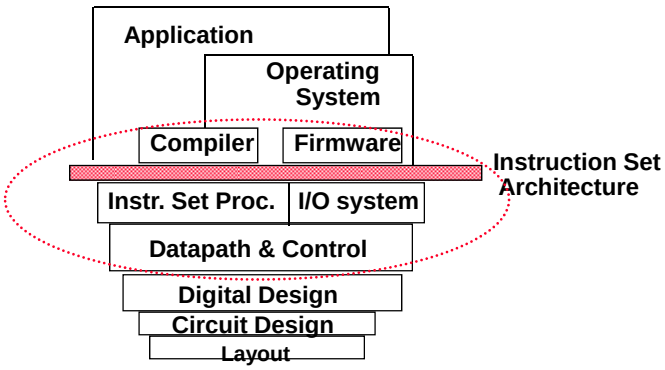
CS4200/5200
Computer Architecture I

Lecture 1 Introduction

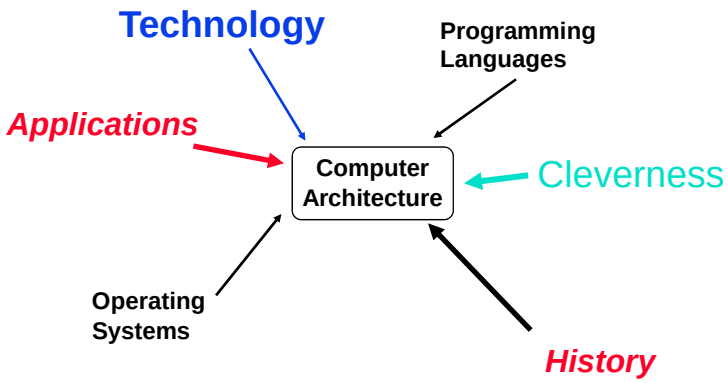
Dr. Xiaobo Zhou
Department of Computer Science

What is "Computer Architecture"?



- Coordination of many *levels of abstraction*
- Under a rapidly changing set of forces
- Design, Measurement, *and* Evaluation

Forces on Computer Architecture



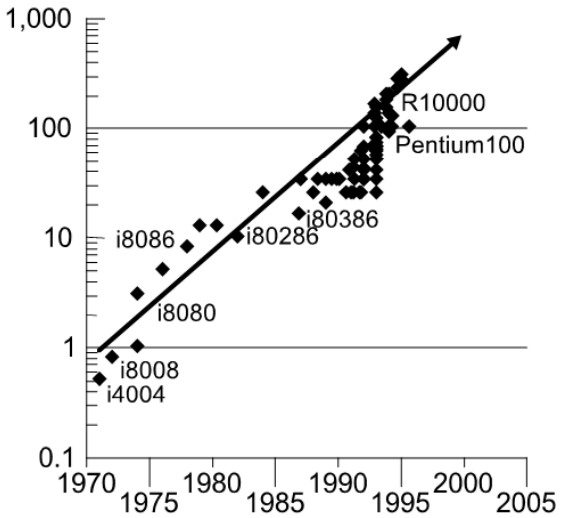
Example: Impact of Applications

- Desktop applications (price-performance)
  - emphasizes performance of integer and FP data types
  - little regard for program (code) size, power consumption
- Server applications (dependability/availability, scalability, throughput)
  - database, file system, web applications, time-sharing
  - FP performance is much less important than integer and character strings
  - little regard for program (code) size, power consumption
- Embedded applications (real-time)
  - Digital signal processors (DSPs) and media processors
  - Value program (code) size and power
    - less memory is cheaper and lower power
    - reduce chip costs: FP instructions may be optional
- CS420/520: architectures for desktops or servers

Classes of Computers

- **Personal Mobile Device (PMD)**
  - **e.g. start phones, tablet computers**
  - **Emphasis on energy efficiency and real-time**
- **Desktop Computing**
  - **Emphasis on price-performance**
- **Servers**
  - **Emphasis on availability, scalability, throughput**
- **Clusters / Warehouse Scale Computers**
  - **Used for "Software as a Service (SaaS)"**
  - **Emphasis on availability, price-performance, power/energy**
  - **Sub-class: Supercomputers, emphasis: floating-point performance and fast internal networks**
- **Embedded Computers**
  - **Emphasis: price**

Technology Trend: Clock rate



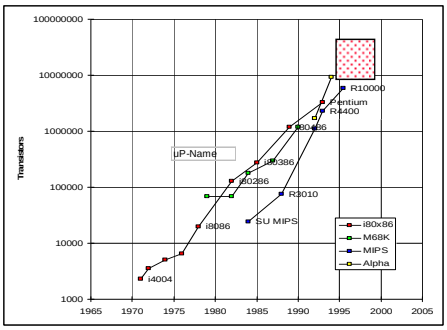
- 20~35% per year ---> today's PC is yesterday's Supercomputer

Technology: Transistors

DRAM chip capacity

DRAM	
Year	Size
1980	64 KB
1983	256 KB
1986	1 MB
1989	4 MB
1992	16 MB
1996	64 MB
1999	256 MB
2002	1 GB

Microprocessor Logic Density



- In ~1985 the single-chip processor (32-bit) and the single-board computer emerged
  - => workstations, personal computers, multiprocessors have been riding this wave since

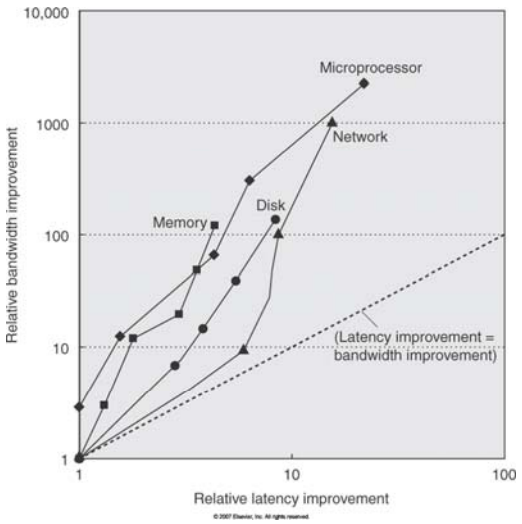
In the 2002+ timeframe, these may well look like mainframes compared single-chip computer (maybe 2+ chips)

Trends in Technology => Dramatic Changes

- Processor
  - IC logic capacity (density x size): about 40% ~ 45% per year (density about 35% per year and die size about 10% per year)
  - clock rate: about 20% ~ 35% per year
- Memory
  - DRAM capacity: about 25%~40% per year (slowing)
  - Memory speed: about 8-10% per year
  - Cost per bit: improves about 25% per year
- Flash capacity: about 50%~60% per year
- Magnetic Disk
  - Capacity: about 40% per year (Speed: about 8-10% per year)
  - Cost per bit: 50-100 times cheaper than DRAM
- Network Bandwidth
  - 10 MB ---------(10 Y)→ 100 Mb ----------(5Y)→ 1 Gb → 10 Gb →
  - US: BW increasing more than 100% per year (optical media) !

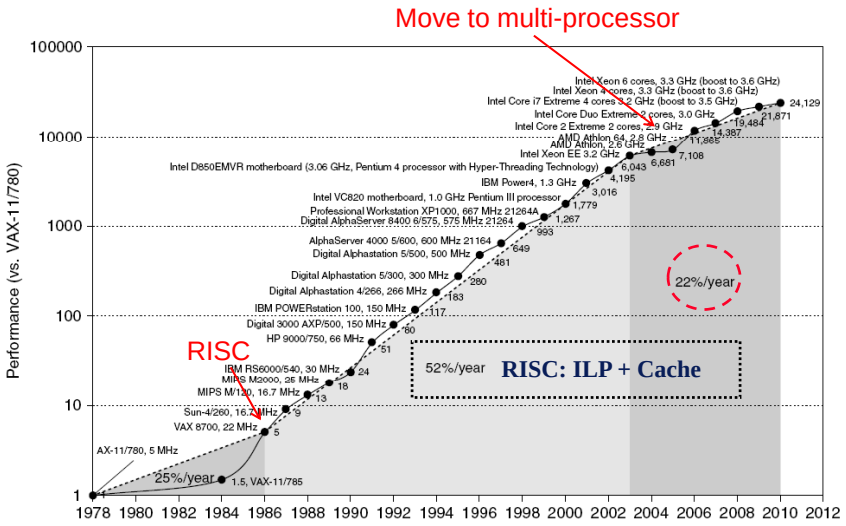
Performance Trends: Bandwidth over Latency

- Bandwidth / throughput
- Latency / response time
- Bandwidth grows by at least the square of the latency improvement

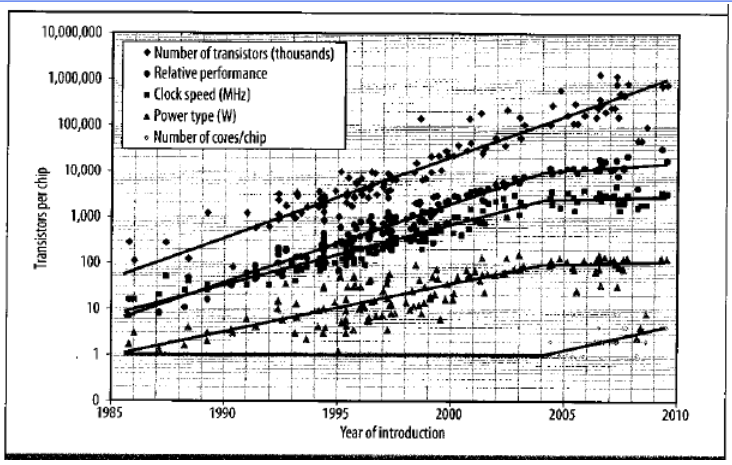


Log-log plot of bandwidth and latency of the relative 6 milestones in Fig 1.9

Processor Performance (SPEC)



Single-Processor Performance (cont.)



° The original Moore's law projection of increasing transistors per chip remains unabated even as single-processor performance has stalled (due to energy and power constraints) [Fuller and Millett, IEEE Computer, 44(1), 2011]

Current Trends in Architecture

° The 25000-fold performance improvement since 1978 allowed programmers today to trade performance for productivity
  - C/C++ → Java/C++, scripting languages like Python and Ruby

° Cannot continue to leverage Instruction-Level parallelism (ILP)
  - Single processor performance improvement ended in 2003

° New models for performance:
  - Data-level parallelism (DLP)
  - Thread-level parallelism (TLP)
  - Request-level parallelism (RLP)

° These require explicit restructuring of the application

What is "Computer Architecture"

Computer Architecture =

Instruction Set Architecture +

Machine Organization + …..(hardware)

Organization: high-level aspects of a computer's design, i.e, memory system, bus structure, CPU design

E.g.: embedded NEC VR 5432 vs. VR 4122, same ISA (M64)
AMD Opteron 64 vs. Intel P4, same ISA (x86)
but different pipelining and memory organizations

Hardware: specifics of a machine, details of logic design, packaging technology

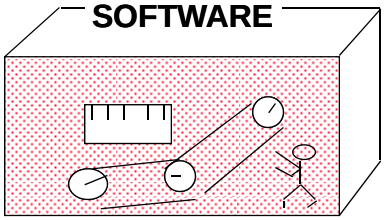
E.g.: P II vs. Celeron, P4 vs. Mobile P4

same ISA & Orga., differ in clock rates, memory systems

Instruction Set Architecture (subset of Computer Arch.)

. . . the attributes of a [computing] system as seen by the programmer, i.e. the conceptual structure and functional behavior, as distinct from the organization of the data flows and controls the logic design, and the physical implementation.

Amdahl, Blaw, and Brooks, 1964



-- Organization of Programmable Storage

-- Data Types & Data Structures: Encodings & Representations

-- Instruction Formats

-- Instruction (or Operation Code) Set

-- Modes of Addressing and Accessing Data Items and Instructions

-- Exceptional Conditions

Computer Architecture's Changing Definition

- 1950s to 1960s: Computer Architecture Course: Computer Arithmetic
- 1970s to mid 1980s: Computer Architecture Course: Instruction Set Design, especially ISA appropriate for compilers
- 1990s~ 2000s: Computer Architecture Course: Design of CPU, memory system, I/O system, Multiprocessors, Networks
- 2010s: Computer Architecture Course: Self adapting systems? Self organizing structures? DNA Systems/Quantum Computing?

Example ISAs (Instruction Set Architectures)

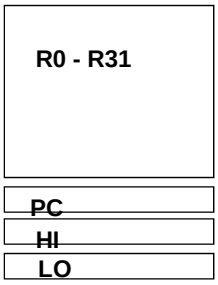
- CISC (Complex Instruction Set Computer)
  - DEC PDP-11
  - VAX
  - Intel (8086,80286,80386, 80486,Pentium, MMX, ...) 1978-96

economic advantages vs. performance
Intel Pentium…CISC?RISC?
Intel IA64 (Itanium)

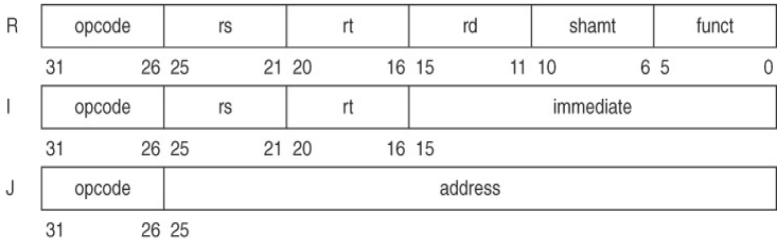
- RISC
  - Digital Alpha (v1, v3) 1992-97
  - HP PA-RISC (v1.1, v2.0) 1986-96
  - Sun Sparc (v8, v9) 1987-95
  - SGI MIPS (MIPS I, II, III, IV, V) 1986-96
  - Power PC
  - Embedded
    - ARM, Hitachi SH, MIPS16, Thumb

MIPS R3000 Instruction Set Architecture

° Instruction Categories
- Computational
- Load/Store
- Jump and Branch
- Floating Point
  - coprocessor
- Memory Management
- Special



Basic instruction formats



Organization

ISA Level

FUs & Interconnect

Logic Designer's View

-- Capabilities & Performance Characteristics of Principal Functional Units

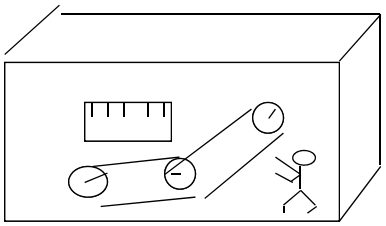
(e.g., Registers, ALU, Shifters, Logic Units, etc.)

-- Ways in which these components are interconnected

-- nature of information flows between components

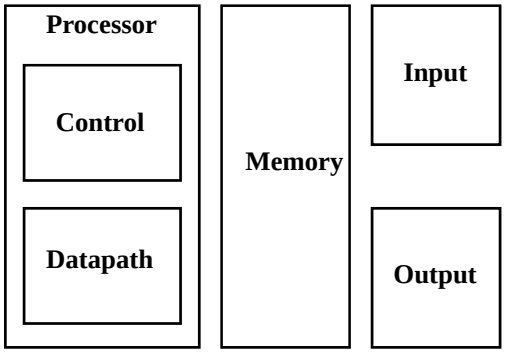
-- logic and means by which such information flow is controlled.

Choreography of FUs to realize the ISA



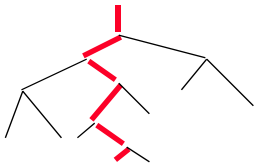
The Big Picture

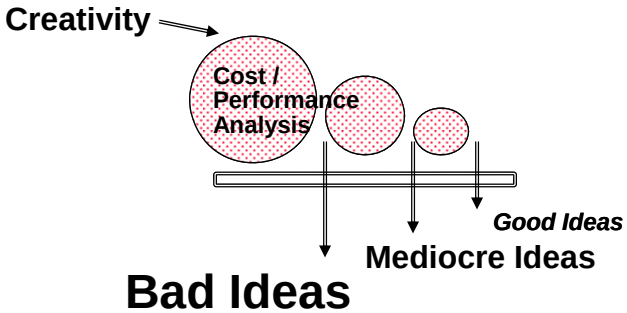
° Since 1946 all computers have had 5 classic components



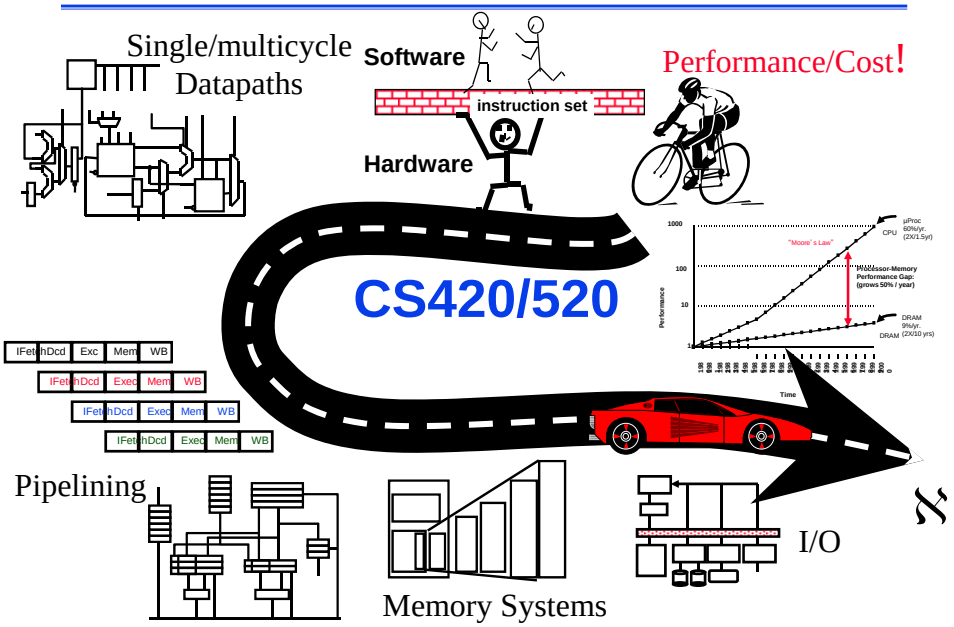
Measurement and Evaluation

Architecture design is an 'iterative process'
-- searching the space of possible designs
-- at all levels of computer systems





Where Are We ??



CSC420/520: Course Overview

Computer Architecture and Engineering

Instruction Set Deign

- **Machine Language**
- **Compiler View**
- **"Computer Architecture"**
- **"Instruction Set Processor"**

"Building Architect"

Computer Hardware Design

- **Machine Implementation**
- **Logic Designer's View**
- **"Processor Architecture"**
- **"Computer Organization"**

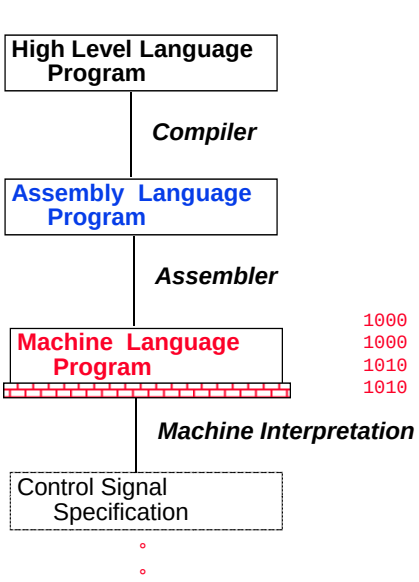
Construction Engineer

Few people design computers! Very few design instruction sets!
Many people design computer components.
Very many people are concerned with computer function, in detail.

CSC420/520:So What's in It for Me?

- In-depth understanding of the inner-workings of modern computers, their evolution, and trade-offs present at the hardware/software boundary.
  - Insight into fast/slow operations that are easy/hard to implementation hardware
- Experience with the *design process* in the context of a large complex (hardware) design.
  - Functional Specification --> Control & Datapath
- Learn how to completely design a correct single processor computer.
  - No magic required to design a computer
- Foundation for students aspiring to work in computer architecture.
- Others: solidifies an intuition about why hardware is as it is.

Levels of Representation (CS216 Review)

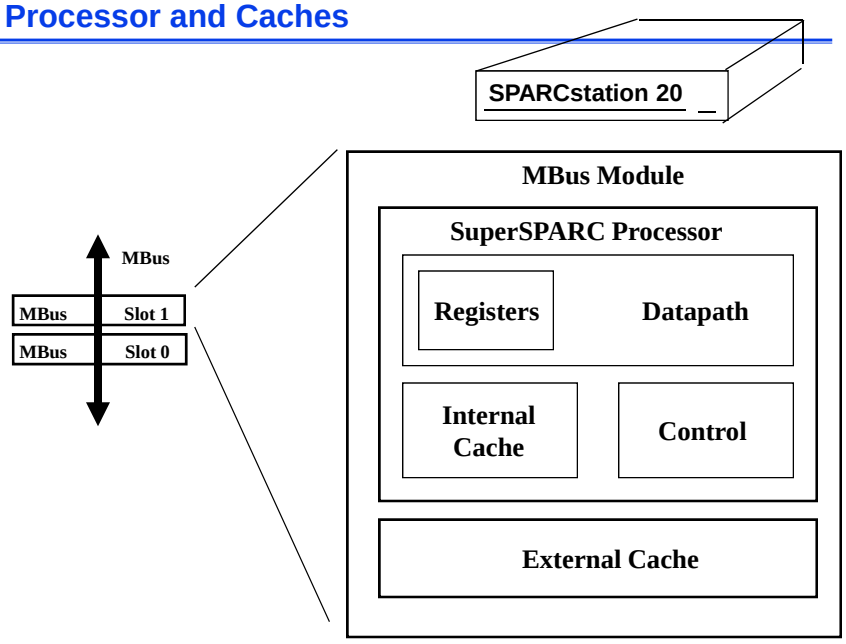



```
temp = v[k];
v[k] = v[k+1];
v[k+1] = temp;
```

```
lw $15,  0($2)
lw $16,  4($2)
sw $16,  0($2)
sw $15,  4($2)
```

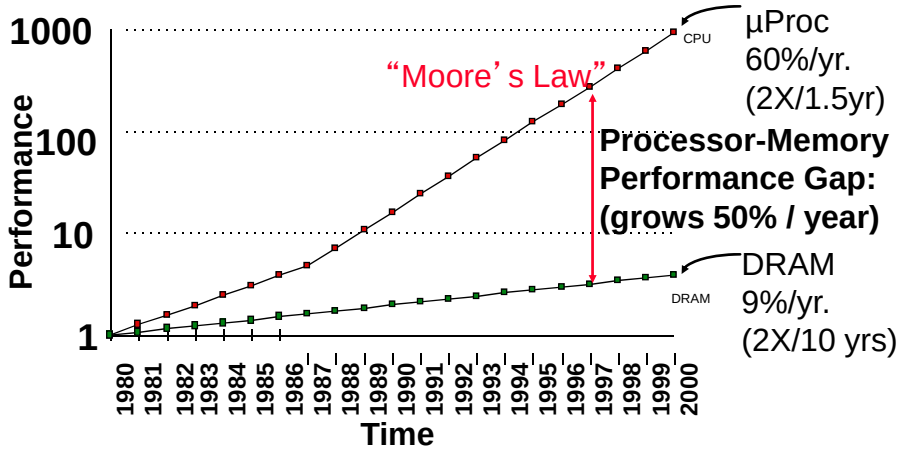
```
1000 1100 0100 1111 0000 0000 0000 0000
1000 1111 0101 1000 0000 1001 1100 0110
1010 1110 1010 1111 0101 1000 0000 1001
1010 1100 0000 1001 1100 0110 1010 1111
```

Processor and Caches



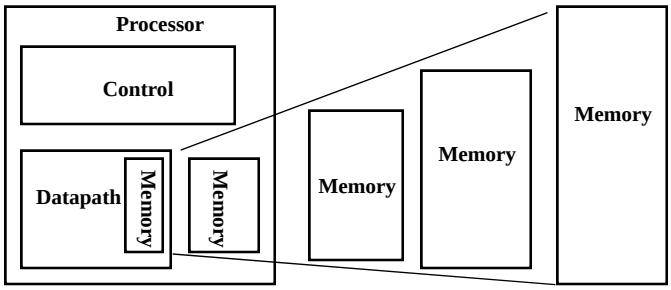
Memory Hierarchy

Processor-DRAM Memory Gap (latency)



An Expanded View of the Memory System

° 1980: no cache in µproc; 1995 2-level cache on chip
(1989 first Intel µproc with a cache on chip)



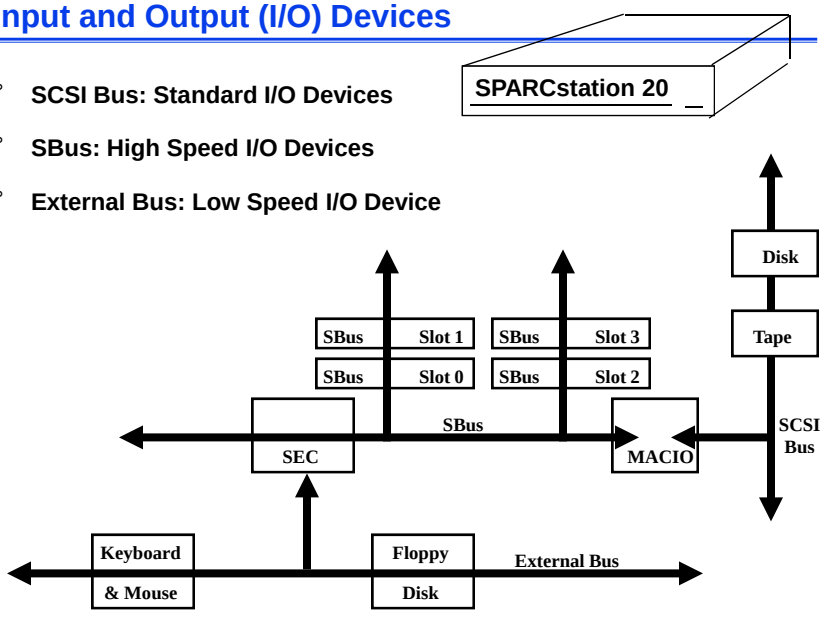
Speed:	Fastest	Slowest
Size:	Smallest	Biggest
Cost:	Highest	Lowest

Input and Output (I/O) Devices

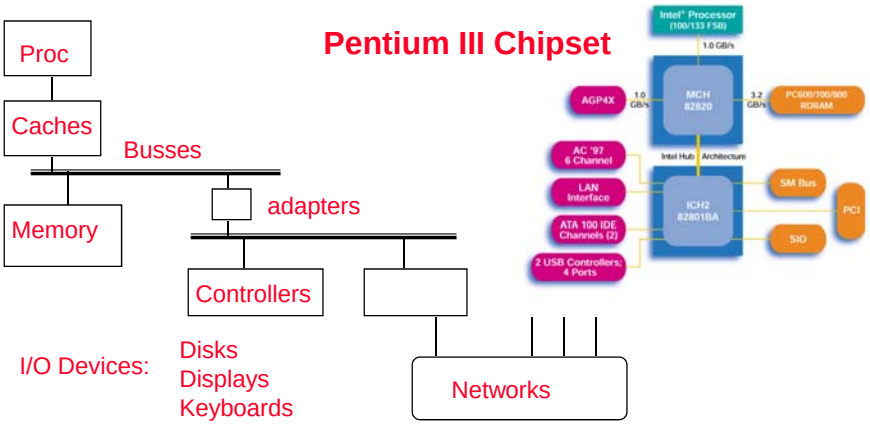
° SCSI Bus: Standard I/O Devices

° SBus: High Speed I/O Devices

° External Bus: Low Speed I/O Device



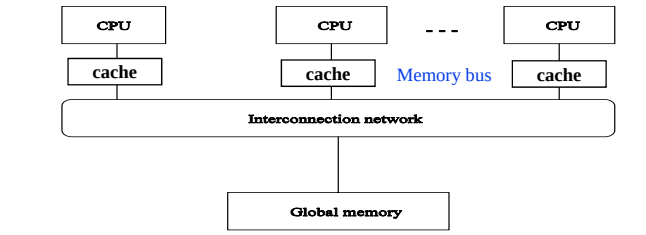
Summary: It's about Communication



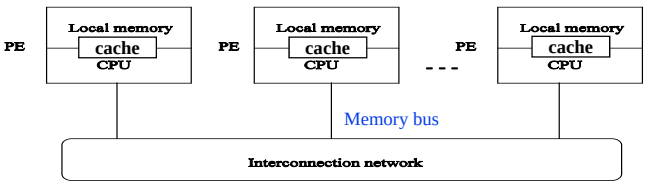
- All have interfaces & organizations

Parallel Architectures

- Multiprocessor: physically shared memory structure
- Multicomputer: physically distributed memory structure.



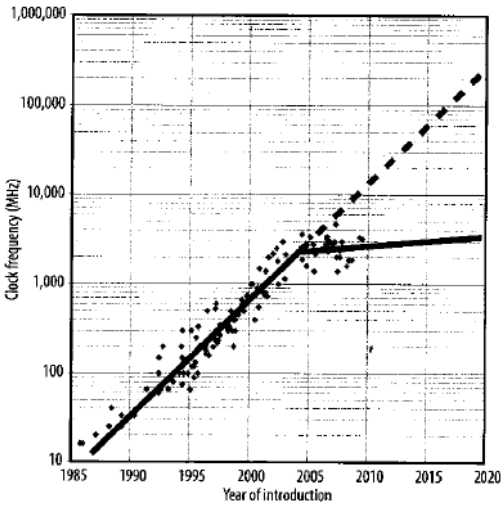
(a)



(b)

Computing performance: game over or next level?

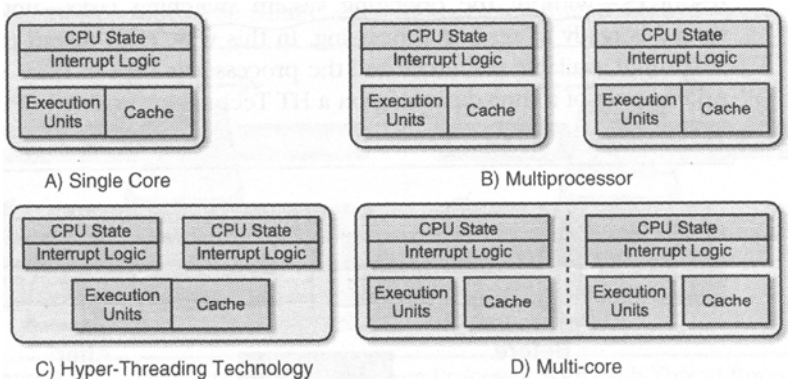
° **The end of dramatic exponential growth n single-processor performance makes the end of the dominance of the single microprocessor in computing. The era of sequential computing must give way to an era in which parallelism holds the forefront. Although important scientific and engineering challenges lie ahead, this is an opportune time for innovation in programming systems and computing architectures. [Fuller and Millett, IEEE Computer, 44(1), 2011]**



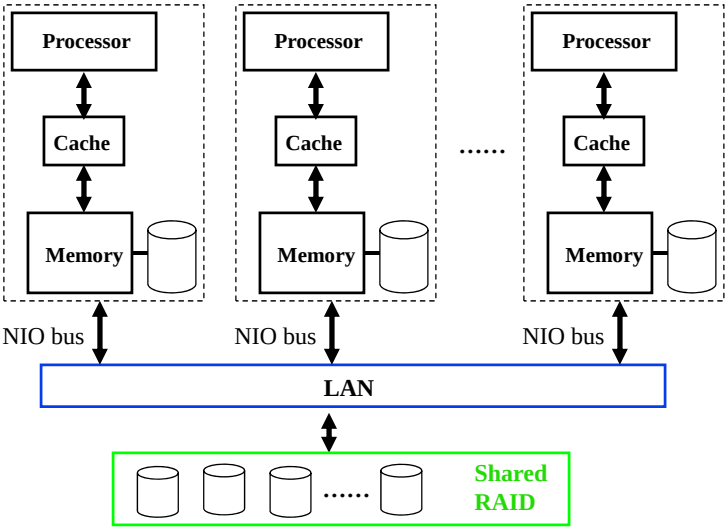
A dashed line represents expectations if single-processor performance had continued its historical trend. [Fuller and Millett, IEEE Computer, 44(1), 2011]

Multi-Processor and Multi-Core

° **Multi-core processors use chip multi-processing (CMP)**
- **Cores are essentially individual processors on a single die**
- **May or may not share on-chip cache**
- **True parallelism**



Cluster Architectures



Modern Data Centers

° **Data centers are the next computing platform**

- **Power consumption, distribution and cooling are major components in a data center's cost breakdown.**

In 2009, we built a university prototype data center for research with a $1.25M equipment grant from the Air Force. The expenditure in racks, 24 HP G6 blade servers, 40 TB HP EVA storage area network with 10Gbps Ethernet and 8Gbps Fibre/iSCSI dual channels and the VMware licenses was about $0.6 million, while the expenditure in three APC InRow RP Air-Cooled and UPS equipments for maximum 40 kWs in the n+1 redundancy design was about another $0.6 million.

Power and Energy

- Green Computing
  - Intel 80386 consumed ~ 2W
  - 3.3 GHz Intel Core i7 consumes 130 W
  - ICT accounts for about 3% of global electricity usage and greenhouse gas, which is about the same as the emissions of airlines. One Google search generated about 7g carbon emission. 200 millions search per day convert to about 70, 000 cars CO2 emission.
  - The International Energy Agency updated a warning in May 2009 that ICT energy use could double by 2022, and triple by 2030. More than half of the energy use and emissions is due to servers and data centers.
- Energy per task is often a better measurement
  - Power is simply energy per unit time: 1 watt = 1 joule per sec
  - Processor A has a 20% higher average power consumption than processor B, but A executes a task in 70% of the time needed by B
  - When is power consumption a more useful measure?

DVFS (Dynamic Voltage-Frequency Scaling)

- Modern microprocessors typically offer a few clock frequencies and voltages in which to operate that se lower power and energy
  - Clock rate can be reduced dynamically to limit power consumption
  - Dynamic power can be greatly reduced by lowering the voltage (^2)

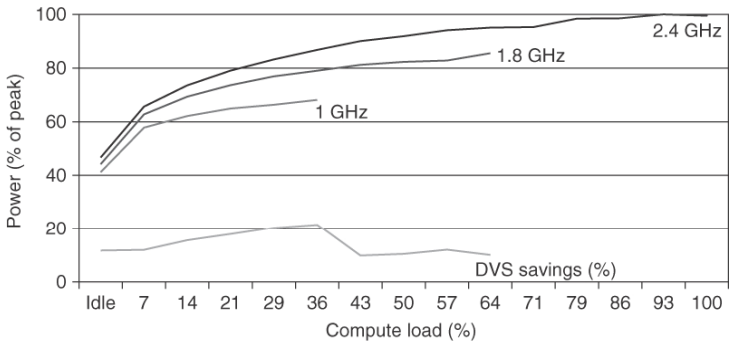


Figure 1.12 Energy savings for a server using an AMD Opteron microprocessor, 8 GB of DRAM, and one ATA disk. At 1.8 GHz, the server can only handle up to two-thirds of the workload without causing service level violations, and, at 1.0 GHz, it can only safely handle one-third of the workload.

Power and Energy

- Performance, Power and Energy
- Problem: Get power in, get power 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 rate can be reduced dynamically to limit power consumption
- Energy per task is often a better measurement
  - Power is simply energy per unit time: 1 watt = 1 joule per sec
  - Processor A has a 20% higher average power consumption than processor B, but A executes a task in 70% of the time needed by B
  - When is power consumption a more useful measure?

Summary

- Trends in Technology and Performance
- Computer Architecture: ISA + Organization + Hardware
- ISA: RISC vs. CISC
- All computers consist of five components
  - Processor: (1) datapath and (2) control
  - (3) Memory
  - (4) Input devices and (5) Output devices
- Not all "memory" are created equally
  - Cache: fast (expensive) memory are placed closer to the processor
  - Main memory: less expensive memory--we can have more
- Input and output (I/O) devices has the messiest organization
  - Wide range of speed: graphics vs. keyboard
  - Wide range of requirements: speed, standard, cost ... etc.
  - Least amount of research (so far)

Reading

- **Reading and Preview**

CA 5: Chapter 1 (or CA 4 - Chapter 1)

CO 4: Chapter 1