EEL 5764 Computer Architecture

Sandip Ray
Department of Electrical and Computer Engineering
University of Florida
sandip@ece.ufl.edu
http://sandip.ece.ufl.edu

Lecture 4-5:

- Technology Trends and Metrics
- Introduction to Memory Hierarchy

Announcements

Class Web page updated with schedules for homework, mid-terms, and projects

Homework Schedule

	Assigned	Due	Type
Problem Set 1	09/03/2019	09/10/2019	Not Graded
Homework 1	09/17/2019	09/24/2019	Graded
Homework 2	09/17/2019	09/24/2019	Graded
Problem Set 2	10/22/2019	10/29/2019	Not Graded
Homework 3	11/05/2019	11/12/2019	Graded
Homework 4	11/15/2019	11/22/2019	Graded

Project Schedule

- Topic Discussion: 09/19/2019
- 1-page Proposal Due: 09/26/2019
- Feedback: 10/04/2019
- Presentation: 11/18/2019 -- 11/22/2019
- Final Report: 12/06/2019

Exam Schedule

- In-class Exam 1: 10/22/2019
- In-class Exam 2: 12/03/2019

Announcements

- 4 graded homeworks, 2 additional problem sets for practice
  - → **Not required to turn in the additional problem sets**
  - → However, if you do turn it in and you're borderline between two grades I will consider your performance in the problem set to see if you can be moved to the higher grade
- Among the 4 graded homeworks, **the one with the lowest score will be dropped**
- Exams will be closed book and closed notes but you'll be allowed 1 both-sided crib sheet of notes of US letter size
- **Exam dates will not be moved.** If you cannot attend the exam for a specific reason you need to talk to me by next week with documentation for your reason

Technology Trends

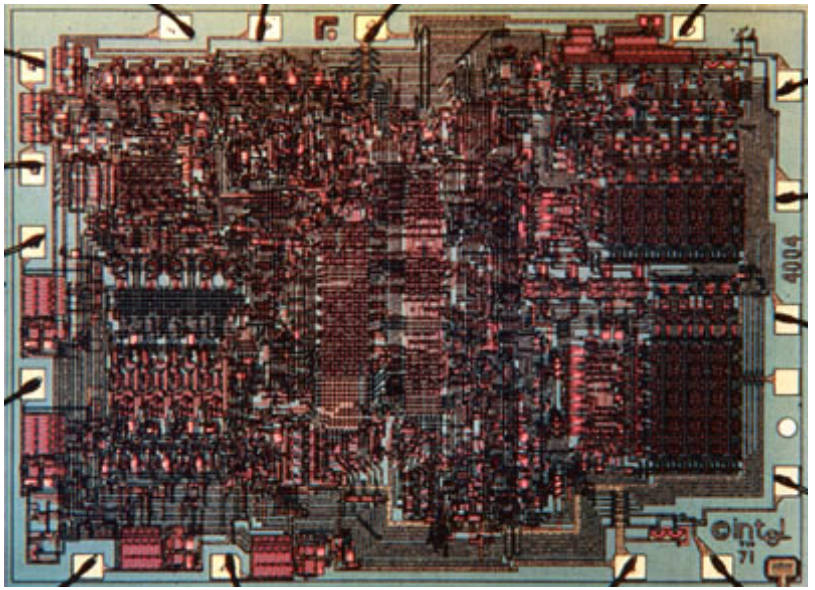
- Integrated circuit technology
  - → Transistor density: +35%/year
  - → Die size: +10-20%/year
  - → Integration overall: +40-55%/year
- DRAM capacity: +25-40%/year (slowing)
  - → Foundation of main memory.
- Flash capacity: +50-60%/year
  - → 15-20X cheaper/bit than DRAM
  - → An order of magnitude slower than DRAM
- Magnetic disk technology: +40%/year
  - → 15-25X cheaper/bit then Flash
  - → 300-500X cheaper/bit than DRAM
  - → Main storage for server or WSC.

Important to design for the next generation of technology!

Current technologies approaching their limits; new technologies being researched

First Microprocessor

- Intel 4004 (1971)
  - Application: calculators
  - Technology: 10000 nm
  - 2300 transistors
  - 13 mm^2
  - 108 KHz
  - 12 Volts
  - 4-bit data
  - Single-cycle datapath



Height of Single-Core Processor

- Intel Pentium4 (2003)
  - Application: desktop/server
  - Technology: 90nm (1/100th of 4004)

  - 55M transistors (20,000x)
  - 101 mm^2 (10x)
  - 3.4 GHz (10,000x)
  - 1.2 Volts (1/10th)

  - 32/64-bit data (16x)
  - 22-stage pipelined datapath
  - 3 instructions per cycle (superscalar)
  - Two levels of on-chip cache
  - data-parallel vector (SIMD) instructions, hyperthreading

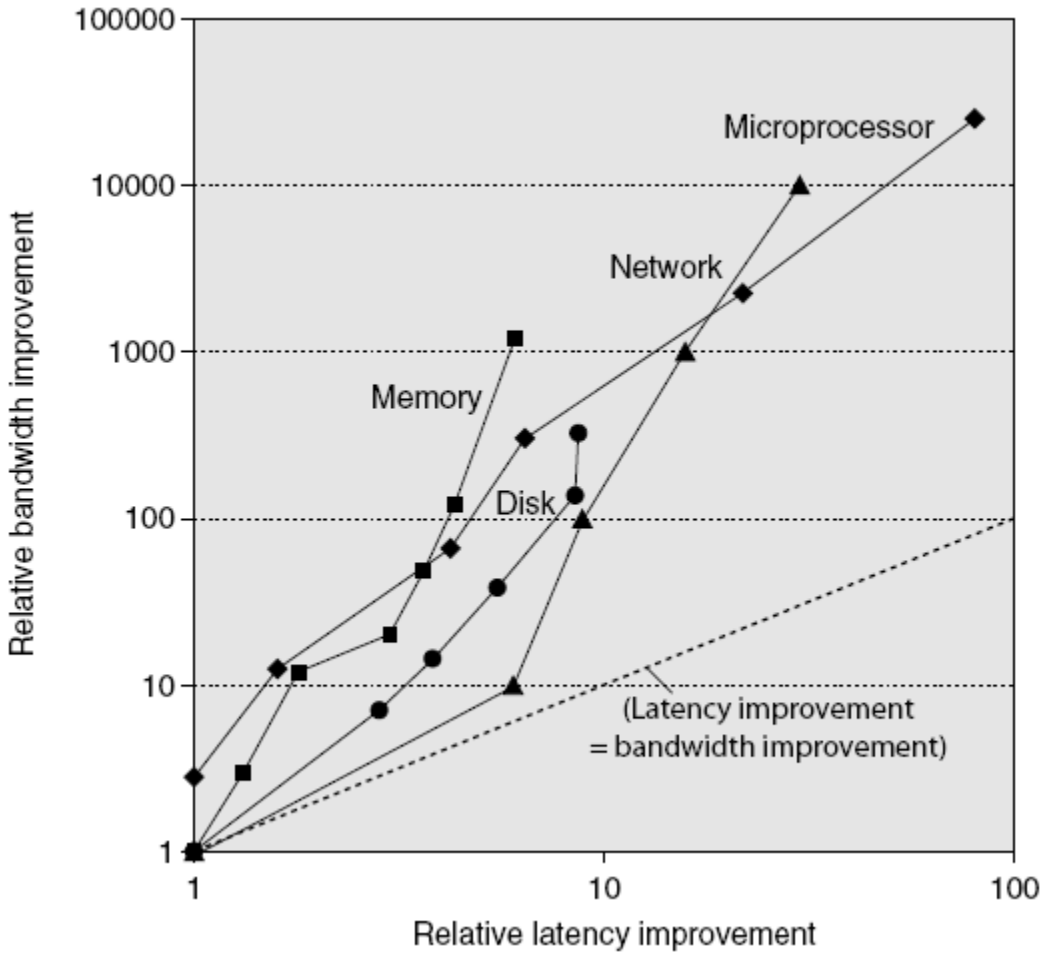
Transistors and Wires

- Transistor feature size
  - →Minimum size of transistor or wire in x or y dimension
  - →10 microns in 1971 to .032 microns in 2011
  - →Now in 2017/2018, seeing 10nm – 7nm.
- Transistor density grows exponentially.
  - →Moore's law – has been slowing down
- Transistor performance scales linearly
- Wire delay does not improve with feature size!
  - →In fact, it is getting worse!
  - →Make on-chip interconnect design an important task!

Bandwidth and Latency

- Bandwidth or throughput
  - →Total work done in a given time
  - →Important for servers and data center operators
- Latency or response time
  - →Time between start and completion of an event
  - →Important for individual users

Bandwidth and Latency



Log-log plot of bandwidth and latency milestones

- Bandwidth
  - → 10,000-25,000X improvement for processors
  - → 300-1200X improvement for memory and disks
- Latency
  - → 30-80X improvement for processors
  - → 6-8X improvement for memory and disks
- Improvement in Bandwidth = square of improvement in latency

Power and Energy

- Problem: Get power in, distribute it, get it out

- Thermal Design Power (TDP)
  - → Characterizes sustained power consumption
  - → Used as target for power supply and cooling system
  - → Lower than peak power, higher than average power consumption

- Clock frequency can be reduced dynamically to limit power consumption

- Energy efficiency is often a better measurement
  - → Power = a design constraint

Dynamic Energy and Power

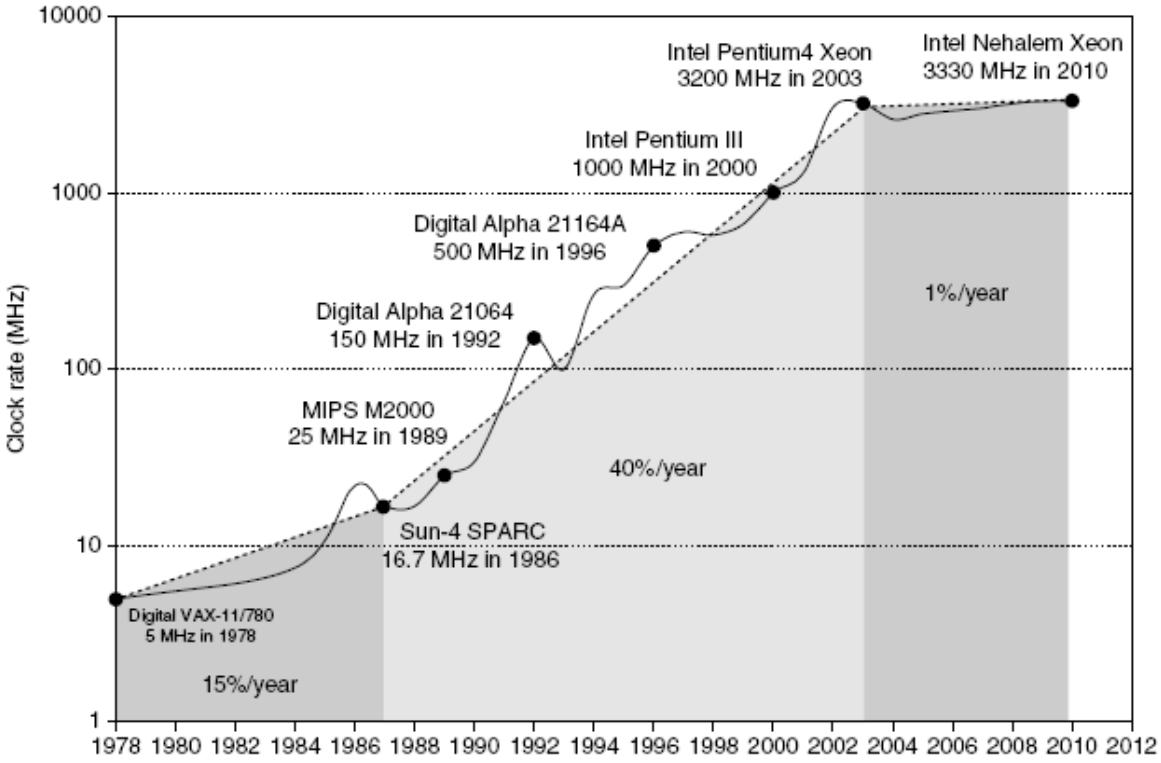
- Dynamic energy
  - → Transistor switch from 0 -> 1 or 1 -> 0
  - → E = ½ x Capacitive load x Voltage^2
- Dynamic power
  - → P = ½ x Capacitive load x Voltage^2 x Frequency switched
- Relation between energy and power consumption

$$P \approx E \times f$$

  - → *P*: power, *E*: energy, *f*: switching frequency
- Reducing clock frequency reduces power, not energy

Power

- Intel 80386 consumed ~2 W
- 3.3 GHz Intel Core i7 consumes 130 W
- Heat must be dissipated from 1.5 x 1.5 cm chip
- This is the limit of what can be cooled by air.
- Performance improvement by increasing *f* is over!



Reducing Power

- Techniques for reducing power:
  - →Turn off components not being used
  - →Dynamic Voltage-Frequency Scaling
  - →Low power state for DRAM, disks
  - →Overclocking, turning off all but one core

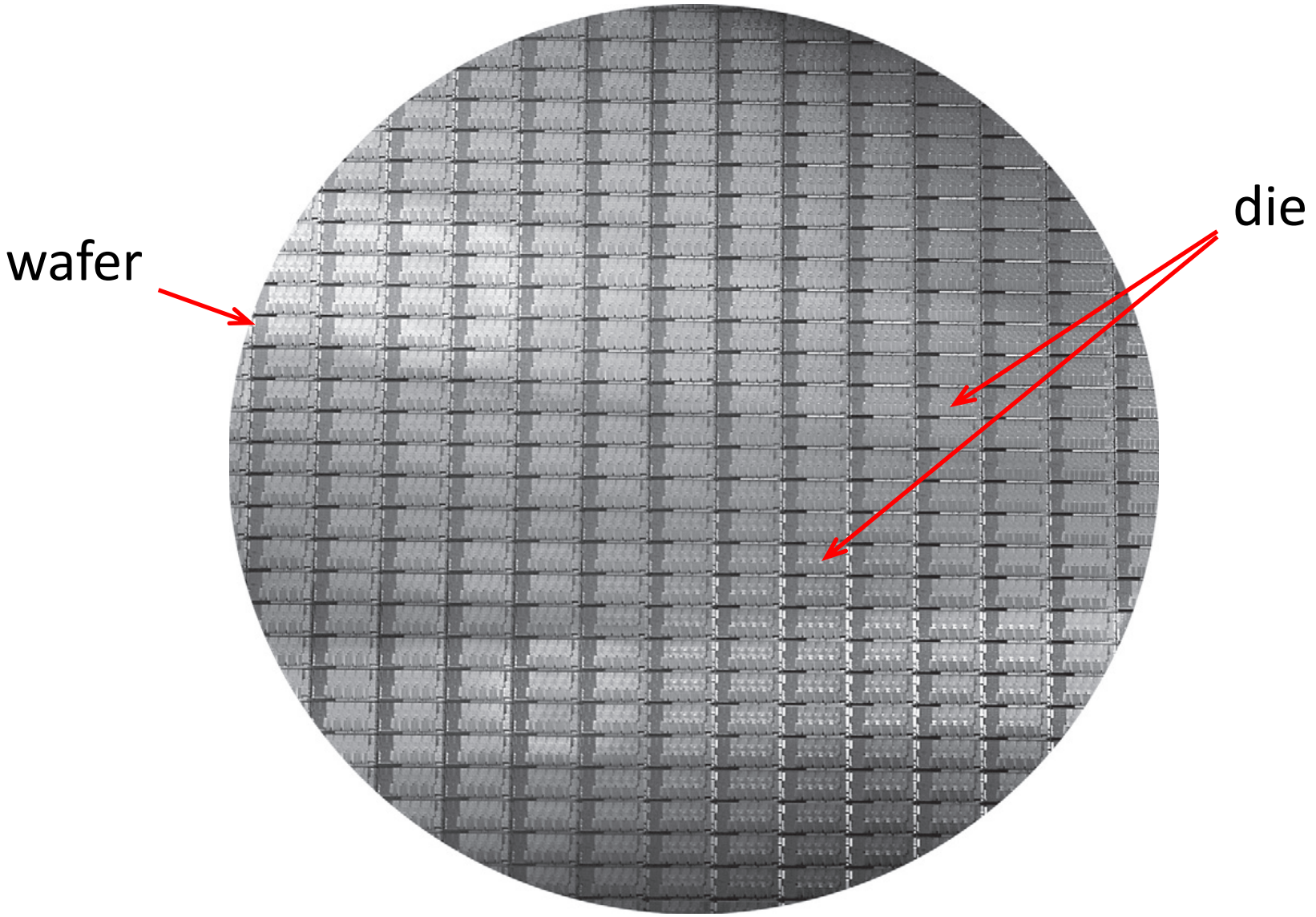
Static Power

- Transistors are not perfect switches.
- Static power consumption
  - → $\text{Current}_{\text{static}}$ x Voltage
  - → Scales with number of transistors
  - → To reduce: power gating
- Static power
  - → Increasing share of overall system power.

Factors Affecting Cost

- Cost: design, manufacturing, material, etc.
- Cost driven down by learning curve
  - →Yield: ratio between good products among all.
- DRAM: price closely tracks cost
  - →Commodity: identical products sold by multiple vendors
- Microprocessors: price depends on volume
  - →10% less for each doubling of volume
- Increase in volume -> reduced unit cost

Integrated Circuit: Wafer and Die



Cost of Integrated Circuit Cost

- Cost of Integrated circuit

$$\text{Cost of integrated circuit} = \frac{\text{Cost of die} + \text{Cost of testing die} + \text{Cost of packaging and final test}}{\text{Final test yield}}$$

$$\text{Cost of die} = \frac{\text{Cost of wafer}}{\text{Dies per wafer} \times \text{Die yield}}$$

$$\text{Dies per wafer} = \frac{\pi \times (\text{Wafer diameter}/2)^2}{\text{Die area}} - \frac{\pi \times \text{Wafer diameter}}{\sqrt{2 \times \text{Die area}}}$$

- Bose-Einstein formula:

$$\text{Die yield} = \text{Wafer yield} \times 1/(1 + \text{Defects per unit area} \times \text{Die area})^N$$

→ Defects per unit area = 0.016-0.057 defects per square cm (2010)

→ N = process-complexity factor = 11.5-15.5 (40 nm, 2010)

Dependability

- As feature size shrinks, computers fail more often.

- Module reliability
  - → Mean time to failure (MTTF)
  - → Mean time to repair (MTTR)
  - → Mean time between failures (MTBF) = MTTF + MTTR
  - → Availability = MTTF / MTBF
  - → = a ratio between service time and total life span

- To improve reliability: redundancy.

Summary

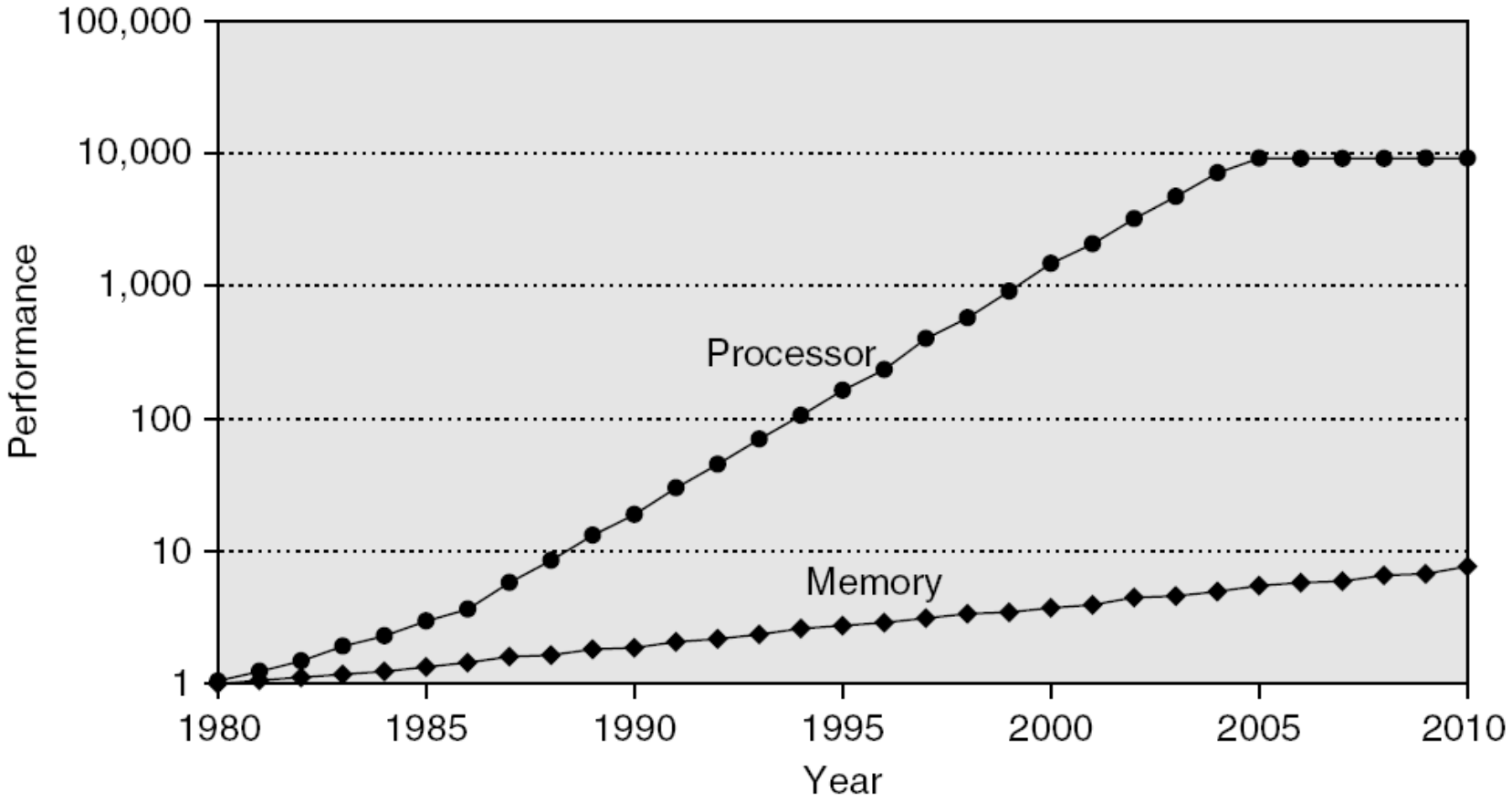
- Computer architecture involves ISA, microarchitecture, HW technologies.
  - → Is about making tradeoffs among design parameters optimized for target applications.
  - → Should be mindful about technology trends, and their impacts on comp. design.
- Classification of comp. arch. wrt different types of parallelism they exploit.
- Reviews of trends of various system parameters.
- Overviews of computer design principles
  - → Exploitation of parallelism, locality, optimization for common cases, etc.

Memory Hierarchy

Memory Technology – Overview

- Static RAM (SRAM)
  - →0.5ns – 2ns, $2000 – $5000 per GB
- Dynamic RAM (DRAM)
  - →20ns – 30ns, $10 – $50 per GB
- Magnetic disk
  - →5ms – 20ms, $0.20 – $2 per GB
- Ideal memory
  - →Access time of SRAM
  - →Capacity and cost/GB of disk

The “Memory Wall”



Processor mem accesses/sec vs DRAM accesses/sec

The "Memory Wall" – A Multi-Core Case

- Aggregate peak bandwidth grows with # cores:
  - → Intel Core i7 can generate two references per core per clock
  - → Four cores and 3.2 GHz clock

    25.6 billion 64-bit data references/second +

    12.8 billion 128-bit instruction references

    = 409.6 GB/s!
- DRAM bandwidth is only 6% of this (25 GB/s)
- Requires:
  - **Multi-port, pipelined caches**
  - **Two levels of cache per core**
  - **Shared third-level cache on chip**

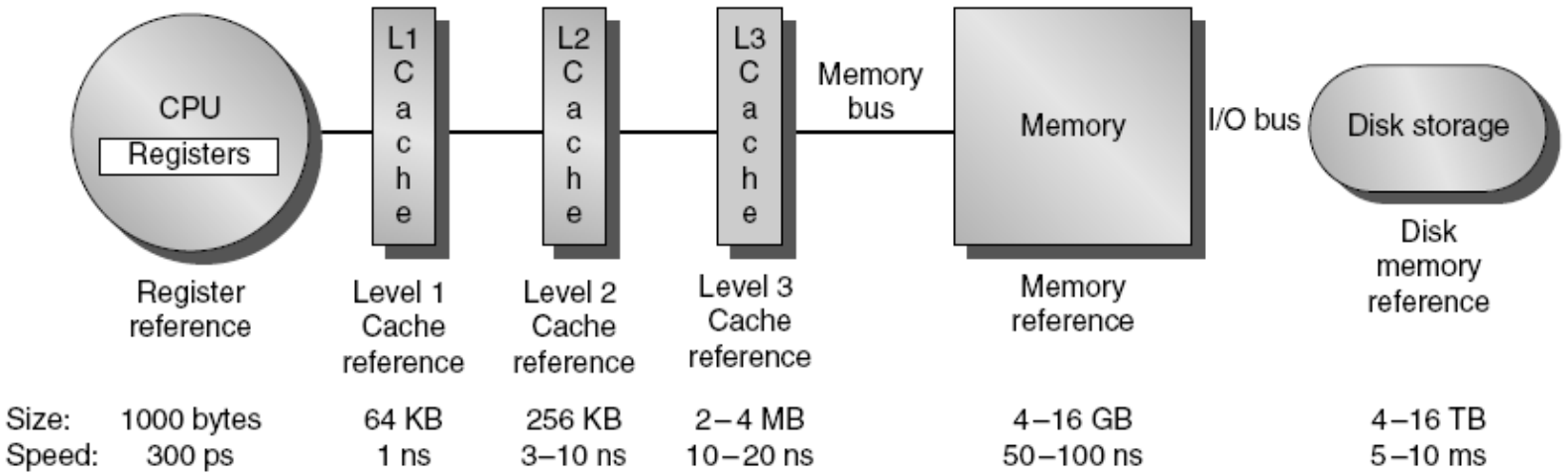
Principle of Locality – Review

- Programs often access a small proportion of their address space at any time
- Temporal locality
  - → Items accessed recently are likely to be accessed again soon
  - → e.g., instructions in a loop, induction variables
- Spatial locality
  - → Items near those accessed recently are likely to be accessed soon
  - → E.g., sequential instruction access, array data

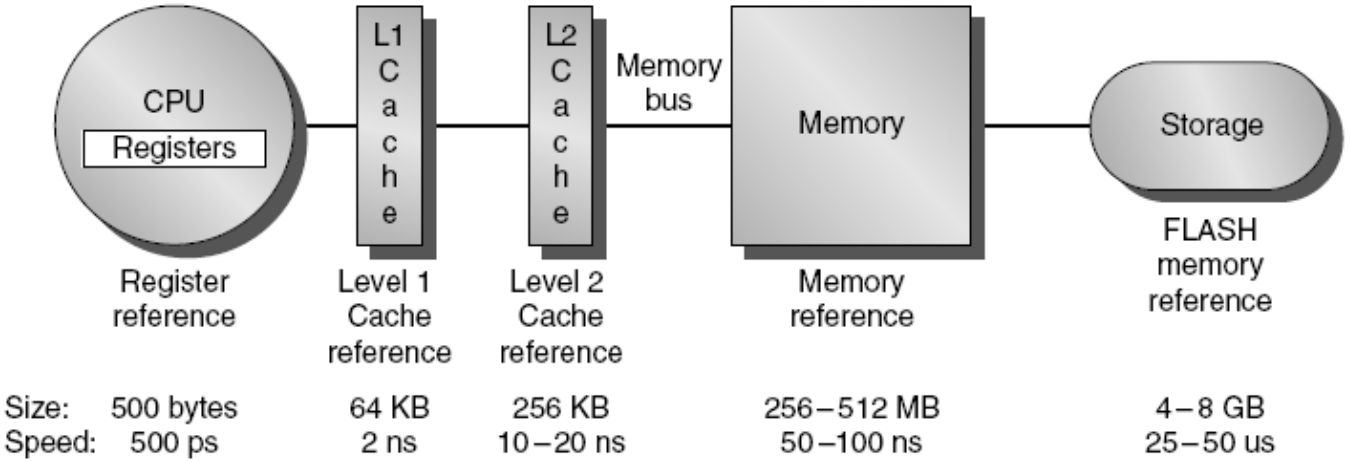
Memory Hierarchy

- Programmers want unlimited amounts of memory with low latency
- Fast memory technology is more expensive per bit than slower memory
- Solution: organize memory system into a hierarchy
  - → Entire addressable memory space available in largest, slowest memory
  - → Incrementally smaller and faster memories, each containing a subset of the memory below it, proceed in steps up toward the processor
- Temporal and spatial locality insures that nearly all references can be found in smaller memories
  - → Gives the illusion of a large, fast memory being presented to the processor

Memory Hierarchy



(a) Memory hierarchy for server



(b) Memory hierarchy for a personal mobile device

Memory Hierarchy Questions

- **Block placement** - where Can a block be placed in the upper level?
- **Block identification** – how to find a block in the upper level?
- **Block replacement** - which block should be replaced on a miss?
- **Write strategy** – how to handle writes?