Introduction

Jin-Soo Kim (jinsookim@skku.edu)
Computer Systems Laboratory
Sungkyunkwan University
http://csl.skku.edu
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
Classes of Computers

- **Desktop computers**
  - General purpose, variety of software
  - Subject to cost/performance tradeoff

- **Server computers**
  - Network based
  - High capacity, performance, reliability
  - Range from small servers to building sized

- **Embedded computers**
  - Hidden as components of systems
  - Stringent power/performance/cost constraints
The Processor Market

![Graph showing the market growth of Cell Phones, PCs, and TVs from 1997 to 2007.](image-url)
Components of a Computer

- **Processor (Datapath/Control), Memory, I/O**
- **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
Modern PC Architecture

Intel® G45 Express Chipset Block Diagram

- Intel® Graphics Media Accelerator X4500HD
  - Intel® Clear Video Technology
    - Full Hardware HD Decode
    - H.264, VC1, MPEG-2
    - HD/SD video post processing
    - Display support: HDMI*, DVI
    - DisplayPort*, HDCP, MEC
  - DirectX® 10 and OpenGL® 2.0
  - API support

- Intel® Core™2 Duo Processor
- Intel® Core™2 Quad Processor

FSB @ 1333MT/s, 64bits

Dual-channel DDR3 @ 1066MHz
64GB support
Max 16GB/s

- DDR2 or DDR3
  - 6.4 GB/s or 8.5 GB/s
- PCIe® x16
- PCI Express® 2.0 Graphics

- 12 Hi-Speed USB 2.0 Ports;
  - Dual EHCI; USB Port Disable
- 6 PCI Express® x1

- Intel® Integrated
  - 10/100/1000 MAC
  - GLC | LLI
- Intel® Gigabit LAN Connect

- 480 Mb/s
  - each
- 500 MB/s
  - each x1

G45 GMCH

ICH10
ICH10R

LPC or SPI

BIOS Support

Intel® High Definition Audio

Intel® Quiet System Technology

6 Serial ATA Ports; eSATA;
Port Disable

Intel® Matrix Storage Technology

Intel® Turbo Memory
with User Pinning

Optional
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
A Safe Place for Data

- **Volatile main memory**
  - Loses instructions and data on power off

- **Non-volatile secondary memory**
  - Magnetic disk
  - Flash, Solid state drives (SSDs)
  - Optical disk (CDROM, DVD)
Networks

- Communication and resource sharing
  - Local area network (LAN): Ethernet
    - Within a building
  - Wide area network (WAN): the Internet
  - Wireless network: WiFi, WiBro, Bluetooth, HSDPA, ...
From Sand to Circuits

Purified silicon ingot

Sand

Slicer

Blank wafers

20 to 40 processing steps

Yield: proportion of working dies per wafer

Patterned wafers

Wafer tester

Tested wafer

Dicer

Tested dies

Bond die to package

Packaged dies

Part tester

Tested packaged dies

Ship to customers

ICE3003: Computer Architecture | Fall 2009 | Jin-Soo Kim (jinsookim@skku.edu)
Processing ICs

1. Start with a partially processed die on a silicon wafer.
2. Deposit oxide layer.
3. Coat with photoresist.
5. Rinse with solvent.
6. Etch with acid.
7. Remove remaining photoresist.

Source: Intel
AMD Opteron X2 Wafer

- 12-inch (300mm) wafer, 117 chips, 90nm

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

Dies per wafer \(\approx \) Wafer area/Die area

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

- 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
Contributor 1: Technology (1)

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

- **Disk**
  - Capacity: about 60% per year

<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>
Memory

- DRAM capacity: about 60% per year (4x every 3 years)
- Memory speed: about 10% per year
Moore’s law (1965)

CPU Transistor Counts 1971-2008 & Moore’s Law

Curve shows Moore’s Law: transistor count doubling every two years

“The number of transistors incorporated in a chip will approximately double every 24 months.”
Gordon Moore, Intel Co-founder
Gordon Moore estimated in 2003 that the number of transistors shipped in a year had reached about 10,000,000,000,000 (10^12). That's about 100 times the number of ants estimated to be in the world.

On the road to a billion transistors per chip, Intel has developed transistors so small that about 200 million of them could fit on the head of each of these pins.

The price per transistor on a chip has dropped dramatically since Intel was founded in 1968. Some people estimate that the price of a transistor is now about the same as that of one printed newspaper character.

In 1978, a commercial flight between New York and Paris cost around $100 and took seven hours. If the principles of Moore's Law had been applied to the airline industry the way they have to the semiconductor industry since 1978, that flight would now cost about a penny and take less than one second.

A chip-making tool under development superimposes magnetically levitated images within a tolerance of ±10,000 the thickness of a human hair — a feat equivalent to driving a car straight for 400 miles while deviating less than one inch.

Because electricity travels a shorter distance in a smaller transistor, smaller transistors mean faster chips. It would take you about 25,000 years to turn a light switch on and off 1.5 trillion times. But Intel has developed transistors that can switch on and off that many times each second.

Source: Intel
Power trends

- In CMOS IC technology,

\[
\text{Power} = \text{Capacitive load} \times \text{Voltage}^2 \times \text{Frequency}
\]
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?
Exploiting parallelism (single processor)

- Instruction level parallelism (ILP)
  - Pipelining
  - Superscalar
  - Out of order execution
  - Branch prediction
  - VLIW (Very Long Instruction Word)

- Data level parallelism (DLP)
  - SIMD instructions (media processing)

- Task level parallelism (TLP)
  - Simultaneous multithreading (Hyperthreading)
  - Multicore (multi-processor)
Superscalar processor examples

- Multiple functional units
Latency and capacity in memory systems

- Low latency access using cache memory
- Capacity increase in main memory
- MMU (Memory Management Unit)


**Multiprocessors**

- **Multicore microprocessors**
  - More than one processor per chip

- **Requires explicitly parallel programming**
  - Compare with instruction level parallelism
    » Hardware executes multiple instructions at once
    » Hidden from the programmer
  - Hard to do
    » Programming for performance
    » Load balancing
    » Optimizing communication and synchronization

- **Think Parallel or Perish!**
Below Your Program

- **Application software**
  - Written in high-level language (HLL)

- **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
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
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, electronics, and quantum dynamics

- **So much to learn!**
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