Introduction

Reading

- Sections 1.1, 1.2, 1.3, 1.4, 1.5, 1.7, 1.8

- Key Ideas
  - Types of systems → implications for computer architecture
  - The impact of technology
  - New rules → the power wall and parallelism
Historical Perspective

- ENIAC built in World War II was the first general purpose computer
  - 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

  Since then, Moore’s Law:
  - Transistor density doubles every 18-24 months
  - Modern version:
    - #cores double every 18-24 months

The Modern Era

- Hundred’s of dollars
- Battery operated
- Internet capable

- Contrast with warehouse scale computing, e.g., Google and Amazon
  - Software as a service
  - Backend for mobile devices
  - Power consumption limited

Internet of Things?
Opening the Box

- Capacitive multitouch LCD screen
- 3.8 V, 25 Watt-hour battery
- Computer board

Inside the Processor

- Apple A5
Inside the Core (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

Reminder

- High-level language
  - Level of abstraction closer to problem domain

- Assembly language
  - Textual representation of instructions

- Hardware representation
  - Encoded instructions and data

How does this work?
Moore's Law

Microprocessor Transistor Counts 1971-2011 & Moore's Law

Goal: Sustain Performance Scaling

Uniprocessor Performance

§1.8 The Sea Change: The Switch to Multiprocessors

Constrained by power, instruction-level parallelism, memory latency
Feature Size

We are currently at 0.032 µm and moving towards 0.022 µm

New Rules: The End of Dennard Scaling

- Voltage is no longer scaling at the same rate
- Slower scaling in power per transistor → increasing power densities

Post Dennard Performance Scaling

\[
\text{Perf} \left( \frac{\text{ops}}{s} \right) = \text{Power} (W) \times \text{Efficiency} \left( \frac{\text{ops}}{\text{joule}} \right)
\]

W. J. Dally, Keynote IITC 2012

---

Power Wall

- In CMOS IC technology

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

\[
\times 30 \quad 5V \to 1V \quad \times 1000
\]
Memory Wall

- "Moore’s Law"
- Processor-Memory Performance Gap: (grows 50% / year)

3D Packaging

- New packaging technologies to increase processor-memory bandwidth
- What problems does this create?

Images from techweekeurope.co.uk
Understanding Cost

- What happens if you simply port a design across technology generations?
- What about design costs?
  - Hardware and software

Integrated Circuit Cost

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

Dies per wafer = \(\frac{\text{Wafer area}}{\text{Die area}}\)

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

- 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
Impact on Design

Tick-Tock Development Model:
Sustained Microprocessor Leadership

Source: Courtesy H.H. Lee, ECE 3055

Average Transistor Cost Per Year

Source:Courtesy H.H. Lee, ECE 3055
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
- Local area network (LAN): Ethernet
  - Within a building
- Wide area network (WAN: the Internet
- Wireless network: WiFi, Bluetooth
Parallelism

- Multicore microprocessors
  - More than one processor per chip

- 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

Multicore, Many Core, and Heterogeneity

- Performance scaling via increasing core count
- The advent of heterogeneous computing

Different instruction sets
Eight Great Ideas

- 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

Concluding Remarks

- New Rules
  - Power and energy efficiency are driving concerns
  - Cost is an exercise in mass production
    - Relationship to instruction set architecture (ISA)?
- Instruction set architecture (ISA)
  - The hardware/software interface is the vehicle for portability and cost management
- Multicore
  - Core scaling vs. frequency scaling
  - Need for parallel programming ➔ need to think parallel!
Study Guide

- Moore’s Law
  - What is it? What are the cost and performance consequences?
- Technology Trends
  - Explain the reason for the shift to power and energy efficient computing
- Understanding Cost
  - What are the major elements of cost?
- Multicore processor
  - Distinguishing features
- Basic Components of a Modern Processor

Glossary

- Energy efficiency
- Dennard Scaling
- Die yield
- Feature size
- Heterogeneity
- Moore’s Law
- Multicore architecture
- Memory Wall
- Performance scaling
- Parallel programming
- Power efficiency
- Power Wall
- Tick-tock development model
- Wafer