CS61C : Machine Structures

Lecture 7.2.2
RAID & Performance

2004-08-05

Kurt Meinz

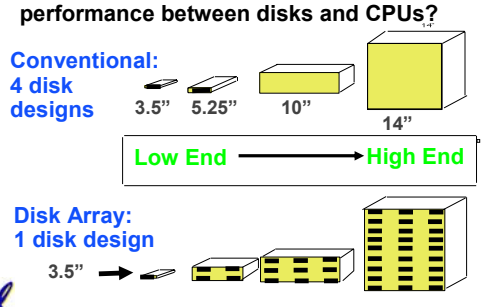
inst.eecs.berkeley.edu/~cs61c

Outline

- RAID
- Performance
- Intro to x86

Use Arrays of Small Disks…

- Katz and Patterson asked in 1987:
  - Can smaller disks be used to close gap in performance between disks and CPUs?



Replace Small Number of Large Disks with Large Number of Small Disks! (1988 Disks)

	IBM 3390K	IBM 3.5" 0061	x70	
Capacity	20 GBytes	320 MBytes	23 GBytes	
Volume	97 cu. ft.	0.1 cu. ft.	11 cu. ft.	9X
Power	3 KW	11 W	1 KW	3X
Data Rate	15 MB/s	1.5 MB/s	120 MB/s	8X
I/O Rate	600 I/Os/s	55 I/Os/s	3900 IOs/s	6X
MTTF	250 KHrs	50 KHrs	??? Hrs	
Cost	$250K	$2K	$150K	

Disk Arrays potentially high performance, high MB per cu. ft., high MB per KW, but what about reliability?

Array Reliability

- Reliability - whether or not a component has failed
  - measured as Mean Time To Failure (MTTF)
  - Reliability of N disks = Reliability of 1 Disk ÷ N (assuming failures independent)
    - 50,000 Hours ÷ 70 disks = 700 hour
  - Disk system MTTF: Drops from 6 years to 1 month!
  - Disk arrays (JBOD) too unreliable to be useful!

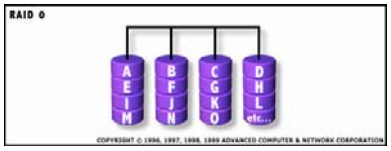
Redundant Arrays of (Inexpensive) Disks

- Files are "striped" across multiple disks
- Redundancy yields high data availability
  - Availability: service still provided to user, even if some components failed
- Disks will still fail
- Contents reconstructed from data redundantly stored in the array
  - ⇒ Capacity penalty to store redundant info
  - ⇒ Bandwidth penalty to update redundant info

Berkeley History, RAID-I

- RAID-I (1989)
  - Consisted of a Sun 4/280 workstation with 128 MB of DRAM, four dual-string SCSI controllers, 28 5.25-inch SCSI disks and specialized disk striping software
- Today RAID is $27 billion dollar industry, 80% nonPC disks sold in RAIDs

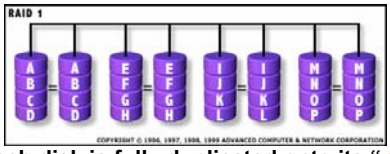
"RAID 0": Striping



- Assume have 4 disks of data for this example, organized in blocks
- Large accesses faster since transfer from several disks at once

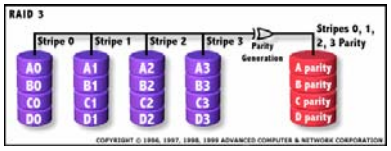
This and next 5 slides from RAID.edu, http://www.acnc.com/04_01_00.html

RAID 1: Mirror



- Each disk is fully duplicated onto its "mirror"
  - Very high availability can be achieved
- Bandwidth reduced on write:
  - 1 Logical write = 2 physical writes
- Most expensive solution: 100% capacity overhead

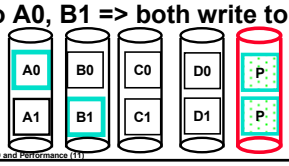
RAID 3: Parity



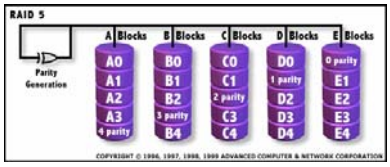
- Parity computed across group to protect against hard disk failures, stored in P disk
- Logically, a single high capacity, high transfer rate disk
- 25% capacity cost for parity in this example vs. 100% for RAID 1 (5 disks vs. 8 disks)

Inspiration for RAID 5

- Small writes (write to one disk):
  - Option 1: read other data disks, create new sum and write to Parity Disk (access all disks)
  - Option 2: since P has old sum, compare old data to new data, add the difference to P: 1 logical write = 2 physical reads + 2 physical writes to 2 disks
- Parity Disk is bottleneck for Small writes: Write to A0, B1 => both write to P disk



RAID 5: Rotated Parity, faster small writes



- Independent writes possible because of interleaved parity
  - Example: write to A0, B1 uses disks 0, 1, 4, 5, so can proceed in parallel
  - Still 1 small write = 4 physical disk accesses

Outline

- RAID
- Performance
- Intro to x86

Performance

- Purchasing Perspective: given a collection of machines (or upgrade options), which has the
  - best performance ?
  - least cost ?
  - best performance / cost ?
- Computer Designer Perspective: faced with design options, which has the
  - best performance improvement ?
  - least cost ?
  - best performance / cost ?
- All require basis for comparison and metric for evaluation
- Solid metrics lead to solid progress!

Two Notions of "Performance"

Plane	DC to Paris	Top Speed	Passen-gers	Throughput (pmph)
Boeing 747	6.5 hours	610 mph	470	286,700
BAD/Sud Concorde	3 hours	1350 mph	132	178,200

- Which has higher performance?
  - Time to deliver 1 passenger?
  - Time to deliver 400 passengers?
- In a computer, time for 1 job called
  Response Time or Execution Time
- In a computer, jobs per day called
  Throughput or Bandwidth

Definitions

- Performance is in units of things per sec
  - bigger is better
- If we are primarily concerned with response time
  - performance(x) = $\frac{1}{\text{execution\_time}(x)}$

" F(ast) is *n* times faster than S(low) " means…

$$n = \frac{\text{performance}(F)}{\text{performance}(S)} = \frac{\text{execution\_time}(S)}{\text{execution\_time}(F)}$$

Example of Response Time v. Throughput

- Time of Concorde vs. Boeing 747?
  - Concord is 6.5 hours / 3 hours = 2.2 times faster
- Throughput of Boeing vs. Concorde?
  - Boeing 747: 286,700 pmph / 178,200 pmph = 1.6 times faster
- Boeing is 1.6 times ("60%") faster in terms of throughput
- Concord is 2.2 times ("120%") faster in terms of flying time (response time)

We will focus primarily on execution time for a single job

Confusing Wording on Performance

- Will (try to) stick to "n times faster"; its less confusing than "m % faster"
- As faster means both increased performance and decreased execution time, to reduce confusion will use "improve performance" or "improve execution time"

What is Time?

- Straightforward definition of time:
  - Total time to complete a task, including disk accesses, memory accesses, I/O activities, operating system overhead, ...
  - "real time", "response time" or "elapsed time"
- Alternative: just time processor (CPU) is working only on your program (since multiple processes running at same time)
  - "CPU execution time" or "CPU time"
  - Often divided into system CPU time (in OS) and user CPU time (in user program)

How to Measure Time?

- User Time ⇒ seconds
- CPU Time: Computers constructed using a clock that runs at a constant rate and determines when events take place in the hardware
  - These discrete time intervals called clock cycles (or informally clocks or cycles)
  - Length of clock period: clock cycle time (e.g., 2 nanoseconds or 2 ns) and clock rate (e.g., 500 megahertz, or 500 MHz), which is the inverse of the clock period; use these!

Measuring Time using Clock Cycles (1/2)

- CPU execution time for program

  = Clock Cycles for a program x Clock Cycle Time

- or

$$= \frac{\text{Clock Cycles for a program}}{\text{Clock Rate}}$$

Measuring Time using Clock Cycles (2/2)

- One way to define clock cycles:

Clock Cycles for program

= Instructions for a program (called "Instruction Count")

x Average Clock cycles Per Instruction (abbreviated "CPI")

- CPI one way to compare two machines with same instruction set, since Instruction Count would be the same

Performance Calculation (1/2)

- CPU execution time for program

  = Clock Cycles for program x Clock Cycle Time

- Substituting for clock cycles:

  CPU execution time for program = (Instruction Count x CPI) x Clock Cycle Time

= Instruction Count x CPI x Clock Cycle Time

Performance Calculation (2/2)

$$\text{CPU time} = \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Cycle}}$$

$$\text{CPU time} = \frac{\cancel{\text{Instructions}}}{\text{Program}} \times \frac{\text{Cycles}}{\cancel{\text{Instruction}}} \times \frac{\text{Seconds}}{\text{Cycle}}$$

$$\text{CPU time} = \frac{\cancel{\text{Instructions}}}{\text{Program}} \times \frac{\cancel{\text{Cycles}}}{\cancel{\text{Instruction}}} \times \frac{\text{Seconds}}{\cancel{\text{Cycle}}}$$

$$\text{CPU time} = \frac{\text{Seconds}}{\text{Program}}$$

- Product of all 3 terms: if missing a term, can't predict time, the real measure of performance

How Calculate the 3 Components?

- Clock Cycle Time: in specification of computer (Clock Rate in advertisements)
- Instruction Count:
  - Count instructions in loop of small program
  - Use simulator to count instructions
  - Hardware counter in spec. register
    - (Pentium II,III,4)
- CPI:
  - Calculate: $\frac{\text{Execution Time / Clock cycle time}}{\text{Instruction Count}}$
  - Hardware counter in special register (PII,III,4)

Calculating CPI Another Way

- First calculate CPI for each individual instruction (add, sub, and, etc.)
- Next calculate frequency of each individual instruction
- Finally multiply these two for each instruction and add them up to get final CPI (the weighted sum)

Example (RISC processor)

Op	$Freq_i$	CPI_i	Prod	(% Time)
ALU	50%	1	.5	(23%)
Load	20%	5	1.0	(45%)
Store	10%	3	.3	(14%)
Branch	20%	2	.4	(18%)
			2.2	

Instruction Mix (Where time spent)

- What if Branch instructions twice as fast?

Example: What about Caches?

- Can Calculate Memory portion of CPI separately
- Miss rates: say L1 cache = 5%, L2 cache = 10%
- Miss penalties: L1 = 5 clock cycles, L2 = 50 clocks
- Assume miss rates, miss penalties same for instruction accesses, loads, and stores
- CPI_{memory}
  = Instruction Frequency * L1 Miss rate * (L1 miss penalty + L2 miss rate * L2 miss penalty) + Data Access Frequency * L1 Miss rate * (L1 miss penalty + L2 miss rate * L2 miss penalty)
  = 100%*5%*(5+10%*50)+(20%+10%)*5%*(5+10%*50)
  = 5%*(10)+(30%)*5%*(10) = 0.5 + 0.15 = 0.65

Overall CPI = 2.2 + 0.65 = 2.85

What Programs Measure for Comparison?

- Ideally run typical programs with typical input before purchase, or before even build machine
  - Called a "workload"; For example:
  - Engineer uses compiler, spreadsheet
  - Author uses word processor, drawing program, compression software
- In some situations it's hard to do
  - Don't have access to machine to "benchmark" before purchase
  - Don't know workload in future

Example Standardized Benchmarks (1/2)

- Standard Performance Evaluation Corporation (SPEC) SPEC CPU2000
  - CINT2000 12 integer (gzip, gcc, crafty, perl, ...)
  - CFP2000 14 floating-point (swim, mesa, art, ...)
  - All relative to base machine Sun 300MHz 256Mb-RAM Ultra5_10, which gets score of 100
  - `www.spec.org/osg/cpu2000/`
  - They measure
    - System speed (SPECint2000)
    - System throughput (SPECint_rate2000)

Example Standardized Benchmarks (2/2)

- SPEC
  - Benchmarks distributed in source code
  - Big Company representatives select workload
    - Sun, HP, IBM, etc.
  - Compiler, machine designers target benchmarks, so try to change every 3 years

Example PC Workload Benchmark

- PCs: Ziff-Davis Benchmark Suite
  - "Business Winstone is a system-level, application-based benchmark that measures a PC's overall performance when running today's top-selling Windows-based 32-bit applications… it doesn't mimic what these packages do; it runs real applications through a series of scripted activities and uses the time a PC takes to complete those activities to produce its performance scores.
  - Also tests for CDs, Content-creation, Audio, 3D graphics, battery life

http://www.etestinglabs.com/benchmarks/

Performance Evaluation

- Good products created when have:
  - Good benchmarks
  - Good ways to summarize performance
- Given sales is a function of performance relative to competition, should invest in improving product as reported by performance summary?
- If benchmarks/summary inadequate, then choose between improving product for real programs vs. improving product to get more sales; Sales almost always wins!

"And in conclusion…"

- Benchmarks
  - Attempt to predict performance
  - Updated every few years
  - Measure everything from simulation of desktop graphics programs to battery life
- Megahertz Myth
  - MHz ≠ performance, it's just one factor
- It's non-trivial to try to help people in developing countries with technology
- Viruses have damaging potential the likes of which we can only imagine.

Outline

- RAID
- Performance
- Intro to x86

MIPS is example of RISC

- RISC = Reduced Instruction Set Computer
  - Term coined at Berkeley, ideas pioneered by IBM, Berkeley, Stanford
- RISC characteristics:
  - Load-store architecture
  - Fixed-length instructions (typically 32 bits)
  - Three-address architecture
- RISC examples: MIPS, SPARC, IBM/Motorola PowerPC, Compaq Alpha, ARM, SH4, HP-PA, ...

MIPS vs. 80386

MIPS	80386
• Address: 32-bit	• 32-bit
• Page size: 4KB	• 4KB
• Data aligned	• Data unaligned
• Destination reg: Left • `add $rd,$rs1,$rs2`	• Right • `add %rs1,%rs2,%rd`
• Regs: $0, $1, ..., $31	• %r0, %r1, ..., %r7
• Reg = 0: $0	• (n.a.)
• Return address: $31	• (n.a.)

MIPS vs. Intel 80x86

- MIPS: "Three-address architecture"
  - Arithmetic-logic specify all 3 operands

    `add $s0,$s1,$s2 # s0=s1+s2`
  - Benefit: fewer instructions ⇒ performance
- x86: "Two-address architecture"
  - Only 2 operands, so the destination is also one of the sources

    `add $s1,$s0 # s0=s0+s1`
  - Often true in C statements: `c += b;`
  - Benefit: smaller instructions ⇒ smaller code

MIPS vs. Intel 80x86

- MIPS: "load-store architecture"
  - Only Load/Store access memory; rest operations register-register; e.g.,

    ```
    lw $t0, 12($gp)
    add $s0,$s0,$t0 # s0=s0+Mem[12+gp]
    ```
  - Benefit: simpler hardware ⇒ easier to pipeline, higher performance
- x86: "register-memory architecture"
  - All operations can have an operand in memory; other operand is a register; e.g.,

    `add 12(%gp),%s0 # s0=s0+Mem[12+gp]`
  - Benefit: fewer instructions ⇒ smaller code

MIPS vs. Intel 80x86

- MIPS: "fixed-length instructions"
  - All instructions same size, e.g., 4 bytes
  - simple hardware ⇒ performance
  - branches can be multiples of 4 bytes
- x86: "variable-length instructions"
  - Instructions are multiple of bytes: 1 to 17;

    ⇒ small code size (30% smaller?)
  - More Recent Performance Benefit: better instruction cache hit rates
  - Instructions can include 8- or 32-bit immediates

Unusual features of 80x86

- 8 32-bit Registers have names; 16-bit 8086 names with "e" prefix:
  - eax, ecx, edx, ebx, esp, ebp, esi, edi
  - 80x86 word is 16 bits, double word is 32 bits
- PC is called `eip` (instruction pointer)
- `leal` (load effective address)
  - Calculate address like a load, but load address into register, not data
  - Load 32-bit address:

    ```
    leal -4000000(%ebp),%esi
     # esi = ebp - 4000000
    ```


Instructions: MIPS vs. 80x86

• `addu, addiu`	• `addl`
• `subu`	• `subl`
• `and,or, xor`	• `andl, orl, xorl`
• `sll, srl, sra`	• `sall, shrl, sarl`
• `lw`	• `movl mem, reg`
• `sw`	• `movl reg, mem`
• `mov`	• `movl reg, reg`
• `li`	• `movl imm, reg`
• `lui`	• `n.a.`

80386 addressing (ALU instructions too)

- base reg + offset (like MIPS)
  - `movl -8000044(%ebp), %eax`
- base reg + index reg (2 regs form addr.)
  - `movl (%eax,%ebx),%edi`
    `# edi = Mem[ebx + eax]`
- scaled reg + index (shift one reg by 1,2)
  - `movl(%eax,%edx,4),%ebx`
    `# ebx = Mem[edx*4 + eax]`
- scaled reg + index + offset
  - `movl 12(%eax,%edx,4),%ebx`
    `# ebx = Mem[edx*4 + eax + 12]`

Branches in 80x86

- Rather than compare registers, x86 uses special 1-bit registers called "condition codes" that are set as a side-effect of ALU operations
  - S - Sign Bit
  - Z - Zero (result is all 0)
  - C - Carry Out
  - P - Parity: set to 1 if even number of ones in rightmost 8 bits of operation
- Conditional Branch instructions then use condition flags for all comparisons: <, <=, >, >=, ==, !=

Branch: MIPS vs. 80x86

• `beq`	• `(cmpl;) je` if previous operation set condition code, then `cmpl` unnecessary
• `bne`	• `(cmpl;) jne`
• `slt; beq`	• `(cmpl;) jlt`
• `slt; bne`	• `(cmpl;) jge`
• `jal`	• `call`
• `jr $31`	• `ret`

While in C/Assembly: 80x86

C
```
while (save[i]==k)
    i = i + j;
```

(i,j,k: %edx,%esi,%ebx)

x86
```
        leal -400(%ebp),%eax
.Loop:  cmpl %ebx,(%eax,%edx,4)
        jne .Exit
        addl %esi,%edx
        j .Loop
.Exit:
```

Note: cmpl replaces sll, add, lw in loop

Unusual features of 80x86

- Memory Stack is part of instruction set
  - `call` places return address onto stack, increments esp (Mem[esp]=eip+6; esp+=4)
  - push places value onto stack, increments esp
  - pop gets value from stack, decrements esp
- `incl, decl` (increment, decrement)

  `incl %edx # edx = edx + 1`
  - Benefit: smaller instructions ⇒ smaller code

Intel Internals

- Hardware below instruction set called "microarchitecture"
- Pentium Pro, Pentium II, Pentium III all based on same microarchitecture (1994)
  - Improved clock rate, increased cache size
- Pentium 4 has new microarchitecture

Dynamic Scheduling in Pentium Pro, II, III

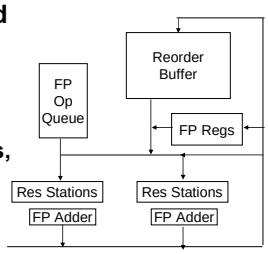
- PPro doesn't pipeline 80x86 instructions
- PPro decode unit translates the Intel instructions into 72-bit "micro-operations" (~ MIPS instructions)
- Takes 1 clock cycle to determine length of 80x86 instructions + 2 more to create the micro-operations
- Most instructions translate to 1 to 4 micro-operations
- 10 stage pipeline for micro-operations

Hardware support

- *Out-of-Order execution*: allow a instructions to execute before branch is resolved ("HW undo")
- When instruction no longer speculative, write results (*instruction commit*)
- Fetch in-order, execute out-of-order, commit in order

Hardware for out of order execution

- Need HW buffer for results of uncommitted instructions: *reorder buffer*
  - Reorder buffer can be operand source
  - Once operand commits, result is found in register
  - Discard results on mispredicted branches or on exceptions



Dynamic Scheduling in Pentium Pro

Max. instructions issued/clock 3

Max. instr. complete exec./clock 5

Max. instr. commited/clock 3

Instructions in reorder buffer 40

2 integer functional units (FU), 1 floating point FU, 1 branch FU, 1 Load FU, 1 Store FU

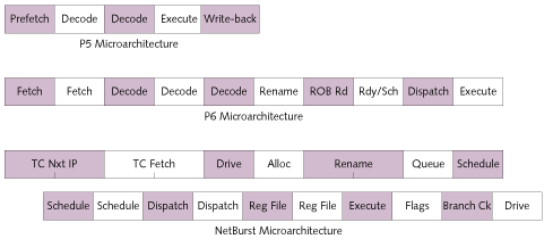
Pentium 4

- Still translate from 80x86 to micro-ops
- P4 has better branch predictor, more FUs
- Clock rates:
  - Pentium III 1 GHz v. Pentium IV 1.5 GHz
  - 10 stage pipeline vs. 20 stage pipeline
- Faster memory bus: 400 MHz v. 133 MHz
- Caches
  - Pentium III: L1I 16KB, L1D 16KB, L2 256 KB
  - Pentium 4: L1I 8 KB, L1D 8 KB, L2 256 KB
  - Block size: PIII 32B v. P4 128B

Pentium 4 features

- Multimedia instructions 128 bits wide vs. 64 bits wide => 144 new instructions
  - When used by programs??
- Instruction Cache holds micro-operations vs. 80x86 instructions
  - no decode stages of 80x86 on cache hit
  - called "trace cache" (TC)
- Using RAMBUS DRAM
  - Bandwidth faster, latency same as SDRAM
  - Cost 3X vs. SDRAM

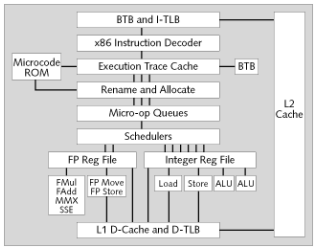
Pentium, Pentium Pro, Pentium 4 Pipeline



- Pentium (P5) = 5 stages
  Pentium Pro, II, III (P6) = 10 stages
  Penitum 4 (NetBurst) = 20 stages

"Pentium 4 (Partially) Previewed," Microprocessor Report, 8/28/00

Block Diagram of Pentium 4 Microarchitecture



- BTB = Branch Target Buffer (branch predictor)
- I-TLB = Instruction TLB, Trace Cache = Instruction cache
- RF = Register File; AGU = Address Generation Unit
- "Double pumped ALU" means ALU clock rate 2X => 2X ALU F.U.s