20**
 - bisection bandwidth, F-89
 - commercial
 - congestion management, F-64 to F-66
 - connectivity, F-62 to F-63
 - cross-company interoperability, F-63 to F-64
 - DECstation 5000 reboots, **F-69**
 - fault tolerance, F-66 to F-69
 - commercial routing/arbitration/switching, **F-56**
 - communication bandwidth, I-3
 - compute-optimized processors vs. receiver overhead, F-88
 - density- vs. SPEC-optimized processors, F-85
 - device example, **F-3**
 - direct vs. high-dimensional, F-92
 - domains, F-3 to F-5, **F-4**
 - Ethernet, F-77 to F-79, **F-78**
 - Ethernet/ATM total time statistics, **F-90**
 - examples, F-70
 - HOL blocking, **F-59**
 - IBM Blue Gene/L, I-43
 - InfiniBand, **F-75**
 - LAN history, F-99 to F-100
 - link bandwidth, F-89
 - memory hierarchy interface, F-87 to F-88
 - mesh network routing, **F-46**
 - MIN vs. direct network costs, F-92
 - multicore single-chip multiprocessor, **364**
- multi-device connections
 - basic considerations, F-20 to F-21
 - effective bandwidth vs. nodes, **F-28**
 - latency vs. nodes, **F-27**
 - performance characterization, F-25 to F-29
 - shared-media networks, F-22 to F-24
 - shared- vs. switched-media networks, **F-22**
 - switched-media networks, F-24
 - topology, routing, arbitration, switching, F-21 to F-22
- multi-device interconnections, shared- vs. switched-media networks, F-24 to F-25
- network-only features, F-94 to F-95
- NIC vs. I/O subsystems, F-90 to F-91
- OCN characteristics, **F-73**
- OCN example, F-70 to F-72
- OCN history, F-103 to F-104
- protection, F-86 to F-87
- routing, F-44 to F-48, **F-54**
- routing/arbitration/switching
 - impact, F-52 to F-55
- SAN characteristics, **F-76**
- software overhead, F-91 to F-92
- speed considerations, F-88
- storage area networks, F-102 to F-103
- switching, F-50 to F-52
- switch microarchitecture, **F-57**
 - basic microarchitecture, F-55 to F-58
 - buffer organizations, F-58 to F-60
 - pipelining, F-60 to F-61, **F-61**
- switch vs. NIC, F-85 to F-86, **F-86**
- system area networks, F-72 to F-74, F-100 to F-102
- system/storage area network, F-74 to F-77
- TCP/IP reliance, F-95
- top-level architecture, **F-71**
- topology, **F-44**
 - basic considerations, F-29 to F-30
- Beneš networks, **F-33**
- centralized switched networks, F-30 to F-34, **F-31**
- direct networks, **F-37**
- distributed switched networks, F-34 to F-40
- performance and costs, **F-40**
- performance effects, F-40 to F-44
- ring network, **F-36**
- two-device interconnections
 - basic considerations, F-5 to F-6
 - effective bandwidth vs. packet size, **F-19**
 - example, **F-6**
 - interface functions, F-6 to F-9
 - performance, F-12 to F-20
 - structure and functions, F-9 to F-12

- virtual channels and throughput, F-93
- WAN example, F-79
- WANs, F-97 to F-99
- wormhole switching performance, F-92 to F-93
- zero-copy protocols, F-91
- Intermittent faults, storage systems, D-11
- Internal fragmentation, virtual memory page size selection, B-47
- Internal Mask Registers, definition, **309**
- International Computer Architecture Symposium (ISCA), L-11 to L-12
- International Mobile Telephony 2000 (IMT-2000), cell phone standards, E-25
- Internet
 - Amazon Web Services, 457
 - array switch, 443
 - cloud computing, 455–456, 461
 - data-intensive applications, 344
 - dependability, 33
 - Google WSC, 464
 - Layer 3 network linkage, **445**
 - Netflix traffic, 460
 - SaaS, 4
 - WSC efficiency, 452
 - WSC memory hierarchy, 445
 - WSCs, 432–433, 435, 437, 439, 446, 453–455
- Internet Archive Cluster
 - container history, L-74 to L-75
 - overview, D-37
 - performance, dependability, cost, D-38 to D-40
 - TB-80 cluster MTTF, D-40 to D-41
 - TB-80 VME rack, **D-38**
- Internet Protocol (IP)
 - internetworking, F-83
 - storage area network history, F-102
 - WAN history, F-98
- Internet Protocol (IP) cores, OCNs, F-3
- Internet Protocol (IP) routers, VOQs, F-60
- Internetworking
 - connection example, **F-80**
 - cost, F-80
 - definition, F-2
 - enabling technologies, F-80 to F-81
 - OSI model layers, F-81, **F-82**
 - protocol-level communication, F-81 to F-82
 - protocol stack, F-83, **F-83**
 - role, **F-81**
 - TCP/IP, F-81, F-83 to F-84
 - TCP/IP headers, **F-84**
- Interprocedural analysis, basic approach, H-10
- Interprocessor communication, large-scale multiprocessors, I-3 to I-6
- Interrupt, *see* Exceptions
- Invalidate protocol
 - directory-based cache coherence protocol example, 382–383
 - example, 359, **360**
 - implementation, 356–357
 - snooping coherence, **355**, 355–356
- Invalid exception, floating-point arithmetic, J-35
- Inverted page table, virtual memory block identification, B-44 to B-45
- I/O bandwidth, definition, D-15
- I/O benchmarks, response time restrictions, **D-18**
- I/O bound workload, Virtual Machines protection, 108
- I/O bus
 - historical background, L-80 to L-81
 - interconnection networks, F-88
 - point-to-point replacement, **D-34**
 - Sony PlayStation 2 Emotion Engine case study, E-15
- I/O cache coherency, basic considerations, 113
- I/O devices
 - address translation, B-38
 - average memory access time, B-17
 - cache coherence enforcement, 354
 - centralized shared-memory multiprocessors, 351
 - future GPU features, 332
 - historical background, L-80 to L-81
 - inclusion, B-34
 - Multimedia SIMD vs. GPUs, 312
 - multiprocessor cost effectiveness, 407
 - performance, D-15 to D-16
 - SANs, F-3 to F-4
 - shared-media networks, F-23
 - switched networks, F-2
 - switch vs. NIC, F-86
 - Virtual Machines impact, 110–111
 - write strategy, B-11
 - Xen VM, 111
- I/O interfaces
 - disk storage, D-4
 - storage area network history, F-102
- I/O latency, shared-memory workloads, 368–369, 371
- I/O network, commercial interconnection network connectivity, F-63
- IOP, *see* I/O processor (IOP)
- I/O processor (IOP)
 - first dynamic scheduling, L-27
 - Sony PlayStation 2 Emotion Engine case study, E-15
- I/O registers, write buffer merging, 87
- I/O subsystems
 - design, D-59 to D-61
 - interconnection network speed, F-88
 - vs. NIC, F-90 to F-91
 - zero-copy protocols, F-91
- I/O systems
 - asynchronous, D-35
 - as black box, **D-23**
 - dirty bits, D-61 to D-64
 - Internet Archive Cluster, *see* Internet Archive Cluster
 - multithreading history, L-34
 - queuing theory, D-23
 - queue calculations, D-29
 - random variable distribution, D-26
 - utilization calculations, D-26
- IP, *see* Intellectual Property (IP) cores; Internet Protocol (IP)
- IPC, *see* Instructions per clock (IPC)
- IPoIB, F-77
- IR, *see* Instruction register (IR)
- ISA, *see* Instruction set architecture (ISA)

- ISCA, *see* International Computer Architecture Symposium (ISCA)
- iSCSI
 NetApp FAS6000 filer, D-42
 storage area network history, F-102
- Issue logic
 ARM Cortex-A8, 233
 ILP, 197
 longer latency pipelines, C-57
 multiple issue processor, **198**
 register renaming *vs.* ROB, 210
 speculation support, 210
- Issue stage
 ID pipe stage, 170
 instruction steps, 174
 MIPS with scoreboard, C-73 to C-74
 out-of-order execution, C-71
 ROB instruction, 186
- Iterative division, floating point, J-27
 to J-31
- J**
- Java benchmarks
 Intel Core i7, 401–405
 SMT on superscalar processors, 230–232
 without SMT, **403–404**
- Java language
 dependence analysis, H-10
 hardware impact on software development, 4
 return address predictors, 206
 SMT, 230–232, 402–405
 SPECjbb, 40
 SPECpower, 52
 virtual functions/methods, A-18
- Java Virtual Machine (JVM)
 early stack architectures, L-17
 IBM, 463
 multicore processor performance, 400
 multithreading-based speedup, **232**
 SPECjbb, 53
- JBOD, *see* RAID 0
- Johnson, Reynold B., L-77
- Jump prediction
 hardware model, 214
 ideal processor, 214
- Jumps
 control flow instructions, 14, A-16, **A-17**, A-21
- GPU conditional branching, 301–302
- MIPS control flow instructions, A-37 to A-38
- MIPS operations, A-35
- return address predictors, 206
- RISC instruction set, C-5
- VAX, K-71 to K-72
- Just-in-time (JIT), L-17
- JVM, *see* Java Virtual Machine (JVM)
- K**
- Kahle, Brewster, L-74
- Kahn, Robert, F-97
- k*-ary *n*-cubes, definition, F-38
- Kendall Square Research KSR-1, L-61
- Kernels
 arithmetic intensity, **286**, 286–287, 327
 benchmarks, 56
 bytes per reference, *vs.* block size, **378**
 caches, 329
 commercial workload, 369–370
 compilers, A-24
 compute bandwidth, 328
 via computing, 327
 EEMBC benchmarks, 38, **E-12**
 FFT, I-7
 FORTRAN, compiler
 vectorization, **G-15**
 FP benchmarks, C-29
 Livermore Fortran kernels, 331
 LU, I-8
 multimedia instructions, A-31
 multiprocessor architecture, 408
 multiprogramming workload, 375–378, **377**
 performance benchmarks, 37, 331
 primitives, A-30
 protecting processes, B-50
 segmented virtual memory, B-51
 SIMD exploitation, 330
 vector, on vector processor and GPU, 334–336
 virtual memory protection, 106
 WSCs, 438
- L**
- L1 caches, *see also* First-level caches
 address translation, B-46
 Alpha 21164 hierarchy, **368**
- ARM Cortex-A8, 116, **116**, **235**
- ARM Cortex-A8 *vs.* A9, **236**
- ARM Cortex-A8 example, **117**
- cache optimization, B-31 to B-33
- case study examples, B-60, B-63 to B-64
- directory-based coherence, 418
- Fermi GPU, 306
- hardware prefetching, 91
- hit time/power reduction, 79–80
- inclusion, 397–398, B-34 to B-35
- Intel Core i7, **118–119**, 121–122, **123**, 124, **124**, 239, 241
- invalidate protocol, **355**, 356–357
- memory consistency, 392
- memory hierarchy, **B-39**
- miss rates, **376–377**
- multiprocessor cache coherence, 352
- multiprogramming workload, 374
- nonblocking cache, 85
- NVIDIA GPU Memory, 304
- Opteron memory, **B-57**
- processor comparison, **242**
- speculative execution, 223
- T1 multithreading unicore
 performance, 228
 virtual memory, B-48 to B-49
- L2 caches, *see also* Second-level caches
 ARM Cortex-A8, 114, **115–116**, **235–236**
- ARM Cortex-A8 example, **117**
- cache optimization, B-31 to B-33, **B-34**
- case study example, B-63 to B-64
- coherency, 352
- commercial workloads, 373
- directory-based coherence, 379, 418–420, 422, 424
- fault detection, 58
- Fermi GPU, 296, 306, 308
- hardware prefetching, 91
- IBM Blue Gene/L, I-42
- inclusion, 397–398, B-35
- Intel Core i7, **118**, 120–122, 124, **124–125**, 239, 241
- invalidation protocol, **355**, 356–357
- and ISA, 241
- memory consistency, 392
- memory hierarchy, **B-39**, **B-48**, **B-57**

- L2 caches (*continued*)
 - multithreading, 225, 228
 - nonblocking cache, 85
 - NVIDIA GPU Memory, 304
 - processor comparison, **242**
 - snooping coherence, 359–361
 - speculation, 223
- L3 caches, *see also* Third-level caches
 - Alpha 21164 hierarchy, **368**
 - coherence, 352
 - commercial workloads, 370, **371**, 374
 - directory-based coherence, 379, 384
 - IBM Blue Gene/L, I-42
 - IBM Power processors, **247**
 - inclusion, 398
 - Intel Core i7, **118**, 121, 124, **124–125**, 239, 241, 403–404
 - invalidation protocol, **355**, 356–357, **360**
 - memory access cycle shift, **372**
 - miss rates, **373**
 - multicore processors, 400–401
 - multithreading, 225
 - nonblocking cache, 83
 - performance/price/power considerations, 52
 - snooping coherence, 359, 361, 363
- LabVIEW, embedded benchmarks, E-13
- Lampson, Butler, F-99
- Lanes
 - GPUs vs. vector architectures, **310**
 - Sequence of SIMD Lane Operations, **292**, **313**
 - SIMD Lane Registers, **309**, **314**
 - SIMD Lanes, 296–297, **297**, 302–303, 308, **309**, 311–312, **314**
 - vector execution time, 269
 - vector instruction set, 271–273
 - Vector Lane Registers, **292**
 - Vector Lanes, **292**, 296–297, 309, 311
- LANs, *see* Local area networks (LANs)
- Large-scale multiprocessors
 - cache coherence implementation
 - deadlock and buffering, I-38 to I-40
 - directory controller, I-40 to I-41
 - DSM multiprocessor, I-36 to I-37
 - overview, I-34 to I-36
 - classification, **I-45**
 - cluster history, L-62 to L-63
 - historical background, L-60 to L-61
 - IBM Blue Gene/L, I-41 to I-44, **I-43 to I-44**
 - interprocessor communication, I-3 to I-6
 - for parallel programming, I-2
 - scientific application performance
 - distributed-memory multiprocessors, I-26 to I-32, **I-28 to I-32**
 - parallel processors, I-33 to I-34
 - symmetric shared-memory multiprocessor, I-21 to I-26, **I-23 to I-25**
 - scientific applications, I-6 to I-12
 - space and relation of classes, **I-46**
 - synchronization mechanisms, I-17 to I-21
 - synchronization performance, I-12 to I-16
- Latency, *see also* Response time
 - advanced directory protocol case study, 425
 - vs. bandwidth, 18–19, **19**
 - barrier synchronization, I-16
 - and cache miss, B-2 to B-3
 - cluster history, L-73
 - communication mechanism, I-3 to I-4
 - definition, D-15
 - deterministic vs. adaptive routing, F-52 to F-55
 - directory coherence, **425**
 - distributed-memory multiprocessors, I-30, **I-32**
 - dynamically scheduled pipelines, C-70 to C-71
 - Flash memory, D-3
 - FP operations, **157**
 - FP pipeline, **C-66**
 - functional units, **C-53**
 - GPU SIMD instructions, 296
 - GPUs vs. vector architectures, 311
 - hazards and forwarding, C-54 to C-58
 - hiding with speculation, 396–397
 - ILP exposure, 157
 - ILP without multithreading, 225
 - ILP for realizable processors, 216–218
 - Intel SCCC, F-70
 - interconnection networks, F-12 to F-20
 - multi-device networks, F-25 to F-29
 - Itanium 2 instructions, **H-41**
 - microarchitectural techniques case study, 247–254
 - MIPS pipeline FP operations, C-52 to C-53
 - misses, single vs. multiple thread executions, **228**
 - multimedia instruction compiler support, A-31
 - NVIDIA GPU Memory structures, 305
 - OCNs vs. SANs, **F-27**
 - out-of-order processors, B-20 to B-21
 - packets, **F-13**, F-14
 - parallel processing, 350
 - performance milestones, **20**
 - pipeline, **C-87**
 - ROB commit, 187
 - routing, F-50
 - routing/arbitration/switching impact, F-52
 - routing comparison, **F-54**
 - SAN example, F-73
 - shared-memory workloads, 368
 - snooping coherence, **414**
 - Sony PlayStation 2 Emotion Engine, E-17
 - Sun T1 multithreading, 226–229
 - switched network topology, F-40 to F-41
 - system area network history, F-101
 - vs. TCP/IP reliance, F-95
 - throughput vs. response time, **D-17**
 - utility computing, L-74
 - vector memory systems, G-9
 - vector start-up, **G-8**
 - WSC efficiency, 450–452
 - WSC memory hierarchy, **443**, 443–444, **444**, 445
 - WSC processor cost-performance, 472–473
 - WSCs vs. datacenters, 456

- Layer 3 network, array and Internet linkage, **445**
- Layer 3 network, WSC memory hierarchy, **445**
- LCA, *see* Least common ancestor (LCA)
- LCD, *see* Liquid crystal display (LCD)
- Learning curve, cost trends, **27**
- Least common ancestor (LCA), routing algorithms, **F-48**
- Least recently used (LRU)
 - AMD Opteron data cache, **B-12, B-14**
 - block replacement, **B-9**
 - memory hierarchy history, **L-11**
 - virtual memory block replacement, **B-45**
- Less than condition code, PowerPC, **K-10 to K-11**
- Level 3, as Content Delivery Network, **460**
- Limit field, IA-32 descriptor table, **B-52**
- Line, memory hierarchy basics, **74**
- Linear speedup
 - cost effectiveness, **407**
 - IBM eServer p5 multiprocessor, **408**
 - multicore processors, **400, 402**
 - performance, **405–406**
- Line locking, embedded systems, **E-4 to E-5**
- Link injection bandwidth
 - calculation, **F-17**
 - interconnection networks, **F-89**
- Link pipelining, definition, **F-16**
- Link reception bandwidth, calculation, **F-17**
- Link register
 - MIPS control flow instructions, **A-37 to A-38**
 - PowerPC instructions, **K-32 to K-33**
 - procedure invocation options, **A-19**
 - synchronization, **389**
- Linpack benchmark
 - cluster history, **L-63**
 - parallel processing debates, **L-58**
 - vector processor example, **267–268**
- VMIPS performance, **G-17 to G-19**
- Linux operating systems
 - Amazon Web Services, **456–457**
 - architecture costs, **2**
 - protection and ISA, **112**
 - RAID benchmarks, **D-22, D-22 to D-23**
 - WSC services, **441**
- Liquid crystal display (LCD), Sanyo VPC-SX500 digital camera, **E-19**
- LISP
 - RISC history, **L-20**
 - SPARC instructions, **K-30**
- Lisp
 - ILP, **215**
 - as MapReduce inspiration, **437**
- Literal addressing mode, basic considerations, **A-10 to A-11**
- Little Endian
 - Intel 80x86, **K-49**
 - interconnection networks, **F-12**
 - memory address interpretation, **A-7**
 - MIPS core extensions, **K-20 to K-21**
 - MIPS data transfers, **A-34**
- Little's law
 - definition, **D-24 to D-25**
 - server utilization calculation, **D-29**
- Livelock, network routing, **F-44**
- Liveness, control dependence, **156**
- Livermore Fortran kernels, performance, **331, L-6**
- LMD, *see* Load memory data (LMD)
- Load instructions
 - control dependences, **155**
 - data hazards requiring stalls, **C-20**
 - dynamic scheduling, **177**
 - ILP, **199, 201**
 - loop-level parallelism, **318**
 - memory port conflict, **C-14**
 - pipelined cache access, **82**
 - RISC instruction set, **C-4 to C-5**
 - Tomasulo's algorithm, **182**
 - VLIW sample code, **252**
- Load interlocks
 - definition, **C-37 to C-39**
 - detection logic, **C-39**
- Load linked
 - locks via coherence, **391**
 - synchronization, **388–389**
- Load locked, synchronization, **388–389**
- Load memory data (LMD), simple MIPS implementation, **C-32 to C-33**
- Load stalls, MIPS R4000 pipeline, **C-67**
- Load-store instruction set architecture
 - basic concept, **C-4 to C-5**
 - IBM 360, **K-87**
 - Intel Core i7, **124**
 - Intel 80x86 operations, **K-62**
 - as ISA, **11**
 - ISA classification, **A-5**
 - MIPS nonaligned data transfers, **K-24, K-26**
 - MIPS operations, **A-35 to A-36, A-36**
 - PowerPC, **K-33**
 - RISC history, **L-19**
 - simple MIPS implementation, **C-32**
 - VMIPS, **265**
- Load/store unit
 - Fermi GPU, **305**
 - ILP hardware model, **215**
 - multiple lanes, **273**
 - Tomasulo's algorithm, **171–173, 182, 197**
 - vector units, **265, 276–277**
- Load upper immediate (LUI), MIPS operations, **A-37**
- Local address space, segmented virtual memory, **B-52**
- Local area networks (LANs)
 - characteristics, **F-4**
 - cross-company interoperability, **F-64**
 - effective bandwidth, **F-18**
 - Ethernet as, **F-77 to F-79**
 - fault tolerance calculations, **F-68**
 - historical overview, **F-99 to F-100**
 - InfiniBand, **F-74**
 - interconnection network domain relationship, **F-4**
 - latency and effective bandwidth, **F-26 to F-28**
 - offload engines, **F-8**
 - packet latency, **F-13, F-14 to F-16**
 - routers/gateways, **F-79**
 - shared-media networks, **F-23**
 - storage area network history, **F-102 to F-103**

- switches, F-29
 - TCP/IP reliance, F-95
 - time of flight, F-13
 - topology, F-30
 - Locality, *see* Principle of locality
 - Local Memory
 - centralized shared-memory architectures, 351
 - definition, **292, 314**
 - distributed shared-memory, 379
 - Fermi GPU, 306
 - Grid mapping, **293**
 - multiprocessor architecture, 348
 - NVIDIA GPU Memory structures, **304, 304–305**
 - SIMD, 315
 - symmetric shared-memory multiprocessors, 363–364
 - Local miss rate, definition, B-31
 - Local node, directory-based cache
 - coherence protocol basics, 382
 - Local optimizations, compilers, A-26
 - Local predictors, tournament
 - predictors, 164–166
 - Local scheduling, ILP, VLIW
 - processor, 194–195
 - Locks
 - via coherence, 389–391
 - hardware primitives, 387
 - large-scale multiprocessor synchronization, I-18 to I-21
 - multiprocessor software development, 409
 - Lock-up free cache, 83
 - Logical units, D-34
 - storage systems, D-34 to D-35
 - Logical volumes, D-34
 - Long displacement addressing, VAX, K-67
 - Long-haul networks, *see* Wide area networks (WANs)
 - Long Instruction Word (LIW)
 - EPIC, L-32
 - multiple-issue processors, L-28, L-30
 - Long integer
 - operand sizes/types, 12
 - SPEC benchmarks, A-14
 - Loop-carried dependences
 - CUDA, 290
 - definition, 315–316
 - dependence distance, H-6
 - dependent computation
 - elimination, 321
 - example calculations, H-4 to H-5
 - GCD, 319
 - loop-level parallelism, H-3
 - as recurrence, 318
 - recurrence form, H-5
 - VMIPS, 268
 - Loop exit predictor, Intel Core i7, 166
 - Loop interchange, compiler
 - optimizations, 88–89
 - Loop-level parallelism
 - definition, 149–150
 - detection and enhancement
 - basic approach, 315–318
 - dependence analysis, H-6 to H-10
 - dependence computation
 - elimination, 321–322
 - dependences, locating, 318–321
 - dependent computation
 - elimination, H-10 to H-12
 - overview, H-2 to H-6
 - history, L-30 to L-31
 - ILP in perfect processor, **215**
 - ILP for realizable processors, 217–218
 - Loop stream detection, Intel Core i7
 - micro-op buffer, 238
 - Loop unrolling
 - basic considerations, 161–162
 - ILP exposure, 157–161
 - ILP limitation studies, 220
 - recurrences, H-12
 - software pipelining, H-12 to H-15, **H-13, H-15**
 - Tomasulo's algorithm, 179, 181–183
 - VLIW processors, **195**
 - Lossless networks
 - definition, F-11 to F-12
 - switch buffer organizations, F-59
 - Lossy networks, definition, F-11 to F-12
 - LRU, *see* Least recently used (LRU)
 - Lucas
 - compiler optimizations, **A-29**
 - data cache misses, B-10
 - LUI, *see* Load upper immediate (LUI)
 - LU kernel
 - characteristics, I-8
 - distributed-memory multiprocessor, **I-32**
 - symmetric shared-memory multiprocessors, I-22, **I-23, I-25**
- ## M
- MAC, *see* Multiply-accumulate (MAC)
 - Machine language programmer, L-17 to L-18
 - Machine memory, Virtual Machines, 110
 - Macro-op fusion, Intel Core i7, 237–238
 - Magnetic storage
 - access time, D-3
 - cost vs. access time, **D-3**
 - historical background, L-77 to L-79
 - Mail servers, benchmarking, D-20
 - Main Memory
 - addressing modes, A-10
 - address translation, B-46
 - arithmetic intensity example, **286, 286–288**
 - block placement, B-44
 - cache function, B-2
 - cache optimization, B-30, B-36
 - coherence protocol, 362
 - definition, **292, 309**
 - DRAM, 17
 - gather-scatter, 329
 - GPU vs. MIMD, 327
 - GPUs and coprocessors, 330
 - GPU threads, 332
 - ILP considerations, 245
 - interlane wiring, 273
 - linear speedups, 407
 - memory hierarchy basics, 76
 - memory hierarchy design, **72**
 - memory mapping, B-42
 - MIPS operations, A-36
 - Multimedia SIMD vs. GPUs, 312
 - multiprocessor cache coherence, 352
 - paging vs. segmentation, **B-43**
 - partitioning, B-50

- Main Memory (*continued*)
 - processor performance
 - calculations, 218–219
 - RISC code size, A-23
 - server energy efficiency, 462
 - symmetric shared-memory
 - multiprocessors, 363
 - vector processor, G-25
 - vs. virtual memory, B-3, B-41
 - virtual memory block
 - identification, B-44 to B-45
 - virtual memory writes, B-45 to B-46
 - VLIW, 196
 - write-back, B-11
 - write process, B-45
- Manufacturing cost
 - chip fabrication case study, 61–62
 - cost trends, 27
 - modern processors, **62**
 - vs. operation cost, 33
- MapReduce
 - cloud computing, 455
 - cost calculations, 458–460, **459**
 - Google usage, **437**
 - reductions, 321
 - WSC batch processing, 437–438
 - WSC cost-performance, 474
- Mark-I, L-3 to L-4, L-6
- Mark-II, L-4
- Mark-III, L-4
- Mark-IV, L-4
- Mask Registers
 - basic operation, 275–276
 - definition, **309**
 - Multimedia SIMD, 283
 - NVIDIA GPU computational
 - structures, 291
 - vector compilers, 303
 - vector vs. GPU, 311
 - VMIPS, 267
- MasPar, L-44
- Massively parallel processors (MPPs)
 - characteristics, **I-45**
 - cluster history, L-62, L-72 to L-73
 - system area network history, F-100 to F-101
- Matrix300 kernel
 - definition, 56
 - prediction buffer, **C-29**
- Matrix multiplication
 - benchmarks, 56
 - LU kernel, I-8
 - multidimensional arrays in vector
 - architectures, 278
- Mauchly, John, L-2 to L-3, L-5, L-19
- Maximum transfer unit, network
 - interfaces, F-7 to F-8
- Maximum vector length (MVL)
 - Multimedia SIMD extensions, 282
 - vector vs. GPU, 311
 - VLRs, 274–275
- M-bus, *see* Memory bus (M-bus)
- McCreight, Ed, F-99
- MCF
 - compiler optimizations, **A-29**
 - data cache misses, B-10
 - Intel Core i7, 240–241
- MCP operating system, L-16
- Mean time between failures (MTBF)
 - fallacies, 56–57
 - RAID, L-79
 - SLA states, 34
- Mean time to failure (MTTF)
 - computer system power
 - consumption case study, 63–64
 - dependability benchmarks, D-21
 - disk arrays, D-6
 - example calculations, 34–35
 - I/O subsystem design, D-59 to D-61
 - RAID reconstruction, D-55 to D-57
 - SLA states, 34
 - TB-80 cluster, D-40 to D-41
 - WSCs vs. servers, 434
- Mean time to repair (MTTR)
 - dependability benchmarks, D-21
 - disk arrays, D-6
 - RAID 6, D-8 to D-9
 - RAID reconstruction, D-56
- Mean time until data loss (MTDL),
 - RAID reconstruction, D-55 to D-57
- Media, interconnection networks, F-9 to F-12
- Media extensions, DSPs, E-10 to E-11
- Mellanox MHEA28-XT, F-76
- Memory access
 - ARM Cortex-A8 example, **117**
 - basic MIPS pipeline, C-36
 - vs. block size, **B-28**
 - cache hit calculation, B-5 to B-6
 - Cray Research T3D, **F-87**
 - data hazards requiring stalls, C-19 to C-21
 - data hazard stall minimization, C-17, C-19
 - distributed-memory
 - multiprocessor, **I-32**
 - exception stopping/restarting, C-46
 - hazards and forwarding, C-56 to C-57
 - instruction set complications, C-49
 - integrated instruction fetch units, 208
 - MIPS data transfers, A-34
 - MIPS exceptions, C-48 to C-49
 - MIPS pipeline control, C-37 to C-39
 - MIPS R4000, C-65
 - multimedia instruction compiler
 - support, A-31
 - pipeline branch issues, **C-40, C-42**
 - RISC classic pipeline, C-7, C-10
 - shared-memory workloads, **372**
 - simple MIPS implementation, C-32 to C-33
 - simple RISC implementation, C-6
 - structural hazards, C-13 to C-14
 - vector architectures, **G-10**
- Memory addressing
 - ALU immediate operands, **A-12**
 - basic considerations, A-11 to A-13
 - compiler-based speculation, H-32
 - displacement values, **A-12**
 - immediate value distribution, **A-13**
 - interpretation, A-7 to A-8
 - ISA, 11
 - vector architectures, **G-10**
- Memory banks, *see also* Banked memory
 - gather-scatter, 280
 - multiprocessor architecture, 347
 - parallelism, 45
 - shared-memory multiprocessors, 363
 - strides, 279
 - vector load/store unit bandwidth, 276–277
 - vector systems, G-9 to G-11
- Memory bus (M-bus)
 - definition, 351

- Google WSC servers, 469
- interconnection networks, F-88
- Memory consistency
 - basic considerations, 392–393
 - cache coherence, 352
 - compiler optimization, 396
 - development of models, L-64
 - directory-based cache coherence
 - protocol basics, 382
 - multiprocessor cache coherency, 353
 - relaxed consistency models, 394–395
 - single-chip multicore processor
 - case study, 412–418
 - speculation to hide latency, 396–397
- Memory-constrained scaling,
 - scientific applications
 - on parallel processors, I-33
- Memory hierarchy
 - address space, B-57 to B-58
 - basic questions, B-6 to B-12
 - block identification, B-7 to B-9
 - block placement issues, B-7
 - block replacement, B-9 to B-10
 - cache optimization
 - basic categories, B-22
 - basic optimizations, **B-40**
 - hit time reduction, B-36 to B-40
 - miss categories, B-23 to B-26
 - miss penalty reduction
 - via multilevel caches, B-30 to B-35
 - read misses vs. writes, B-35 to B-36
 - miss rate reduction
 - via associativity, B-28 to B-30
 - via block size, B-26 to B-28
 - via cache size, B-28
 - pipelined cache access, 82
 - cache performance, B-3 to B-6
 - average memory access time, B-17 to B-20
 - basic considerations, B-16
 - basic equations, **B-22**
 - example calculation, B-16
 - out-of-order processors, B-20 to B-22
 - case studies, B-60 to B-67
 - development, L-9 to L-12
 - inclusion, 397–398
 - interconnection network
 - protection, F-87 to F-88
 - levels in slow down, **B-3**
 - Opteron data cache example, B-12 to B-15, **B-13**
 - Opteron L1/L2, **B-57**
 - OS and page size, B-58
 - overview, **B-39**
 - Pentium vs. Opteron protection, B-57
 - processor examples, B-3
 - process protection, B-50
 - terminology, B-2 to B-3
 - virtual memory
 - basic considerations, B-40 to B-44, B-48 to B-49
 - basic questions, B-44 to B-46
 - fast address translation, B-46
 - overview, **B-48**
 - paged example, B-54 to B-57
 - page size selection, B-46 to B-47
 - segmented example, B-51 to B-54
 - write strategy, B-10 to B-12
 - WSCs, **443**, 443–446, **444**
- Memory hierarchy design
 - access times, **77**
 - Alpha 21264 floorplan, **143**
 - ARM Cortex-A8 example, 114–117, **115–117**
 - cache coherency, 112–113
 - cache optimization
 - case study, 131–133
 - compiler-controlled
 - prefetching, 92–95
 - compiler optimizations, 87–90
 - critical word first, 86–87
 - energy consumption, **81**
 - hardware instruction
 - prefetching, 91–92, **92**
 - multibanked caches, 85–86, **86**
 - nonblocking caches, 83–85, **84**
 - overview, 78–79
 - pipelined cache access, 82
 - techniques overview, **96**
 - way prediction, 81–82
 - write buffer merging, 87, **88**
 - cache performance prediction, 125–126
 - cache size and misses per instruction, **126**
 - DDR2 SDRAM timing diagram, **139**
 - highly parallel memory systems, 133–136
 - high memory bandwidth, 126
 - instruction miss benchmarks, **127**
 - instruction simulation, 126
 - Intel Core i7, 117–124, **119**, **123–125**
 - Intel Core i7 three-level cache hierarchy, **118**
 - Intel Core i7 TLB structure, **118**
 - Intel 80x86 virtualization issues, **128**
 - memory basics, 74–78
 - overview, 72–74
 - protection and ISA, 112
 - server vs. PMD, **72**
 - system call virtualization/paravirtualization
 - performance, **141**
 - virtual machine monitor, 108–109
 - Virtual Machines ISA support, 109–110
 - Virtual Machines protection, 107–108
 - Virtual Machines and virtual memory and I/O, 110–111
 - virtual memory protection, 105–107
 - VMM on nonvirtualizable ISA, 128–129
 - Xen VM example, 111
 - Memory Interface Unit
 - NVIDIA GPU ISA, 300
 - vector processor example, **310**
 - Memoryless, definition, D-28
 - Memory mapping
 - memory hierarchy, B-48 to B-49
 - segmented virtual memory, B-52
 - TLBs, 323
 - virtual memory definition, B-42
 - Memory-memory instruction set
 - architecture, ISA
 - classification, A-3, A-5
 - Memory protection
 - control dependence, 155
 - Pentium vs. Opteron, B-57
 - processes, B-50

- Memory protection (*continued*)
 - safe calls, B-54
 - segmented virtual memory
 - example, B-51 to B-54
 - virtual memory, B-41
- Memory stall cycles
 - average memory access time, B-17
 - definition, B-4 to B-5
 - miss rate calculation, B-6
 - out-of-order processors, B-20 to B-21
 - performance equations, **B-22**
- Memory system
 - cache optimization, B-36
 - coherency, 352–353
 - commercial workloads, 367, 369–371
 - computer architecture, 15
 - C program evaluation, **134–135**
 - dependability enhancement, 104–105
 - distributed shared-memory, 379, 418
 - gather-scatter, 280
 - GDRAMs, 323
 - GPUs, 332
 - ILP, 245
 - hardware vs. software speculation, 221–222
 - speculative execution, 222–223
 - Intel Core i7, 237, 242
 - latency, B-21
 - MIPS, C-33
 - multiprocessor architecture, 347
 - multiprocessor cache coherence, 352
 - multiprogramming workload, 377–378
 - page size changes, B-58
 - price/performance/power considerations, 53
 - RISC, C-7
 - Roofline model, 286
 - shared-memory multiprocessors, 363
 - SMT, 399–400
 - stride handling, 279
 - T1 multithreading uncore
 - performance, 227
 - vector architectures, G-9 to G-11
 - vector chaining, G-11
 - vector processors, 271, 277
 - virtual, B-43, B-46
- Memory technology basics
 - DRAM, **98**, 98–100, **99**
 - DRAM and DIMM characteristics, **101**
 - DRAM performance, 100–102
 - Flash memory, 102–104
 - overview, 96–97
 - performance trends, **20**
 - SDRAM power consumption, 102, **103**
 - SRAM, 97–98
- Mesh interface unit (MIU), Intel
 - SCCC, F-70
- Mesh network
 - characteristics, **F-73**
 - deadlock, F-47
 - dimension-order routing, F-47 to F-48
 - OCN history, F-104
 - routing example, **F-46**
- Mesh topology
 - characteristics, F-36
 - direct networks, **F-37**
 - NEWS communication, F-42 to F-43
- MESI, *see* Modified-Exclusive-Shared-Invalid (MESI) protocol
- Message ID, packet header, F-8, F-16
- Message-passing communication
 - historical background, L-60 to L-61
 - large-scale multiprocessors, I-5 to I-6
- Message Passing Interface (MPI)
 - function, F-8
 - InfiniBand, **F-77**
 - lack in shared-memory multiprocessors, I-5
- Messages
 - adaptive routing, F-93 to F-94
 - coherence maintenance, **381**
 - InfiniBand, F-76
 - interconnection networks, F-6 to F-9
 - zero-copy protocols, F-91
- MFLOPS, *see* Millions of floating-point operations per second (MFLOPS)
- Microarchitecture
 - as architecture component, 15–16
 - ARM Cortex-A8, 241
 - Cray X1, G-21 to G-22
 - data hazards, 168
 - ILP exploitation, 197
 - Intel Core i7, 236–237
 - Nehalem, **411**
 - OCNs, F-3
 - out-of-order example, **253**
 - PTX vs. x86, 298
 - switches, *see* Switch
 - microarchitecture techniques case study, 247–254
- Microbenchmarks
 - disk array deconstruction, D-51 to D-55
 - disk deconstruction, D-48 to D-51
- Microfusion, Intel Core i7 micro-op buffer, 238
- Microinstructions
 - complications, C-50 to C-51
 - x86, 298
- Micro-ops
 - Intel Core i7, **237**, 238–240, **239**
 - processor clock rates, 244
- Microprocessor overview
 - clock rate trends, **24**
 - cost trends, 27–28
 - desktop computers, 6
 - embedded computers, 8–9
 - energy and power, 23–26
 - inside disks, D-4
 - integrated circuit improvements, 2
 - and Moore's law, 3–4
 - performance trends, 19–20, **20**
 - power and energy system trends, 21–23
 - recent advances, L-33 to L-34
 - technology trends, 18
- Microprocessor without Interlocked Pipeline Stages, *see* MIPS (Microprocessor without Interlocked Pipeline Stages)
- Microsoft
 - cloud computing, 455
 - containers, L-74
 - Intel support, 245
 - WSCs, 464–465
- Microsoft Azure, 456, L-74
- Microsoft DirectX, L-51 to L-52
- Microsoft Windows
 - benchmarks, 38

- multithreading, 223
- RAID benchmarks, **D-22**, D-22 to D-23
- time/volume/commoditization
 - impact, 28
 - WSC workloads, 441
- Microsoft Windows 2008 Server
 - real-world considerations, 52–55
 - SPECpower benchmark, 463
- Microsoft Xbox, L-51
- Migration, cache coherent
 - multiprocessors, 354
- Millions of floating-point operations per second (MFLOPS)
 - early performance measures, L-7
 - parallel processing debates, L-57 to L-58
 - SIMD computer history, L-55
 - SIMD supercomputer
 - development, L-43
 - vector performance measures,
 - G-15 to G-16
- MIMD (Multiple Instruction Streams, Multiple Data Streams)
 - and Amdahl's law, 406–407
 - definition, 10
 - early computers, L-56
 - first vector computers, L-46, L-48
 - GPU programming, 289
 - GPUs vs. vector architectures, 310
 - with Multimedia SIMD, vs. GPU, 324–330
 - multiprocessor architecture, 346–348
 - speedup via parallelism, **263**
 - TLP, basic considerations, 344–345
- Minicomputers, replacement by
 - microprocessors, 3–4
- Minniespec benchmarks
 - ARM Cortex-A8, **116**, 235
 - ARM Cortex-A8 memory, 115–116
- MINs, *see* Multistage interconnection networks (MINs)
- MIPS (Microprocessor without Interlocked Pipeline Stages)
 - addressing modes, 11–12
 - basic pipeline, C-34 to C-36
 - branch predictor correlation, 163
 - cache performance, B-6
 - conditional branches, K-11
 - conditional instructions, H-27
 - control flow instructions, 14
 - data dependences, 151
 - data hazards, 169
 - dynamic scheduling with
 - Tomasulo's algorithm, 171, 173
 - early pipelined CPUs, L-26
 - embedded systems, E-15
 - encoding, 14
 - exceptions, **C-48**, C-48 to C-49
 - exception stopping/restarting, C-46 to C-47
 - features, **K-44**
 - FP pipeline performance, C-60 to C-61, **C-62**
 - FP unit with Tomasulo's algorithm, **173**
 - hazard checks, C-71
 - ILP, 149
 - ILP exposure, 157–158
 - ILP hardware model, 215
 - instruction execution issues, K-81
 - instruction formats, core
 - instructions, K-6
 - instruction set complications, C-49 to C-51
 - ISA class, 11
 - ISA example
 - addressing modes for data transfer, A-34
 - arithmetic/logical instructions, **A-37**
 - basic considerations, A-32 to A-33
 - control flow instructions, A-37 to A-38, **A-38**
 - data types, A-34
 - dynamic instruction mix, **A-41**, A-41 to A-42, **A-42**
 - FP operations, A-38 to A-39
 - instruction format, **A-35**
 - load-store instructions, **A-36**
 - MIPS operations, A-35 to A-37
 - registers, A-34
 - usage, A-39
 - Livermore Fortran kernel
 - performance, **331**
 - memory addressing, 11
 - multicycle operations
 - basic considerations, C-51 to C-54
 - hazards and forwarding, C-54 to C-58
 - precise exceptions, C-58 to C-60
 - multimedia support, K-19
 - multiple-issue processor history, L-29
 - operands, 12
 - performance measurement history,
 - L-6 to L-7
 - pipeline branch issues, C-39 to C-42
 - pipeline control, C-36 to C-39
 - pipe stage, **C-37**
 - processor performance
 - calculations, 218–219
 - registers and usage conventions, **12**
 - RISC code size, A-23
 - RISC history, L-19
 - RISC instruction set lineage, **K-43**
 - as RISC systems, **K-4**
 - scoreboard components, **C-76**
 - scoreboarding, C-72
 - scoreboarding steps, **C-73**, C-73 to C-74
 - simple implementation, C-31 to C-34, **C-34**
 - Sony PlayStation 2 Emotion Engine, E-17
 - unaligned word read instructions, **K-26**
 - unpipelined functional units, **C-52**
 - vs. VAX, K-65 to K-66, **K-75**, **K-82**
 - write strategy, B-10
- MIPS16
 - addressing modes, **K-6**
 - arithmetic/logical instructions, **K-24**
 - characteristics, **K-4**
 - constant extension, **K-9**
 - data transfer instructions, **K-23**
 - embedded instruction format, **K-8**
 - instructions, K-14 to K-16
 - multiply-accumulate, **K-20**
 - RISC code size, A-23
 - unique instructions, K-40 to K-42
- MIPS32, vs. VAX sort, **K-80**
- MIPS64
 - addressing modes, **K-5**
 - arithmetic/logical instructions, **K-11**

- MIPS64 (*continued*)
 - conditional branches, **K-17**
 - constant extension, **K-9**
 - conventions, **K-13**
 - data transfer instructions, **K-10**
 - FP instructions, **K-23**
 - instruction list, K-26 to K-27
 - instruction set architecture formats, **14**
 - instruction subset, **13, A-40**
 - in MIPS R4000, C-61
 - nonaligned data transfers, K-24 to K-26
 - RISC instruction set, C-4
- MIPS2000, instruction benchmarks, **K-82**
- MIPS 3010, chip layout, **J-59**
- MIPS core
 - compare and conditional branch, K-9 to K-16
 - equivalent RISC instructions
 - arithmetic/logical, **K-11**
 - arithmetic/logical instructions, **K-15**
 - common extensions, K-19 to K-24
 - control instructions, **K-12, K-16**
 - conventions, **K-16**
 - data transfers, **K-10**
 - embedded RISC data transfers, **K-14**
 - FP instructions, **K-13**
 - instruction formats, K-9
- MIPS M2000, L-21, **L-21**
- MIPS MDMX
 - characteristics, **K-18**
 - multimedia support, K-18
- MIPS R2000, L-20
- MIPS R3000
 - integer arithmetic, J-12
 - integer overflow, **J-11**
- MIPS R3010
 - arithmetic functions, J-58 to J-61
 - chip comparison, **J-58**
 - floating-point exceptions, J-35
- MIPS R4000
 - early pipelined CPUs, L-27
 - FP pipeline, C-65 to C-67, **C-66**
 - integer pipeline, **C-63**
 - pipeline overview, C-61 to C-65
 - pipeline performance, C-67 to C-70
 - pipeline structure, C-62 to C-63
- MIPS R8000, precise exceptions, C-59
- MIPS R10000, 81
 - latency hiding, 397
 - precise exceptions, C-59
- Misalignment, memory address
 - interpretation, A-7 to A-8, **A-8**
- MISD, *see* Multiple Instruction Streams, Single Data Stream
- Misprediction rate
 - branch-prediction buffers, **C-29**
 - predictors on SPEC89, **166**
 - profile-based predictor, **C-27**
 - SPECCPU2006 benchmarks, **167**
- Mispredictions
 - ARM Cortex-A8, **232, 235**
 - branch predictors, 164–167, 240, C-28
 - branch-target buffers, 205
 - hardware-based speculation, 190
 - hardware vs. software speculation, 221
 - integer vs. FP programs, **212**
 - Intel Core i7, **237**
 - prediction buffers, **C-29**
 - static branch prediction, C-26 to C-27
- Misses per instruction
 - application/OS statistics, **B-59**
 - cache performance, B-5 to B-6
 - cache protocols, 359
 - cache size effect, **126**
 - L3 cache block size, **371**
 - memory hierarchy basics, 75
 - performance impact calculations, B-18
 - shared-memory workloads, **372**
 - SPEC benchmarks, **127**
 - strided access-TLB interactions, 323
- Miss penalty
 - average memory access time, B-16 to B-17
 - cache optimization, 79, B-35 to B-36
 - cache performance, B-4, B-21
 - compiler-controlled prefetching, 92–95
 - critical word first, 86–87
 - hardware prefetching, 91–92
- ILP speculative execution, 223
- memory hierarchy basics, 75–76
- nonblocking cache, 83
- out-of-order processors, B-20 to B-22
- processor performance
 - calculations, 218–219
- reduction via multilevel caches, B-30 to B-35
- write buffer merging, 87
- Miss rate
 - AMD Opteron data cache, B-15
 - ARM Cortex-A8, **116**
 - average memory access time, B-16 to B-17, **B-29**
 - basic categories, B-23
 - vs. block size, **B-27**
 - cache optimization, 79
 - and associativity, B-28 to B-30
 - and block size, B-26 to B-28
 - and cache size, B-28
 - cache performance, B-4
 - and cache size, **B-24 to B-25**
 - compiler-controlled prefetching, 92–95
 - compiler optimizations, 87–90
 - early IBM computers, L-10 to L-11
 - example calculations, B-6, B-31 to B-32
 - hardware prefetching, 91–92
 - Intel Core i7, **123, 125, 241**
 - memory hierarchy basics, 75–76
 - multilevel caches, **B-33**
 - processor performance
 - calculations, 218–219
 - scientific workloads
 - distributed-memory multiprocessors, **I-28 to I-30**
 - symmetric shared-memory multiprocessors, I-22, **I-23 to I-25**
 - shared-memory multiprocessing workload, **376, 376–377**
 - shared-memory workload, 370–373
 - single vs. multiple thread executions, **228**
 - Sun T1 multithreading unicycle performance, 228
 - vs. virtual addressed cache size, **B-37**

- MIT Raw, characteristics, **F-73**
- Mitsubishi M32R
 - addressing modes, **K-6**
 - arithmetic/logical instructions, **K-24**
 - characteristics, **K-4**
 - condition codes, **K-14**
 - constant extension, **K-9**
 - data transfer instructions, **K-23**
 - embedded instruction format, **K-8**
 - multiply-accumulate, **K-20**
 - unique instructions, **K-39** to **K-40**
- MIU, *see* Mesh interface unit (MIU)
- Mixed cache
 - AMD Opteron example, **B-15**
 - commercial workload, **373**
- Mixer, radio receiver, **E-23**
- Miya, Eugene, **L-65**
- M/M/1 model
 - example, **D-32**, **D-32** to **D-33**
 - overview, **D-30**
 - RAID performance prediction, **D-57**
 - sample calculations, **D-33**
- M/M/2 model, RAID performance prediction, **D-57**
- MMX, *see* Multimedia Extensions (MMX)
- Mobile clients
 - data usage, **3**
 - GPU features, **324**
 - vs. server GPUs, **323–330**
- Modified-Exclusive-Shared-Invalid (MESI) protocol, characteristics, **362**
- Modified-Owned-Exclusive-Shared-Invalid (MOESI) protocol, characteristics, **362**
- Modified state
 - coherence protocol, **362**
 - directory-based cache coherence protocol basics, **380**
 - large-scale multiprocessor cache coherence, **I-35**
 - snooping coherence protocol, **358–359**
- Modula-3, integer division/remainder, **J-12**
- Module availability, definition, **34**
- Module reliability, definition, **34**
- MOESI, *see* Modified-Owned-Exclusive-Shared-Invalid (MOESI) protocol
- Moore's law
 - DRAM, **100**
 - flawed architectures, **A-45**
 - interconnection networks, **F-70**
 - and microprocessor dominance, **3–4**
 - point-to-point links and switches, **D-34**
 - RISC, **A-3**
 - RISC history, **L-22**
 - software importance, **55**
 - switch size, **F-29**
 - technology trends, **17**
- Mortar shot graphs, multiprocessor performance measurement, **405–406**
- Motion JPEG encoder, Sanyo VPC-SX500 digital camera, **E-19**
- Motorola 68000
 - characteristics, **K-42**
 - memory protection, **L-10**
- Motorola 68882, floating-point precisions, **J-33**
- Move address, **VAX**, **K-70**
- MPEG
 - Multimedia SIMD Extensions history, **L-49**
 - multimedia support, **K-17**
 - Sanyo VPC-SX500 digital camera, **E-19**
 - Sony PlayStation 2 Emotion Engine, **E-17**
- MPI, *see* Message Passing Interface (MPI)
- MPPs, *see* Massively parallel processors (MPPs)
- MSP, *see* Multi-Streaming Processor (MSP)
- MTBF, *see* Mean time between failures (MTBF)
- MTDL, *see* Mean time until data loss (MTDL)
- MTTF, *see* Mean time to failure (MTTF)
- MTTR, *see* Mean time to repair (MTTR)
- Multibanked caches
 - cache optimization, **85–86**
 - example, **86**
- Multichip modules, OCNs, **F-3**
- Multicomputers
 - cluster history, **L-63**
 - definition, **345**, **L-59**
- historical background, **L-64** to **L-65**
- Multicore processors
 - architecture goals/requirements, **15**
 - cache coherence, **361–362**
 - centralized shared-memory multiprocessor structure, **347**
 - Cray X1E, **G-24**
 - directory-based cache coherence, **380**
 - directory-based coherence, **381**, **419**
 - DSM architecture, **348**, **379**
 - multichip
 - cache and memory states, **419**
 - with DSM, **419**
 - multiprocessors, **345**
 - OCN history, **F-104**
 - performance, **400–401**, **401**
 - performance gains, **398–400**
 - performance milestones, **20**
 - single-chip case study, **412–418**
 - and SMT, **404–405**
 - snooping cache coherence implementation, **365**
 - SPEC benchmarks, **402**
 - uniform memory access, **364**
 - write invalidate protocol implementation, **356–357**
- Multics protection software, **L-9**
- Multicycle operations, MIPS pipeline
 - basic considerations, **C-51** to **C-54**
 - hazards and forwarding, **C-54** to **C-58**
 - precise exceptions, **C-58** to **C-60**
- Multidimensional arrays
 - dependences, **318**
 - in vector architectures, **278–279**
- Multiflow processor, **L-30**, **L-32**
- Multigrid methods, Ocean application, **I-9** to **I-10**
- Multilevel caches
 - cache optimizations, **B-22**
 - centralized shared-memory architectures, **351**
 - memory hierarchy basics, **76**
 - memory hierarchy history, **L-11**
 - miss penalty reduction, **B-30** to **B-35**
 - miss rate vs. cache size, **B-33**

- Multilevel caches (*continued*)
 - Multimedia SIMD vs. GPU, 312
 - performance equations, **B-22**
 - purpose, 397
 - write process, B-11
- Multilevel exclusion, definition, B-35
- Multilevel inclusion
 - definition, 397, B-34
 - implementation, 397
 - memory hierarchy history, L-11
- Multimedia applications
 - desktop processor support, **E-11**
 - GPUs, 288
 - ISA support, A-46
 - MIPS FP operations, A-39
 - vector architectures, 267
- Multimedia Extensions (MMX)
 - compiler support, A-31
 - desktop RISCs, **K-18**
 - desktop/server RISCs, K-16 to K-19
 - SIMD history, 262, L-50
 - vs. vector architectures, 282–283
- Multimedia instructions
 - ARM Cortex-A8, 236
 - compiler support, A-31 to A-32
- Multimedia SIMD Extensions
 - basic considerations, 262, 282–284
 - compiler support, A-31
 - DLP, 322
 - DSPs, E-11
 - vs. GPUs, **312**
 - historical background, L-49 to L-50
 - MIMD, vs. GPU, 324–330
 - parallelism classes, 10
 - programming, 285
 - Roofline visual performance model, 285–288, **287**
 - 256-bit-wide operations, **282**
 - vs. vector, 263–264
- Multimedia user interfaces, PMDs, 6
- Multimode fiber, interconnection networks, F-9
- Multipass array multiplier, example, **J-51**
- Multiple Instruction Streams, Multiple Data Streams, *see* MIMD (Multiple Instruction Streams, Multiple Data Streams)
- Multiple Instruction Streams, Single Data Stream (MISD), definition, 10
- Multiple-issue processors
 - basic VLIW approach, 193–196
 - with dynamic scheduling and speculation, 197–202
 - early development, L-28 to L-30
 - instruction fetch bandwidth, 202–203
 - integrated instruction fetch units, 207
 - loop unrolling, 162
 - microarchitectural techniques case study, 247–254
 - primary approaches, **194**
 - SMT, 224, 226
 - with speculation, **198**
 - Tomasulo's algorithm, 183
- Multiple lanes technique
 - vector instruction set, 271–273
 - vector performance, G-7 to G-9
 - vector performance calculations, G-8
- Multiple paths, ILP limitation studies, 220
- Multiple-precision addition, J-13
- Multiply-accumulate (MAC)
 - DSP, E-5
 - embedded RISCs, **K-20**
 - TI TMS320C55 DSP, E-8
- Multiply operations
 - chip comparison, J-61
 - floating point
 - denormals, J-20 to J-21
 - examples, **J-19**
 - multiplication, J-17 to J-20
 - precision, J-21
 - rounding, **J-18**, J-19
 - integer arithmetic
 - array multiplier, **J-50**
 - Booth recoding, **J-49**
 - even/odd array, **J-52**
 - issues, J-11
 - with many adders, J-50 to J-54
 - multipass array multiplier, **J-51**
 - n*-bit unsigned integers, **J-4**
 - Radix-2, J-4 to J-7
 - signed-digit addition table, **J-54**
 - with single adder, J-47 to J-49, **J-48**
 - Wallace tree, **J-53**
 - integer shifting over zeros, J-45 to J-47
 - PA-RISC instructions, K-34 to K-35
 - unfinished instructions, **179**
- Multiprocessor basics
 - architectural issues and approaches, 346–348
 - architecture goals/requirements, 15
 - architecture and software development, 407–409
 - basic hardware primitives, 387–389
 - cache coherence, 352–353
 - coining of term, L-59
 - communication calculations, 350
 - computer categories, 10
 - consistency models, 395
 - definition, 345
 - early machines, L-56
 - embedded systems, E-14 to E-15
 - fallacies, 55
 - locks via coherence, 389–391
 - low-to-high-end roles, 344–345
 - parallel processing challenges, 349–351
 - for performance gains, 398–400
 - performance trends, 21
 - point-to-point example, **413**
 - shared-memory, *see* Shared-memory multiprocessors
 - SMP, 345, 350, 354–355, 363–364
 - streaming Multiprocessor, **292**, **307**, **313–314**
- Multiprocessor history
 - bus-based coherent multiprocessors, L-59 to L-60
 - clusters, L-62 to L-64
 - early computers, L-56
 - large-scale multiprocessors, L-60 to L-61
 - parallel processing debates, L-56 to L-58
 - recent advances and developments, L-58 to L-60
 - SIMD computers, L-55 to L-56
 - synchronization and consistency models, L-64
 - virtual memory, L-64

- Multiprogramming
 - definition, 345
 - multithreading, 224
 - performance, 36
 - shared-memory workload
 - performance, 375–378, **377**
 - shared-memory workloads, 374–375
 - software optimization, 408
 - virtual memory-based protection, 105–106, B-49
 - workload execution time, **375**
- Multistage interconnection networks (MINs)
 - bidirectional, F-33 to F-34
 - crossbar switch calculations, F-31 to F-32
 - vs. direct network costs, F-92
 - example, **F-31**
 - self-routing, F-48
 - system area network history, F-100 to F-101
 - topology, F-30 to F-31, F-38 to F-39
- Multistage switch fabrics, topology, F-30
- Multi-Streaming Processor (MSP)
 - Cray X1, G-21 to G-23, **G-22**, G-23 to G-24
 - Cray X1E, G-24
 - first vector computers, L-46
- Multithreaded SIMD Processor
 - block diagram, **294**
 - definition, **292, 309, 313–314**
 - Fermi GPU architectural innovations, 305–308
 - Fermi GPU block diagram, **307**
 - Fermi GTX 480 GPU floorplan, **295, 295–296**
 - GPU programming, 289–290
 - GPUs vs. vector architectures, **310, 310–311**
 - Grid mapping, **293**
 - NVIDIA GPU computational structures, 291
 - NVIDIA GPU Memory structures, **304, 304–305**
 - Roofline model, **326**
- Multithreaded vector processor
 - definition, **292**
 - Fermi GPU comparison, 305
- Multithreading
 - coarse-grained, 224–226
 - definition and types, 223–225
 - fine-grained, 224–226
 - GPU programming, 289
 - historical background, L-34 to L-35
 - ILP, 223–232
 - memory hierarchy basics, 75–76
 - parallel benchmarks, **231, 231–232**
 - for performance gains, 398–400
 - SMT, *see* Simultaneous multithreading (SMT)
 - Sun T1 effectiveness, 226–229
- MVAPICH, F-77
- MVL, *see* Maximum vector length (MVL)
- MXP processor, components, E-14
- Myrinet SAN, F-67
 - characteristics, **F-76**
 - cluster history, L-62 to L-63, L-73
 - routing algorithms, F-48
 - switch vs. NIC, F-86
 - system area network history, F-100
- N**
- NAK, *see* Negative acknowledge (NAK)
- Name dependences
 - ILP, 152–153
 - locating dependences, 318–319
 - loop-level parallelism, 315
 - scoreboarding, C-79
 - Tomasulo's algorithm, 171–172
- Nameplate power rating, WSCs, 449
- NaN (Not a Number), J-14, J-16, J-21, J-34
- NAND Flash, definition, 103
- NAS, *see* Network attached storage (NAS)
- NAS Parallel Benchmarks
 - InfiniBand, F-76
 - vector processor history, G-28
- National Science Foundation, WAN history, F-98
- Natural parallelism
 - embedded systems, E-15
 - multiprocessor importance, 344
 - multithreading, 223
- n*-bit adder, carry-lookahead, **J-38**
- n*-bit number representation, J-7 to J-10
- n*-bit unsigned integer division, **J-4**
- N-body algorithms, Barnes
 - application, I-8 to I-9
- NBS DYSEAC, L-81
- N-cube topology, characteristics, F-36
- NEC Earth Simulator, peak performance, **58**
- NEC SX/2, L-45, L-47
- NEC SX/5, L-46, L-48
- NEC SX/6, L-46, L-48
- NEC SX-8, L-46, L-48
- NEC SX-9
 - first vector computers, L-49
 - Roofline model, 286–288, **287**
- NEC VR 4122, embedded benchmarks, E-13
- Negative acknowledge (NAK)
 - cache coherence, I-38 to I-39
 - directory controller, I-40 to I-41
 - DSM multiprocessor cache coherence, I-37
- Negative condition code, MIPS core, K-9 to K-16
- Negative-first routing, F-48
- Nested page tables, 129
- NetApp, *see* Network Appliance (NetApp)
- Netflix, AWS, 460
- Netscape, F-98
- Network Appliance (NetApp)
 - FAS6000 filer, D-41 to D-42
 - NFS benchmarking, **D-20**
 - RAID, D-9
 - RAID row-diagonal parity, **D-9**
- Network attached storage (NAS)
 - block servers vs. filers, D-35
 - WSCs, 442
- Network bandwidth, interconnection network, F-18
- Network-Based Computer Laboratory (Ohio State), F-76, **F-77**
- Network buffers, network interfaces, F-7 to F-8
- Network fabric, switched-media networks, F-24
- Network File System (NFS)
 - benchmarking, D-20, **D-20**
 - block servers vs. filers, D-35
 - interconnection networks, F-89
 - server benchmarks, 40
 - TCP/IP, F-81

- Networking costs, WSC *vs.*
 - datacenters, 455
 - Network injection bandwidth
 - interconnection network, F-18
 - multi-device interconnection
 - networks, F-26
 - Network interface
 - fault tolerance, F-67
 - functions, F-6 to F-7
 - message composition/processing,
 - F-6 to F-9
 - Network interface card (NIC)
 - functions, F-8
 - Google WSC servers, 469
 - vs.* I/O subsystem, F-90 to F-91
 - storage area network history,
 - F-102
 - vs.* switches, F-85 to F-86, **F-86**
 - zero-copy protocols, F-91
 - Network layer, definition, **F-82**
 - Network nodes
 - direct network topology, **F-37**
 - distributed switched networks,
 - F-34 to F-36
 - Network on chip (NoC),
 - characteristics, F-3
 - Network ports, interconnection
 - network topology, F-29
 - Network protocol layer,
 - interconnection
 - networks, F-10
 - Network reception bandwidth,
 - interconnection
 - network, F-18
 - Network reconfiguration
 - commercial interconnection
 - networks, F-66
 - fault tolerance, F-67
 - switch *vs.* NIC, F-86
 - Network technology, *see also*
 - Interconnection
 - networks
 - Google WSC, 469
 - performance trends, 19–20
 - personal computers, F-2
 - trends, 18
 - WSC bottleneck, 461
 - WSC goals/requirements, 433
 - Network of Workstations, L-62, L-73
 - NEWS communication, *see*
 - North-East-West-South
 - communication
 - Newton's iteration, J-27 to J-30
 - NFS, *see* Network File System (NFS)
 - NIC, *see* Network interface card (NIC)
 - Nicely, Thomas, J-64
 - NMOS, DRAM, **99**
 - NoC, *see* Network on chip (NoC)
 - Nodes
 - coherence maintenance, **381**
 - communication bandwidth, I-3
 - direct network topology, **F-37**
 - directory-based cache coherence,
 - 380**
 - distributed switched networks,
 - F-34 to F-36
 - IBM Blue Gene/L, I-42 to I-44
 - IBM Blue Gene/L 3D torus
 - network, F-73
 - network topology performance and
 - costs, **F-40**
 - in parallel, **336**
 - points-to analysis, H-9
 - Nokia cell phone, circuit board, **E-24**
 - Nonaligned data transfers, MIPS64,
 - K-24 to K-26
 - Nonatomic operations
 - cache coherence, 361
 - directory protocol, 386
 - Nonbinding prefetch, cache
 - optimization, 93
 - Nonblocking caches
 - cache optimization, 83–85,
 - 131–133
 - effectiveness, **84**
 - ILP speculative execution,
 - 222–223
 - Intel Core i7, **118**
 - memory hierarchy history, L-11
 - Nonblocking crossbar, centralized
 - switched networks, F-32
 - to F-33
 - Nonfaulting prefetches, cache
 - optimization, 92
 - Nonrestoring division, J-5, **J-6**
 - Nonuniform memory access
 - (NUMA)
 - DSM as, 348
 - large-scale multiprocessor history,
 - L-61
 - snooping limitations, 363–364
 - Non-unit strides
 - multidimensional arrays in vector
 - architectures, 278–279
 - vector processor, **310**, 310–311,
 - G-25
 - North-East-West-South
 - communication,
 - network topology
 - calculations, F-41 to
 - F-43
 - North-last routing, F-48
 - Not a Number (NaN), J-14, J-16, J-21,
 - J-34
 - Notifications, interconnection
 - networks, F-10
 - NOW project, L-73
 - No-write allocate
 - definition, B-11
 - example calculation, B-12
 - NSFNET, F-98
 - NTSC/PAL encoder, Sanyo
 - VPC-SX500 digital
 - camera, E-19
 - Nullification, PA-RISC instructions,
 - K-33 to K-34
 - Nullifying branch, branch delay slots,
 - C-24 to C-25
 - NUMA, *see* Nonuniform memory
 - access (NUMA)
 - NVIDIA GeForce, L-51
 - NVIDIA systems
 - fine-grained multithreading, 224
 - GPU comparisons, 323–330,
 - 325**
 - GPU computational structures,
 - 291–297
 - GPU computing history, L-52
 - GPU ISA, 298–300
 - GPU Memory structures, **304**,
 - 304–305
 - GPU programming, 289
 - graphics pipeline history, L-51
 - scalable GPUs, L-51
 - terminology, 313–315
 - N*-way set associative
 - block placement, B-7
 - conflict misses, B-23
 - memory hierarchy basics, 74
 - TLBs, B-49
 - NYU Ultracomputer, L-60
- O**
- Observed performance, fallacies, 57
 - Occupancy, communication
 - bandwidth, I-3

- Ocean application
 - characteristics, I-9 to I-10
 - distributed-memory
 - multiprocessor, **I-32**
 - distributed-memory
 - multiprocessors, I-30
 - example calculations, I-11 to I-12
 - miss rates, **I-28**
 - symmetric shared-memory
 - multiprocessors, **I-23**
- OCNs, *see* On-chip networks (OCNs)
- Offline reconstruction, RAID, D-55
- Offload engines
 - network interfaces, F-8
 - TCP/IP reliance, F-95
- Offset
 - addressing modes, 12
 - AMD64 paged virtual memory, B-55
 - block identification, B-7 to B-8
 - cache optimization, B-38
 - call gates, B-54
 - control flow instructions, A-18
 - directory-based cache coherence
 - protocols, 381–382
 - example, B-9
 - gather-scatter, 280
 - IA-32 segment, B-53
 - instruction decode, C-5 to C-6
 - main memory, B-44
 - memory mapping, B-52
 - MIPS, C-32
 - MIPS control flow instructions, A-37 to A-38
 - misaligned addresses, A-8
 - Opteron data cache, B-13 to B-14
 - pipelining, **C-42**
 - PTX instructions, 300
 - RISC, C-4 to C-6
 - RISC instruction set, C-4
 - TLB, B-46
 - Tomasulo's approach, 176
 - virtual memory, B-43 to B-44, B-49, B-55 to B-56
- OLTP, *see* On-Line Transaction Processing (OLTP)
- Omega
 - example, **F-31**
 - packet blocking, F-32
 - topology, F-30
- OMNETPP, Intel Core i7, 240–241
- On-chip cache
 - optimization, 79
 - SRAM, 98–99
- On-chip memory, embedded systems, E-4 to E-5
- On-chip networks (OCNs)
 - basic considerations, F-3
 - commercial implementations, **F-73**
 - commercial interconnection
 - networks, F-63
 - cross-company interoperability, F-64
 - DOR, F-46
 - effective bandwidth, F-18, **F-28**
 - example system, F-70 to F-72
 - historical overview, F-103 to F-104
 - interconnection network domain
 - relationship, **F-4**
 - interconnection network speed, F-88
 - latency and effective bandwidth, F-26 to F-28
 - latency vs. nodes, **F-27**
 - link bandwidth, F-89
 - packet latency, **F-13**, F-14 to F-16
 - switch microarchitecture, F-57
 - time of flight, F-13
 - topology, F-30
 - wormhole switching, F-51
- One's complement, J-7
- One-way conflict misses, definition, B-23
- Online reconstruction, RAID, D-55
- On-Line Transaction Processing (OLTP)
 - commercial workload, **369, 371**
 - server benchmarks, 41
 - shared-memory workloads, 368–370, 373–374
 - storage system benchmarks, D-18
- OpenCL
 - GPU programming, 289
 - GPU terminology, **292**, 313–315
 - NVIDIA terminology, 291
 - processor comparisons, 323
- OpenGL, L-51
- Open source software
 - Amazon Web Services, 457
 - WSCs, 437
 - Xen VMM, *see* Xen virtual machine
- Open Systems Interconnect (OSI)
 - Ethernet, F-78 to F-79
 - layer definitions, **F-82**
- Operand addressing mode, Intel 80x86, **K-59**, K-59 to K-60
- Operand delivery stage, Itanium 2, H-42
- Operands
 - DSP, E-6
 - forwarding, **C-19**
 - instruction set encoding, A-21 to A-22
 - Intel 80x86, **K-59**
 - ISA, 12
 - ISA classification, A-3 to A-4
 - MIPS data types, A-34
 - MIPS pipeline, C-71
 - MIPS pipeline FP operations, C-52 to C-53
 - NVIDIA GPU ISA, 298
 - per ALU instruction example, **A-6**
 - TMS320C55 DSP, **E-6**
 - type and size, A-13 to A-14
 - VAX, K-66 to K-68, **K-68**
 - vector execution time, 268–269
- Operating systems (general)
 - address translation, B-38
 - and architecture development, 2
 - communication performance, F-8
 - disk access scheduling, D-44 to D-45, **D-45**
 - memory protection performance, B-58
 - miss statistics, **B-59**
 - multiprocessor software
 - development, 408
 - and page size, B-58
 - segmented virtual memory, B-54
 - server benchmarks, 40
 - shared-memory workloads, 374–378
 - storage systems, D-35
- Operational costs
 - basic considerations, 33
 - WSCs, 434, 438, 452, 456, 472
- Operational expenditures (OPEX)
 - WSC costs, 452–455, **454**
 - WSC TCO case study, 476–478
- Operation faults, storage systems, D-11
- Operator dependability, disks, D-13 to D-15

- OPEX, *see* Operational expenditures (OPEX)
- Optical media, interconnection networks, F-9
- Oracle database
 - commercial workload, 368
 - miss statistics, **B-59**
 - multithreading benchmarks, **232**
 - single-threaded benchmarks, **243**
 - WSC services, 441
- Ordering, and deadlock, F-47
- Organization
 - buffer, switch microarchitecture, F-58 to F-60
 - cache, performance impact, B-19
 - cache blocks, B-7 to B-8
 - cache optimization, B-19
 - coherence extensions, 362
 - computer architecture, 11, 15–16
 - DRAM, **98**
 - MIPS pipeline, **C-37**
 - multiple-issue processor, 197, **198**
 - Opteron data cache, B-12 to B-13, **B-13**
 - pipelines, 152
 - processor history, 2–3
 - processor performance equation, 49
 - shared-memory multiprocessors, 346
 - Sony PlayStation Emotion Engine, **E-18**
 - TLB, B-46
- Orthogonality, compiler
 - writing-architecture relationship, A-30
- OSI, *see* Open Systems Interconnect (OSI)
- Out-of-order completion
 - data hazards, 169
 - MIPS pipeline, C-71
 - MIPS R10000 sequential consistency, 397
 - precise exceptions, C-58
- Out-of-order execution
 - and cache miss, B-2 to B-3
 - cache performance, B-21
 - data hazards, 169–170
 - hardware-based execution, 184
 - ILP, 245
 - memory hierarchy, B-2 to B-3
 - microarchitectural techniques case study, 247–254
 - MIPS pipeline, C-71
 - miss penalty, B-20 to B-22
 - performance milestones, **20**
 - power/DLP issues, 322
 - processor comparisons, 323
 - R10000, 397
 - SMT, 246
 - Tomasulo's algorithm, 183
- Out-of-order processors
 - DLP, 322
 - Intel Core i7, 236
 - memory hierarchy history, L-11
 - multithreading, 226
 - vector architecture, 267
- Out-of-order write, dynamic scheduling, 171
- Output buffered switch
 - HOL blocking, F-60
 - microarchitecture, F-57, **F-57**
 - organizations, F-58 to F-59
 - pipelined version, **F-61**
- Output dependence
 - compiler history, L-30 to L-31
 - definition, 152–153
 - dynamic scheduling, 169–171, C-72
 - finding, H-7 to H-8
 - loop-level parallelism calculations, 320
 - MIPS scoreboarding, C-79
- Overclocking
 - microprocessors, 26
 - processor performance equation, 52
- Overflow, integer arithmetic, J-8, J-10 to J-11, **J-11**
- Overflow condition code, MIPS core, K-9 to K-16
- Overhead
 - adaptive routing, F-93 to F-94
 - Amdahl's law, F-91
 - communication latency, I-4
 - interconnection networks, F-88, F-91 to F-92
 - OCNs vs. SANs, **F-27**
 - vs. peak performance, 331
 - shared-memory communication, I-5
 - sorting case study, D-64 to D-67
 - time of flight, F-14
 - vector processor, **G-4**
- Overlapping triplets
 - historical background, J-63
 - integer multiplication, J-49
- Oversubscription
 - array switch, 443
 - Google WSC, 469
 - WSC architecture, 441, 461
- P**
- Packed decimal, definition, A-14
- Packet discarding, congestion management, F-65
- Packets
 - ATM, F-79
 - bidirectional rings, F-35 to F-36
 - centralized switched networks, F-32
 - effective bandwidth vs. packet size, **F-19**
 - format example, **F-7**
 - IBM Blue Gene/L 3D torus network, F-73
 - InfiniBand, **F-75**, F-76
 - interconnection networks, multi-device networks, F-25
 - latency issues, F-12, **F-13**
 - lossless vs. lossy networks, F-11 to F-12
 - network interfaces, F-8 to F-9
 - network routing, F-44
 - routing/arbitration/switching impact, F-52
 - switched network topology, F-40
 - switching, F-51
 - switch microarchitecture, F-57 to F-58
 - switch microarchitecture
 - pipelining, F-60 to F-61
 - TI TMS320C6x DSP, **E-10**
 - topology, F-21
 - virtual channels and throughput, F-93
- Packet transport, interconnection networks, F-9 to F-12
- Page coloring, definition, B-38
- Paged segments, characteristics, B-43 to B-44
- Paged virtual memory
 - Opteron example, B-54 to B-57
 - protection, 106
 - vs. segmented, **B-43**

- Page faults
 - cache optimization, A-46
 - exceptions, C-43 to C-44
 - hardware-based speculation, 188
 - and memory hierarchy, B-3
 - MIPS exceptions, C-48
 - Multimedia SIMD Extensions, 284
 - stopping/restarting execution, C-46
 - virtual memory definition, B-42
 - virtual memory miss, B-45
- Page offset
 - cache optimization, B-38
 - main memory, B-44
 - TLB, B-46
 - virtual memory, B-43, B-49, B-55 to B-56
- Pages
 - definition, B-3
 - vs. segments, **B-43**
 - size selection, B-46 to B-47
 - virtual memory definition, B-42 to B-43
 - virtual memory fast address translation, B-46
- Page size
 - cache optimization, B-38
 - definition, B-56
 - memory hierarchy example, B-39, B-48
 - and OS, B-58
 - OS determination, B-58
 - paged virtual memory, B-55
 - selection, B-46 to B-47
 - virtual memory, B-44
- Page Table Entry (PTE)
 - AMD64 paged virtual memory, B-56
 - IA-32 equivalent, B-52
 - Intel Core i7, 120
 - main memory block, B-44 to B-45
 - paged virtual memory, B-56
 - TLB, B-47
- Page tables
 - address translation, B-46 to B-47
 - AMD64 paged virtual memory, B-55 to B-56
 - descriptor tables as, B-52
 - IA-32 segment descriptors, B-53
 - main memory block, B-44 to B-45
 - multiprocessor software development, 407–409
 - multithreading, 224
 - protection process, B-50
 - segmented virtual memory, B-51
 - virtual memory block identification, B-44
 - virtual-to-physical address mapping, **B-45**
- Paired single operations, DSP media extensions, E-11
- Palt, definition, B-3
- Papadopolous, Greg, L-74
- Parallelism
 - cache optimization, 79
 - challenges, 349–351
 - classes, 9–10
 - computer design principles, 44–45
 - dependence analysis, H-8
 - DLP, *see* Data-level parallelism (DLP)
 - Ethernet, F-78
 - exploitation statically, H-2
 - exposing with hardware support, H-23 to H-27
 - global code scheduling, H-15 to H-23, **H-16**
 - IA-64 instruction format, H-34 to H-35
 - ILP, *see* Instruction-level parallelism (ILP)
 - loop-level, 149–150, 215, 217–218, 315–322
 - MIPS scoreboarding, C-77 to C-78
 - multiprocessors, 345
 - natural, 223, 344
 - request-level, 4–5, 9, 345, 434
 - RISC development, 2
 - software pipelining, H-12 to H-15
 - for speedup, **263**
 - superblock scheduling, H-21 to H-23, **H-22**
 - task-level, 9
 - TLP, *see* Thread-level parallelism (TLP)
 - trace scheduling, H-19 to H-21, **H-20**
 - vs. window size, **217**
 - WSCs vs. servers, 433–434
- Parallel processors
 - areas of debate, L-56 to L-58
 - bus-based coherent multiprocessor history, L-59 to L-60
 - cluster history, L-62 to L-64
 - early computers, L-56
 - large-scale multiprocessor history, L-60 to L-61
 - recent advances and developments, L-58 to L-60
 - scientific applications, I-33 to I-34
 - SIMD computer history, L-55 to L-56
 - synchronization and consistency models, L-64
 - virtual memory history, L-64
- Parallel programming
 - computation communication, I-10 to I-12
 - with large-scale multiprocessors, I-2
- Parallel Thread Execution (PTX)
 - basic GPU thread instructions, **299**
 - GPU conditional branching, 300–303
 - GPUs vs. vector architectures, 308
 - NVIDIA GPU ISA, 298–300
 - NVIDIA GPU Memory structures, 305
- Parallel Thread Execution (PTX) Instruction
 - CUDA Thread, 300
 - definition, **292, 309, 313**
 - GPU conditional branching, 302–303
 - GPU terms, 308
 - NVIDIA GPU ISA, 298, 300
- Paravirtualization
 - system call performance, **141**
 - Xen VM, 111
- Parity
 - dirty bits, D-61 to D-64
 - fault detection, 58
 - memory dependability, 104–105
 - WSC memory, 473–474
- PARSEC benchmarks
 - Intel Core i7, 401–405
 - SMT on superscalar processors, 230–232, **231**
 - speedup without SMT, **403–404**
- Partial disk failure, dirty bits, D-61 to D-64
- Partial store order, relaxed consistency models, 395
- Partitioned add operation, DSP media extensions, E-10
- Partitioning
 - Multimedia SIMD Extensions, 282
 - virtual memory protection, B-50
 - WSC memory hierarchy, 445

- Pascal programs
 - compiler types and classes, **A-28**
 - integer division/remainder, **J-12**
- Pattern, disk array deconstruction, D-51
- Payload
 - messages, F-6
 - packet format, **F-7**
- p* bits, J-21 to J-23, J-25, J-36 to J-37
- PC, *see* Program counter (PC)
- PCI bus, historical background, L-81
- PCIe, *see* PCI-Express (PCIe)
- PCI-Express (PCIe), F-29, F-63
 - storage area network history, F-102 to F-103
- PCI-X, F-29
 - storage area network history, F-102
- PCI-X 2.0, F-63
- PCMCIA slot, Sony PlayStation 2
 - Emotion Engine case study, E-15
- PC-relative addressing mode, VAX, K-67
- PDP-11, L-10, L-17 to L-19, L-56
- PDU, *see* Power distribution unit (PDU)
- Peak performance
 - Cray X1E, G-24
 - DAXPY on VMIPS, G-21
 - DLP, 322
 - fallacies, 57–58
 - multiple lanes, 273
 - multiprocessor scaled programs, **58**
 - Roofline model, **287**
 - vector architectures, 331
 - VMIPS on DAXPY, G-17
 - WSC operational costs, 434
- Peer-to-peer
 - internetworking, F-81 to F-82
 - wireless networks, E-22
- Pegasus, L-16
- PennySort competition, D-66
- Perfect Club benchmarks
 - vector architecture programming, **281**, 281–282
 - vector processor history, G-28
- Perfect processor, ILP hardware
 - model, 214–215, **215**
- Perfect-shuffle exchange,
 - interconnection network topology, F-30 to F-31
- Performability, RAID reconstruction, D-55 to D-57
- Performance, *see also* Peak performance
 - advanced directory protocol case study, 420–426
 - ARM Cortex-A8, 233–236, **234**
 - ARM Cortex-A8 memory, 115–117
 - bandwidth vs. latency, 18–19
 - benchmarks, 37–41
 - branch penalty reduction, C-22
 - branch schemes, C-25 to C-26
 - cache basics, B-3 to B-6
 - cache performance
 - average memory access time, B-16 to B-20
 - basic considerations, B-3 to B-6, B-16
 - basic equations, **B-22**
 - basic optimizations, **B-40**
 - example calculation, B-16 to B-17
 - out-of-order processors, B-20 to B-22
- compiler optimization impact, A-27
- cost-performance
 - extensive pipelining, C-80 to C-81
 - WSC Flash memory, 474–475
 - WSC goals/requirements, 433
 - WSC hardware inactivity, 474
 - WSC processors, 472–473
- CUDA, 290–291
- desktop benchmarks, 38–40
- directory-based coherence case study, 418–420
- dirty bits, D-61 to D-64
- disk array deconstruction, D-51 to D-55
- disk deconstruction, D-48 to D-51
- DRAM, 100–102
- embedded computers, 9, **E-13 to E-14**
- Google server benchmarks, 439–441
- hardware fallacies, 56
- high-performance computing, 432, 435–436, B-10
- historical milestones, **20**
- ILP exploitation, 201
- ILP for realizable processors, 216–218
- Intel Core i7, 239–241, **240**, 401–405
- Intel Core i7 memory, 122–124
- interconnection networks
 - bandwidth considerations, F-89
 - multi-device networks, F-25 to F-29
 - routing/arbitration/switching impact, F-52 to F-55
 - two-device networks, F-12 to F-20
- Internet Archive Cluster, D-38 to D-40
- interprocessor communication, I-3 to I-6
- I/O devices, D-15 to D-16
- I/O subsystem design, D-59 to D-61
- I/O system design/evaluation, D-36
- ISA, 241–243
- Itanium 2, H-43
- large-scale multiprocessors
 - scientific applications
 - distributed-memory multiprocessors, I-26 to I-32, **I-28 to I-30, I-32**
 - parallel processors, I-33 to I-34
 - symmetric shared-memory multiprocessor, I-21 to I-26, **I-23 to I-25**
 - synchronization, I-12 to I-16
- MapReduce, 438
- measurement, reporting,
 - summarization, 36–37
- memory consistency models, 393
- memory hierarchy design, **73**
- memory hierarchy and OS, B-58
- memory threads, GPUs, 332
- MIPS FP pipeline, C-60 to C-61
- MIPS M2000 vs. VAX 8700, **K-82**
- MIPS R4000 pipeline, C-67 to C-70, **C-68**
- multicore processors, 400–401, **401**
- multiprocessing/multithreading, 398–400
- multiprocessors, measurement
 - issues, 405–406

- multiprocessor software
 - development, 408–409
- network topologies, **F-40**, F-40 to F-44
- observed, 57
- peak
 - DLP, 322
 - fallacies, 57–58
 - multiple lanes, 273
 - Roofline model, **287**
 - vector architectures, 331
 - WSC operational costs, 434
- pipelines with stalls, C-12 to C-13
- pipelining basics, C-10 to C-11
- processors, historical growth, 2–3, **3**
- quantitative measures, L-6 to L-7
- real-time, PMDs, 6
- real-world server considerations, 52–55
- results reporting, 41
- results summarization, 41–43, **43**
- RISC classic pipeline, C-7
- server benchmarks, 40–41
- as server characteristic, 7
- single-chip multicore processor
 - case study, 412–418
- single-thread, 399
 - processor benchmarks, **243**
- software development, 4
- software overhead issues, F-91 to F-92
- sorting case study, D-64 to D-67
- speculation cost, 211
- Sun T1 multithreading unicore, 227–229
- superlinear, 406
- switch microarchitecture
 - pipelining, F-60 to F-61
- symmetric shared-memory
 - multiprocessors, 366–378
 - scientific workloads, I-21 to I-26, **I-23**
- system call virtualization/paravirtualization, **141**
- transistors, scaling, 19–21
- vector, and memory bandwidth, 332
- vector add instruction, **272**
- vector kernel implementation, 334–336
- vector processor, G-2 to G-7
 - DAXPY on VMIPS, G-19 to G-21
 - sparse matrices, G-12 to G-14
 - start-up and multiple lanes, G-7 to G-9
- vector processors
 - chaining, G-11 to G-12
 - chaining/unchaining, **G-12**
- vector vs. scalar, 331–332
- VMIPS on Linpack, G-17 to G-19
- wormhole switching, F-92 to F-93
- Permanent failure, commercial
 - interconnection networks, F-66
- Permanent faults, storage systems, D-11
- Personal computers
 - LANs, F-4
 - networks, F-2
 - PCIe, F-29
- Personal mobile device (PMD)
 - characteristics, 6
 - as computer class, **5**
 - embedded computers, 8–9
 - Flash memory, 18
 - integrated circuit cost trends, 28
 - ISA performance and efficiency prediction, 241–243
 - memory hierarchy basics, 78
 - memory hierarchy design, **72**
 - power and energy, 25
 - processor comparison, **242**
- PetaBox GB2000, Internet Archive Cluster, D-37
- Phase-ordering problem, compiler structure, A-26
- Phits, *see* Physical transfer units (phits)
- Physical addresses
 - address translation, B-46
 - AMD Opteron data cache, B-12 to B-13
 - ARM Cortex-A8, **115**
 - directory-based cache coherence protocol basics, 382
 - main memory block, B-44
 - memory hierarchy, B-48 to B-49
 - memory hierarchy basics, 77–78
 - memory mapping, B-52
 - paged virtual memory, B-55 to B-56
- page table-based mapping, **B-45**
- safe calls, B-54
- segmented virtual memory, B-51
- sharing/protection, B-52
- translation, B-36 to B-39
- virtual memory definition, B-42
- Physical cache, definition, B-36 to B-37
- Physical channels, F-47
- Physical layer, definition, **F-82**
- Physical memory
 - centralized shared-memory multiprocessors, 347
 - directory-based cache coherence, 354
 - future GPU features, 332
 - GPU conditional branching, 303
 - main memory block, B-44
 - memory hierarchy basics, B-41 to B-42
 - multiprocessors, 345
 - paged virtual memory, B-56
 - processor comparison, 323
 - segmented virtual memory, B-51
 - unified, 333
 - Virtual Machines, 110
- Physical transfer units (phits), F-60
- Physical volumes, D-34
- PID, *see* Process-identifier (PID) tags
- Pin-out bandwidth, topology, F-39
- Pipeline bubble, stall as, C-13
- Pipeline cycles per instruction
 - basic equation, 148
 - ILP, 149
 - processor performance calculations, 218–219
 - R4000 performance, C-68 to C-69
- Pipelined circuit switching, F-50
- Pipelined CPUs, early versions, L-26 to L-27
- Pipeline delays
 - ARM Cortex-A8, **235**
 - definition, 228
 - fine-grained multithreading, 227
 - instruction set complications, C-50
 - multiple branch speculation, 211
 - Sun T1 multithreading unicore performance, 227–228
- Pipeline interlock
 - data dependences, 151
 - data hazards requiring stalls, C-20
 - MIPS R4000, C-65
 - MIPS vs. VMIPS, 268

- Pipeline latches
 - ALU, **C-40**
 - definition, C-35
 - R4000, C-60
 - stopping/restarting execution, C-47
- Pipeline organization
 - dependences, 152
 - MIPS, **C-37**
- Pipeline registers
 - branch hazard stall, **C-42**
 - data hazards, C-57
 - data hazard stalls, C-17 to C-20
 - definition, C-35
 - example, **C-9**
 - MIPS, C-36 to C-39
 - MIPS extension, C-53
 - PC as, C-35
 - pipelining performance issues, C-10
 - RISC processor, C-8, C-10
- Pipeline scheduling
 - basic considerations, 161–162
 - vs. dynamic scheduling, 168–169
 - ILP exploitation, 197
 - ILP exposure, 157–161
 - microarchitectural techniques case study, 247–254
 - MIPS R4000, C-64
- Pipeline stall cycles
 - branch scheme performance, C-25
 - pipeline performance, C-12 to C-13
- Pipelining
 - branch cost reduction, C-26
 - branch hazards, C-21 to C-26
 - branch issues, C-39 to C-42
 - branch penalty reduction, C-22 to C-25
 - branch-prediction buffers, C-27 to C-30, **C-29**
 - branch scheme performance, C-25 to C-26
 - cache access, 82
 - case studies, C-82 to C-88
 - classic stages for RISC, C-6 to C-10
 - compiler scheduling, L-31
 - concept, C-2 to C-3
 - cost-performance, C-80 to C-81
 - data hazards, C-16 to C-21
 - definition, C-2
 - dynamically scheduled pipelines, C-70 to C-80
 - example, **C-8**
 - exception stopping/restarting, C-46 to C-47
 - exception types and requirements, C-43 to C-46
 - execution sequences, C-80
 - floating-point addition speedup, J-25
 - graphics pipeline history, L-51
 - hazard classes, C-11
 - hazard detection, **C-38**
 - implementation difficulties, C-43 to C-49
 - independent FP operations, **C-54**
 - instruction set complications, C-49 to C-51
 - interconnection networks, F-12
 - latencies, **C-87**
 - MIPS, C-34 to C-36
 - MIPS control, C-36 to C-39
 - MIPS exceptions, **C-48**, C-48 to C-49
 - MIPS FP performance, C-60 to C-61
 - MIPS multicycle operations
 - basic considerations, C-51 to C-54
 - hazards and forwarding, C-54 to C-58
 - precise exceptions, C-58 to C-60
 - MIPS R4000
 - FP pipeline, C-65 to C-67, **C-67**
 - overview, C-61 to C-65
 - pipeline performance, C-67 to C-70
 - pipeline structure, C-62 to C-63
 - multiple outstanding FP operations, **C-54**
 - performance issues, C-10 to C-11
 - performance with stalls, C-12 to C-13
 - predicted-not-taken scheme, **C-22**
 - RISC instruction set, C-4 to C-5, C-70
 - simple implementation, C-30 to C-43, **C-34**
 - simple RISC, C-5 to C-6, **C-7**
 - static branch prediction, C-26 to C-27
 - structural hazards, C-13 to C-16, **C-15**
 - switch microarchitecture, F-60 to F-61
 - unoptimized code, C-81
- Pipe segment, definition, C-3
- Pipe stage
 - branch prediction, C-28
 - data hazards, C-16
 - definition, C-3
 - dynamic scheduling, C-71
 - FP pipeline, C-66
 - integrated instruction fetch units, 207
 - MIPS, C-34 to C-35, **C-37**, C-49
 - MIPS extension, C-53
 - MIPS R4000, **C-62**
 - out-of-order execution, 170
 - pipeline stalls, C-13
 - pipelining performance issues, C-10
 - register additions, **C-35**
 - RISC processor, C-7
 - stopping/restarting execution, C-46
 - WAW, 153
- pjbb2005 benchmark
 - Intel Core i7, 402
 - SMT on superscalar processors, 230–232, **231**
- PLA, early computer arithmetic, J-65
- PMD, *see* Personal mobile device (PMD)
- Points-to analysis, basic approach, H-9
- Point-to-point links
 - bus replacement, **D-34**
 - Ethernet, F-79
 - storage systems, D-34
 - switched-media networks, F-24
- Point-to-point multiprocessor, example, **413**
- Point-to-point networks
 - directory-based coherence, 418
 - directory protocol, 421–422
 - SMP limitations, 363–364
- Poison bits, compiler-based speculation, H-28, H-30
- Poisson, Siméon, D-28
- Poisson distribution
 - basic equation, D-28
 - random variables, D-26 to D-34
- Polycyclic scheduling, L-30
- Portable computers
 - interconnection networks, F-85
 - processor comparison, **242**
- Port number, network interfaces, F-7

- Position independence, control flow
 - instruction addressing modes, A-17
- Power
 - distribution for servers, **490**
 - distribution overview, **447**
 - and DLP, 322
 - first-level caches, 79–80
 - Google server benchmarks, 439–441
 - Google WSC, 465–468
 - PMDs, 6
 - real-world server considerations, 52–55
 - WSC infrastructure, 447
 - WSC power modes, 472
 - WSC resource allocation case study, 478–479
 - WSC TCO case study, 476–478
- Power consumption, *see also* Energy efficiency
 - cache optimization, **96**
 - cache size and associativity, **81**
 - case study, 63–64
 - computer components, **63**
 - DDR3 SDRAM, **103**
 - disks, D-5
 - embedded benchmarks, E-13
 - GPUs vs. vector architectures, 311
 - interconnection networks, F-85
 - ISA performance and efficiency prediction, 242–243
 - microprocessor, 23–26
 - SDRAMs, 102
 - SMT on superscalar processors, 230–231
 - speculation, 210–211
 - system trends, 21–23
 - TI TMS320C55 DSP, E-8
 - WSCs, 450
- Power distribution unit (PDU), WSC
 - infrastructure, 447
- Power failure
 - exceptions, C-43 to C-44, C-46
 - utilities, 435
 - WSC storage, 442
- Power gating, transistors, 26
- Power modes, WSCs, 472
- PowerPC
 - addressing modes, **K-5**
 - AltiVec multimedia instruction compiler support, A-31
 - ALU, K-5
 - arithmetic/logical instructions, **K-11**
 - branches, K-21
 - cluster history, L-63
 - conditional branches, **K-17**
 - conditional instructions, H-27
 - condition codes, K-10 to K-11
 - consistency model, 395
 - constant extension, **K-9**
 - conventions, **K-13**
 - data transfer instructions, **K-10**
 - features, **K-44**
 - FP instructions, **K-23**
 - IBM Blue Gene/L, I-41 to I-42
 - multimedia compiler support, A-31, K-17
 - precise exceptions, C-59
 - RISC architecture, A-2
 - RISC code size, A-23
 - as RISC systems, **K-4**
 - unique instructions, K-32 to K-33
- PowerPC ActiveC
 - characteristics, **K-18**
 - multimedia support, K-19
- PowerPC AltiVec, multimedia support, **E-11**
- Power-performance
 - low-power servers, **477**
 - servers, **54**
- Power Supply Units (PSUs),
 - efficiency ratings, **462**
- Power utilization effectiveness (PUE)
 - datacenter comparison, **451**
 - Google WSC, **468**
 - Google WSC containers, 464–465
 - WSC, 450–452
 - WSCs vs. datacenters, 456
 - WSC server energy efficiency, 462
- Precise exceptions
 - definition, C-47
 - dynamic scheduling, 170
 - hardware-based speculation, 187–188, 221
 - instruction set complications, C-49
 - maintaining, C-58 to C-60
 - MIPS exceptions, C-48
- Precisions, floating-point arithmetic, J-33 to J-34
- Predicated instructions
 - exposing parallelism, H-23 to H-27
 - IA-64, H-38 to H-40
- Predicate Registers
 - definition, **309**
 - GPU conditional branching, 300–301
 - IA-64, H-34
 - NVIDIA GPU ISA, 298
 - vectors vs. GPUs, 311
- Predication, TI TMS320C6x DSP, E-10
- Predicted-not-taken scheme
 - branch penalty reduction, **C-22**, C-22 to C-23
 - MIPS R4000 pipeline, C-64
- Predictions, *see also* Mispredictions
 - address aliasing, 213–214, 216
 - branch
 - correlation, 162–164
 - cost reduction, 162–167, C-26
 - dynamic, C-27 to C-30
 - ideal processor, 214
 - ILP exploitation, 201
 - instruction fetch bandwidth, 205
 - integrated instruction fetch units, 207
 - Intel Core i7, 166–167, 239–241
 - static, C-26 to C-27
 - branch-prediction buffers, C-27 to C-30, **C-29**
 - jump prediction, 214
 - PMDs, 6
 - return address buffer, **207**
 - 2-bit scheme, **C-28**
 - value prediction, 202, 212–213
- Prefetching
 - integrated instruction fetch units, 208
 - Intel Core i7, 122, **123–124**
 - Itanium 2, H-42
 - MIPS core extensions, K-20
 - NVIDIA GPU Memory structures, 305
 - parallel processing challenges, 351
- Prefix, Intel 80x86 integer operations, K-51
- Presentation layer, definition, **F-82**
- Present bit, IA-32 descriptor table, B-52
- Price vs. cost, 32–33
- Price-performance ratio
 - cost trends, 28
 - Dell PowerEdge servers, **53**
 - desktop computers, 6
 - processor comparisons, 55
 - WSCs, 8, 441

Primitives

- architect-compiler writer
 - relationship, A-30
- basic hardware types, 387–389
- compiler writing-architecture
 - relationship, A-30
- CUDA Thread, 289
- dependent computation
 - elimination, 321
- GPU vs. MIMD, 329
- locks via coherence, 391
- operand types and sizes, A-14 to A-15
- PA-RISC instructions, K-34 to K-35
- synchronization, 394, L-64

Principle of locality

- bidirectional MINs, F-33 to F-34
- cache optimization, B-26
- cache performance, B-3 to B-4
- coining of term, L-11
- commercial workload, 373
- computer design principles, 45
- definition, 45, B-2
- lock accesses, 390
- LRU, B-9
- memory accesses, 332, B-46
- memory hierarchy design, 72
- multilevel application, B-2
- multiprogramming workload, 375
- scientific workloads on symmetric
 - shared-memory
 - multiprocessors, I-25
- stride, 278
- WSC bottleneck, 461
- WSC efficiency, 450

Private data

- cache protocols, 359
- centralized shared-memory
 - multiprocessors, 351–352

Private Memory

- definition, **292, 314**
- NVIDIA GPU Memory structures, **304**

Private variables, NVIDIA GPU

- Memory, 304

Procedure calls

- compiler structure, A-25 to A-26
- control flow instructions, **A-17**, A-19 to A-21
- dependence analysis, 321

- high-level instruction set, A-42 to A-43

- IA-64 register model, H-33
- invocation options, A-19
- ISAs, 14
- MIPS control flow instructions, A-38
- return address predictors, 206
- VAX, **B-73 to B-74**, K-71 to K-72
- VAX vs. MIPS, **K-75**
- VAX swap, B-74 to B-75

Process concept

- definition, 106, B-49
- protection schemes, B-50

Process-identifier (PID) tags, cache

- addressing, B-37 to B-38

Process IDs, Virtual Machines, 110

Processor consistency

- latency hiding with speculation, 396–397
- relaxed consistency models, 395

Processor cycles

- cache performance, B-4
- definition, C-3
- memory banks, 277
- multithreading, 224

Processor-dependent optimizations

- compilers, A-26
- performance impact, A-27
- types, **A-28**

Processor-intensive benchmarks,

- desktop performance, 38

Processor performance

- and average memory access time, B-17 to B-20
- vs. cache performance, B-16
- clock rate trends, **24**
- desktop benchmarks, 38, 40
- historical trends, **3**, 3–4
- multiprocessors, 347
- uniprocessors, 344

Processor performance equation,

- computer design principles, 48–52

Processor speed

- and clock rate, 244
- and CPI, 244
- snooping cache coherence, 364

Process switch

- definition, 106, B-49
- miss rate vs. virtual addressing, B-37

- multithreading, 224

PID, **B-37**

- virtual memory-based protection, B-49 to B-50

Producer-server model, response time and throughput, **D-16**

Productivity

- CUDA, 290–291
- NVIDIA programmers, 289
- software development, 4
- virtual memory and programming, B-41
- WSC, 450

Profile-based predictor, misprediction rate, **C-27**

Program counter (PC)

- addressing modes, A-10
- ARM Cortex-A8, **234**
- branch hazards, C-21
- branch-target buffers, **203**, 203–204, 206
- control flow instruction addressing modes, A-17
- dynamic branch prediction, C-27 to C-28
- exception stopping/restarting, C-46 to C-47
- GPU conditional branching, 303
- Intel Core i7, 120
- M32R instructions, K-39
- MIPS control flow instructions, A-38
- multithreading, 223–224
- pipeline branch issues, C-39 to C-41
- pipe stages, **C-35**
- precise exceptions, C-59 to C-60
- RISC classic pipeline, C-8
- RISC instruction set, C-5
- simple MIPS implementation, C-31 to C-33
- TLP, 344
- virtual memory protection, 106

Program counter-relative addressing control flow instructions, A-17 to A-18, A-21

definition, A-10

MIPS instruction format, A-35

Programming models

- CUDA, 300, 310, 315
- GPUs, 288–291
- latency in consistency models, 397

- memory consistency, 393
- Multimedia SIMD architectures, 285
- vector architectures, 280–282
- WSCs, 436–441
- Programming primitive, CUDA Thread, 289
- Program order
 - cache coherence, 353
 - control dependences, 154–155
 - data hazards, 153
 - dynamic scheduling, 168–169, 174
 - hardware-based speculation, 192
 - ILP exploitation, 200
 - name dependences, 152–153
 - Tomasulo's approach, 182
- Protection schemes
 - control dependence, 155
 - development, L-9 to L-12
 - and ISA, 112
 - network interfaces, F-7
 - network user access, F-86 to F-87
 - Pentium vs. Opteron, B-57
 - processes, B-50
 - safe calls, B-54
 - segmented virtual memory
 - example, B-51 to B-54
 - Virtual Machines, 107–108
 - virtual memory, 105–107, B-41
- Protocol deadlock, routing, F-44
- Protocol stack
 - example, **F-83**
 - internetworking, F-83
- Pseudo-least recently used (LRU)
 - block replacement, B-9 to B-10
 - Intel Core i7, **118**
- PSUs, *see* Power Supply Units (PSUs)
- PTE, *see* Page Table Entry (PTE)
- PTX, *see* Parallel Thread Execution (PTX)
- PUE, *see* Power utilization
 - effectiveness (PUE)
- Python language, hardware impact on
 - software development, 4

Q

- QCDOD, L-64
- QoS, *see* Quality of service (QoS)
- QsNet^{II}, F-63, **F-76**
- Quadrants SAN, F-67, F-100 to F-101
- Quality of service (QoS)
 - dependability benchmarks, D-21

- WAN history, F-98
- Quantitative performance measures,
 - development, L-6 to L-7
- Queue
 - definition, D-24
 - waiting time calculations, D-28 to D-29
- Queue discipline, definition, D-26
- Queueing locks, large-scale
 - multiprocessor synchronization, I-18 to I-21
- Queueing theory
 - basic assumptions, D-30
 - Little's law, D-24 to D-25
 - M/M/1 model, D-31 to D-33, **D-32**
 - overview, D-23 to D-26
 - RAID performance prediction, D-57 to D-59
 - single-server model, **D-25**
- Quickpath (Intel Xeon), cache coherence, 361

R

- Race-to-halt, definition, 26
- Rack units (U), WSC architecture, 441
- Radio frequency amplifier, radio receiver, **E-23**
- Radio receiver, components, **E-23**
- Radio waves, wireless networks, E-21
- Radix-2 multiplication/division, J-4 to J-7, **J-6, J-55**
- Radix-4 multiplication/division, J-48 to J-49, **J-49, J-56 to J-57, J-60 to J-61**
- Radix-8 multiplication, J-49
- RAID (Redundant array of inexpensive disks)
 - data replication, 439
 - dependability benchmarks, D-21, **D-22**
 - disk array deconstruction case study, D-51, **D-55**
 - disk deconstruction case study, D-48
 - hardware dependability, D-15
 - historical background, L-79 to L-80
 - I/O subsystem design, D-59 to D-61
 - logical units, D-35
 - memory dependability, 104
 - NetApp FAS6000 filer, D-41 to D-42
 - overview, D-6 to D-8, **D-7**
 - performance prediction, D-57 to D-59
 - reconstruction case study, D-55 to D-57
 - row-diagonal parity, **D-9**
 - WSC storage, 442
- RAID 0, definition, D-6
- RAID 1
 - definition, D-6
 - historical background, L-79
- RAID 2
 - definition, D-6
 - historical background, L-79
- RAID 3
 - definition, D-7
 - historical background, L-79 to L-80
- RAID 4
 - definition, D-7
 - historical background, L-79 to L-80
- RAID 5
 - definition, D-8
 - historical background, L-79 to L-80
- RAID 6
 - characteristics, D-8 to D-9
 - hardware dependability, D-15
- RAID 10, D-8
- RAM (random access memory), switch microarchitecture, F-57
- RAMAC-350 (Random Access Method of Accounting Control), L-77 to L-78, L-80 to L-81
- Random Access Method of Accounting Control, L-77 to L-78
- Random replacement
 - cache misses, B-10
 - definition, B-9
- Random variables, distribution, D-26 to D-34
- RAR, *see* Read after read (RAR)
- RAS, *see* Row access strobe (RAS)
- RAW, *see* Read after write (RAW)
- Ray casting (RC)
 - GPU comparisons, 329
 - throughput computing kernel, **327**

- RDMA, *see* Remote direct memory access (RDMA)
- Read after read (RAR), absence of
 - data hazard, 154
- Read after write (RAW)
 - data hazards, 153
 - dynamic scheduling with
 - Tomasulo's algorithm, 170–171
 - first vector computers, L-45
 - hazards, stalls, **C-55**
 - hazards and forwarding, C-55 to C-57
 - instruction set complications, C-50
 - microarchitectural techniques case study, 253
 - MIPS FP pipeline performance, C-60 to C-61
 - MIPS pipeline control, C-37 to C-38
 - MIPS pipeline FP operations, C-53
 - MIPS scoreboarding, C-74
 - ROB, 192
 - TI TMS320C55 DSP, E-8
 - Tomasulo's algorithm, 182
 - unoptimized code, C-81
- Read miss
 - AMD Opteron data cache, B-14
 - cache coherence, 357, **358**, 359–361
 - coherence extensions, 362
 - directory-based cache coherence
 - protocol example, 380, 382–386
 - memory hierarchy basics, 76–77
 - memory stall clock cycles, B-4
 - miss penalty reduction, B-35 to B-36
 - Opteron data cache, B-14
 - vs.* write-through, B-11
- Read operands stage
 - ID pipe stage, 170
 - MIPS scoreboarding, C-74 to C-75
 - out-of-order execution, C-71
- Realizable processors, ILP limitations, 216–220
- Real memory, Virtual Machines, 110
- Real-time constraints, definition, E-2
- Real-time performance, PMDs, 6
- Real-time performance requirement,
 - definition, E-3
- Real-time processing, embedded systems, E-3 to E-5
- Rearrangeably nonblocking,
 - centralized switched networks, F-32 to F-33
- Receiving overhead
 - communication latency, I-3 to I-4
 - interconnection networks, F-88
 - OCNs *vs.* SANs, **F-27**
 - time of flight, F-14
- RECEN, *see* Regional explicit congestion notification (RECEN)
- Reconfiguration deadlock, routing, F-44
- Reconstruction, RAID, D-55 to D-57
- Recovery time, vector processor, G-8
- Recurrences
 - basic approach, H-11
 - loop-carried dependences, H-5
- Red-black Gauss-Seidel, Ocean
 - application, I-9 to I-10
- Reduced Instruction Set Computer,
 - see* RISC (Reduced Instruction Set Computer)
- Reductions
 - commercial workloads, 371
 - cost trends, 28
 - loop-level parallelism
 - dependences, 321
 - multiprogramming workloads, **377**
 - T1 multithreading uncore
 - performance, 227
 - WSCs, 438
- Redundancy
 - Amdahl's law, 48
 - chip fabrication cost case study, 61–62
 - computer system power
 - consumption case study, 63–64
 - index checks, B-8
 - integrated circuit cost, 32
 - integrated circuit failure, 35
 - simple MIPS implementation, C-33
 - WSC, 433, 435, 439
 - WSC bottleneck, 461
 - WSC storage, 442
- Redundant array of inexpensive disks,
 - see* RAID (Redundant array of inexpensive disks)
- Redundant multiplication, integers, J-48
- Redundant power supplies, example calculations, 35
- Reference bit
 - memory hierarchy, B-52
 - virtual memory block replacement, B-45
- Regional explicit congestion
 - notification (RECEN), congestion management, F-66
- Register addressing mode
 - MIPS, 12
 - VAX, K-67
- Register allocation
 - compilers, 396, A-26 to A-29
 - VAX sort, K-76
 - VAX swap, K-72
- Register deferred addressing, VAX, K-67
- Register definition, **314**
- Register fetch (RF)
 - MIPS data path, C-34
 - MIPS R4000, C-63
 - pipeline branches, C-41
 - simple MIPS implementation, C-31
 - simple RISC implementation, C-5 to C-6
- Register file
 - data hazards, C-16, **C-18**, C-20
 - dynamic scheduling, 172, **173**, 175, 177–178
 - Fermi GPU, 306
 - field, 176
 - hardware-based speculation, 184
 - longer latency pipelines, C-55 to C-57
 - MIPS exceptions, C-49
 - MIPS implementation, C-31, C-33
 - MIPS R4000, **C-64**
 - MIPS scoreboarding, C-75
 - Multimedia SIMD Extensions, 282, 285
 - multiple lanes, 272, **273**
 - multithreading, 224
 - OCNs, F-3
 - precise exceptions, C-59
 - RISC classic pipeline, C-7 to C-8
 - RISC instruction set, C-5 to C-6
 - scoreboarding, C-73, C-75

- speculation support, 208
- structural hazards, C-13
- Tomasulo's algorithm, **180**, 182
- vector architecture, 264
- VMIPS, 265, 308
- Register indirect addressing mode,
 - Intel 80x86, K-47
- Register management,
 - software-pipelined loops, H-14
- Register-memory instruction set architecture
 - architect-compiler writer relationship, A-30
 - dynamic scheduling, 171
 - Intel 80x86, K-52
 - ISA classification, 11, A-3 to A-6
- Register prefetch, cache optimization, 92
- Register renaming
 - dynamic scheduling, 169–172
 - hardware vs. software speculation, 222
 - ideal processor, 214
 - ILP hardware model, 214
 - ILP limitations, 213, 216–217
 - ILP for realizable processors, 216
 - instruction delivery and speculation, 202
 - microarchitectural techniques case study, 247–254
 - name dependences, 153
 - vs. ROB, 208–210
 - ROB instruction, 186
 - sample code, **250**
 - SMT, 225
 - speculation, 208–210
 - superscalar code, **251**
 - Tomasulo's algorithm, 183
 - WAW/WAR hazards, 220
- Register result status, MIPS scoreboard, C-76
- Registers
 - DSP examples, **E-6**
 - IA-64, H-33 to H-34
 - instructions and hazards, **C-17**
 - Intel 80x86, K-47 to K-49, **K-48**
 - network interface functions, F-7
 - pipe stages, **C-35**
 - PowerPC, K-10 to K-11
 - VAX swap, B-74 to B-75
- Register stack engine, IA-64, H-34
- Register tag example, **177**
- Register windows, SPARC
 - instructions, K-29 to K-30
- Regularity
 - bidirectional MINs, F-33 to F-34
 - compiler writing-architecture relationship, A-30
- Relative speedup, multiprocessor performance, 406
- Relaxed consistency models
 - basic considerations, 394–395
 - compiler optimization, 396
 - WSC storage software, 439
- Release consistency, relaxed consistency models, 395
- Reliability
 - Amdahl's law calculations, 56
 - commercial interconnection networks, F-66
 - example calculations, 48
 - I/O subsystem design, D-59 to D-61
 - modules, SLAs, 34
 - MTTF, 57
 - redundant power supplies, 34–35
 - storage systems, D-44
 - transistor scaling, 21
- Relocation, virtual memory, B-42
- Remainder, floating point, J-31 to J-32
- Remington-Rand, L-5
- Remote direct memory access (RDMA), InfiniBand, F-76
- Remote node, directory-based cache coherence protocol
 - basics, 381–382
- Reorder buffer (ROB)
 - compiler-based speculation, H-31
 - dependent instructions, **199**
 - dynamic scheduling, 175
 - FP unit with Tomasulo's algorithm, **185**
 - hardware-based speculation, 184–192
 - ILP exploitation, 199–200
 - ILP limitations, 216
 - Intel Core i7, 238
 - vs. register renaming, 208–210
- Repeat interval, MIPS pipeline FP operations, C-52 to C-53
- Replication
 - cache coherent multiprocessors, 354
 - centralized shared-memory architectures, 351–352
 - coherence enforcement, 354
 - R4000 performance, C-70
 - RAID storage servers, 439
 - TLP, 344
 - virtual memory, B-48 to B-49
 - WSCs, 438
- Reply, messages, F-6
- Reproducibility, performance results reporting, 41
- Request
 - messages, F-6
 - switch microarchitecture, F-58
- Requested protection level, segmented virtual memory, B-54
- Request-level parallelism (RLP)
 - basic characteristics, 345
 - definition, 9
 - from ILP, 4–5
 - MIMD, 10
 - multicore processors, 400
 - multiprocessors, 345
 - parallelism advantages, 44
 - server benchmarks, 40
 - WSCs, 434, 436
- Request phase, arbitration, F-49
- Request-reply deadlock, routing, F-44
- Reservation stations
 - dependent instructions, 199–200
 - dynamic scheduling, 178
 - example, **177**
 - fields, 176
 - hardware-based speculation, 184, 186, 189–191
 - ILP exploitation, 197, 199–200
 - Intel Core i7, 238–240
 - loop iteration example, **181**
 - microarchitectural techniques case study, 253–254
 - speculation, 208–209
 - Tomasulo's algorithm, 172, **173**, 174–176, 179, **180**, 180–182
- Resource allocation
 - computer design principles, 45
 - WSC case study, 478–479
- Resource sparing, commercial interconnection networks, F-66

- Response time, *see also* Latency
 I/O benchmarks, **D-18**
 performance considerations, 36
 performance trends, 18–19
 producer-server model, **D-16**
 server benchmarks, 40–41
 storage systems, D-16 to D-18
 vs. throughput, **D-17**
 user experience, 4
 WSCs, 450
- Responsiveness
 PMDs, 6
 as server characteristic, 7
- Restartable pipeline
 definition, C-45
 exceptions, C-46 to C-47
- Restorations, SLA states, 34
- Restoring division, J-5, **J-6**
- Resume events
 control dependences, 156
 exceptions, C-45 to C-46
 hardware-based speculation, 188
- Return address predictors
 instruction fetch bandwidth, 206–207
 prediction accuracy, **207**
- Returns
 Amdahl's law, 47
 cache coherence, 352–353
 compiler technology and
 architectural decisions, A-28
 control flow instructions, 14, **A-17**, A-21
 hardware primitives, 388
 Intel 80x86 integer operations, K-51
 invocation options, A-19
 procedure invocation options, A-19
 return address predictors, 206
- Reverse path, cell phones, E-24
- RF, *see* Register fetch (RF)
- Rings
 characteristics, **F-73**
 NEWS communication, F-42
 OCN history, F-104
 process protection, B-50
 topology, F-35 to F-36, **F-36**
- Ripple-carry adder, J-3, **J-3**, **J-42**
 chip comparison, J-60
- Ripple-carry addition, J-2 to J-3
- RISC (Reduced Instruction Set Computer)
 addressing modes, K-5 to K-6
 Alpha-unique instructions, K-27 to K-29
 architecture flaws vs. success, A-45
 ARM-unique instructions, K-36 to K-37
 basic concept, C-4 to C-5
 basic systems, K-3 to K-5
 cache performance, B-6
 classic pipeline stages, C-6 to C-10
 code size, A-23 to A-24
 compiler history, L-31
 desktop/server systems, **K-4**
 instruction formats, **K-7**
 multimedia extensions, K-16 to K-19
 desktop systems
 addressing modes, **K-5**
 arithmetic/logical instructions, **K-11**, **K-22**
 conditional branches, **K-17**
 constant extension, **K-9**
 control instructions, **K-12**
 conventions, **K-13**
 data transfer instructions, **K-10**, **K-21**
 features, **K-44**
 FP instructions, K-13, **K-23**
 multimedia extensions, **K-18**
 development, 2
 early pipelined CPUs, L-26
 embedded systems, **K-4**
 addressing modes, **K-6**
 arithmetic/logical instructions, **K-15**, **K-24**
 conditional branches, **K-17**
 constant extension, **K-9**
 control instructions, **K-16**
 conventions, **K-16**
 data transfers, **K-14**, **K-23**
 DSP extensions, K-19
 instruction formats, **K-8**
 multiply-accumulate, **K-20**
 historical background, L-19 to L-21
 instruction formats, K-5 to K-6
 instruction set lineage, **K-43**
 ISA performance and efficiency
 prediction, 241
 M32R-unique instructions, K-39 to K-40
 MIPS16-unique instructions, K-40 to K-42
 MIPS64-unique instructions, K-24 to K-27
 MIPS core common extensions, K-19 to K-24
 MIPS M2000 vs. VAX 8700, **L-21**
 Multimedia SIMD Extensions
 history, L-49 to L-50
 operations, 12
 PA-RISC-unique, K-33 to K-35
 pipelining efficiency, C-70
 PowerPC-unique instructions, K-32 to K-33
 Sanyo VPC-SX500 digital camera, E-19
 simple implementation, C-5 to C-6
 simple pipeline, **C-7**
 SPARC-unique instructions, K-29 to K-32
 Sun T1 multithreading, 226–227
 SuperH-unique instructions, K-38 to K-39
 Thumb-unique instructions, K-37 to K-38
 vector processor history, G-26
 Virtual Machines ISA support, 109
 Virtual Machines and virtual memory and I/O, 110
- RISC-I, L-19 to L-20
 RISC-II, L-19 to L-20
 RLP, *see* Request-level parallelism (RLP)
- ROB, *see* Reorder buffer (ROB)
- Roofline model
 GPU performance, **326**
 memory bandwidth, 332
 Multimedia SIMD Extensions, 285–288, **287**
- Round digit, J-18
- Rounding modes, J-14, J-17 to J-19, **J-18**, **J-20**
 FP precisions, J-34
 fused multiply-add, J-33
- Round-robin (RR)
 arbitration, F-49
 IBM 360, K-85 to K-86
 InfiniBand, F-74
- Routers
 BARRNet, **F-80**

Ethernet, F-79
 Routing algorithm
 commercial interconnection
 networks, **F-56**
 fault tolerance, F-67
 implementation, F-57
 Intel SCCC, F-70
 interconnection networks, F-21 to
 F-22, **F-27**, F-44 to F-48
 mesh network, **F-46**
 network impact, F-52 to F-55
 OCN history, F-104
 and overhead, F-93 to F-94
 SAN characteristics, **F-76**
 switched-media networks, F-24
 switch microarchitecture
 pipelining, F-61
 system area network history, F-100
 Row access strobe (RAS), DRAM, 98
 Row-diagonal parity
 example, **D-9**
 RAID, D-9
 Row major order, blocking, 89
 RR, *see* Round-robin (RR)
 RS format instructions, IBM 360,
 K-87
 Ruby on Rails, hardware impact on
 software development, 4
 RX format instructions, IBM 360,
 K-86 to K-87

S

S3, *see* Amazon Simple Storage
 Service (S3)
 SaaS, *see* Software as a Service (SaaS)
 Sandy Bridge dies, wafer example, **31**
 SANs, *see* System/storage area
 networks (SANs)
 Sanyo digital cameras, SOC, **E-20**
 Sanyo VPC-SX500 digital camera,
 embedded system case
 study, E-19
 SAS, *see* Serial Attach SCSI (SAS)
 drive
 SASI, L-81
 SATA (Serial Advanced Technology
 Attachment) disks
 Google WSC servers, 469
 NetApp FAS6000 filer, D-42
 power consumption, D-5
 RAID 6, D-8
 vs. SAS drives, **D-5**

 storage area network history, F-103
 Saturating arithmetic, DSP media
 extensions, E-11
 Saturating operations, definition, K-18
 to K-19
 SAXPY, GPU raw/relative
 performance, **328**
 Scalability
 cloud computing, 460
 coherence issues, 378–379
 Fermi GPU, 295
 Java benchmarks, 402
 multicore processors, 400
 multiprocessing, 344, 395
 parallelism, 44
 as server characteristic, 7
 transistor performance and wires,
 19–21
 WSCs, 8, 438
 WSCs vs. servers, 434
 Scalable GPUs, historical background,
 L-50 to L-51
 Scalar expansion, loop-level parallelism
 dependences, 321
 Scalar Processors, *see also*
 Superscalar processors
 definition, **292, 309**
 early pipelined CPUs, L-26 to L-27
 lane considerations, **273**
 Multimedia SIMD/GPU
 comparisons, 312
 NVIDIA GPU, 291
 prefetch units, 277
 vs. vector, 311, G-19
 vector performance, 331–332
 Scalar registers
 Cray X1, G-21 to G-22
 GPUs vs. vector architectures, 311
 loop-level parallelism
 dependences, 321–322
 Multimedia SIMD vs. GPUs, 312
 sample renaming code, 251
 vector vs. GPU, 311
 vs. vector performance, 331–332
 VMIPS, 265–266
 Scaled addressing, VAX, K-67
 Scaled speedup, Amdahl's law and
 parallel computers,
 406–407
 Scaling
 Amdahl's law and parallel
 computers, 406–407

 cloud computing, 456
 computation-to-communication
 ratios, **I-11**
 DVFS, 25, 52, 467
 dynamic voltage-frequency, 25,
 52, 467
 Intel Core i7, 404
 interconnection network speed, F-88
 multicore vs. single-core, 402
 processor performance trends, 3
 scientific applications on parallel
 processing, I-34
 shared- vs. switched-media
 networks, F-25
 transistor performance and wires,
 19–21
 VMIPS, 267
 Scan Line Interleave (SLI), scalable
 GPUs, L-51
 SCCC, *see* Intel Single-Chip Cloud
 Computing (SCCC)
 Schorr, Herb, L-28
 Scientific applications
 Barnes, I-8 to I-9
 basic characteristics, I-6 to I-7
 cluster history, L-62
 distributed-memory
 multiprocessors, I-26 to
 I-32, **I-28 to I-32**
 FFT kernel, I-7
 LU kernel, I-8
 Ocean, I-9 to I-10
 parallel processors, I-33 to I-34
 parallel program computation/
 communication, I-10 to
 I-12, **I-11**
 parallel programming, I-2
 symmetric shared-memory
 multiprocessors, I-21 to
 I-26, **I-23 to I-25**
 Scoreboarding
 ARM Cortex-A8, 233, **234**
 components, **C-76**
 definition, 170
 dynamic scheduling, 171, 175
 and dynamic scheduling, C-71 to
 C-80
 example calculations, C-77
 MIPS structure, **C-73**
 NVIDIA GPU, 296
 results tables, **C-78 to C-79**
 SIMD thread scheduler, 296

- Scripting languages, software
 - development impact, 4
- SCSI (Small Computer System Interface)
 - Berkeley's Tertiary Disk project, D-12
 - dependability benchmarks, D-21
 - disk storage, D-4
 - historical background, L-80 to L-81
 - I/O subsystem design, D-59
 - RAID reconstruction, D-56
 - storage area network history, F-102
- SDRAM, *see* Synchronous dynamic random-access memory (SDRAM)
- SDRWAVE, J-62
- Second-level caches, *see also* L2 caches
 - ARM Cortex-A8, 114
 - ILP, 245
 - Intel Core i7, 121
 - interconnection network, F-87
 - Itanium 2, **H-41**
 - memory hierarchy, B-48 to B-49
 - miss penalty calculations, B-33 to B-34
 - miss penalty reduction, B-30 to B-35
 - miss rate calculations, B-31 to B-35
 - and relative execution time, **B-34**
 - speculation, 210
 - SRAM, 99
- Secure Virtual Machine (SVM), 129
- Seek distance
 - storage disks, **D-46**
 - system comparison, **D-47**
- Seek time, storage disks, **D-46**
- Segment basics
 - Intel 80x86, **K-50**
 - vs.* page, **B-43**
 - virtual memory definition, B-42 to B-43
- Segment descriptor, IA-32 processor, B-52, **B-53**
- Segmented virtual memory
 - bounds checking, B-52
 - Intel Pentium protection, B-51 to B-54
 - memory mapping, B-52
 - vs.* paged, **B-43**
- safe calls, B-54
- sharing and protection, B-52 to B-53
- Self-correction, Newton's algorithm, J-28 to J-29
- Self-draining pipelines, L-29
- Self-routing, MINs, F-48
- Semantic clash, high-level instruction set, A-41
- Semantic gap, high-level instruction set, A-39
- Semiconductors
 - DRAM technology, 17
 - Flash memory, 18
 - GPU *vs.* MIMD, 325
 - manufacturing, 3–4
- Sending overhead
 - communication latency, I-3 to I-4
 - OCNs *vs.* SANs, **F-27**
 - time of flight, F-14
- Sense-reversing barrier
 - code example, **I-15, I-21**
 - large-scale multiprocessor synchronization, I-14
- Sequence of SIMD Lane Operations, definition, **292, 313**
- Sequency number, packet header, F-8
- Sequential consistency
 - latency hiding with speculation, 396–397
 - programmer's viewpoint, 394
 - relaxed consistency models, 394–395
 - requirements and implementation, 392–393
- Sequential interleaving, multibanked caches, 86, **86**
- Sequent Symmetry, L-59
- Serial Advanced Technology
 - Attachment disks, *see* SATA (Serial Advanced Technology
 - Technology Attachment) disks
- Serial Attach SCSI (SAS) drive
 - historical background, L-81
 - power consumption, D-5
 - vs.* SATA drives, **D-5**
- Serialization
 - barrier synchronization, I-16
 - coherence enforcement, 354
 - directory-based cache coherence, 382
- DSM multiprocessor cache
 - coherence, I-37
- hardware primitives, 387
- multiprocessor cache coherency, 353
- page tables, 408
- snooping coherence protocols, 356
- write invalidate protocol
 - implementation, 356
- Serpentine recording, L-77
- Serve-longest-queue (SLQ) scheme, arbitration, F-49
- ServerNet interconnection network,
 - fault tolerance, F-66 to F-67
- Servers, *see also* Warehouse-scale computers (WSCs)
 - as computer class, **5**
 - cost calculations, **454, 454–455**
 - definition, D-24
 - energy savings, **25**
 - Google WSC, **440, 467, 468–469**
 - GPU features, **324**
 - memory hierarchy design, **72**
 - vs.* mobile GPUs, 323–330
 - multiprocessor importance, 344
 - outage/anomaly statistics, 435
 - performance benchmarks, 40–41
 - power calculations, 463
 - power distribution example, **490**
 - power-performance benchmarks, **54, 439–441**
 - power-performance modes, **477**
 - real-world examples, 52–55
 - RISC systems
 - addressing modes and instruction formats, K-5 to K-6
 - examples, K-3, **K-4**
 - instruction formats, **K-7**
 - multimedia extensions, K-16 to K-19
 - single-server model, **D-25**
 - system characteristics, **E-4**
 - workload demands, 439
 - WSC *vs.* datacenters, 455–456
 - WSC data transfer, 446
 - WSC energy efficiency, 462–464
 - vs.* WSC facility costs, 472
 - WSC memory hierarchy, 444
 - WSC resource allocation case study, 478–479

- vs. WSCs, 432–434
- WSC TCO case study, 476–478
- Server side Java operations per second (ssj_ops)
 - example calculations, 439
 - power-performance, **54**
 - real-world considerations, 52–55
- Server utilization
 - calculation, D-28 to D-29
 - queuing theory, D-25
- Service accomplishment, SLAs, 34
- Service Health Dashboard, AWS, 457
- Service interruption, SLAs, 34
- Service level agreements (SLAs)
 - Amazon Web Services, 457
 - and dependability, 33
 - WSC efficiency, 452
- Service level objectives (SLOs)
 - and dependability, 33
 - WSC efficiency, 452
- Session layer, definition, **F-82**
- Set associativity
 - and access time, **77**
 - address parts, **B-9**
 - AMD Opteron data cache, B-12 to B-14
 - ARM Cortex-A8, 114
 - block placement, B-7 to B-8
 - cache block, B-7
 - cache misses, 83–84, **B-10**
 - cache optimization, 79–80, B-33 to B-35, B-38 to B-40
 - commercial workload, **371**
 - energy consumption, **81**
 - memory access times, **77**
 - memory hierarchy basics, 74, 76
 - nonblocking cache, 84
 - performance equations, **B-22**
 - pipelined cache access, 82
 - way prediction, 81
- Set basics
 - block replacement, B-9 to B-10
 - definition, B-7
- Set-on-less-than instructions (SLT)
 - MIPS16, K-14 to K-15
 - MIPS conditional branches, K-11 to K-12
- Settle time, D-46
- SFF, *see* Small form factor (SFF) disk
- SFS benchmark, NFS, D-20
- SGI, *see* Silicon Graphics systems (SGI)
- Shadow page table, Virtual Machines, 110
- Sharding, WSC memory hierarchy, 445
- Shared-media networks
 - effective bandwidth vs. nodes, **F-28**
 - exampl, **F-22**
 - latency and effective bandwidth, F-26 to F-28
 - multiple device connections, F-22 to F-24
 - vs. switched-media networks, F-24 to F-25
- Shared Memory
 - definition, **292, 314**
 - directory-based cache coherence, 418–420
 - DSM, 347–348, **348**, 354–355, 378–380
 - invalidate protocols, 356–357
 - SMP/DSM definition, 348
 - terminology comparison, 315
- Shared-memory communication, large-scale multiprocessors, I-5
- Shared-memory multiprocessors
 - basic considerations, 351–352
 - basic structure, 346–347
 - cache coherence, 352–353
 - cache coherence enforcement, 354–355
 - cache coherence example, 357–362
 - cache coherence extensions, 362–363
 - data caching, 351–352
 - definition, L-63
 - historical background, L-60 to L-61
 - invalidate protocol
 - implementation, 356–357
 - limitations, 363–364
 - performance, 366–378
 - single-chip multicore case study, 412–418
 - SMP and snooping limitations, 363–364
 - snooping coherence
 - implementation, 365–366
- snooping coherence protocols, 355–356
- WSCs, 435, 441
- Shared-memory synchronization, MIPS core extensions, K-21
- Shared state
 - cache block, 357, 359
 - cache coherence, **360**
 - cache miss calculations, 366–367
 - coherence extensions, 362
 - directory-based cache coherence
 - protocol basics, 380, 385
 - private cache, 358
- Sharing addition, segmented virtual memory, B-52 to B-53
- Shear algorithms, disk array deconstruction, D-51 to D-52, **D-52 to D-54**
- Shifting over zeros, integer multiplication/division, J-45 to J-47
- Short-circuiting, *see* Forwarding
- SI format instructions, IBM 360, K-87
- Signals, definition, E-2
- Signal-to-noise ratio (SNR), wireless networks, E-21
- Signed-digit representation
 - example, **J-54**
 - integer multiplication, J-53
- Signed number arithmetic, J-7 to J-10
- Sign-extended offset, RISC, C-4 to C-5
- Significand, J-15
- Sign magnitude, J-7
- Silicon Graphics 4D/240, L-59
- Silicon Graphics Altix, **F-76**, L-63
- Silicon Graphics Challenge, L-60
- Silicon Graphics Origin, L-61, L-63
- Silicon Graphics systems (SGI)
 - economies of scale, 456
 - miss statistics, **B-59**
 - multiprocessor software
 - development, 407–409
 - vector processor history, G-27
- SIMD (Single Instruction Stream, Multiple Data Stream)
 - definition, 10
 - Fermi GPU architectural innovations, 305–308
 - GPU conditional branching, 301

SIMD (*continued*)

- GPU examples, **325**
- GPU programming, 289–290
- GPUs vs. vector architectures, 308–**309**
- historical overview, L-55 to L-56
- loop-level parallelism, 150
- MapReduce, 438
- memory bandwidth, 332
- multimedia extensions, *see* Multimedia SIMD Extensions
- multiprocessor architecture, 346
- multithreaded, *see* Multithreaded SIMD Processor
- NVIDIA GPU computational structures, 291
- NVIDIA GPU ISA, 300
- power/DLP issues, 322
- speedup via parallelism, **263**
- supercomputer development, L-43 to L-44
- system area network history, F-100
- Thread Block mapping, **293**
- TI 320C6x DSP, E-9

SIMD Instruction

- CUDA Thread, 303
- definition, **292, 313**
- DSP media extensions, E-10
- function, 150, 291
- GPU Memory structures, **304**
- GPUs, 300, 305
- Grid mapping, **293**
- IBM Blue Gene/L, I-42
- Intel AVX, 438
- multimedia architecture programming, 285
- multimedia extensions, 282–285, 312
- multimedia instruction compilers, A-31 to A-32
- Multithreaded SIMD Processor block diagram, **294**
- PTX, 301
- Sony PlayStation 2, **E-16**
- Thread of SIMD Instructions, 295–296
- thread scheduling, 296–297, **297**, 305
- vector architectures as superset, 263–264
- vector/GPU comparison, 308

- Vector Registers, **309**
- SIMD Lane Registers, definition, **309, 314**

SIMD Lanes

- definition, **292, 296, 309**
- DLP, 322
- Fermi GPU, 305, 307
- GPU, 296–297, 300, 324
- GPU conditional branching, 302–303
- GPUs vs. vector architectures, 308, **310, 311**
- instruction scheduling, **297**
- multimedia extensions, 285
- Multimedia SIMD vs. GPUs, 312, 315
- multithreaded processor, 294
- NVIDIA GPU Memory, 304
- synchronization marker, 301
- vector vs. GPU, 308, 311

SIMD Processors, *see also*

- Multithreaded SIMD Processor
- block diagram, **294**
- definition, **292, 309, 313–314**
- dependent computation elimination, 321
- design, 333
- Fermi GPU, 296, 305–308
- Fermi GTX 480 GPU floorplan, **295, 295–296**
- GPU conditional branching, 302
- GPU vs. MIMD, 329
- GPU programming, 289–290
- GPUs vs. vector architectures, **310, 310–311**
- Grid mapping, **293**
- Multimedia SIMD vs. GPU, 312
- multiprocessor architecture, 346
- NVIDIA GPU computational structures, 291
- NVIDIA GPU Memory structures, 304–305
- processor comparisons, 324
- Roofline model, **287, 326**
- system area network history, F-100

SIMD Thread

- GPU conditional branching, 301–302
- Grid mapping, **293**
- Multithreaded SIMD processor, 294

- NVIDIA GPU, 296
- NVIDIA GPU ISA, 298
- NVIDIA GPU Memory structures, 305
- scheduling example, **297**
- vector vs. GPU, 308
- vector processor, 310

SIMD Thread Scheduler

- definition, **292, 314**
- example, 297
- Fermi GPU, 295, 305–307, **306**
- GPU, 296

SIMT (Single Instruction, Multiple Thread)

- GPU programming, 289
- vs. SIMD, **314**
- Warp, **313**

Simultaneous multithreading (SMT)

- characteristics, 226
- definition, 224–225
- historical background, L-34 to L-35
- IBM eServer p5 575, **399**
- ideal processors, 215
- Intel Core i7, 117–118, 239–241
- Java and PARSEC workloads, **403–404**

- multicore performance/energy efficiency, 402–405
- multiprocessing/
 - multithreading-based performance, 398–400
 - multithreading history, L-35
 - superscalar processors, 230–232

Single-extended precision

- floating-point arithmetic, J-33 to J-34

Single Instruction, Multiple Thread, *see* SIMT (Single Instruction, Multiple Thread)**Single Instruction Stream, Multiple Data Stream, *see* SIMD (Single Instruction Stream, Multiple Data Stream)****Single Instruction Stream, Single Data Stream, *see* SISD (Single Instruction Stream, Single Data Stream)**

- Single-level cache hierarchy, miss rates *vs.* cache size, B-33
- Single-precision floating point arithmetic, J-33 to J-34
- GPU examples, 325
- GPU *vs.* MIMD, 328
- MIPS data types, A-34
- MIPS operations, A-36
- Multimedia SIMD Extensions, 283
- operand sizes/types, 12, A-13
- as operand type, A-13 to A-14
- representation, J-15 to J-16
- Single-Streaming Processor (SSP)
 - Cray X1, G-21 to G-24
 - Cray X1E, G-24
- Single-thread (ST) performance
 - IBM eServer p5 575, 399, **399**
 - Intel Core i7, 239
 - ISA, 242
 - processor comparison, **243**
- SISD (Single Instruction Stream, Single Data Stream), 10
 - SIMD computer history, L-55
- Skippy algorithm
 - disk deconstruction, D-49
 - sample results, **D-50**
- SLAs, *see* Service level agreements (SLAs)
- SLI, *see* Scan Line Interleave (SLI)
- SLOs, *see* Service level objectives (SLOs)
- SLQ, *see* Serve-longest-queue (SLQ) scheme
- SLT, *see* Set-on-less-than instructions (SLT)
- SM, *see* Distributed shared memory (DSM)
- Small Computer System Interface, *see* SCSI (Small Computer System Interface)
- Small form factor (SFF) disk, L-79
- Smalltalk, SPARC instructions, K-30
- Smart interface cards, *vs.* smart switches, F-85 to F-86
- Smartphones
 - ARM Cortex-A8, 114
 - mobile *vs.* server GPUs, 323–324
- Smart switches, *vs.* smart interface cards, F-85 to F-86
- SMP, *see* Symmetric multiprocessors (SMP)
- SMT, *see* Simultaneous multithreading (SMT)
- Snooping cache coherence
 - basic considerations, 355–356
 - controller transitions, **421**
 - definition, 354–355
 - directory-based, 381, 386, 420–421
 - example, 357–362
 - implementation, 365–366
 - large-scale multiprocessor history, L-61
 - large-scale multiprocessors, I-34 to I-35
 - latencies, **414**
 - limitations, 363–364
 - sample types, **L-59**
 - single-chip multicore processor
 - case study, 412–418
 - symmetric shared-memory machines, 366
- SNR, *see* Signal-to-noise ratio (SNR)
- SoC, *see* System-on-chip (SoC)
- Soft errors, definition, 104
- Soft real-time
 - definition, E-3
 - PMDs, 6
- Software as a Service (SaaS)
 - clusters/WSCs, 8
 - software development, 4
 - WSCs, 438
 - WSCs *vs.* servers, 433–434
- Software development
 - multiprocessor architecture issues, 407–409
 - performance *vs.* productivity, 4
 - WSC efficiency, 450–452
- Software pipelining
 - example calculations, H-13 to H-14
 - loops, execution pattern, **H-15**
 - technique, H-12 to H-15, **H-13**
- Software prefetching, cache optimization, 131–133
- Software speculation
 - definition, 156
 - vs.* hardware speculation, 221–222
 - VLIW, 196
- Software technology
 - ILP approaches, 148
 - large-scale multiprocessors, I-6
- large-scale multiprocessor synchronization, I-17 to I-18
- network interfaces, F-7
- vs.* TCP/IP reliance, F-95
- Virtual Machines protection, 108
- WSC running service, 434–435
- Solaris, RAID benchmarks, **D-22**, D-22 to D-23
- Solid-state disks (SSDs)
 - processor performance/price/power, 52
 - server energy efficiency, 462
 - WSC cost-performance, 474–475
- Sonic Smart Interconnect, OCNs, F-3
- Sony PlayStation 2
 - block diagram, **E-16**
 - embedded multiprocessors, E-14
 - Emotion Engine case study, E-15 to E-18
 - Emotion Engine organization, **E-18**
- Sorting, case study, D-64 to D-67
- Sort primitive, GPU *vs.* MIMD, 329
- Sort procedure, VAX
 - bubble sort, **K-76**
 - example code, K-77 to K-79
 - vs.* MIPS32, **K-80**
 - register allocation, K-76
- Source routing, basic concept, F-48
- SPARCLE processor, L-34
- Sparse matrices
 - loop-level parallelism
 - dependences, 318–319
 - vector architectures, 279–280, G-12 to G-14
 - vector execution time, 271
 - vector mask registers, 275
- Spatial locality
 - coining of term, L-11
 - definition, 45, B-2
 - memory hierarchy design, 72
- SPEC benchmarks
 - branch predictor correlation, 162–164
 - desktop performance, 38–40
 - early performance measures, L-7
 - evolution, **39**
 - fallacies, 56
 - operands, A-14
 - performance, 38
 - performance results reporting, 41

- SPEC benchmarks (*continued*)
 - processor performance growth, 3
 - static branch prediction, C-26 to C-27
 - storage systems, D-20 to D-21
 - tournament predictors, 164
 - two-bit predictors, **165**
 - vector processor history, G-28
- SPEC89 benchmarks
 - branch-prediction buffers, C-28 to C-30, **C-30**
 - MIPS FP pipeline performance, **C-61 to C-62**
 - misprediction rates, **166**
 - tournament predictors, 165–166
 - VAX 8700 vs. MIPS M2000, **K-82**
- SPEC92 benchmarks
 - hardware vs. software speculation, 221
 - ILP hardware model, **215**
 - MIPS R4000 performance, C-68 to C-69, **C-69**
 - misprediction rate, **C-27**
- SPEC95 benchmarks
 - return address predictors, 206–207, **207**
 - way prediction, 82
- SPEC2000 benchmarks
 - ARM Cortex-A8 memory, 115–116
 - cache performance prediction, 125–126
 - cache size and misses per instruction, **126**
 - compiler optimizations, **A-29**
 - compulsory miss rate, B-23
 - data reference sizes, **A-44**
 - hardware prefetching, 91
 - instruction misses, **127**
- SPEC2006 benchmarks, evolution, **39**
- SPECCPU2000 benchmarks
 - displacement addressing mode, **A-12**
 - Intel Core i7, 122
 - server benchmarks, 40
- SPECCPU2006 benchmarks
 - branch predictors, **167**
 - Intel Core i7, **123–124, 240, 240–241**
 - ISA performance and efficiency prediction, 241
 - Virtual Machines protection, 108
- SPECfp benchmarks
 - hardware prefetching, 91
 - interconnection network, F-87
 - ISA performance and efficiency prediction, 241–242
 - Itanium 2, **H-43**
 - MIPS FP pipeline performance, C-60 to C-61
 - nonblocking caches, **84**
 - tournament predictors, 164
- SPECfp92 benchmarks
 - Intel 80x86 vs. DLX, **K-63**
 - Intel 80x86 instruction lengths, **K-60**
 - Intel 80x86 instruction mix, **K-61**
 - Intel 80x86 operand type distribution, **K-59**
 - nonblocking cache, 83
- SPECfp2000 benchmarks
 - hardware prefetching, **92**
 - MIPS dynamic instruction mix, **A-42**
 - Sun Ultra 5 execution times, **43**
- SPECfp2006 benchmarks
 - Intel processor clock rates, **244**
 - nonblocking cache, 83
- SPECfpRate benchmarks
 - multicore processor performance, 400
 - multiprocessor cost effectiveness, 407
 - SMT, 398–400
 - SMT on superscalar processors, 230
- SPEChpc96 benchmark, vector
 - processor history, G-28
- Special-purpose machines
 - historical background, L-4 to L-5
 - SIMD computer history, L-56
- Special-purpose register
 - compiler writing-architecture relationship, A-30
 - ISA classification, A-3
 - VMIPS, 267
- Special values
 - floating point, J-14 to J-15
 - representation, **J-16**
- SPECINT benchmarks
 - hardware prefetching, **92**
 - interconnection network, F-87
 - ISA performance and efficiency prediction, 241–242
 - Itanium 2, **H-43**
 - nonblocking caches, **84**
- SPECInt92 benchmarks
 - Intel 80x86 vs. DLX, **K-63**
 - Intel 80x86 instruction lengths, **K-60**
 - Intel 80x86 instruction mix, **K-62**
 - Intel 80x86 operand type distribution, **K-59**
 - nonblocking cache, 83
- SPECInt95 benchmarks,
 - interconnection networks, F-88
- SPECINT2000 benchmarks, MIPS
 - dynamic instruction mix, **A-41**
- SPECINT2006 benchmarks
 - Intel processor clock rates, **244**
 - nonblocking cache, 83
- SPECIntRate benchmark
 - multicore processor performance, 400
 - multiprocessor cost effectiveness, 407
 - SMT, 398–400
 - SMT on superscalar processors, 230
- SPEC Java Business Benchmark (JBB)
 - multicore processor performance, 400
 - multicore processors, **402**
 - multiprocessing/
 - multithreading-based performance, 398
 - server, 40
 - Sun T1 multithreading uncore performance, 227–229, **229**
- SPECJVM98 benchmarks, ISA
 - performance and efficiency prediction, 241
- SPECMail benchmark, characteristics, D-20
- SPEC-optimized processors, vs.
 - density-optimized, F-85
- SPECPower benchmarks
 - Google server benchmarks, 439–440, **440**
 - multicore processor performance, 400

- real-world server considerations, 52–55
- WSCs, **463**
- WSC server energy efficiency, 462–463
- SPECRate benchmarks
 - Intel Core i7, **402**
 - multicore processor performance, 400
 - multiprocessor cost effectiveness, 407
 - server benchmarks, 40
- SPECRate2000 benchmarks, SMT, 398–400
- SPECRatios
 - execution time examples, **43**
 - geometric means calculations, 43–44
- SPECSFS benchmarks
 - example, **D-20**
 - servers, 40
- Speculation, *see also* Hardware-based speculation; Software speculation
 - advantages/disadvantages, 210–211
 - compilers, *see* Compiler speculation
 - concept origins, L-29 to L-30
 - and energy efficiency, 211–212
 - FP unit with Tomasulo's algorithm, **185**
 - hardware vs. software, 221–222
 - IA-64, H-38 to H-40
 - ILP studies, L-32 to L-33
 - Intel Core i7, **123–124**
 - latency hiding in consistency models, 396–397
 - memory reference, hardware support, H-32
 - and memory system, 222–223
 - microarchitectural techniques case study, 247–254
 - multiple branches, 211
 - register renaming vs. ROB, 208–210
- SPECvirt_Sc2010 benchmarks, server, 40
- SPECWeb benchmarks
 - characteristics, D-20
 - dependability, D-21
 - parallelism, 44
 - server benchmarks, 40
- SPECWeb99 benchmarks
 - multiprocessing/
 - multithreading-based performance, 398
 - Sun T1 multithreading uncore performance, 227, **229**
- Speedup
 - Amdahl's law, 46–47
 - floating-point addition, J-25 to J-26
 - integer addition
 - carry-lookahead, J-37 to J-41
 - carry-lookahead circuit, **J-38**
 - carry-lookahead tree, **J-40 to J-41**
 - carry-lookahead tree adder, **J-41**
 - carry-select adder, **J-43**, J-43 to J-44, **J-44**
 - carry-skip adder, J-41 to J-43, **J-42**
 - overview, J-37
 - integer division
 - radix-2 division, **J-55**
 - radix-4 division, **J-56**
 - radix-4 SRT division, **J-57**
 - with single adder, J-54 to J-58
 - integer multiplication
 - array multiplier, **J-50**
 - Booth recoding, **J-49**
 - even/odd array, **J-52**
 - with many adders, J-50 to J-54
 - multipass array multiplier, **J-51**
 - signed-digit addition table, **J-54**
 - with single adder, J-47 to J-49, **J-48**
 - Wallace tree, **J-53**
 - integer multiplication/division,
 - shifting over zeros, J-45 to J-47
 - integer SRT division, J-45 to J-46, **J-46**
 - linear, 405–407
 - via parallelism, **263**
 - pipeline with stalls, C-12 to C-13
 - relative, 406
 - scaled, 406–407
 - switch buffer organizations, F-58 to F-59
 - true, 406
- Sperry-Rand, L-4 to L-5
- Spin locks
 - via coherence, 389–390
 - large-scale multiprocessor synchronization
 - barrier synchronization, I-16
 - exponential back-off, **I-17**
- SPLASH parallel benchmarks, SMT
 - on superscalar processors, 230
- Split, GPU vs. MIMD, 329
- SPRAM, Sony PlayStation 2 Emotion Engine organization, **E-18**
- Sprowl, Bob, F-99
- Squared coefficient of variance, D-27
- SRAM, *see* Static random-access memory (SRAM)
- SRT division
 - chip comparison, J-60 to J-61
 - complications, J-45 to J-46
 - early computer arithmetic, J-65
 - example, **J-46**
 - historical background, J-63
 - integers, with adder, J-55 to J-57
 - radix-4, J-56, **J-57**
- SSDs, *see* Solid-state disks (SSDs)
- SSE, *see* Intel Streaming SIMD Extension (SSE)
- SS format instructions, IBM 360, K-85 to K-88
- ssj_ops, *see* Server side Java operations per second (ssj_ops)
- SSP, *see* Single-Streaming Processor (SSP)
- Stack architecture
 - and compiler technology, A-27
 - flaws vs. success, A-44 to A-45
 - historical background, L-16 to L-17
 - Intel 80x86, **K-48**, **K-52**, **K-54**
 - operands, A-3 to A-4
- Stack frame, VAX, K-71
- Stack pointer, VAX, K-71
- Stack or Thread Local Storage,
 - definition, **292**
- Stale copy, cache coherency, 112
- Stall cycles
 - advanced directory protocol case study, 424
 - average memory access time, B-17
 - branch hazards, C-21
 - branch scheme performance, C-25

- Stall cycles (*continued*)
 - definition, B-4 to B-5
 - example calculation, B-31
 - loop unrolling, 161
 - MIPS FP pipeline performance, C-60
 - miss rate calculation, B-6
 - out-of-order processors, B-20 to B-21
 - performance equations, **B-22**
 - pipeline performance, C-12 to C-13
 - single-chip multicore
 - multiprocessor case study, 414–418
 - structural hazards, C-15
- Stalls
 - AMD Opteron data cache, B-15
 - ARM Cortex-A8, **235**, 235–236
 - branch hazards, **C-42**
 - data hazard minimization, C-16 to C-19, **C-18**
 - data hazards requiring, C-19 to C-21
 - delayed branch, **C-65**
 - Intel Core i7, 239–241
 - microarchitectural techniques case study, 252
 - MIPS FP pipeline performance, C-60 to C-61, **C-61 to C-62**
 - MIPS pipeline multicycle operations, C-51
 - MIPS R4000, **C-64**, **C-67**, C-67 to C-69, **C-69**
 - miss rate calculations, B-31 to B-32
 - necessity, **C-21**
 - nonblocking cache, 84
 - pipeline performance, C-12 to C-13
 - from RAW hazards, FP code, **C-55**
 - structural hazard, **C-15**
 - VLIV sample code, **252**
 - VMIPS, 268
- Standardization, commercial
 - interconnection networks, F-63 to F-64
- Stardent-1500, Livermore Fortran kernels, **331**
- Start-up overhead, *vs.* peak performance, 331
- Start-up time
 - DAXPY on VMIPS, G-20
 - memory banks, 276
 - page size selection, B-47
 - peak performance, 331
 - vector architectures, 331, G-4, **G-4**, **G-8**
 - vector convoys, **G-4**
 - vector execution time, 270–271
 - vector performance, G-2
 - vector performance measures, G-16
 - vector processor, G-7 to G-9, G-25
 - VMIPS, **G-5**
- State transition diagram
 - director *vs.* cache, **385**
 - directory-based cache coherence, **383**
- Statically based exploitation, ILP, H-2
- Static power
 - basic equation, 26
 - SMT, 231
- Static random-access memory (SRAM)
 - characteristics, 97–98
 - dependability, 104
 - fault detection pitfalls, 58
 - power, 26
 - vector memory systems, G-9
 - vector processor, G-25
 - yield, 32
- Static scheduling
 - definition, C-71
 - ILP, 192–196
 - and unoptimized code, C-81
- Sticky bit, J-18
- Stop & Go, *see* Xon/Xoff
- Storage area networks
 - dependability benchmarks, D-21 to D-23, **D-22**
 - historical overview, F-102 to F-103
 - I/O system as black box, **D-23**
- Storage systems
 - asynchronous I/O and OSes, D-35
 - Berkeley's Tertiary Disk project, D-12
 - block servers *vs.* filers, D-34 to D-35
 - bus replacement, D-34
 - component failure, D-43
 - computer system availability, D-43 to D-44, **D-44**
- dependability benchmarks, D-21 to D-23
- dirty bits, D-61 to D-64
- disk array deconstruction case study, D-51 to D-55, **D-52 to D-55**
- disk arrays, D-6 to D-10
- disk deconstruction case study, D-48 to D-51, **D-50**
- disk power, D-5
- disk seeks, D-45 to D-47
- disk storage, D-2 to D-5
- file system benchmarking, **D-20**, D-20 to D-21
- Internet Archive Cluster, *see* Internet Archive Cluster
- I/O performance, D-15 to D-16
- I/O subsystem design, D-59 to D-61
- I/O system design/evaluation, D-36 to D-37
- mail server benchmarking, D-20 to D-21
- NetApp FAS6000 filer, D-41 to D-42
- operator dependability, D-13 to D-15
- OS-scheduled disk access, D-44 to D-45, **D-45**
- point-to-point links, D-34, **D-34**
- queue I/O request calculations, D-29
- queuing theory, D-23 to D-34
- RAID performance prediction, D-57 to D-59
- RAID reconstruction case study, D-55 to D-57
- real faults and failures, D-6 to D-10
- reliability, D-44
- response time restrictions for benchmarks, **D-18**
- seek distance comparison, **D-47**
- seek time *vs.* distance, **D-46**
- server utilization calculation, D-28 to D-29
- sorting case study, D-64 to D-67
- Tandem Computers, D-12 to D-13
- throughput *vs.* response time, **D-16**, D-16 to D-18, **D-17**
- TP benchmarks, D-18 to D-19

- transactions components, **D-17**
- web server benchmarking, D-20 to D-21
- WSC vs. datacenter costs, 455
- WSCs, 442–443
- Store conditional
 - locks via coherence, 391
 - synchronization, 388–389
- Store-and-forward packet switching, F-51
- Store instructions, *see also* Load-store instruction set architecture
 - definition, C-4
 - instruction execution, 186
 - ISA, 11, A-3
 - MIPS, A-33, **A-36**
 - NVIDIA GPU ISA, 298
 - Opteron data cache, B-15
 - RISC instruction set, C-4 to C-6, C-10
 - vector architectures, 310
- Streaming Multiprocessor
 - definition, **292, 313–314**
 - Fermi GPU, 307
- Strecker, William, K-65
- Strided accesses
 - Multimedia SIMD Extensions, 283
 - Roofline model, **287**
 - TLB interaction, 323
- Strided addressing, *see also* Unit stride addressing
 - multimedia instruction compiler support, A-31 to A-32
- Strides
 - gather-scatter, 280
 - highly parallel memory systems, 133
 - multidimensional arrays in vector architectures, 278–279
 - NVIDIA GPU ISA, 300
 - vector memory systems, G-10 to G-11
 - VMIPS, **266**
- String operations, Intel 80x86, K-51, **K-53**
- Stripe, disk array deconstruction, D-51
- Stripping
 - disk arrays, D-6
 - RAID, D-9
- Strip-Mined Vector Loop
 - convoys, G-5
- DAXPY on VMIPS, G-20
- definition, **292**
- multidimensional arrays, 278
- Thread Block comparison, 294
- vector-length registers, 274
- Strip mining
 - DAXPY on VMIPS, G-20
 - GPU conditional branching, 303
 - GPUs vs. vector architectures, 311
 - NVIDIA GPU, 291
 - vector, **275**
 - VLRs, 274–275
- Strong scaling, Amdahl's law and parallel computers, 407
- Structural hazards
 - basic considerations, C-13 to C-16
 - definition, C-11
 - MIPS pipeline, C-71
 - MIPS scoreboarding, C-78 to C-79
 - pipeline stall, **C-15**
 - vector execution time, 268–269
- Structural stalls, MIPS R4000
 - pipeline, C-68 to C-69
- Subset property, and inclusion, 397
- Summary overflow condition code, PowerPC, K-10 to K-11
- Sun Microsystems
 - cache optimization, B-38
 - fault detection pitfalls, 58
 - memory dependability, 104
- Sun Microsystems Enterprise, L-60
- Sun Microsystems Niagara (T1/T2) processors
 - characteristics, **227**
 - CPI and IPC, **399**
 - fine-grained multithreading, 224, **225**, 226–229
 - manufacturing cost, **62**
 - multicore processor performance, 400–401
 - multiprocessing/
 - multithreading-based performance, 398–400
 - multithreading history, L-34
 - T1 multithreading uncore
 - performance, 227–229
- Sun Microsystems SPARC
 - addressing modes, **K-5**
 - ALU operands, **A-6**
 - arithmetic/logical instructions, **K-11, K-31**
 - branch conditions, A-19
 - conditional branches, K-10, **K-17**
 - conditional instructions, H-27
 - constant extension, **K-9**
 - conventions, **K-13**
 - data transfer instructions, **K-10**
 - fast traps, K-30
 - features, **K-44**
 - FP instructions, **K-23**
 - instruction list, K-31 to K-32
 - integer arithmetic, J-12
 - integer overflow, **J-11**
 - ISA, A-2
 - LISP, K-30
 - MIPS core extensions, K-22 to K-23
 - overlapped integer/FP operations, K-31
 - precise exceptions, C-60
 - register windows, K-29 to K-30
 - RISC history, L-20
 - as RISC system, **K-4**
 - Smalltalk, K-30
 - synchronization history, L-64
 - unique instructions, K-29 to K-32
- Sun Microsystems SPARCCenter, L-60
- Sun Microsystems SPARCstation-2, F-88
- Sun Microsystems SPARCstation-20, F-88
- Sun Microsystems SPARC V8,
 - floating-point precisions, J-33
- Sun Microsystems SPARC VIS
 - characteristics, **K-18**
 - multimedia support, **E-11, K-18**
- Sun Microsystems Ultra 5,
 - SPECfp2000 execution times, **43**
- Sun Microsystems UltraSPARC, L-62, L-73
- Sun Microsystems UltraSPARC T1 processor,
 - characteristics, **F-73**
- Sun Modular Datacenter, L-74 to L-75
- Superblock scheduling
 - basic process, H-21 to H-23
 - compiler history, L-31
 - example, **H-22**
- Supercomputers
 - commercial interconnection networks, F-63
 - direct network topology, **F-37**

- Supercomputers (*continued*)
 - low-dimensional topologies, F-100
 - SAN characteristics, **F-76**
 - SIMD, development, L-43 to L-44
 - vs. WSCs, 8
- Superlinear performance,
 - multiprocessors, 406
- Superpipelining
 - definition, C-61
 - performance histories, **20**
- Superscalar processors
 - coining of term, L-29
 - ideal processors, 214–215
 - ILP, 192–197, 246
 - studies, L-32
 - microarchitectural techniques case study, 250–251
 - multithreading support, **225**
 - recent advances, L-33 to L-34
 - register renaming code, **251**
 - rename table and register substitution logic, **251**
 - SMT, 230–232
 - VMIPS, 267
- Superscalar registers, sample renaming code, 251
- Supervisor process, virtual memory protection, 106
- Sussenguth, Ed, L-28
- Sutherland, Ivan, L-34
- SVM, *see* Secure Virtual Machine (SVM)
- Swap procedure, VAX
 - code example, K-72, K-74
 - full procedure, K-75 to K-76
 - overview, K-72 to K-76
 - register allocation, K-72
 - register preservation, B-74 to B-75
- Swim, data cache misses, B-10
- Switched-media networks
 - basic characteristics, F-24
 - vs. buses, F-2
 - effective bandwidth vs. nodes, **F-28**
 - example, **F-22**
 - latency and effective bandwidth, F-26 to F-28
 - vs. shared-media networks, F-24 to F-25
- Switched networks
 - centralized, F-30 to F-34
 - DOR, F-46
 - OCN history, F-104
 - topology, F-40
- Switches
 - array, WSCs, 443–444
 - Beneš networks, **F-33**
 - context, 307, B-49
 - early LANs and WANs, F-29
 - Ethernet switches, 16, **20**, 53, 441–444, 464–465, 469
 - interconnecting node calculations, F-35
 - vs. NIC, F-85 to F-86, **F-86**
 - process switch, 224, B-37, B-49 to B-50
 - storage systems, D-34
 - switched-media networks, F-24
 - WSC hierarchy, 441–442, **442**
 - WSC infrastructure, 446
 - WSC network bottleneck, 461
- Switch fabric, switched-media networks, F-24
- Switching
 - commercial interconnection networks, **F-56**
 - interconnection networks, F-22, **F-27**, F-50 to F-52
 - network impact, F-52 to F-55
 - performance considerations, F-92 to F-93
 - SAN characteristics, **F-76**
 - switched-media networks, F-24
 - system area network history, F-100
- Switch microarchitecture
 - basic microarchitecture, F-55 to F-58
 - buffer organizations, F-58 to F-60
 - enhancements, F-62
 - HOL blocking, **F-59**
 - input-output-buffered switch, **F-57**
 - pipelining, F-60 to F-61, **F-61**
- Switch ports
 - centralized switched networks, F-30
 - interconnection network topology, F-29
- Switch statements
 - control flow instruction addressing modes, A-18
 - GPU, 301
- Syllable, IA-64, H-35
- Symbolic loop unrolling, software pipelining, H-12 to H-15, **H-13**
- Symmetric multiprocessors (SMP)
 - characteristics, **I-45**
 - communication calculations, 350
 - directory-based cache coherence, 354
 - first vector computers, L-47, L-49
 - limitations, 363–364
 - snooping coherence protocols, 354–355
 - system area network history, F-101
 - TLP, 345
- Symmetric shared-memory
 - multiprocessors, *see* also Centralized shared-memory multiprocessors
 - data caching, 351–352
 - limitations, 363–364
 - performance
 - commercial workload, 367–369
 - commercial workload measurement, 369–374
 - multiprogramming and OS workload, 374–378
 - overview, 366–367
 - scientific workloads, I-21 to I-26, **I-23 to I-25**
- Synapse N + 1, L-59
- Synchronization
 - AltaVista search, 369
 - basic considerations, 386–387
 - basic hardware primitives, 387–389
 - consistency models, 395–396
 - cost, 403
 - Cray X1, G-23
 - definition, 375
 - GPU comparisons, 329
 - GPU conditional branching, 300–303
 - historical background, L-64
 - large-scale multiprocessors
 - barrier synchronization, I-13 to I-16, **I-14**, **I-16**
 - challenges, I-12 to I-16
 - hardware primitives, I-18 to I-21
 - sense-reversing barrier, **I-21**
 - software implementations, I-17 to I-18
 - tree-based barriers, **I-19**
 - locks via coherence, 389–391

message-passing communication, I-5

MIMD, 10

MIPS core extensions, K-21

programmer's viewpoint, 393–394

PTX instruction set, 298–299

relaxed consistency models, 394–395

single-chip multicore processor
case study, 412–418

vector vs. GPU, 311

VLIW, 196

WSCs, 434

Synchronous dynamic random-access memory (SDRAM)
ARM Cortex-A8, 117

DRAM, **99**

vs. Flash memory, 103

IBM Blue Gene/L, I-42

Intel Core i7, 121

performance, 100

power consumption, 102, **103**

SDRAM timing diagram, **139**

Synchronous event, exception
requirements, C-44 to C-45

Synchronous I/O, definition, D-35

Synonyms
address translation, B-38

dependability, 34

Synthetic benchmarks
definition, 37

typical program fallacy, A-43

System area networks, historical
overview, F-100 to F-102

System calls
CUDA Thread, 297

multiprogrammed workload, 378

virtualization/paravirtualization
performance, **141**

virtual memory protection, 106

System interface controller (SIF), Intel
SCCC, F-70

System-on-chip (SoC)
cell phone, E-24

cross-company interoperability, F-64

embedded systems, E-3

Sanyo digital cameras, **E-20**

Sanyo VPC-SX500 digital camera, E-19

shared-media networks, F-23

System Performance and Evaluation
Cooperative (SPEC),
see SPEC benchmarks

System Processor
definition, **309**

DLP, 262, 322

Fermi GPU, 306

GPU issues, 330

GPU programming, 288–289

NVIDIA GPU ISA, 298

NVIDIA GPU Memory, 305

processor comparisons, 323–324

synchronization, 329

vector vs. GPU, 311–312

System response time, transactions,
D-16, **D-17**

Systems on a chip (SOC), cost trends,
28

System/storage area networks (SANs)
characteristics, F-3 to F-4

communication protocols, F-8

congestion management, F-65

cross-company interoperability, F-64

effective bandwidth, F-18

example system, F-72 to F-74

fat trees, F-34

fault tolerance, F-67

InfiniBand example, F-74 to F-77

interconnection network domain
relationship, **F-4**

LAN history, F-99

latency and effective bandwidth,
F-26 to F-28

latency vs. nodes, **F-27**

packet latency, **F-13**, F-14 to F-16

routing algorithms, F-48

software overhead, F-91

TCP/IP reliance, F-95

time of flight, F-13

topology, F-30

System Virtual Machines, definition,
107

T

Tag

AMD Opteron data cache, B-12 to B-14

ARM Cortex-A8, **115**

cache optimization, 79–80

dynamic scheduling, **177**

invalidate protocols, 357

memory hierarchy basics, 74

memory hierarchy basics, 77–78

virtual memory fast address
translation, B-46

write strategy, B-10

Tag check (TC)
MIPS R4000, C-63

R4000 pipeline, B-62 to B-63

R4000 pipeline structure, **C-63**

write process, B-10

Tag fields
block identification, B-8

dynamic scheduling, **173**, 175

Tail duplication, superblock
scheduling, H-21

Tailgating, definition, G-20

Tandem Computers
cluster history, L-62, L-72

faults, **D-14**

overview, D-12 to D-13

Target address
branch hazards, C-21, **C-42**

branch penalty reduction, C-22 to C-23

branch-target buffer, 206

control flow instructions, A-17 to A-18

GPU conditional branching, 301

Intel Core i7 branch predictor, 166

MIPS control flow instructions,
A-38

MIPS implementation, C-32

MIPS pipeline, C-36, **C-37**

MIPS R4000, C-25

pipeline branches, C-39

RISC instruction set, C-5

Target channel adapters (TCAs),
switch vs. NIC, F-86

Target instructions
branch delay slot scheduling, C-24

as branch-target buffer variation,
206

GPU conditional branching, 301

Task-level parallelism (TLP),
definition, 9

TB, *see* Translation buffer (TB)

TB-80 VME rack
example, **D-38**

MTTF calculation, D-40 to D-41

TC, *see* Tag check (TC)

TCAs, *see* Target channel adapters (TCAs)

- TCO, *see* Total Cost of Ownership (TCO)
- TCP, *see* Transmission Control Protocol (TCP)
- TCP/IP, *see* Transmission Control Protocol/Internet Protocol (TCP/IP)
- TDMA, *see* Time division multiple access (TDMA)
- TDP, *see* Thermal design power (TDP)
- Technology trends
 - basic considerations, 17–18
 - performance, 18–19
- Teleconferencing, multimedia support, K-17
- Temporal locality
 - blocking, 89–90
 - cache optimization, B-26
 - coining of term, L-11
 - definition, 45, B-2
 - memory hierarchy design, 72
- TERA processor, L-34
- Terminate events
 - exceptions, C-45 to C-46
 - hardware-based speculation, 188
 - loop unrolling, 161
- Tertiary Disk project
 - failure statistics, **D-13**
 - overview, D-12
 - system log, **D-43**
- Test-and-set operation,
 - synchronization, 388
- Texas Instruments 8847
 - arithmetic functions, J-58 to J-61
 - chip comparison, **J-58**
 - chip layout, **J-59**
- Texas Instruments ASC
 - first vector computers, L-44
 - peak performance vs. start-up overhead, 331
- TFLOPS, parallel processing debates, L-57 to L-58
- TFT, *see* Thin-film transistor (TFT)
- Thacker, Chuck, F-99
- Thermal design power (TDP), power trends, 22
- Thin-film transistor (TFT), Sanyo VPC-SX500 digital camera, E-19
- Thinking Machines, L-44, L-56
- Thinking Multiprocessors CM-5, L-60
- Think time, transactions, D-16, **D-17**
- Third-level caches, *see also* L3 caches
 - ILP, 245
 - interconnection network, F-87
 - SRAM, 98–99
- Thrash, memory hierarchy, B-25
- Thread Block
 - CUDA Threads, 297, 300, 303
 - definition, **292, 313**
 - Fermi GTX 480 GPU flooplan, 295
 - function, 294
 - GPU hardware levels, 296
 - GPU Memory performance, 332
 - GPU programming, 289–290
 - Grid mapping, **293**
 - mapping example, **293**
 - multithreaded SIMD Processor, 294
 - NVIDIA GPU computational structures, 291
 - NVIDIA GPU Memory structures, **304**
 - PTX Instructions, 298
- Thread Block Scheduler
 - definition, **292, 309, 313–314**
 - Fermi GTX 480 GPU flooplan, 295
 - function, 294, 311
 - GPU, 296
 - Grid mapping, **293**
 - multithreaded SIMD Processor, 294
- Thread-level parallelism (TLP)
 - advanced directory protocol case study, 420–426
- Amdahl's law and parallel computers, 406–407
- centralized shared-memory multiprocessors
 - basic considerations, 351–352
 - cache coherence, 352–353
 - cache coherence enforcement, 354–355
 - cache coherence example, 357–362
 - cache coherence extensions, 362–363
 - invalidate protocol implementation, 356–357
 - SMP and snooping limitations, 363–364
 - snooping coherence implementation, 365–366
 - snooping coherence protocols, 355–356
- definition, 9
- directory-based cache coherence
 - case study, 418–420
 - protocol basics, 380–382
 - protocol example, 382–386
- DSM and directory-based coherence, 378–380
- embedded systems, E-15
- IBM Power7, 215
- from ILP, 4–5
- inclusion, 397–398
- Intel Core i7 performance/energy efficiency, 401–405
- memory consistency models
 - basic considerations, 392–393
 - compiler optimization, 396
 - programming viewpoint, 393–394
 - relaxed consistency models, 394–395
 - speculation to hide latency, 396–397
- MIMDs, 344–345
- multicore processor performance, 400–401
- multicore processors and SMT, 404–405
- multiprocessing/
 - multithreading-based performance, 398–400
- multiprocessor architecture, 346–348
- multiprocessor cost effectiveness, 407
- multiprocessor performance, 405–406
- multiprocessor software
 - development, 407–409
- vs. multithreading, 223–224
- multithreading history, L-34 to L-35
- parallel processing challenges, 349–351
- single-chip multicore processor
 - case study, 412–418
- Sun T1 multithreading, 226–229
- symmetric shared-memory multiprocessor
 - performance
 - commercial workload, 367–369
 - commercial workload measurement, 369–374

- multiprogramming and OS workload, 374–378
 - overview, 366–367
 - synchronization
 - basic considerations, 386–387
 - basic hardware primitives, 387–389
 - locks via coherence, 389–391
- Thread Processor
 - definition, **292**, **314**
 - GPU, 315
- Thread Processor Registers, definition, **292**
- Thread Scheduler in a Multithreaded CPU, definition, **292**
- Thread of SIMD Instructions
 - characteristics, 295–296
 - CUDA Thread, 303
 - definition, **292**, **313**
 - Grid mapping, **293**
 - lane recognition, 300
 - scheduling example, **297**
 - terminology comparison, **314**
 - vector/GPU comparison, 308–**309**
- Thread of Vector Instructions, definition, **292**
- Three-dimensional space, direct networks, F-38
- Three-level cache hierarchy
 - commercial workloads, 368
 - ILP, 245
 - Intel Core i7, 118, **118**
- Throttling, packets, F-10
- Throughput, *see also* Bandwidth
 - definition, C-3, F-13
 - disk storage, **D-4**
 - Google WSC, 470
 - ILP, 245
 - instruction fetch bandwidth, 202
 - Intel Core i7, 236–237
 - kernel characteristics, **327**
 - memory banks, 276
 - multiple lanes, 271
 - parallelism, 44
 - performance considerations, 36
 - performance trends, 18–19
 - pipelining basics, C-10
 - precise exceptions, C-60
 - producer-server model, **D-16**
 - vs.* response time, **D-17**
 - routing comparison, **F-54**
 - server benchmarks, 40–41
 - servers, 7
 - storage systems, D-16 to D-18
 - uniprocessors, TLP
 - basic considerations, 223–226
 - fine-grained multithreading on Sun T1, 226–229
 - superscalar SMT, 230–232
 - and virtual channels, F-93
 - WSCs, 434
- Ticks
 - cache coherence, **391**
 - processor performance equation, 48–49
- Tilera TILE-Gx processors, OCNs, F-3
- Time-cost relationship, components, 27–28
- Time division multiple access (TDMA), cell phones, E-25
- Time of flight
 - communication latency, I-3 to I-4
 - interconnection networks, F-13
- Timing independent, L-17 to L-18
- TI TMS320C6x DSP
 - architecture, **E-9**
 - characteristics, E-8 to E-10
 - instruction packet, **E-10**
- TI TMS320C55 DSP
 - architecture, E-7
 - characteristics, E-7 to E-8
 - data operands, **E-6**
- TLB, *see* Translation lookaside buffer (TLB)
- TLP, *see* Task-level parallelism (TLP); Thread-level parallelism (TLP)
- Tomasulo's algorithm
 - advantages, 177–178
 - dynamic scheduling, 170–176
 - FP unit, **185**
 - loop-based example, 179, 181–183
 - MIP FP unit, **173**
 - register renaming *vs.* ROB, 209
 - step details, 178, **180**
- TOP500, L-58
- Top Of Stack (TOS) register, ISA operands, **A-4**
- Topology
 - Ben's networks, **F-33**
 - centralized switched networks, F-30 to F-34, **F-31**
 - definition, F-29
 - direct networks, **F-37**
 - distributed switched networks, F-34 to F-40
 - interconnection networks, F-21 to F-22, **F-44**
 - basic considerations, F-29 to F-30
 - fault tolerance, F-67
 - network performance and cost, **F-40**
 - network performance effects, F-40 to F-44
 - rings, **F-36**
 - routing/arbitration/switching impact, F-52
 - system area network history, F-100 to F-101
- Torus networks
 - characteristics, F-36
 - commercial interconnection networks, F-63
 - direct networks, **F-37**
 - fault tolerance, F-67
 - IBM Blue Gene/L, F-72 to F-74
 - NEWS communication, F-43
 - routing comparison, **F-54**
 - system area network history, F-102
- TOS, *see* Top Of Stack (TOS) register
- Total Cost of Ownership (TCO), WSC case study, 476–479
- Total store ordering, relaxed consistency models, 395
- Tournament predictors
 - early schemes, L-27 to L-28
 - ILP for realizable processors, 216
 - local/global predictor combinations, 164–166
- Toy programs, performance benchmarks, 37
- TP, *see* Transaction-processing (TP)
- TPC, *see* Transaction Processing Council (TPC)
- Trace compaction, basic process, H-19
- Trace scheduling
 - basic approach, H-19 to H-21
 - overview, **H-20**
- Trace selection, definition, H-19
- Tradebeans benchmark, SMT on superscalar processors, 230
- Traffic intensity, queuing theory, D-25

- Trailer
 - messages, F-6
 - packet format, **F-7**
- Transaction components, D-16, **D-17**, I-38 to I-39
- Transaction-processing (TP)
 - server benchmarks, 41
 - storage system benchmarks, D-18 to D-19
- Transaction Processing Council (TPC)
 - benchmarks overview, D-18 to D-19, **D-19**
 - parallelism, 44
 - performance results reporting, 41
 - server benchmarks, 41
 - TPC-B, shared-memory workloads, 368
 - TPC-C
 - file system benchmarking, D-20
 - IBM eServer p5 processor, **409**
 - multiprocessing/
 - multithreading-based performance, 398
 - multiprocessor cost
 - effectiveness, 407
 - single vs. multiple thread executions, **228**
 - Sun T1 multithreading uncore performance, 227–229, **229**
 - WSC services, 441
 - TPC-D, shared-memory workloads, 368–369
 - TPC-E, shared-memory workloads, 368–369
 - Transfers, *see also* Data transfers
 - as early control flow instruction definition, A-16
 - Transforms, DSP, E-5
 - Transient failure, commercial interconnection networks, F-66
 - Transient faults, storage systems, D-11
 - Transistors
 - clock rate considerations, 244
 - dependability, 33–36
 - energy and power, 23–26
 - ILP, 245
 - performance scaling, 19–21
 - processor comparisons, 324
 - processor trends, 2
 - RISC instructions, A-3
 - shrinking, 55
 - static power, 26
 - technology trends, 17–18
 - Translation buffer (TB)
 - virtual memory block
 - identification, B-45
 - virtual memory fast address translation, B-46
 - Translation lookaside buffer (TLB)
 - address translation, B-39
 - AMD64 paged virtual memory, B-56 to B-57
 - ARM Cortex-A8, 114–115
 - cache optimization, 80, B-37
 - coining of term, L-9
 - Intel Core i7, **118**, 120–121
 - interconnection network
 - protection, F-86
 - memory hierarchy, B-48 to B-49
 - memory hierarchy basics, 78
 - MIPS64 instructions, K-27
 - Opteron, **B-47**
 - Opteron memory hierarchy, **B-57**
 - RISC code size, A-23
 - shared-memory workloads, 369–370
 - speculation advantages/
 - disadvantages, 210–211
 - strided access interactions, 323
 - Virtual Machines, 110
 - virtual memory block
 - identification, B-45
 - virtual memory fast address translation, B-46
 - virtual memory page size selection, B-47
 - virtual memory protection, 106–107
 - Transmission Control Protocol (TCP),
 - congestion management, F-65
 - Transmission Control Protocol/
 Internet Protocol (TCP/
 IP)
 - ATM, F-79
 - headers, **F-84**
 - internetworking, F-81, F-83 to F-84, F-89
 - reliance on, F-95
 - WAN history, F-98
 - Transmission speed, interconnection
 - network performance, F-13
 - Transmission time
 - communication latency, I-3 to I-4
 - time of flight, F-13 to F-14
 - Transport latency
 - time of flight, F-14
 - topology, F-35 to F-36
 - Transport layer, definition, **F-82**
 - Transputer, F-100
 - Tree-based barrier, large-scale
 - multiprocessor synchronization, **I-19**
 - Tree height reduction, definition, H-11
 - Trees, MINs with nonblocking, F-34
 - Trellis codes, definition, E-7
 - TRIPS Edge processor, F-63
 - characteristics, **F-73**
 - Trojan horses
 - definition, B-51
 - segmented virtual memory, B-53
 - True dependence
 - finding, H-7 to H-8
 - loop-level parallelism calculations, 320
 - vs. name dependence, 153
 - True sharing misses
 - commercial workloads, 371, 373
 - definition, 366–367
 - multiprogramming workloads, 377
 - True speedup, multiprocessor
 - performance, 406
 - TSMC, Stratton, F-3
 - TSS operating system, L-9
 - Turbo mode
 - hardware enhancements, 56
 - microprocessors, 26
 - Turing, Alan, L-4, L-19
 - Turn Model routing algorithm,
 - example calculations, F-47 to F-48
 - Two-level branch predictors
 - branch costs, 163
 - Intel Core i7, 166
 - tournament predictors, 165
 - Two-level cache hierarchy
 - cache optimization, B-31
 - ILP, 245
 - Two's complement, J-7 to J-8
 - Two-way conflict misses, definition, B-23

- Two-way set associativity
 - ARM Cortex-A8, 233
 - cache block placement, B-7, **B-8**
 - cache miss rates, **B-24**
 - cache miss rates vs. size, B-33
 - cache optimization, B-38
 - cache organization calculations, B-19 to B-20
 - commercial workload, 370–373, **371**
 - multiprogramming workload, 374–375
 - nonblocking cache, 84
 - Opteron data cache, B-13 to B-14
 - 2:1 cache rule of thumb, B-29
 - virtual to cache access scenario, **B-39**
- TX-2, L-34, L-49
- “Typical” program, instruction set considerations, A-43
- U**
- U, *see* Rack units (U)
- Ultron, DECstation 5000 reboots, **F-69**
- UMA, *see* Uniform memory access (UMA)
- Unbiased exponent, J-15
- Uncached state, directory-based cache coherence protocol
 - basics, 380, 384–386
- Unconditional branches
 - branch folding, 206
 - branch-prediction schemes, C-25 to C-26
 - VAX, K-71
- Underflow
 - floating-point arithmetic, J-36 to J-37, J-62
 - gradual, J-15
- Unicasting, shared-media networks, F-24
- Unicode character
 - MIPS data types, A-34
 - operand sizes/types, 12
 - popularity, A-14
- Unified cache
 - AMD Opteron example, **B-15**
 - performance, B-16 to B-17
- Uniform memory access (UMA)
 - multicore single-chip multiprocessor, **364**
 - SMP, 346–348
- Uninterruptible instruction
 - hardware primitives, 388
 - synchronization, 386
- Uninterruptible power supply (UPS)
 - Google WSC, 467
 - WSC calculations, 435
 - WSC infrastructure, 447
- Uniprocessors
 - cache protocols, 359
 - development views, 344
 - linear speedups, 407
 - memory hierarchy design, 73
 - memory system coherency, 353, **358**
 - misses, 371, 373
 - multiprogramming workload, 376–377
 - multithreading
 - basic considerations, 223–226
 - fine-grained on T1, 226–229
 - simultaneous, on superscalars, 230–232
 - parallel vs. sequential programs, 405–406
 - processor performance trends, 3–4, 344
 - SISD, 10
 - software development, 407–408
- Unit stride addressing
 - gather-scatter, 280
 - GPU vs. MIMD with Multimedia SIMD, 327
 - GPUs vs. vector architectures, 310
 - multimedia instruction compiler support, A-31
 - NVIDIA GPU ISA, 300
 - Roofline model, **287**
- UNIVAC I, L-5
- UNIX systems
 - architecture costs, 2
 - block servers vs. filers, D-35
 - cache optimization, B-38
 - floating point remainder, J-32
 - miss statistics, **B-59**
 - multiprocessor software
 - development, 408
 - multiprogramming workload, 374
 - seek distance comparison, **D-47**
 - vector processor history, G-26
- Unpacked decimal, A-14, J-16
- Unshielded twisted pair (UTP), LAN history, F-99
- Up*/down* routing
 - definition, F-48
 - fault tolerance, F-67
- UPS, *see* Uninterruptible power supply (UPS)
- USB, Sony PlayStation 2 Emotion Engine case study, E-15
- Use bit
 - address translation, B-46
 - segmented virtual memory, B-52
 - virtual memory block replacement, B-45
- User-level communication, definition, F-8
- User maskable events, definition, C-45 to C-46
- User nonmaskable events, definition, C-45
- User-requested events, exception requirements, C-45
- Utility computing, 455–461, L-73 to L-74
- Utilization
 - I/O system calculations, D-26
 - queuing theory, D-25
- UTP, *see* Unshielded twisted pair (UTP)
- V**
- Valid bit
 - address translation, B-46
 - block identification, B-7
 - Opteron data cache, B-14
 - paged virtual memory, B-56
 - segmented virtual memory, B-52
 - snooping, 357
 - symmetric shared-memory multiprocessors, 366
- Value prediction
 - definition, 202
 - hardware-based speculation, 192
 - ILP, 212–213, 220
 - speculation, 208
- VAPI, InfiniBand, F-77
- Variable length encoding
 - control flow instruction branches, A-18
 - instruction sets, **A-22**
 - ISAs, 14
- Variables
 - and compiler technology, A-27 to A-29

Variables (*continued*)

- CUDA, 289
- Fermi GPU, 306
- ISA, A-5, **A-12**
- locks via coherence, 389
- loop-level parallelism, 316
- memory consistency, 392
- NVIDIA GPU Memory, 304–305
- procedure invocation options,
 - A-19
- random, distribution, D-26 to D-34
- register allocation, A-26 to A-27
- in registers, A-5
- synchronization, 375
- TLP programmer's viewpoint, 394

VCs, *see* Virtual channels (VCs)

Vector architectures

- computer development, L-44 to L-49
- definition, 9
- DLP
 - basic considerations, 264
 - definition terms, **309**
 - gather/scatter operations, 279–280
 - multidimensional arrays, 278–279
 - multiple lanes, 271–273
 - programming, 280–282
 - vector execution time, 268–271
 - vector-length registers, 274–275
 - vector load/store unit
 - bandwidth, 276–277
 - vector-mask registers, 275–276
 - vector processor example, 267–268
 - VMIPS, 264–267
- GPU conditional branching, 303
- vs. GPUs, 308–312
- mapping examples, **293**
- memory systems, G-9 to G-11
- multimedia instruction compiler support, A-31
- vs. Multimedia SIMD Extensions, 282
- peak performance vs. start-up overhead, 331
- power/DLP issues, 322
- vs. scalar performance, 331–332
- start-up latency and dead time, **G-8**
- strided access-TLB interactions, 323

- vector-register characteristics, **G-3**

Vector Functional Unit

- vector add instruction, 272–273
- vector execution time, 269
- vector sequence chimes, 270
- VMIPS, 264

Vector Instruction

- definition, **292, 309**
- DLP, 322
- Fermi GPU, 305
- gather-scatter, 280
- instruction-level parallelism, 150
- mask registers, 275–276
- Multimedia SIMD Extensions, 282
- multiple lanes, 271–273
- Thread of Vector Instructions, **292**
- vector execution time, 269
- vector vs. GPU, 308, 311
- vector processor example, 268
- VMIPS, 265–267, **266**

Vectorizable Loop

- characteristics, 268
- definition, 268, **292, 313**
- Grid mapping, **293**
- Livermore Fortran kernel
 - performance, 331
- mapping example, **293**
- NVIDIA GPU computational structures, 291

Vectorized code

- multimedia compiler support, A-31
- vector architecture programming, 280–282
- vector execution time, 271
- VMIPS, 268

Vectorized Loop, *see also* Body of Vectorized Loop

- definition, **309**
- GPU Memory structure, **304**
- vs. Grid, 291, 308
- mask registers, 275
- NVIDIA GPU, 295
- vector vs. GPU, 308

Vectorizing compilers

- effectiveness, G-14 to G-15
- FORTTRAN test kernels, **G-15**
- sparse matrices, G-12 to G-13

Vector Lane Registers, definition, **292**

Vector Lanes

- control processor, 311
- definition, **292, 309**
- SIMD Processor, 296–297, **297**

Vector-length register (VLR)

- basic operation, 274–275
- performance, G-5
- VMIPS, 267

Vector load/store unit

- memory banks, 276–277
- VMIPS, 265

Vector loops

- NVIDIA GPU, 294
- processor example, 267
- strip-mining, 303
- vector vs. GPU, 311
- vector-length registers, 274–275
- vector-mask registers, 275–276

Vector-mask control, characteristics, 275–276

Vector-mask registers

- basic operation, 275–276
- Cray X1, G-21 to G-22
- VMIPS, 267

Vector Processor

- caches, 305
- compiler vectorization, **281**
- Cray X1
 - MSP modules, **G-22**
 - overview, G-21 to G-23
- Cray X1E, G-24
- definition, **292, 309**
- DLP processors, 322
- DSP media extensions, E-10
- example, 267–268
- execution time, **G-7**
- functional units, **272**
- gather-scatter, 280
- vs. GPUs, 276
- historical background, G-26
- loop-level parallelism, 150
- loop unrolling, 196
- measures, G-15 to G-16
- memory banks, 277
- and multiple lanes, **273, 310**
- multiprocessor architecture, 346
- NVIDIA GPU computational structures, 291
- overview, G-25 to G-26
- peak performance focus, 331
- performance, G-2 to G-7
 - start-up and multiple lanes, G-7 to G-9
- performance comparison, **58**
- performance enhancement chaining, G-11 to G-12

- DAXPY on VMIPS, G-19 to G-21
- sparse matrices, G-12 to G-14
- PTX, 301
- Roofline model, 286–287, **287**
- vs. scalar processor, 311, 331, 333, G-19
- vs. SIMD Processor, 294–296
- Sony PlayStation 2 Emotion Engine, E-17 to E-18
- start-up overhead, **G-4**
- stride, 278
- strip mining, **275**
- vector execution time, 269–271
- vector/GPU comparison, 308
- vector kernel implementation, 334–336
- VMIPS, 264–265
- VMIPS on DAXPY, G-17
- VMIPS on Linpack, G-17 to G-19
- Vector Registers
 - definition, **309**
 - execution time, 269, 271
 - gather-scatter, 280
 - multimedia compiler support, A-31
 - Multimedia SIMD Extensions, 282
 - multiple lanes, 271–273
 - NVIDIA GPU, 297
 - NVIDIA GPU ISA, 298
 - performance/bandwidth trade-offs, 332
 - processor example, 267
 - strides, 278–279
 - vector vs. GPU, 308, 311
 - VMIPS, 264–267, **266**
- Very-large-scale integration (VLSI)
 - early computer arithmetic, J-63
 - interconnection network topology, F-29
 - RISC history, L-20
 - Wallace tree, J-53
- Very Long Instruction Word (VLIW)
 - clock rates, 244
 - compiler scheduling, L-31
 - EPIC, L-32
 - IA-64, H-33 to H-34
 - ILP, 193–196
 - loop-level parallelism, 315
 - M32R, K-39 to K-40
 - multiple-issue processors, **194**, L-28 to L-30
 - multithreading history, L-34
 - sample code, **252**
 - TI 320C6x DSP, E-8 to E-10
- VGA controller, L-51
- Video
 - Amazon Web Services, 460
 - application trends, 4
 - PMDs, 6
 - WSCs, 8, 432, 437, 439
- Video games, multimedia support, K-17
- VI interface, L-73
- Virtual address
 - address translation, B-46
 - AMD64 paged virtual memory, **B-55**
 - AMD Opteron data cache, B-12 to B-13
 - ARM Cortex-A8, **115**
 - cache optimization, B-36 to B-39
 - GPU conditional branching, 303
 - Intel Core i7, 120
 - mapping to physical, B-45
 - memory hierarchy, **B-39, B-48**, B-48 to B-49
 - memory hierarchy basics, 77–78
 - miss rate vs. cache size, **B-37**
 - Opteron mapping, **B-55**
 - Opteron memory management, B-55 to B-56
 - and page size, B-58
 - page table-based mapping, **B-45**
 - translation, B-36 to B-39
 - virtual memory, B-42, B-49
- Virtual address space
 - example, **B-41**
 - main memory block, B-44
- Virtual caches
 - definition, B-36 to B-37
 - issues with, B-38
- Virtual channels (VCs), F-47
 - HOL blocking, **F-59**
 - Intel SCCC, F-70
 - routing comparison, **F-54**
 - switching, F-51 to F-52
 - switch microarchitecture
 - pipelining, F-61
 - system area network history, F-101
 - and throughput, F-93
- Virtual cut-through switching, F-51
- Virtual functions, control flow instructions, A-18
- Virtualizable architecture
 - Intel 80x86 issues, **128**
 - system call performance, **141**
 - Virtual Machines support, 109
 - VMM implementation, 128–129
- Virtualizable GPUs, future technology, 333
- Virtual machine monitor (VMM)
 - characteristics, 108
 - nonvirtualizable ISA, 126, 128–129
 - requirements, 108–109
 - Virtual Machines ISA support, 109–110
 - Xen VM, 111
- Virtual Machines (VMs)
 - Amazon Web Services, 456–457
 - cloud computing costs, 471
 - early IBM work, L-10
 - ISA support, 109–110
 - protection, 107–108
 - protection and ISA, 112
 - server benchmarks, 40
 - and virtual memory and I/O, 110–111
 - WSCs, 436
 - Xen VM, 111
- Virtual memory
 - basic considerations, B-40 to B-44, B-48 to B-49
 - basic questions, B-44 to B-46
 - block identification, B-44 to B-45
 - block placement, B-44
 - block replacement, B-45
 - vs. caches, B-42 to B-43
 - classes, B-43
 - definition, B-3
 - fast address translation, B-46
 - Multimedia SIMD Extensions, 284
 - multithreading, 224
 - paged example, B-54 to B-57
 - page size selection, B-46 to B-47
 - parameter ranges, **B-42**
 - Pentium vs. Opteron protection, B-57
 - protection, 105–107
 - segmented example, B-51 to B-54
 - strided access-TLB interactions, 323
 - terminology, B-42
 - Virtual Machines impact, 110–111
 - writes, B-45 to B-46
- Virtual methods, control flow instructions, A-18

- Virtual output queues (VOQs), switch microarchitecture, F-60
 - VLIW, *see* Very Long Instruction Word (VLIW)
 - VLR, *see* Vector-length register (VLR)
 - VLSI, *see* Very-large-scale integration (VLSI)
 - VMCS, *see* Virtual Machine Control State (VMCS)
 - VME rack
 - example, **D-38**
 - Internet Archive Cluster, D-37
 - VMIPS
 - basic structure, **265**
 - DAXPY, G-18 to G-20
 - DLP, 265–267
 - double-precision FP operations, **266**
 - enhanced, DAXPY performance, G-19 to G-21
 - gather/scatter operations, 280
 - ISA components, 264–265
 - multidimensional arrays, 278–279
 - Multimedia SIMD Extensions, 282
 - multiple lanes, 271–272
 - peak performance on DAXPY, G-17
 - performance, G-4
 - performance on Linpack, G-17 to G-19
 - sparse matrices, G-13
 - start-up penalties, **G-5**
 - vector execution time, 269–270, G-6 to G-7
 - vector vs. GPU, 308
 - vector-length registers, 274
 - vector load/store unit bandwidth, 276
 - vector performance measures, G-16
 - vector processor example, 267–268
 - VLR, 274
 - VMM, *see* Virtual machine monitor (VMM)
 - VMs, *see* Virtual Machines (VMs)
 - Voltage regulator controller (VRC), Intel SCCC, F-70
 - Voltage regulator modules (VRMs), WSC server energy efficiency, 462
 - Volume-cost relationship, components, 27–28
 - Von Neumann, John, L-2 to L-6
 - Von Neumann computer, L-3
 - Voodoo2, L-51
 - VOQs, *see* Virtual output queues (VOQs)
 - VRC, *see* Voltage regulator controller (VRC)
 - VRMs, *see* Voltage regulator modules (VRMs)
- W**
- Wafers
 - example, **31**
 - integrated circuit cost trends, 28–32
 - Wafer yield
 - chip costs, 32
 - definition, 30
 - Waiting line, definition, D-24
 - Wait time, shared-media networks, F-23
 - Wallace tree
 - example, J-53, **J-53**
 - historical background, J-63
 - Wall-clock time
 - execution time, 36
 - scientific applications on parallel processors, I-33
 - WANs, *see* Wide area networks (WANs)
 - WAR, *see* Write after read (WAR)
 - Warehouse-scale computers (WSCs)
 - Amazon Web Services, 456–461
 - basic concept, 432
 - characteristics, 8
 - cloud computing, 455–461
 - cloud computing providers, 471–472
 - cluster history, L-72 to L-73
 - computer architecture
 - array switch, 443
 - basic considerations, 441–442
 - memory hierarchy, **443**, 443–446, **444**
 - storage, 442–443
 - as computer class, **5**
 - computer cluster forerunners, 435–436
 - cost-performance, 472–473
 - costs, 452–455, **453–454**
 - definition, 345
 - and ECC memory, 473–474
 - efficiency measurement, 450–452
 - facility capital costs, 472
 - Flash memory, 474–475
 - Google
 - containers, 464–465
 - cooling and power, 465–468
 - monitoring and repairing, 469–470
 - PUE, **468**
 - server, **467**
 - servers, 468–469
 - MapReduce, 437–438
 - network as bottleneck, 461
 - physical infrastructure and costs, 446–450
 - power modes, 472
 - programming models and workloads, 436–441
 - query response-time curve, **482**
 - relaxed consistency, 439
 - resource allocation, 478–479
 - server energy efficiency, 462–464
 - vs. servers, 432–434
 - SPECPower benchmarks, **463**
 - switch hierarchy, 441–442, **442**
 - TCO case study, 476–478
- Warp, L-31
 - definition, **292**, **313**
 - terminology comparison, **314**
- Warp Scheduler
 - definition, **292**, **314**
 - Multithreaded SIMD Processor, 294
- Wavelength division multiplexing (WDM), WAN history, F-98
- WAW, *see* Write after write (WAW)
- Way prediction, cache optimization, 81–82
- Way selection, 82
- WB, *see* Write-back cycle (WB)
- WCET, *see* Worst-case execution time (WCET)
- WDM, *see* Wavelength division multiplexing (WDM)
- Weak ordering, relaxed consistency models, 395
- Weak scaling, Amdahl's law and parallel computers, 406–407

- Web index search, shared-memory workloads, 369
- Web servers
 - benchmarking, D-20 to D-21
 - dependability benchmarks, D-21
 - ILP for realizable processors, 218
 - performance benchmarks, 40
 - WAN history, F-98
- Weighted arithmetic mean time, D-27
- Weitek 3364
 - arithmetic functions, J-58 to J-61
 - chip comparison, **J-58**
 - chip layout, **J-60**
- West-first routing, F-47 to F-48
- Wet-bulb temperature
 - Google WSC, 466
 - WSC cooling systems, 449
- Whirlwind project, L-4
- Wide area networks (WANs)
 - ATM, F-79
 - characteristics, F-4
 - cross-company interoperability, F-64
 - effective bandwidth, F-18
 - fault tolerance, F-68
 - historical overview, F-97 to F-99
 - InfiniBand, F-74
 - interconnection network domain relationship, **F-4**
 - latency and effective bandwidth, F-26 to F-28
 - offload engines, F-8
 - packet latency, **F-13**, F-14 to F-16
 - routers/gateways, F-79
 - switches, F-29
 - switching, F-51
 - time of flight, F-13
 - topology, F-30
- Wilkes, Maurice, L-3
- Winchester, L-78
- Window
 - latency, B-21
 - processor performance calculations, 218
 - scoreboarding definition, C-78
 - TCP/IP headers, **F-84**
- Windowing, congestion management, F-65
- Window size
 - ILP limitations, 221
 - ILP for realizable processors, 216–217
 - vs. parallelism, **217**
- Windows operating systems, *see* Microsoft Windows
- Wireless networks
 - basic challenges, **E-21**
 - and cell phones, E-21 to E-22
- Wires
 - energy and power, 23
 - scaling, 19–21
- Within instruction exceptions
 - definition, C-45
 - instruction set complications, C-50
 - stopping/restarting execution, C-46
- Word count, definition, **B-53**
- Word displacement addressing, VAX, K-67
- Word offset, MIPS, C-32
- Words
 - aligned/misaligned addresses, **A-8**
 - AMD Opteron data cache, B-15
 - DSP, E-6
 - Intel 80x86, K-50
 - memory address interpretation, A-7 to A-8
 - MIPS data transfers, A-34
 - MIPS data types, A-34
 - MIPS unaligned reads, **K-26**
 - operand sizes/types, 12
 - as operand type, A-13 to A-14
 - VAX, K-70
- Working set effect, definition, I-24
- Workloads
 - execution time, 37
 - Google search, 439
 - Java and PARSEC without SMT, **403–404**
 - RAID performance prediction, D-57 to D-59
 - symmetric shared-memory multiprocessor performance, 367–374, I-21 to I-26
 - WSC goals/requirements, 433
 - WSC resource allocation case study, 478–479
 - WSCs, 436–441
- Wormhole switching, F-51, F-88
 - performance issues, F-92 to F-93
 - system area network history, F-101
- Worst-case execution time (WCET), definition, E-4
- Write after read (WAR)
 - data hazards, 153–154, 169
- dynamic scheduling with
 - Tomasulo's algorithm, 170–171
- hazards and forwarding, C-55
- ILP limitation studies, 220
- MIPS scoreboarding, C-72, C-74 to C-75, C-79
- multiple-issue processors, L-28
- register renaming vs. ROB, 208
- ROB, 192
- TI TMS320C55 DSP, E-8
- Tomasulo's advantages, 177–178
- Tomasulo's algorithm, 182–183
- Write after write (WAW)
 - data hazards, 153, 169
 - dynamic scheduling with
 - Tomasulo's algorithm, 170–171
- execution sequences, C-80
- hazards and forwarding, C-55 to C-58
- ILP limitation studies, 220
- microarchitectural techniques case study, 253
- MIPS FP pipeline performance, C-60 to C-61
- MIPS scoreboarding, C-74, C-79
- multiple-issue processors, L-28
- register renaming vs. ROB, 208
- ROB, 192
- Tomasulo's advantages, 177–178
- Write allocate
 - AMD Opteron data cache, B-12
 - definition, B-11
 - example calculation, B-12
- Write-back cache
 - AMD Opteron example, B-12, B-14
 - coherence maintenance, **381**
 - coherency, 359
 - definition, B-11
 - directory-based cache coherence, 383, 386
 - Flash memory, 474
 - FP register file, **C-56**
 - invalidate protocols, 355–357, **360**
 - memory hierarchy basics, 75
 - snooping coherence, **355**, 356–357, 359
- Write-back cycle (WB)
 - basic MIPS pipeline, C-36
 - data hazard stall minimization, C-17

- Write-back cycle (*continued*)
 - execution sequences, C-80
 - hazards and forwarding, C-55 to C-56
 - MIPS exceptions, C-49
 - MIPS pipeline, **C-52**
 - MIPS pipeline control, C-39
 - MIPS R4000, C-63, C-65
 - MIPS scoreboarding, C-74
 - pipeline branch issues, **C-40**
 - RISC classic pipeline, C-7 to C-8, C-10
 - simple MIPS implementation, C-33
 - simple RISC implementation, C-6
 - Write broadcast protocol, definition, 356
 - Write buffer
 - AMD Opteron data cache, B-14
 - Intel Core i7, **118**, 121
 - invalidate protocol, 356
 - memory consistency, 393
 - memory hierarchy basics, 75
 - miss penalty reduction, 87, B-32, B-35 to B-36
 - write merging example, **88**
 - write strategy, B-11
 - Write hit
 - cache coherence, **358**
 - directory-based coherence, 424
 - single-chip multicore multiprocessor, 414
 - snooping coherence, 359
 - write process, B-11
 - Write invalidate protocol
 - directory-based cache coherence protocol example, 382–383
 - example, 359, **360**
 - implementation, 356–357
 - snooping coherence, 355–356
 - Write merging
 - example, **88**
 - miss penalty reduction, 87
 - Write miss
 - AMD Opteron data cache, B-12, B-14
 - cache coherence, **358**, **359**, **360**, 361
 - definition, 385
 - directory-based cache coherence, 380–383, 385–386
 - example calculation, B-12
 - locks via coherence, 390
 - memory hierarchy basics, 76–77
 - memory stall clock cycles, B-4
 - Opteron data cache, B-12, B-14
 - snooping cache coherence, 365
 - write process, B-11 to B-12
 - write speed calculations, 393
 - Write result stage
 - data hazards, 154
 - dynamic scheduling, 174–175
 - hardware-based speculation, 192
 - instruction steps, 175
 - ROB instruction, 186
 - scoreboarding, C-74 to C-75, C-78 to C-80
 - status table examples, C-77
 - Tomasulo's algorithm, 178, **180**, 190
 - Write serialization
 - hardware primitives, 387
 - multiprocessor cache coherence, 353
 - snooping coherence, 356
 - Write stall, definition, B-11
 - Write strategy
 - memory hierarchy considerations, B-6, B-10 to B-12
 - virtual memory, B-45 to B-46
 - Write-through cache
 - average memory access time, B-16
 - coherency, **352**
 - invalidate protocol, 356
 - memory hierarchy basics, 74–75
 - miss penalties, B-32
 - optimization, B-35
 - snooping coherence, 359
 - write process, B-11 to B-12
 - Write update protocol, definition, 356
 - WSCs, *see* Warehouse-scale computers (WSCs)
- ## X
- XBox, L-51
 - Xen Virtual Machine
 - Amazon Web Services, 456–457
 - characteristics, 111
 - Xerox Palo Alto Research Center, LAN history, F-99
 - XIMD architecture, L-34
 - Xon/Xoff, interconnection networks, F-10, F-17
- ## Y
- Yahoo!, WSCs, 465
 - Yield
 - chip fabrication, 61–62
 - cost trends, 27–32
 - Fermi GTX 480, 324
- ## Z
- Z-80 microcontroller, cell phones, E-24
 - Zero condition code, MIPS core, K-9 to K-16
 - Zero-copy protocols
 - definition, F-8
 - message copying issues, F-91
 - Zero-load latency, Intel SCCC, F-70
 - Zuse, Konrad, L-4 to L-5
 - Zynga, FarmVille, 460