

## Index

### A

ACPI (Advanced Configuration & Power Interface), 342

activity list, 166

address prediction. *See* prefetching

address translation. *See* TLB

Alewife, 307

aliasing in branch prediction, 136, 151  
destructive, 152

neutral, 152

Alpha 21064, 78, 83

Alpha 21164, 78–83, 170, 229

branch misprediction penalty, 83

branch predictor, 80

front-end, 80–82

load-store instruction, 82

scoreboard, 81, 184

slotting, 80

Alpha 21264, 136, 144, 170, 202

branch predictor, 157–158

memory dependence prediction, 190

Alpha 21364, 310

AMD Athlon, 164, 165, 182, 234

AMD K5, 136

AMD K7, 165

AMD Opteron, 182

Amdahl's law, 11, 262, 296

arithmetic mean, 16

associative memory, 50

atomic operation, 283

average memory access time, 56

### B

back-end, 75

bandwidth, 245

barrier synchronization, 289

centralized, 290

sense reversal, 290

Belady's algorithm, 67

benchmarks, 12–14

Dhrystone, 13

kernels, 13

Linpack, 13

Livermore loops, 13

SPEC (System Performance Evaluation Corporation), 13

Whetstone, 13

BHT, 142, 144, 147

bit

bias, 152

dirty (in cache), 54

dirty (in paging systems), 61

full-empty, 287

invalid (in register), 315

poison, 114

predecoded, 80, 164

ready, 93, 183

temporal, 236

valid (in cache), 55

valid (in paging systems), 61

bogus address, 146

bogus value, 315

Boolean vector, 275

BPB. *See* Branch Prediction Buffer

brainiacs, 77

Branch History Table. *See* BHT

branch instruction frequencies,

131

branch misprediction, 166

misfetch, 148

penalty, 149

branch misprediction repair, 144

for global predictor, 144

for local history registers, 145

speculative, 143

branch prediction

correlated, 138

global schemes, 138

local schemes, 138

not taken, 40

static, 132

Branch Prediction Buffer, 136

- branch predictor
  - agree, 152
  - anatomy of, 130–131, 196
  - bimodal, 136, 138
  - global register, 140
  - gshare, 140–142, 153
  - hybrid, 155
  - in Alpha 21264. *See* Alpha 21264
  - in IBM Power4, 241
  - indirect, 148
  - loop predictor, 151
  - skewed, 153
  - tournament, 155
  - two-level, 139, 142
  - YAGS, 154
- branch predictor update
  - bimodal predictor, 137
  - global register, 139
  - PHT, 139
- Branch Target Buffer. *See* BTB
- BTB, 105, 146
  - combined BTB-BHT, 147
  - combined BTB-PHT, 137
  - decoupled BTB-PHT, 146
  - integrated BTB-PHT, 146, 148
  - integrated with instruction cache, 160
- bubble, 37
- bus bridge, 55
- busy-waiting, 283
- Butterfly network, 267
- bypassing, 120, 177
- C**
- cache access, 209
  - with physical index and physical tags, 210
  - with TLB access in parallel, 210, 211
  - with virtual index and physical tags, 212
  - with virtual index and virtual tags, 211
- cache block. *cache line*
- cache coherence, 269
  - 2-bit protocol, 279
  - 3-state basic snoopy protocol, 271–273
  - directory protocols, 270, 322
  - Dir<sub>i</sub>B protocols, 279
  - Dir<sub>i</sub>NB protocols, 280
  - full directory protocol, 275–278
  - home node, 276
  - MESI protocol, 274–275
  - MOSI protocol, 275
  - remote node, 276
  - snoopy protocols, 270
  - write-invalidate protocol, 270
  - write-update protocol, 270
- cache hit detection, 52, 237
  - in direct-mapped cache, 51
  - in fully-associative cache, 50
  - in set-associative cache, 51
  - MRU-based, 237
  - two step, 237
- cache miss
  - 3C, 54
  - capacity, 55
  - coherence, 55, 275
  - cold. *See* compulsory
  - compulsory, 54
  - conflict, 55
- cache organization
  - column-associative. *See* column-associative cache
  - direct-mapped, 51
  - fully-associative, 50
  - generic, 49
  - geometry, 52
  - hierarchy, 208
  - L1 D-cache, 208
  - L1 I-cache, 208
  - L2 unified, 208
  - L3, 208, 240
  - lock-up free. *See* lock-up free cache
  - non-blocking. *See* lock-up free
  - non-uniform access, 240
  - sector. *See* sector cache
  - set-associative, 51
  - skewed-associative, 215
  - victim. *See* victim cache
- cache performance
  - associativity, impact of, 57, 208
  - capacity, impact of, 57, 208
  - hit rate, 56
  - instruction bandwidth, 159
  - line size, impact of, 58, 159
  - local hit rate (L2), 56
  - miss rate, 56
  - size, impact of. *See* associativity
- cache pollution, 219, 223
- cache tiling, 218
- cache write strategies, 53–54
  - copy-back. *See* write-back
  - store-through. *See* write-through
  - write-allocate, 54
  - write-around, 54
  - write-back, 54
  - write-through, 54
- CDC 6600, 85–88
  - PPUs (Peripheral Processor Units), 305
  - scoreboard, 86–87
- chip multiprocessing. *See* CMP
- CISC, 44, 103, 122, 163
- cluster, 201
- CMP, 111, 241, 260
  - challenges for, 346–348
  - characteristics of general-purpose, 319
  - design parameters, 346
- coarse-grained multithreading, 306, 309
- code reordering, 217
  - basic block reordering, 217
  - procedure reordering, 217
- collapsing buffer, 160

- colliding load, 194  
column access strobe (CAS), 247  
column-associative cache, 213, 215  
commit. *See* pipeline stage commit  
common data bus (CDB), 96, 98  
compare-and-swap, 283  
Complex Instruction Set Computer. *See* CISC  
conditional move, 81, 113  
conflicting load, 194  
Connection Machine CM-1, 262  
Connection Machine CM-2, 269  
content-addressable memory (CAM). *See*  
    associative memory  
context (in multithreading), 307  
context-switch, 62  
control flow (in interconnection networks), 269  
Convex Exemplar, 267  
CPI (cycles per instruction), 7  
Cray 3TD, 269  
critical instruction, 198  
critical section, 281, 290  
critical word first, 231  
cross-bar, 267, 320  
cycle time, 6, 9, 75
- D**  
dance-hall architecture, 263, 320  
data dependencies, 35  
    name, 89  
    output (WAW), 35  
    read after write (RAW), 35, 184  
    write after read (WAR), 35  
data-flow machine, 318  
DDR SDRAM (double data rate), 248  
dead lines, 227  
deadlock, 280, 281, 290  
Denelcor HEP, 288  
DIMM (Dual Inline Memory Modules), 248  
Direct Rambus, 251  
dispatch. *See* pipeline stage dispatch  
distributed shared-memory. *See* NUMA  
DMA (direct memory access), 244, 320, 326  
DRAM, 246  
    access time, 47, 247  
    cycle time, 247  
    precharge, 247  
drowsy caches, 341  
dynamic memory disambiguation, 188, 317  
dynamic random access memory. *See* DRAM  
dynamic voltage and frequency scaling (DVFS),  
    339–340
- E**  
EDO DRAM, 247  
efficiency, 12  
embedded system, 324  
EPIC, 111, 303  
error-correcting codes (ECC), 252  
ESDRAM, 247
- exceptions, 40, 166  
    precise, 41  
exchange-and-swap, 283  
exclusive caching. *See* multilevel exclusion  
execution time, 6, 14, 75, 339, 345  
Explicitly Parallel Instruction Computing. *See*  
    EPIC
- F**  
false sharing, 281  
fast page mode, 247  
fence, 293  
fetch-and-add, 283, 290  
fetch-and- $\Theta$ , 283  
fine-grained multithreading, 305, 309, 320  
floating-point, 44  
    addition, 45  
    biased exponent, 45  
    IEEE standard representation, 45  
    mantissa (normalized), 45  
    multiplication, 46  
    Not a Number, 45  
Flynn's taxonomy, 261  
FO4, 343  
forwarding (in pipeline), 36, 321  
fragmentation, 61  
front-end, 75  
functional unit, 44, 76, 178
- G**  
geometric mean, 17
- H**  
Hamming codes, 252  
hardware scouting, 315  
harmonic mean, 16  
hazards  
    control, 35, 196  
    data, 35, 196  
    structural, 35, 196  
hit under miss, 229  
hot spot (in directory protocols), 285  
hot spot (in interconnection networks),  
    290  
hot spot (in thermal management), 336  
HP 7100, 216  
HP 7200, 216  
HP PA 1700, 229  
HP PA-RISC, 294  
hypercube, 268, 269  
hypertexting, 110, 310
- I**  
I/O bus, 55  
IBM Cell Multiprocessor, 324–327  
IBM Power4, 239, 267, 280  
    cache hierarchy, 241–244  
IBM Power5, 241, 267, 280, 310  
IBM Power6, 241

- IBM PowerPC, 170, 178, 284
    - multithreaded variation, 308
  - IBM PowerPC, 137, 147, 213, 266
  - IBM SP, 267
  - IBM System/360 Model 85, 239
  - IBM System/360 Model 91, 95–96
  - if-conversion, 113
  - Illiac IV, 262
  - ILP, 76, 111, 303
  - indirect branch, 131
  - instruction buffer, 80, 95
    - prefetch, 159
  - instruction execution cycle, 2
  - instruction set architecture, 1, 9, 19, 75, 164
  - instruction window, 93, 178
  - instruction-level parallelism. *See* ILP
  - Intel Centrino Duo, 342
  - Intel Core, 102, 103, 111, 164, 165, 170, 182, 191, 310
    - branch misprediction penalty, 77
  - Intel Core Duo, 322
    - snoopy protocol, 324
  - Intel Core2, 164
  - Intel IA-32 architecture, 102, 283
  - Intel IA-64 architecture, 115, 284
  - Intel Itanium, 115–118, 293
    - bundles, 115
  - Intel IXP 2800, 329–330
  - Intel P6 microarchitecture, 103–109, 169
    - AGU, 184
    - back-end, 106–109
    - branch misprediction, 107
    - BTB-PHT, 147
    - decode, 103, 105, 164
    - instruction buffer, 105
    - instruction window, 178
    - load speculation, 109
    - memory dependence prediction, 189
    - Memory Order Buffer, 109
    - microinstruction sequencer, 105, 111
    - Register Alias Table (RAT), 106
    - μops, 103
  - Intel Pentium, 103
  - Intel Pentium 4, 102, 103, 106, 110, 170, 182, 310
    - branch misprediction penalty, 77
    - trace cache, 163
  - Intel Pentium II, 102, 103
  - Intel Pentium III, 102, 103, 106, 110, 142, 170, 182
    - branch misprediction penalty, 77
  - Intel Pentium M, 103, 110, 151, 170, 342
  - Intel Pentium Pro, 103, 109
  - interconnection network, 264
    - centralized, 264
    - decentralized, 264
    - direct, 267
    - indirect, 267
    - multistage, 267
  - interleaving, 248
  - interrupts, 40
  - inverted index structure, 238
  - IPC, 7, 75
    - defined as instructions per cycle, 7
  - ISA. *See* instruction set architecture
  - issue. *See* pipeline stage issue
  - issue slot, 149
- L**
- latency, 7, 30
    - inter-branch instructions, 131
    - main memory, 208, 245
    - variable for large caches, 240
  - line and way prediction, 160
  - load bypassing, 187
  - load forwarding, 187
  - load predictor table, 190
  - load speculation, 108, 185, 194
  - load, non-binding, 220
  - load-locked, 284
  - load-store architecture, 31
  - load-store window, 187
  - locality
    - principle of, 48
    - spatial, 48
    - temporal, 48
    - value, 197
  - lock, 282
    - acquire, 283
    - contention, 285
    - release, 283
  - lock-up free cache, 108, 209, 220, 229–231, 273, 307
  - look-ahead program counter, 225
  - loop stream mode, 164
  - loop unrolling, 221
- M**
- main memory. *See* latency, *See* memory bank, *See* memory bandwidth
  - Markov-based prefetching, 226
  - memory bandwidth
    - improvements, 248
  - memory bank, 248
  - memory bus, 54, 55, 220
  - memory controller, 240, 245
  - memory dependence prediction, 185, 189
  - memory hierarchy. *See* cache, *See* main memory
  - memory wall, 47, 208
  - mesh, 267
  - MFLOPS (Millions of Floating-point Operations per Second), 9
  - microprogramming, 111
    - horizontal, 111
    - vertical, 111
  - MIMD, 261, 262
  - MIPS (Millions of Instructions Per Second), 8
  - MIPS R10000, 137, 165, 170, 178
  - MIPS R6000, 211
  - missing status holding register. *See* MSHR

## Index

365

- MLI, 232, 240, 280  
   for single processors, 232–234  
   partial MLI, 281, 322
- MMX instructions, 102, 294, 295
- Moore's law, 4, 23, 77, 111, 245, 260, 319
- MRU policy (Most Recently Used), 210
- MSHR, 229–230
- multicores, 319  
   cache hierarchy, 323  
   monolithic, 322
- multilevel exclusion, 234
- multilevel inclusion property. *See* MLI
- multimedia data characteristics, 294
- multimedia instructions, 294
- Multiple Instruction Multiple Data. *See* MIMD
- multiprocessing, 260, 303  
   performance improvement with, 314
- multiprogramming, 60, 303
- multithreading, 78, 303, 304  
   performance improvement with, 314  
   speculative, 317
- N**
- Netburst. *See* Intel Pentium,
- network processor, 328
- non-uniform memory access. *See* NUMA
- NUMA, 263, 267, 276
- O**
- OBL. *See* sequential prefetching
- out-of-order. *See* superscalar
- P**
- page coloring, 211
- page fault, 62
- page table, 61  
   inverted, 238
- page table entry (PTE), 61
- page, in virtual memory, 60
- Pal (Privileged Access Library), 64
- parallel processing, 260, 303
- parity checking, 252
- Pattern History Table. *See* PHT
- performance, 9
- phases, 21
- PHT, 136, 139
- PID, 64, 307
- pipeline  
   balanced, 30  
   buffering, 31  
   depth, 75  
   drain, 83  
   flush, 31  
   optimal depth, 345  
   registers, 32, 76  
   width, 75
- pipeline stage, 30  
   address generation (AG), 42  
   broadcast, 182  
   commit, 76, 91  
   dispatch, 94, 180  
   execution (EX), 32  
   instruction decode (ID), 32  
   instruction fetch (IF), 32  
   issue, 81, 93, 94, 180  
   memory access (Mem), 32  
   renaming, 90, 91  
   select, 182  
   thread selection, 321  
   wakeup, 182  
   write back (WB), 32
- pipelining, 303  
   defined, 30
- power consumption. *See* power dissipation
- power dissipation, 5, 110, 336  
   dynamic, 5, 336  
   leakage, 6, 337  
   static, 6
- power management, 336
- predication, 113, 156
- prefetch buffer, 219, 227, 243
- prefetch instructions, 220
- prefetching, 218–219, 229, 273  
   accuracy, 219  
   address correlation. *See* Markov-based  
     prefetching  
   coverage, 219  
   for pointer-based structures, 226  
   scheduled region, 228  
   sequential. *See* sequential prefetching  
   state-less. *See* state-less prefetching  
   timeliness, 219, 220, 225  
   via software. *See* software prefetching
- process, 303
- process ID number. *See* PID
- processor consistency, 293
- producer-consumer, 282, 291
- promotion strategy, 241
- purge, 56
- Q**
- queuing locks, 286–287
- R**
- Rambus, 251
- RAW. *See* data dependencies
- reassembling buffer, 248
- Reduced Instruction Set Computer. *See* RISC
- reference prediction table (RPT),  
   224
- register  
   global in branch prediction, 139  
   history. *See* global  
   linked, 284  
   local history, 142
- register renaming. *See* also pipeline stage  
   renaming  
   extended register scheme, 168

- register renaming (*cont.*)
    - monolithic scheme, 166–168
    - ROB-based, 89–90
  - register stacking, 119
  - registers
    - architectural, 90, 165
    - logical, 90
    - physical, 90
  - relaxed consistency memory models, 229, 292
  - reorder buffer. *See* ROB
  - replacement algorithm (in cache), 53
    - least-recently-used. *See* LRU
    - LRU, 53, 236
    - MRU-based, 235
    - optimal, 236
    - tree-based LRU approximation, 235, 244
    - victim line, 53
  - reservation station, 93, 178
  - resource adaptation, 341–342
  - resource hibernation, 340
  - resume buffer, 151
  - return address stack, 150
  - ring, 266, 267, 327
  - RISC, 31, 78, 326
  - ROB, 91, 96, 166
  - rotation time, 65
  - row access strobe (RAS), 246
  - run-ahead execution, 315
- S**
- saturating counter, 153, 190
    - 2-bit, 133, 135
    - 3-bit, 135
  - scheduling policy, 183
  - SCI (Scalable Coherent Interface), 280
  - scoreboard. *See* Alpha 21164, CDC 6600
  - SDRAM (Synchronous DRAM), 247
  - sector cache, 239, 243, 244
  - seek time, 65
  - segment, 61
  - select, 180
  - sequential consistency, 291
  - sequential prefetching, 222
    - in IBM Power4, 243, 244
    - nextline, 222
    - one-block-look-ahead (OBL), 222
    - prefetch on miss, 222
  - sequential process, 304
  - shared memory. *See* NUMA, UMA
  - shared-bus, 263, 265, 270
    - contention, 265
  - signature, 227
  - SIMD, 261, 262, 269, 295, 327
  - simulators
    - execution-driven, 19
    - trace-driven, 19
  - Simultaneous MultiThreading. *See* SMT
  - single assignment rule, 318
  - Single Instruction Multiple Data. *See* SIMD
  - slack of instruction, 200
  - SMP, 263, 266
  - SMT, 110, 241, 310–313
    - performance improvement with, 314
  - software pipelining, 118
  - software prefetching, 220–222
  - SPEC CPU2000, 14–15
  - SPEC CPU2006, 15
  - speculation. *See* branch prediction, load speculation, prefetching
  - speed demons, 77
  - speedup, 10, 76
    - linear, 262
    - superlinear, 11
  - spin lock, 283, 284
  - split-transaction bus, 250, 266, 273
  - SRAM (static random access memory), 246
  - SSE instructions, 102, 294
  - stack property, 67
  - stall (in pipeline), 37
  - state
    - process, 31, 41
  - state-less prefetching, 228
  - steering of instructions, 202
  - store buffer, 185
  - store instruction with store buffer, 186
  - store sets, 191–194
  - store-conditional, 284
  - stream buffers, 223
    - extension, 224
  - streaming processor, 262
  - stride, 223
  - stride prefetching, 223–225
  - Sun Niagara, 267, 280, 320–322
  - Sun Sparc, 283
  - Sun Sparc, 35, 136
  - Sun UltraSparc, 136, 212, 295
  - Sun Visual Instruction Set (VIS), 294
  - superpipelined, 79
  - superscalar, 75, 303
    - dynamic. *See* out-of-order
    - in-order, 76
    - m-way, 75, 201
    - out-of-order, 76, 309
    - static. *See* in-order
  - Symmetric MultiProcessor. *See* SMP
  - synchronization, 281
  - synonym problem, 211
- T**
- target address prediction. *See* BTB
  - Tera MTA, 288, 305
  - test-and-set, 283
  - test-and-test-and-set, 285
  - thermal management, 336
  - thread-level parallelism, 303
  - throughput, 7, 30, 313

TLB, 62, 185, 228

hit, 63

miss, 64

TLB-slice, 211

Tomasulo's algorithm, 165

in IBM System/360 Model 91, 96–98, 163

in P6 microarchitecture, 107

torus, 268

trace cache, 110, 161–163

transactional memory, 289

transfer time (to/from disk), 65

Translation Look-Aside Buffers.

*See* TLB

## U

uniform memory access (UMA), 263

## V

value prediction, 196–198

vector processor, 262

Very Long Instruction Word. *See* VLIW

victim cache, 215–216

virtual cache, 211

VLIW, 111, 262

von Neumann model, 2, 29

## W

wakeup, 180

wakeup-select, 181

speculative, 182

Wallace tree, 46

WAR. *See* data dependencies

WAW. *See* data dependencies

way prediction, 210

weak ordering, 293

weighted arithmetic mean, 17

wire delays, 343–344

wire length, 76

write buffer, 54, 229

write-allocate. *See* cache write strategies

write-around. *See* cache write strategies

write-back. *See* cache write strategies

write-through. *See* cache write strategies

write-validate, 232

μops fusion, 111