Introduction

Reading

• Sections 1.1, 1.2, 1.3, 1.4, 1.7, 1.8

• Key Ideas
  ❖ Types of systems → implications for computer architecture
  ❖ The impact of technology
  ❖ New rules → the power wall and parallelism
 Architecture

The Modern Era

- Hundred’s of $$
- 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

www.3g.co.uk  blogs.intel.com
Technology

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
Moore’s Law

Goal: Sustain Performance Scaling

Metric: Ops/sec

From wikipedia.org
Memory Wall

“Moore’s Law for Logic vs. Memory”

µProc 60%/yr.

Processor-Memory Performance Gap:
(grows 50% / year)

DRAM 7%/yr.

Time

Metrics:
Bandwidth: bytes/sec
Latency: secs

Moore's Law in Action

Cost Reduction On Track

\[ \frac{\text{Cost per Transistor}}{\text{Area}} \times \frac{\text{Area}}{\text{Transistors}} = \frac{\text{Cost}}{\text{Transistors}} \]
3D Packaging

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

Images from techweekeurope.co.uk

---

Energy

---
New Rules: The End of Dennard Scaling

- Voltage is no longer scaling at the same rate
  - Feature size
- Slower scaling in power per transistor \( \rightarrow \) 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

New Metrics
- Power efficiency
- MIPS/watt
- Energy Efficiency
- Ops/joule

• In CMOS IC technology

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

\[ \times 30 \quad 5V \rightarrow 1V \quad \times 1000 \]

Concurrency
Concurrency & Parallelism

• Multicore microprocessors
  ◦ More than one processor per chip

• Parallel programming
  ◦ Compare with instruction level parallelism
    o Hardware executes multiple instructions at once
    o Hidden from the programmer
  ◦ Hard to do
    o Programming for performance
    o Load balancing
    o Optimizing communication and synchronization

```assembly
try: add $t0,$zero,$s4 ; copy exchange value
    ll $t1,0($s1) ; load linked
    sc $t0,0($s1) ; store conditional
    beq $t0,$zero,try ; branch store fails
    add $s4,$zero,$t1 ; put load value in $s4
```

Multicore, Many Core, and Heterogeneity

NVIDIA Kepler

• Performance scaling via increasing core count

Intel Ivy Bridge

• The advent of heterogeneous computing

AMD Trinity
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
• Performance scaling
• Dennard Scaling
• Parallel programming
• Feature size
• Power efficiency
• Heterogeneity
• Power Wall
• Moore’s Law
• Multicore architecture
• Memory Wall

(22)