CSE 431 Computer Architecture Fall 2015

Chapter 1: Abstractions Technology, and Performance

Mary Jane Irwin (www.cse.psu.edu/~mji)

[Adapted from *Computer Organization and Design, 5th Edition*, Patterson & Hennessy, © 2014, Morgan Kaufmann]

Course Administration

- Instructor: Mary Jane Irwin mji@cse.psu.edu
  348C IST Bldg
  Office Hrs: T 10-11:30am & W 1-2:30pm
- TA: Jing Chen jxc669@psu.edu
  339 IST Bldg (Office Hours)
  Office Hrs: M & F 9-10:30am
- Labs: Accounts on machines in 218 IST (Dells running RedHat Linux)
- URL: Angel
- Text: Required: *Computer Org and Design*, 5rd Ed., Patterson & Hennessy, ©2014
- Slides: Hard copy handed out in class; pdf on Angel after lecture

Grading Information

- Grade determinates
  - First Exam ~27.5%
    - Tuesday, October 6th, 20:15 to 22:15, Location: 22 Deike
  - Second Exam ~27.5%
    - Tuesday, November 17th, 20:15 to 22:15, Location: 22 Deike
  - Homeworks and Final Project (6) ~35%
    - To be submitted on Angel by 23:55 on the due date. No late assignments will be accepted.
  - Class participation & on-line (Angel) quizzes ~10%
- Let me know about exam conflicts ASAP
- Grades will be posted on Angel
  - Must submit email request for change of grade after discussions with the TA (Homeworks/Quizzes) or instructor (Exams)
  - November 30th deadline for filing grade corrections; no requests for grade changes will be accepted after this date

Course Structure & Schedule

- Lecture: 8:00 to 9:25am Tuesdays and Thursdays
  - August 25th to November 17th (Final Project (HW6) due Monday, December 14th)
  - 30 x 75 minute classes ~ 26 x 85 minute classes
- Design focused class
  - Simulation of architecture alternatives using SimpleScalar
- Lectures:
  - 2 weeks review of the MIPS ISA and basic architecture
  - 2 weeks scalar pipelined datapath design issues
  - 2 week memory hierarchies and memory design issues
  - 2.5 weeks superscalar datapath design issues
  - 1 weeks storage and I/O design issues
  - 2.5 weeks multiprocessor/multicore design issues
  - 1 week exams

Course Content

- Memory hierarchy and design, CPU design, pipelining, multiprocessor architecture.
  - "This course will introduce students to the architecture-level design issues of a computer system. They will apply their knowledge of digital logic design to explore the high-level interaction of the individual computer system hardware components. Concepts of sequential and parallel architecture including the interaction of different memory components, their layout and placement, communication among multiple processors, effects of pipelining, and performance issues, will be covered. Students will apply these concepts by studying and evaluating the merits and demerits of selected computer system architectures."
    - To learn what determines the capabilities and performance of computer systems and to understand the interactions between the computer's architecture and its software so that future software designers (compiler writers, operating system designers, database programmers, application programmers, …) can achieve the best cost-performance trade-offs and so that future computer architects understand the effects of their design choices on software.

What You Should Know – 271, 331 (, & 311)

- Basic logic design & machine organization
  - logical minimization, FSMs, component design
  - processor, memory, I/O
- Create, assemble, run, debug programs in an assembly language
  - MIPS preferred
- Create, simulate, and debug hardware structures in a hardware description language
  - VHDL or verilog
- Create, compile, and run C (C++, Java) programs
- Create, organize, and edit files and run programs on Unix/Linux

Aside: Some Basic Definitions

- Kilobyte – 2^{10} or 1,024 bytes
- Megabyte– 2^{20} or 1,048,576 bytes
  - sometimes “rounded” to 10^6 or 1,000,000 bytes
- Gigabyte – 2^{30} or 1,073,741,824 bytes
  - sometimes rounded to 10^9 or 1,000,000,000 bytes
- Terabyte – 2^{40} or 1,099,511,627,776 bytes
  - sometimes rounded to 10^{12} or 1,000,000,000,000 bytes
- Petabyte – 2^{50} or 1024 terabytes
  - sometimes rounded to 10^{15} or 1,000,000,000,000,000 bytes
- Exabyte – 2^{60} or 1024 petabytes
  - Sometimes rounded to 10^{18} or 1,000,000,000,000,000,000 bytes

Aside: Binary Notation for Common Sizes

- Resolved the ambiguity between 2^X and 10^Y by adding a binary notation for the common sizes

Decimal Notation		Binary Notation		% Larger
KB	10^3	KiB	2^{10}	2%
MB	10^6	MiB	2^{20}	5%
GB	10^9	GiB	2^{30}	7%
TB	10^{12}	TiB	2^{40}	10%
PB	10^{15}	PiB	2^{50}	13%
EB	10^{18}	EiB	2^{60}	15%
ZB	10^{21}	ZiB	2^{70}	18%
YB	10^{24}	YiB	2^{80}	21%

Classes of Computers

- Servers/Clouds/Data Centers/Supercomputers
  - Multiple, simultaneous users
  - Network based, terabytes of memory, petabytes of storage
  - High capacity, performance, reliability/availability, security
  - Range from small servers to building sized
- Desktop/laptop (PC)/tablet computers
  - Single user
  - General purpose, variety of software/applications
  - Subject to cost/performance/power/reliability tradeoff
- Embedded computers (processors)
  - Hidden as components of systems, used for one predetermined application (or small set of applications)
  - Stringent power/performance/cost constraints

Embedded Processor Characteristics

The largest class of computers spanning the widest range of applications and performance

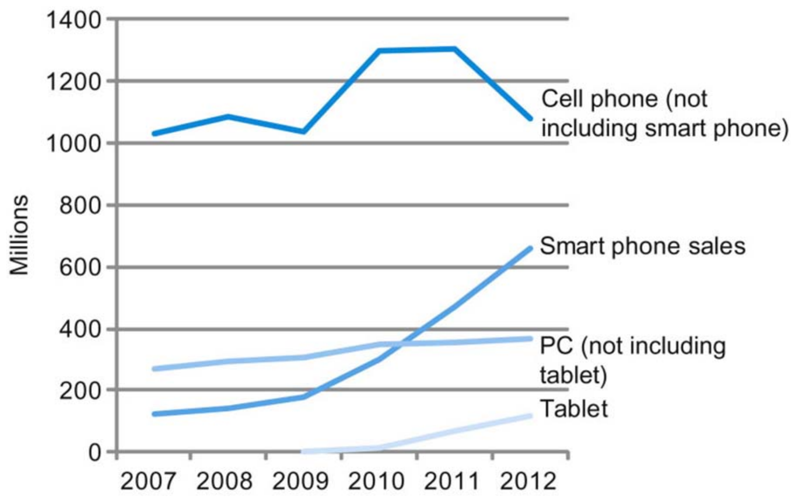
- Often have minimum performance requirements. Example?
- Often have stringent limitations on cost. Example?
- Often have stringent limitations on power consumption. Example?
- Often have low tolerance for failure. Example?

The PostPC Era

- Personal mobile devices (PMDs)
  - Battery operated, touch screen (no mouse, no keyboard)
  - Connects to the Internet, download “apps”
  - A few hundreds of dollars (or less) in cost
  - Smart phones, tablets, electronic glasses, cameras, …
- Cloud computing
  - Warehouse Scale Computers (WSC)
  - Software as a Service (SaaS) deployed via the Cloud
  - Portion of software run on a PMD and a portion runs in the Cloud
  - Amazon, Google, …

Growth in PMDs (Personal Mobile Devices)

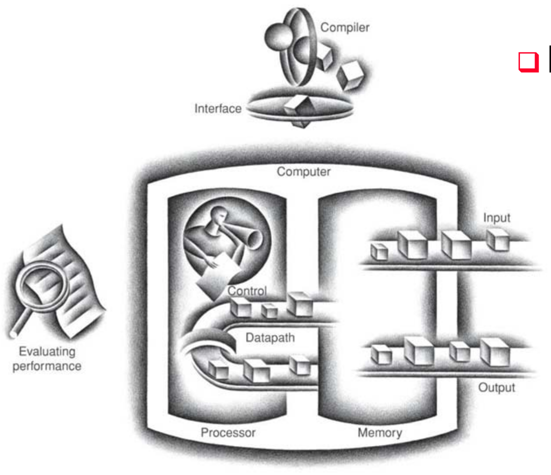
PMDs growth >> PC growth



Eight Great Ideas in Computer Architecture

- Design for ***Moore's Law***
- Use ***abstraction*** to simplify design
- Make the ***common case fast***
- Performance *via* ***parallelism***
- Performance *via* ***pipelining***
- Performance *via* ***prediction***
- ***Hierarchy*** of memories
- ***Dependability*** *via* redundancy

The Five Classic Components



- Input/output includes
  - User-interface devices (Display, keyboard, mouse)
  - Storage devices (Hard disk, CD/DVD, flash)
  - Network adapters (For communicating with other computers)

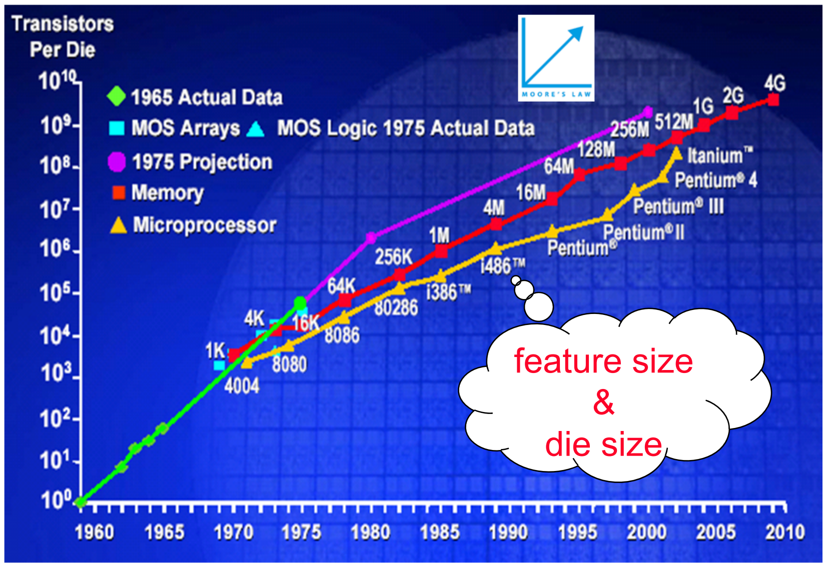
datapath + control = processor (CPU)

Instruction Set Architecture (ISA)

- **ISA**, or simply architecture – the abstract interface between the hardware and the lowest level software that encompasses all the information necessary to write a machine language program, including instructions, registers, memory access, I/O, …
  - Enables implementations of varying cost and performance to run identical software

- **ABI** (application binary interface) – the user portion of the instruction set (the ISA) plus the operating system interfaces used by application programmers
  - Defines a standard for binary portability across computers

Moore's Law: 2X transistors / "2 years"

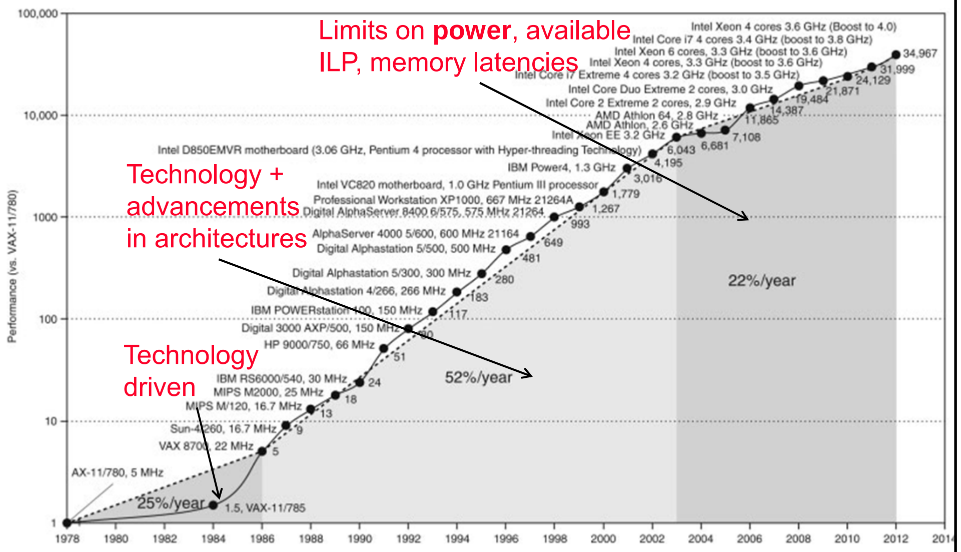


Technology Scaling Road Map (ITRS)

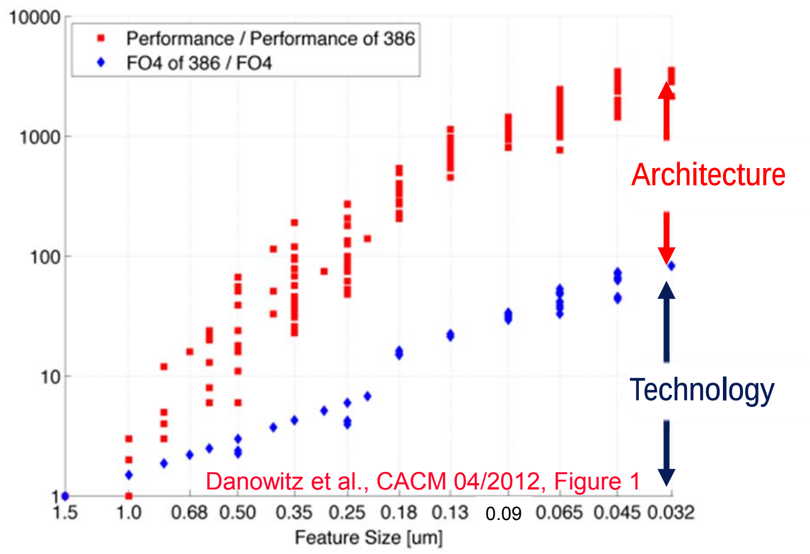
Year	2008	2010	2012	2014	2016
Feature size (nm)	45	32	22	14	12?
Intg. Capacity (BT)	0.5	1	2	4	8

- Fun facts about 45nm transistors
  - 30 million can fit on the head of a pin
  - You could fit more than 2,000 across the width of a human hair
  - If car prices had fallen at the same rate as the price of a single transistor has since 1968, a new car today would cost about 1 cent

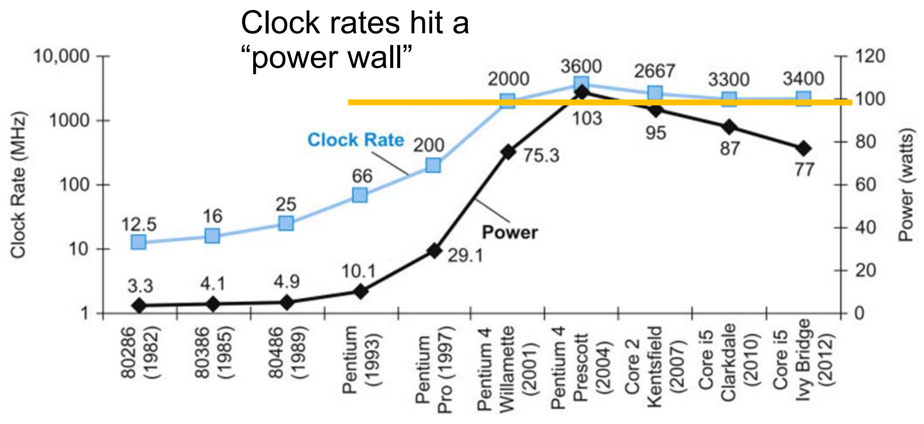
Growth in Processor Performance (SPECint)



Technology + Architecture = Performance



But What Happened to Clock Rates and Why?



- In CMOS IC technology

$$\text{Power} = \text{Capacitive load} \times \text{Voltage}^2 \times \text{Frequency}$$

??? $5V \rightarrow 1V$ x272

"For the P6, success criteria included performance above a certain level and failure criteria included power dissipation above some threshold."

Bob Colwell, Pentium Chronicles

Reducing Power

- Suppose a new CPU has
  - 85% of the capacitive load of the previous generation
  - 15% voltage reduction, 15% slower clock

$$\frac{P_{new}}{P_{old}} = \frac{(C_{old} \times 0.85) \times (V_{old} \times 0.85)^2 \times (F_{old} \times 0.85)}{C_{old} \times V_{old}{}^2 \times F_{old}} = 0.52$$

- We have hit the power wall
  - We can't reduce the supply voltage much further (Dennard scaling is over), or the capacitive load
  - We can't remove more heat without new cooling technologies (e.g., liquid cooled)
- How can we increase the performance while lowering (or keeping the same) clock rate?

The Move to Multicore Processors

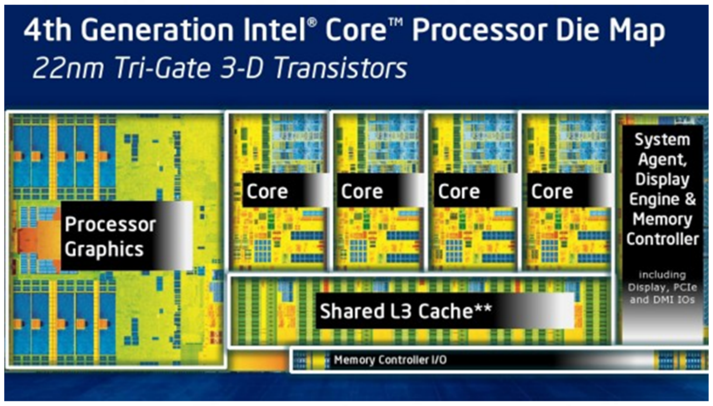
- The power challenge has forced a change in the design of microprocessors
  - Since 2002 the rate of improvement in the response time of programs on desktop computers slowed from a factor of 1.5 per year to less than a factor of 1.2 per year
- As of 2006 all server companies were shipping microprocessors with multiple cores per chip (processor)

Product	AMD Opteron X	Intel i7 Haswell	IBM Power 7+
Release date	2013	2013	2012
Technology	28nm bulk	22nm FinFET	32nm SOI
Cores/Clock	4/2.0 GHz	4/3.5GHz	8/4.4 GHz
Power (TDP)	22W	84W	~120 W

- Plan of record was to double the number of cores per chip per generation (about every two years)

E.g., Intel's Core i7 (Haswell)

- 22nm/FinFET technology, ~1.4 billion transistors, 4 cores/8 threads, 8MBs of shared L3 cache


4th Generation Intel® Core™ Processor Die Map
22nm Tri-Gate 3-D Transistors
Processor Graphics
Core
Core
Core
Core
System Agent, Display Engine & Memory Controller
including Display, PCIe and DMI IOs
Shared L3 Cache**
Memory Controller I/O

http://en.wikipedia.org/wiki/Haswell_%28microarchitecture%29

Courtesy, Intel ®

Multicore Performance Issues

- Private L1 caches, private or shared L2, … LL caches?
  - Best performance depends upon the applications and how much information they "share" in the cache, or how much they conflict in the cache
- Contention for memory controller(s) and port(s) to DRAM
- Requires explicitly parallel programming (multiple (parallel) threads for one application) – CSE 597D, (Fa15)
  - Compare with instruction level parallelism (ILP) where the hardware executes multiple instructions at once (so hidden from the programmer)
  - Parallel programming for performance is hard to do
    - Load balancing across cores
    - Cache sharing/contention, contention for DRAM controller(s)
    - Have to optimize for thread communication and synchronization

Defining Performance

- Response time (execution time) – how long does it take to do a task
  - Important to individual users
- Throughput (bandwidth) – number of tasks completed per unit time
  - Important to data center managers
- How are response time and throughput affected by
  1. Replacing the core with a faster version ?
  2. Adding more cores ?
- Our focus, for now, will be response time

"Never let an engineer get away with simply presenting the data. Always insist that he or she lead off with the conclusions to which the data led."

Bob Colwell, Pentium Chronicles

Relative Performance

- To maximize performance, need to minimize execution time

$$\text{performance}_X = 1 / \text{execution\_time}_X$$

If computer X is n times faster than computer Y, then

$$\frac{\text{performance}_X}{\text{performance}_Y} = \frac{\text{execution\_time}_Y}{\text{execution\_time}_X} = n$$

- Decreasing response time almost always improves throughput

Relative Performance Example

- If computer A runs a program in 10 seconds and computer B runs the same program in 15 seconds, how much faster is A than B?

We know that A is n times faster than B if

$$\frac{\text{performance}_A}{\text{performance}_B} = \frac{\text{execution\_time}_B}{\text{execution\_time}_A} = n$$

The performance ratio is $\frac{15}{10} = 1.5$

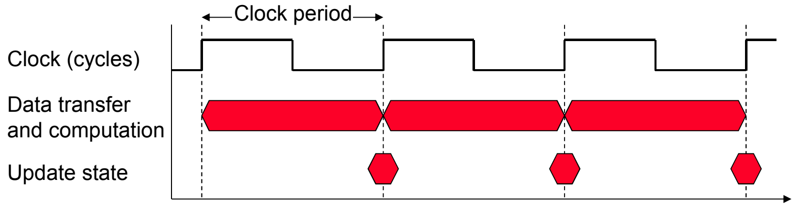
So A is 1.5 times faster than B (or A is 50% faster than B!)

Measuring Execution Time

- Elapsed time
  - Total response time, including all aspects
    - Processing, I/O, OS overhead, idle time
  - Determines system performance
- CPU time
  - Time spent processing a given job
    - Discounts I/O time, other jobs' shares
  - Comprises user CPU time and system CPU time
  - Different programs are affected differently by CPU and system performance

CPU Clocking

- Operation of digital hardware governed by a constant-rate clock

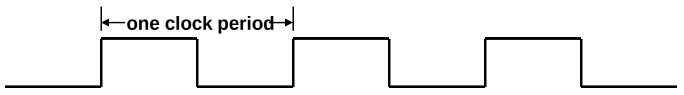


- Clock period (cycle): duration of a clock cycle
  - E.g., 250ps = 0.25ns = 250×10^{-12}s
- Clock frequency (rate): cycles per second
  - E.g., 4.0GHz = 4000MHz = 4.0×10^{9}Hz

Aside: Machine Clock Rate

- Clock rate (clock cycles per second in MHz or GHz) is inverse of clock cycle time (clock period)

CC = 1 / CR



10 nsec clock cycle	=>	100 MHz clock rate
5 nsec clock cycle	=>	200 MHz clock rate
2 nsec clock cycle	=>	500 MHz clock rate
1 nsec (10^{-9}) clock cycle	**=>**	**1 GHz (10^{9}) clock rate**
500 psec (10^{-12}) clock cycle	**=>**	**2 GHz clock rate**
250 psec clock cycle	=>	4 GHz clock rate
200 psec clock cycle	=>	5 GHz clock rate

Performance Factors

- CPU execution time (CPU time) – time the CPU spends working on a task (not including time waiting for I/O or running other programs)

$$\text{CPU execution time for a program} = \text{\# CPU clock cycles for a program} \times \text{clock cycle time}$$

or

$$\text{CPU execution time for a program} = \frac{\text{\# CPU clock cycles for a program}}{\text{clock rate}}$$

- Can improve performance by reducing either the length of the clock cycle (increasing clock rate) or reducing the number of clock cycles required for a program
- The architect must often trade off clock rate against the number of clock cycles for a program

Improving Performance Example

- A program runs on computer A with a 2 GHz clock in 10 seconds. What clock rate must computer B run at to run this program in 6 seconds? Unfortunately, to accomplish this, computer B will require 1.2 times as many clock cycles as computer A to run the program.

$$\text{CPU time}_A = \frac{\text{CPU clock cycles}_A}{\text{clock rate}_A}$$

$$\text{CPU clock cycles}_A = 10 \text{ sec} \times 2\text{x}10^9 \text{ cycles/sec} = 20 \times 10^9 \text{ clock cycles}$$

$$\text{CPU time}_B = \frac{1.2 \times 20 \times 10^9 \text{ clock cycles}}{\text{clock rate}_B}$$

$$\text{clock rate}_B = \frac{1.2 \times 20 \times 10^9 \text{ cycles}}{6 \text{ seconds}} = 4 \text{ GHz}$$

Instruction Performance

- Not all instructions take the same amount of time to execute
  - One way to think about execution time is that it equals the number of instructions executed multiplied by the average time per instruction

$$\text{\# CPU clock cycles for a program} = \text{\# Instructions for a program} \times \text{Average clock cycles per instruction}$$

- Clock cycles per instruction (CPI) – the average number of clock cycles each instruction takes to execute
  - A way to compare two different implementations of the same ISA

	Computer_A	Computer_B
Avg CPI	2	1.2

THE Performance Equation

- Our basic performance equation is then

$$\text{CPU time} = \text{Instruction\_count} \times \text{CPI} \times \text{clock\_cycle}$$

or

$$\text{CPU time} = \frac{\text{Instruction\_count} \times \text{CPI}}{\text{clock\_rate}}$$

- This equation separates the three key factors that affect performance
  - Can measure the CPU execution time by running the program
  - The clock rate is usually given
  - Can measure overall instruction count by using profilers/ simulators without knowing all of the implementation details
  - CPI varies by instruction type and ISA implementation for which we must know the implementation details

Using the Performance Equation

- Computers A and B implement the same ISA. Computer A has a clock cycle time of 250 ps and an average CPI of 2.0 for some program and computer B has a clock cycle time of 500 ps and an average CPI of 1.2 for the same program. Which computer is faster and by how much?

Each computer executes the same number of instructions, *I*, so

$$\text{CPU time}_A = I \times 2.0 \times 250 \text{ ps} = 500 \times I \text{ ps}$$

$$\text{CPU time}_B = I \times 1.2 \times 500 \text{ ps} = 600 \times I \text{ ps}$$

Clearly, A is faster … by the ratio of execution times

$$\frac{\text{performance}_A}{\text{performance}_B} = \frac{\text{execution\_time}_B}{\text{execution\_time}_A} = \frac{600 \times I \text{ ps}}{500 \times I \text{ ps}} = 1.2$$

Average (Effective) CPI

- Computing the overall average CPI is done by looking at the different types of instructions and their individual cycle counts and averaging

$$\text{Overall effective CPI} = \sum_{i=1}^{n} (\text{CPI}_i \times \text{IC}_i)$$

  - Where IC_i is the count (percentage) of the number of instructions of class i executed
  - CPI_i is the number of clock cycles per instruction for that instruction class
  - n is the number of instruction classes

- The overall effective CPI varies by instruction mix – a measure of the dynamic frequency of instructions for one or many programs

A Simple Performance Tradeoff Example

Op	Freq	CPI_i	Freq x CPI_i			
ALU	50%	1	.5	.5	.5	.25
Load	20%	5	1.0	.4	1.0	1.0
Store	10%	3	.3	.3	.3	.3
Branch	20%	2	.4	.4	.2	.4
Average (Effective) CPI			$\sum =$ 2.2	1.6	2.0	1.95

- How much faster would the machine be if a better data cache reduced the average load time to 2 cycles?
  CPU time new = 1.6 x IC x CC so 2.2/1.6 means 37.5% faster
- How does this compare with using branch prediction to shave a cycle off the branch time?
  CPU time new = 2.0 x IC x CC so 2.2/2.0 means 10% faster
- What if two ALU instructions could be executed at once?
  CPU time new = 1.95 x IC x CC so 2.2/1.95 means 12.8% faster

Determinates of CPU Performance

CPU time = Instruction_count x CPI x clock_cycle

	Instruction_ count	CPI	clock_cycle
Algorithm	X	X	
Programming language	X	X	
Compiler	X	X	
ISA	X	X	X
Core organization		X	X
Technology			X

Workloads and Benchmarks

- Benchmarks – a set of programs that form a “workload” specifically chosen to measure performance
- SPEC (System Performance Evaluation Cooperative) creates standard sets of benchmarks starting with SPEC89. SPEC CPU2006 consists of 12 integer benchmarks (CINT2006) and 17 floating-point benchmarks (CFP2006).

www.spec.org

- There are also benchmark collections for power workloads (SPECpower_ssj2008), for mail workloads (SPECmail2008), for multimedia workloads (mediabench), …

SPEC CINT2006 on Intel i7 (CR = 2.66GHz)

Name	ICx10^9	CPI	ExTime (sec)	RefTime (sec)	SPEC ratio
perl	2,252	0.60	508	9,770	19.2
bzip2	2,390	0.70	629	9,650	**15.4**
gcc	794	1.20	358	8,050	22.5
mcf	**221**	**2.66**	221	9,120	41.2
go	1,274	1.10	527	10,490	19.9
hmmer	**2,616**	0.60	590	9,330	**15.8**
sjeng	1,948	0.80	586	12,100	20.7
libquantum	659	**0.44**	109	20,720	**190.0**
h264avc	**3,793**	**0.50**	713	22,130	31.0
omnetpp	**367**	**2.10**	290	6,250	21.5
astar	1,250	1.00	470	7,020	**14.9**
xalancbmk	1,045	0.70	275	6,900	25.1
Geometric Mean					**25.7**

Comparing and Summarizing Performance

- How do we summarize the performance for a benchmark set with a single number?
  - First the execution times are normalized giving the "SPEC ratio" (bigger is faster, i.e., SPEC ratio is the inverse of execution time)
  - The SPEC ratios are then "averaged" using the geometric mean (GM)

$$\text{GM} = \sqrt[n]{\prod_{i=1}^{n} \text{SPEC ratio}_i}$$

- Guiding principle in reporting performance measurements is reproducibility – list everything another experimenter would need to duplicate the experiment (version of the operating system, compiler settings, input set used, specific computer configuration (clock rate, cache sizes and speed, memory size and speed, etc.))

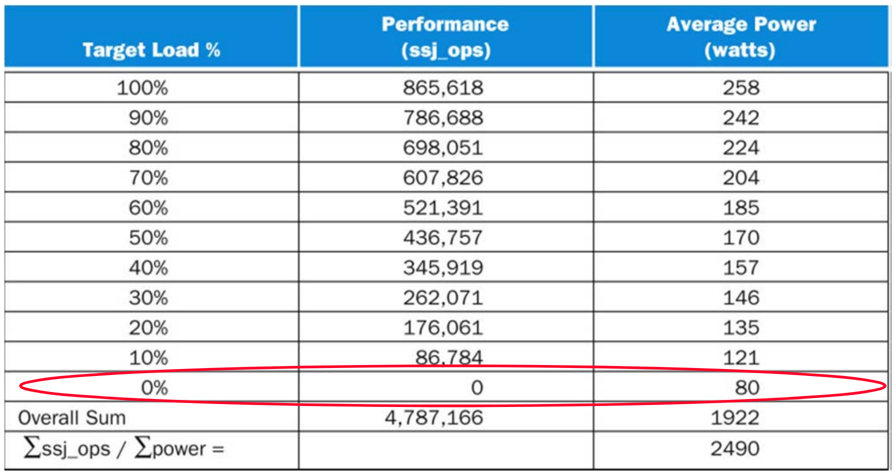
SPEC Power Benchmarks

- Power consumption of a server under different workload levels (divided into 10% increments)
  - Performance: ssj_ops/sec
  - Energy: joules
  - Power: Watts (joules/sec)

$$\text{Overall ssj\_ops per Watt} = \left(\sum_{i=0}^{10} \text{ssj\_ops}_i\right) \Big/ \left(\sum_{i=0}^{10} \text{power}_i\right)$$

http://www.zdnet.com/blog/ou/spec-launches-standardized-energy-efficiency-benchmark/927

SPECpower_ssj2008 on 2.66GHz Intel Xeon



Target Load %	Performance (ssj_ops)	Average Power (watts)
100%	865,618	258
90%	786,688	242
80%	698,051	224
70%	607,826	204
60%	521,391	185
50%	436,757	170
40%	345,919	157
30%	262,071	146
20%	176,061	135
10%	86,784	121
0%	0	80
Overall Sum	4,787,166	1922
$\sum$ssj_ops / $\sum$power =		2490

Fallacy: Lowest Power at (Near) Idle

- Look back at the Intel Xeon benchmark
  - At 100% load: 258W
  - At 50% load: 170W (66%)
  - At 10% load: 121W (47%)
- A Google data center
  - Mostly operates at a 10% to 50% load
  - At 100% load less than 1% of the time
- Data centers should be designed to make the power proportional to the load
  - "Energy Proportional Computing," Barroso and Hölzle, IEEE Computer Magazine, Dec 2007
  - Try to design servers so that they consume 10% of peak power when running at 10% workload levels, etc.

Pittfall: Amdahl's Law

- Used to determine the maximum expected improvement to overall system performance when only part of the system is improved

$$T_{\text{improved}} = \frac{T_{\text{affected}}}{\text{improvement factor}} + T_{\text{unaffected}}$$

- How much faster must the multiplier be to get a 2x performance improvement overall if multiples account for 20 seconds of the 100 second run time?

- Corollary: Make the common case fast

Pitfall: Using MIPS as a Performance Metric

- MIPS: Millions of Instructions Per Second
- Faster machines have higher MIPS rating

$$\text{MIPS} = \frac{\text{Instruction count}}{\text{Execution time} \times 10^6}$$
$$= \frac{\text{Instruction count}}{\frac{\text{Instruction count} \times \text{CPI}}{\text{Clock rate}} \times 10^6} = \frac{\text{Clock rate}}{\text{CPI} \times 10^6}$$

  - What companies advertise (and easy to calculate), but doesn't account for
    - Differences in ISAs between computers
    - Differences in complexity between instructions

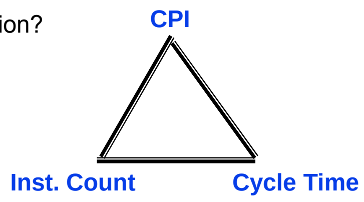
- MIPS varies between programs on the same computer (thus, a computer cannot have a single MIPS rating) !

Summary: Evaluating ISAs

- Design-time Metrics
  - Can it be implemented, in how long, at what cost (size, power)?
  - Can it be programmed? Ease of compilation?
- Static Metrics
  - How many bytes does the program occupy in memory?
- Dynamic Metrics
  - How many instructions are executed? How many bytes does the corefetch to execute the program?
  - How many clocks are required per instruction?
  - How "lean" (fast) a clock is practical?

Best Metric: Time to execute the program!

depends on the instructions set, the processor organization, and compilation techniques.



Next Week's Material and Reminders

- Next week
  - MIPS ISA review and MIPS arithmetic
    - Reading assignment – PH, Chapters 2 and 3
- TA will schedule a SimpleScalar evening tutorial session(s) in the lab, watch Angel for details
- Reminders
  - HW1 due Sep 3rd
  - Qz1 will come on line Aug 27th and close Sep 7th
  - First evening midterm exam scheduled
    - Tuesday, October 6th, 20:15 to 22:15, Location 22 Deike
    - Please let me know ASAP (via email) if you have a conflict