Introduction

Hwansoo Han
The Computer Revolution

- Progress in computer technology
  - Underpinned by Moore’s Law

- Makes novel applications feasible
  - Computers in automobiles
  - Cell phones
  - Human genome project
  - World Wide Web
  - Search Engines

- Computers are pervasive
What You Will Learn

- How programs are translated into the machine language and how the hardware executes them
- The hardware/software interface
- What determines program performance and how it can be improved
- How hardware designers improve performance
- What is parallel processing
Understanding Performance

- **Algorithm**
  - Determines number of operations executed
- **Programming language, compiler, architecture**
  - Determine number of machine instructions executed per operation
- **Processor and memory system**
  - Determine how fast instructions are executed
- **I/O system (including OS)**
  - Determines how fast I/O operations are executed
Below Your Program

- **Application software**
  - Written in high-level language

- **System software**
  - Compiler: translates HLL code to machine code
  - Operating System: service code
    - Handling input/output
    - Managing memory and storage
    - Scheduling tasks & sharing resources

- **Hardware**
  - Processor, memory, I/O controllers
Levels of Program Code

- **High-level language**
  - Level of abstraction closer to problem domain
  - Provides for productivity and portability

- **Assembly language**
  - Textual representation of instructions

- **Hardware representation**
  - Binary digits (bits)
  - Encoded instructions and data
Components of a Computer

- Same components for all kinds of computer
  - Desktop, server, embedded
- Input/output includes
  - User-interface devices
    - Display, keyboard, mouse
  - Storage devices
    - Hard disk, CD/DVD, flash
- Network adapters
  - For communicating with other computers
Opening the Box

Hard drive  Processor  Fan with cover  Spot for memory DIMMs  Spot for battery  Motherboard  Fan with cover  DVD drive
Inside the Processor (CPU)

- Datapath
  - Performs operations on data

- Control
  - Sequences datapath, memory, ...

- Cache memory
  - Small fast SRAM memory for immediate access to data
Inside the Processor

- AMD Barcelona: 4 processor cores
Abstractions in Computer Systems

- Abstraction helps us deal with complexity
  - Hide lower-level detail

- Instruction set architecture (ISA)
  - The hardware/software interface

- Application binary interface (ABI)
  - The ISA plus system software interface

- Implementation
  - The details underlying and interface
A Safe Place for Data

- Volatile main memory
  - Loses instructions and data when power off
- Non-volatile secondary memory
  - Magnetic disk
  - Flash memory
  - Optical disk (CDROM, DVD)
Networks

- Communication and resource sharing
  - System area network (SAN): Infiniband
  - Local area network (LAN): Ethernet - within a building
  - Wide area network (WAN): the Internet

- Wireless network: WiFi, Bluetooth
Manufacturing ICs

- **Yield**
  - Proportion of working dies per wafer

![Diagram of IC manufacturing process]

1. Silicon ingot
2. Slicer
3. Blank wafers
4. 20 to 40 processing steps
5. Patterned wafers

- Bond die to package
- Tested dies
- Tested wafer
- Wafer tester
- Packaged dies
- Tested packaged dies
- Part tester
- Ship to customers
AMD Opteron X2 Wafer

- X2: 300mm wafer, 117 chips, 90nm technology
- X4: 45nm technology
Integrated Circuit Cost

Cost per die = \frac{\text{Cost per wafer}}{\text{Dies per wafer} \times \text{Yield}}

\text{Dies per wafer} \approx \frac{\text{Wafer area}}{\text{Die area}}

\text{Yield} = \frac{1}{(1 + (\text{Defects per area} \times \frac{\text{Die area}}{2}))^2}

- Nonlinear relation to area and defect rate
  - Wafer cost and area are fixed
  - Defect rate determined by manufacturing process
  - Die area determined by architecture and circuit design
Uniprocessor Performance

Constrained by power, instruction-level parallelism, memory latency

- Architecture-driven
- Technology-driven
Contributor 1: Technology

- **Processor**
  - Logic capacity: about 30% per year
  - Clock rate: about 20% per year

- **Memory**
  - DRAM capacity: about 60% per year (4x every 3 years)
  - Memory speed: about 10% per year
  - Cost per bit: improves about 25% per year

- **Disk**
  - Capacity: about 60% per year
Technology-driven Improvement (1)

- Moore's law
  - The number of transistors per integrated circuit would double every 18 months

Gordon Moore (co-founder of Intel)
## Technology-driven Improvement (2)

- Electronics technology continues to evolve
  - Increased capacity and performance
  - Reduced cost

### Technology Evolution

<table>
<thead>
<tr>
<th>Year</th>
<th>Technology</th>
<th>Relative performance/cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>1951</td>
<td>Vacuum tube</td>
<td>1</td>
</tr>
<tr>
<td>1965</td>
<td>Transistor</td>
<td>35</td>
</tr>
<tr>
<td>1975</td>
<td>Integrated circuit (IC)</td>
<td>900</td>
</tr>
<tr>
<td>1995</td>
<td>Very large scale IC (VLSI)</td>
<td>2,400,000</td>
</tr>
<tr>
<td>2005</td>
<td>Ultra large scale IC</td>
<td>6,200,000,000</td>
</tr>
</tbody>
</table>

### DRAM Capacity

![DRAM Capacity Chart](chart.png)

- **Year of introduction:**
  - 1976: 16K
  - 1979: 64K
  - 1983: 256K
  - 1986: 1M
  - 1988: 4M
  - 1993: 16M
  - 1996: 64M
  - 1998: 256M
  - 2002: 512M
  - 2008: 1G

- **Kbit capacity:**
  - 1976: 16K
  - 1979: 64K
  - 1983: 256K
  - 1986: 1M
  - 1988: 4M
  - 1993: 16M
  - 1996: 64M
  - 1998: 256M
  - 2002: 512M
  - 2008: 1G
In CMOS IC technology

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

- \times 40
- \times 1000
- 5V → 1V
Reducing Power

- Suppose a new CPU has
  - 85% of capacitive load of old CPU
  - 15% voltage and 15% frequency reduction

\[
\frac{P_{\text{new}}}{P_{\text{old}}} = \frac{C_{\text{old}} \times 0.85 \times (V_{\text{old}} \times 0.85)^2 \times F_{\text{old}} \times 0.85}{C_{\text{old}} \times V_{\text{old}}^2 \times F_{\text{old}}} = 0.85^4 = 0.52
\]

- The power wall
  - We can’t reduce voltage further
  - We can’t remove more heat
  - How else can we improve performance?
Contributor 2: Architecture

- Exploiting parallelism in processing
  - Instruction level parallelism (ILP)
    - Pipelining
    - Superscalar
    - Out of order execution
    - Branch prediction
    - VLIW (software approach)
  - Data level parallelism (DLP)
    - SIMD instructions (media processing)
  - Task level parallelism (TLP)
    - Simultaneous multithreading (SMT) - hyperthreading
    - Multicore (multi-processors)

- Improving latency and capacity in memory system
  - Low latency access using cache memory
  - Capacity increase in main memory
Superscalar

- Multiple functional units
  - Multiple integer units
  - Multiple floating point units
Multiprocessors

- Multicore microprocessors
  - More than one processor core per chip

- Requires explicitly parallel programming
  - Instruction level parallelism implements parallelism implicitly
    - Hardware executes multiple instructions at once
    - Hidden from the programmer
  - Explicit parallel programming is hard to do
    - Programming for performance
    - Load balancing
    - Optimizing communication and synchronization
How Do Computers Work?

- Need to understand abstractions such as:
  - Applications software
  - Systems software
  - Assembly Language
  - Machine Language
  - Architectural Issues: i.e., Caches, Virtual Memory, Pipelining
  - Sequential logic, finite state machines
  - Combinational logic, arithmetic circuits
  - Boolean logic, 1s and 0s
  - Transistors used to build logic gates (CMOS)
  - Semiconductors/Silicon used to build transistors
  - Properties of atoms, electrons, and quantum dynamics

- So much to learn!
Historical Perspective

- ENIAC built in World War II
  - The first general purpose computer by John Mauchly & Presper Eckert
  - Used for computing artillery firing tables
  - 80 feet long by 8.5 feet high and several feet wide
  - Each of the twenty 10 digit registers was 2 feet long
  - Used 18,000 vacuum tubes
  - Performed 1900 additions per second

Moore’s Law:
Transistor capacity doubles every 18-24 months