Why do we worry about performance?

As a consumer:

An application might need a certain level of performance (DOOM) - want to be assured that your machine can handle it (here interested in absolute performance),

would like to maximize the use of your dollar (here interested in performance/$).

As a company selling computer systems:

You compete on the basis of absolute performance, performance per unit price, and increasingly, performance per unit power consumption.

As a designer:

Faced with many design alternatives to meet design objectives. The way to choose the best design alternative is to analyze and measure each one with regards to performance, cost, and power consumption.

How do we characterize computer performance?
Two Notions of “Performance”

<table>
<thead>
<tr>
<th>Plane</th>
<th>DC to Paris</th>
<th>Top Speed</th>
<th>Passengers</th>
<th>Throughput (passengers x mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boeing 747</td>
<td>6.5 hours</td>
<td>610 mph</td>
<td>470</td>
<td>286,700</td>
</tr>
<tr>
<td>BAD/Sud Concorde</td>
<td>3 hours</td>
<td>1350 mph</td>
<td>132</td>
<td>178,200</td>
</tr>
</tbody>
</table>

Which has higher performance?

Interested in time to deliver 1 passenger?
Or, interested in delivering as many passengers per day as possible?

In a computer, time for one task called

**Response Time** or **Execution Time**

In a computer, tasks per unit time called

**Throughput** or **Bandwidth**
Definitions

Performance is characterized as the number of actions per unit time (throughput) or as time per action (response time).

“F is n times faster than S” means:

\[ n = \frac{\text{performance}(F)}{\text{performance}(S)} \]

If we are concerned with response time:

\[ \text{performance}(x) = \frac{1}{\text{execution\_time}(x)} \]

and, “F is n times faster than S” means:

\[ n = \frac{\text{execution\_time}(S)}{\text{execution\_time}(F)} \]
Example of Response Time vs. 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)

In cs61c we will focus primarily on response time not throughput.
Confusing Wording on Performance

• We will use “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)
Measuring performance

• 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 its hard to do
  • Don’t have access to machine to “benchmark” before purchase
  • Don’t know workload in future
Benchmarks

- Obviously, apparent speed of processor depends on code used to test it

- Need industry standards so that different processors can be fairly compared

- Companies exist that create these benchmarks: “typical” code used to evaluate systems

- Need to be changed every 2 or 3 years since designers could (and do!) target for these standard benchmarks
Example Standardized Benchmarks (1/2)

- Standard Performance Evaluation Corporation (SPEC) SPEC CPU2000
  - SPECINT2000 12 integer (gzip, gcc, crafty, perl, ...)
  - SPECFP2000 14 floating-point (swim, mesa, art, ...)
    www.spec.org/osg/cpu2000/

- They measure
  - System speed (SPECint2000)
  - System throughput (SPECint_rate2000)

- Benchmarks distributed in source code
- Big Company representatives select workload
  - Sun, HP, IBM, etc.
### SPEC CPU2000 - SPECfp and SPECint example results

Absolute Values and Values Relative to the IBM eServer pSeries 690Turbo (1.3 GHz CPU).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Machine</th>
<th>SPECfp</th>
<th>SPECint</th>
<th>Test</th>
<th>Relative Values (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Date</td>
<td>SPECfp</td>
</tr>
<tr>
<td>1</td>
<td>HP Integrity RX2600/Itanium2,1500MHz</td>
<td>2119.0</td>
<td>1322.0</td>
<td>Jun-03</td>
<td>167%</td>
</tr>
<tr>
<td>2</td>
<td>HP Integrity RX5670/Itanium2,1500MHz</td>
<td>2108.0</td>
<td>1312.0</td>
<td>Jun-03</td>
<td>167%</td>
</tr>
<tr>
<td>3</td>
<td>HP ZX6000/Itanium2, 1.5GHz,6MB-L3, H</td>
<td>2106.0</td>
<td>1315.0</td>
<td>Jun-03</td>
<td>166%</td>
</tr>
<tr>
<td>4</td>
<td>SGI Altix 3000 (1.5GHz, Itanium 2)</td>
<td>2100.0</td>
<td>1243.0</td>
<td>Jul-03</td>
<td>166%</td>
</tr>
<tr>
<td>5</td>
<td>NovaScale 4040 Itanium2/1500</td>
<td>2015.0</td>
<td>1107.0</td>
<td>Jul-03</td>
<td>159%</td>
</tr>
<tr>
<td>6</td>
<td>Dell PowerEdge 3250/1.5 GHz Itanium2</td>
<td>1875.0</td>
<td>1099.0</td>
<td>Jul-03</td>
<td>148%</td>
</tr>
<tr>
<td>7</td>
<td>ION Computers I2X2 (1.4GHz Itanium2)</td>
<td>1817.0</td>
<td>926.0</td>
<td>Jul-03</td>
<td>144%</td>
</tr>
<tr>
<td>8</td>
<td>SGI Altix 3000 (1.3GHz, Itanium 2)</td>
<td>1783.0</td>
<td>875.0</td>
<td>Jun-03</td>
<td>141%</td>
</tr>
</tbody>
</table>
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/
Beyond Benchmarks

• Benchmarks are good tools for empirically measuring computer system performance.

• Need more analysis to understand what factors contribute to performance and how to effect them.

• Start with “clock rate”:
  • Processor manufactures compete with clock-rate, usually in the popular press.
  • Unfortunately its not the whole story, in fact sometimes it’s the wrong story.
Clock Rate

- Computers (and other digital systems) are 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, cycles, or ticks**).
  - 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.
  - Typical clock period these days for high-end processors is <1ns for >1GHz rate.
Measuring Time using Clock Cycles (1/2)

CPU execution time for program

\[ \text{CPU execution time} = \text{Clock Cycles for a program} \times \text{Clock Cycle Time} \]

or

\[ \text{CPU execution time} = \frac{\text{Clock Cycles for a program}}{\text{Clock Rate}} \]
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)

**CPU time** = \( \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Cycle}} \)

- **CPU time** = \( \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Cycle}} \)
- **CPU time** = \( \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Cycle}} \)
- **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: \( \frac{\text{Execution Time}}{\text{Clock cycle time}} \) \( \frac{\text{Instruction Count}}{} \)  
  • 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)

<table>
<thead>
<tr>
<th>Op</th>
<th>Freq$_i$</th>
<th>CPI$_i$</th>
<th>Prod</th>
<th>(% Time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALU</td>
<td>50%</td>
<td>1</td>
<td>.5</td>
<td>(23%)</td>
</tr>
<tr>
<td>Load</td>
<td>20%</td>
<td>5</td>
<td>1.0</td>
<td>(45%)</td>
</tr>
<tr>
<td>Store</td>
<td>10%</td>
<td>3</td>
<td>.3</td>
<td>(14%)</td>
</tr>
<tr>
<td>Branch</td>
<td>20%</td>
<td>2</td>
<td>.4</td>
<td>(18%)</td>
</tr>
</tbody>
</table>

**Instruction Mix** (Where time spent)

- What if Branch instructions twice as fast?
Back to CPU time Formula

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

From the point of view of a computer designer, what factors influence each of these terms?
It’s the Formula that changed the world!

• Computer Architecture Circa 1980: Digital Equipment Corp. VAX

• Very elaborate instruction set:
  
  OPCODE operand1 operand2 …

  • Operands can be in memory or registers and accessed through a variety (~20) addressing modes.

  • Any addressing mode could be used for any operand for any opcode.

• Variable length instruction encoding:
  
  - Example: integer add: 3-19 bytes
VAX – a CISC (complex instruction set)

• Example addressing modes:
  Register          r4
  Base (displacement) [r4 + offset]
  Immediate         0xffff0101
  Pc-relative        [pc] + offset
  Deferred (indirect) [ [r3 + offset] ]
  Index (scaled)    [r3 + r4*8]
  Auto-increment    (r4)+

• Example opcodes:
  push              push an item onto a stack
  insque            insert an item onto a queue
  aobleq op1 op2    increment op1 and branch if op1=op2
  Special call/ret  handle arguments, stack setup
  poly             take x and a pointer to a set of coefficients, computes a + bx + cx^2 + dx^3+…
RISC (reduced instruction set computer)

Looked like this was going to be the future of computer architecture…. until, armed with “the formula”, Patterson and others argued for RISC.

• Contrast VAX to MIPS:
  • Fixed length instruction encodings
  • One memory addressing mode
  • Memory addressing limited to ld/st
  • Limited opcodes

• Overall the RISC philosophy is to optimize the machine for the commonly used instructions and emulate the others if needed.
CISC implementation issues

- The complex instruction set meant lots of complex circuits ⇒ slow clock.
- The complex instructions have high CPI.
  - Inherently many steps
  - Difficult to reduce because of complexity
- The potential for low instruction/program. However, in practice difficult for a compiler to take advantage of all the complexity.
- Therefore, versus RISC, higher sec/cyc, much higher CPI, somewhat lower inst/program not enough to make up the difference ⇒ RISC wins out.
- All new architectures since then have been RISC.

CPU time = Instructions per Program x Cycles per Instruction x Seconds per Cycle
“And in conclusion…” 1/2

- Latency v. Throughput

- Performance doesn’t depend on any single factor: need to know Instruction Count, Clocks Per Instruction and Clock Rate to get valid estimations

$$\text{CPU time} = \frac{\text{Instructions}}{\text{Program}} \times \frac{\text{Cycles}}{\text{Instruction}} \times \frac{\text{Seconds}}{\text{Cycle}}$$
“And in conclusion…” 2/2

• 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
### Bonus Slide: Winstone 99 (W99) Results

<table>
<thead>
<tr>
<th>Company</th>
<th>Processor</th>
<th>Price</th>
<th>Clock</th>
<th>W99</th>
</tr>
</thead>
<tbody>
<tr>
<td>emachines</td>
<td>Cyrix MII</td>
<td>$653</td>
<td>250</td>
<td>14.5</td>
</tr>
<tr>
<td>CompUSA</td>
<td>Intel Celeron</td>
<td>$764</td>
<td>400</td>
<td>18.0</td>
</tr>
<tr>
<td>Compaq</td>
<td>AMD K6-2</td>
<td>$902</td>
<td>350</td>
<td>15.4</td>
</tr>
<tr>
<td>HP</td>
<td>Intel Celeron</td>
<td>$1,070</td>
<td>366</td>
<td>17.6</td>
</tr>
<tr>
<td>Compaq</td>
<td>AMD K6-2</td>
<td>$1,453</td>
<td>450</td>
<td>17.9</td>
</tr>
<tr>
<td>Compaq</td>
<td>AMD K6-3</td>
<td>$1,479</td>
<td>400</td>
<td>22.3</td>
</tr>
<tr>
<td>HP</td>
<td>Intel Pentium II</td>
<td>$1,483</td>
<td>400</td>
<td>18.9</td>
</tr>
<tr>
<td>NEC</td>
<td>Intel Pentium III</td>
<td>$1,680</td>
<td>400</td>
<td>22.0</td>
</tr>
</tbody>
</table>

- **Note:** 2 Compaq Machines using K6-2 v. 6-3:
  - **K6-2 Clock Rate is 1.125 times faster**, but
  - **K6-3 Winstone 99 rating is 1.25 times faster!**