CSCI 4717/5717
Computer Architecture

Topic: Performance
Reading: Stallings, Section 2.2

Performance from User's Point of View

Types of applications that require performance:

- Image processing
- Handwriting and speech recognition
- Video conferencing
- Multimedia development
- Multimedia playback
- Simulations
- Artificial intelligence

Real-World Applications

- Gaming/entertainment
- Weather forecasting
- Oceanography
- Seismic/petroleum exploration
- Medical research and diagnosis
- Aerodynamics and structure analysis
- Nuclear physics
- Military/defense
- Interfaces for disabled
- Socio-economics

Original Architecture

- Basic building blocks are the same as IAS computer from 60 years ago.
- Not one component, however, has been left unexamined in terms of squeezing out more performance.
- Design and implementation has become extremely sophisticated.
- This course examines techniques for achieving maximum performance

Measuring performance

- The benefits of a new or modified design cannot be determined without having a way to measure the difference
- An increase in a machine's performance is viewed in one of two (competing) ways:
  - Reduced response time to an individual job "do stuff faster"
  - Increase in overall throughput "do more stuff"

Other measures of performance

- Cost
  - Cost of designing SW
  - Purchase cost of hardware
  - Purchase of components such as peripherals
- Compatibility
- S/W availability
- Maintainability

Effect of Improved Technology

Of the following technological improvements, which increases throughput, reduces response time, or both?

- Faster clock cycle time
- Multiple processors for separate tasks
- Parallel processing of array or vector-type problems

Effects of Moore's Law

The doubling of the number of transistors on a single chip every 18 months has had some effects on the application of technology:

- Costs have fallen dramatically since chip prices have not changed substantially since Moore made his prediction
- Tighter packaging has allowed for shorter electrical paths and therefore faster execution
- Smaller packaging has allowed for more applications in more environments
- Reduction in power and cooling requirements which also helps with portability
- Solder connections are not as reliable, therefore, with more functions on a single chip, there are fewer unreliable solder connections

Effects of Moore's Law (continued)

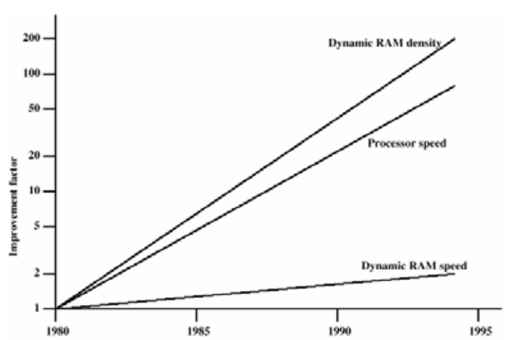
As technology allows for higher levels of performance, processor designers must come up with ways to use it.

- Keeping all parts of the processor busy
  - Coordinating multiple pipelines
  - Improved branch prediction
- Parallel processors
- Optimizing execution
  - Real-time analysis of code to "re-order"execution
  - Speculative execution of code
- Incorporating multiple functions on single chip

Performance Mismatch

- Experienced significant improvement
  - Processor speed
  - Memory capacity
- Experienced only minor improvement
  - Memory speed
  - Bus rates
  - I/O device performance

DRAM and Processor Characteristics



Effects of Performance Mismatch

- Processor stalls – "wait states"
- Fewer DRAMs are needed per system reducing opportunity for parallel transfers
- I/O device performance improvements are offset by greater demands, e.g., video capture.

DRAM Solutions

- Increase number of bits retrieved at one time
  - Make DRAM “wider” rather than “deeper”
- Change DRAM interface
  - Add third level of cache
- Reduce frequency of main memory access
  - More complex cache and cache on chip
- Increase interconnection bandwidth
  - High speed buses
  - Hierarchy of buses

I/O Solutions

- Caching and buffering schemes
- Higher speed interfaces
- Distributed processors
- Imposing physical restrictions on peripherals
  - Distance
  - Number of devices on a bus

Changes Affect Entire System

Design is more than making a component go faster. Designer must also:

- Assess how the change affects the system as a whole
- Investigate a wider number of performance measurements, i.e., be careful when using narrowly defined test/benchmark data

Johnson City to New York City

- Walking (3 miles/hour) -- Distance: 620.11 miles
  Estimated Time: 8 days, 14 hours, 40 minutes
- Bicycle -- Total Distance: 620.11 miles
  Estimated Time: 3 days, 5 hours, 30 minutes
- Bus -- 09:35p to 11:20a
  Estimated Time: 13 hours, 45 minutes
- Driving -- Distance: 620.11 miles
  Estimated Time: 10 hours, 43 minutes
- Flying -- TRI 6:10 am to Charlotte, then NC to LaGuardia 10:47 am
  Estimated Time: 3 hours and 42 minutes
- Drive to Charlotte (166.92 miles; Estimated Time: 4 hours), Fly from Charlotte 7:50 am to LaGuardia 9:47 am (1 hour and 57 minutes)

Other considerations

- Car rental in NY
- Stress of driving that long
- Parking your own car in NY
- Differences in ticket prices
- Fear of flying
- Parking fees at airports

Computer Example

- Which is faster for a short (1K to 5K) data transfer, 56K modem or 2400 BAUD modem?
- Issues to consider
  - Data transfer time
  - Synchronization
  - Disconnect time

Computer Example (continued)

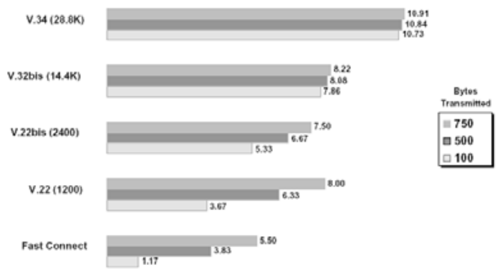


Image source: Arc Electronics, "Quick Connection modems," on-line: http://www.arcelect.com/Transaction_based_quick_connect_modem_tutorial.htm, last visited: August 31, 2005

In-Class Exercise

- So when it comes to computer performance, what are we going to measure?
- Break into groups and brainstorm on the following performance issues:
  - What items should we measure?
  - What are the units of these measures?
  - What applications rely on these measures?
  - What laboratory methods can we use to take these measures?

Memory Access Times

- Measured in nanoseconds or bandwidth (bits per second)
- Write speed (time to reliably store data)
- Read speed (time to reliable retrieve data)

Instruction Execution - MIPS

- MIPS – Millions of instructions per second
- Affected by two things
- How many cycles it takes to complete an instruction
- Clock rate, from which cycle duration is calculated
- Advantage -- this measurement can be used to determine the speed of your program since you know how many instructions each part of your program contains

Problems with MIPS

- Comparing different processors (e.g., RISC machines to non-RISC machines) is useless since measurement is based on instruction set
- Different instructions use different number of cycles
- Example: Using floating point instructions has a lower MIPS rating than using a floating point function based on integer instructions

Instruction Execution – MFLOPS

- MFLOPS – Millions of floating point operations per second
- Better than MIPS only when primary application is something that requires a great deal of floating point instructions
- Still not completely balanced between different processor architectures