Computer Architecture
----A Quantitative Approach

陈文智

浙江大学计算机学院

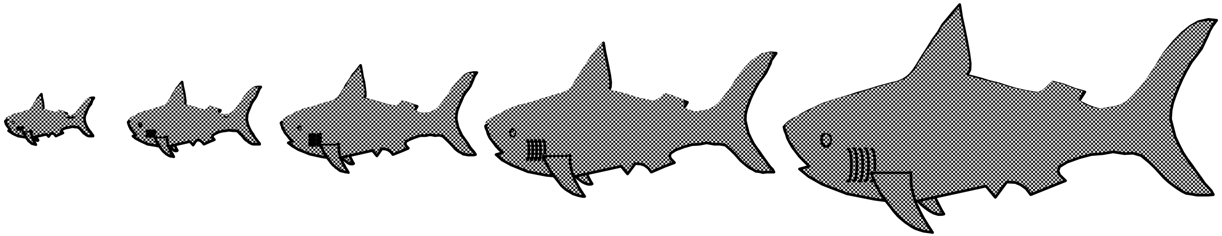
chenwz@zju.edu.cn

Topics in Chapter 1

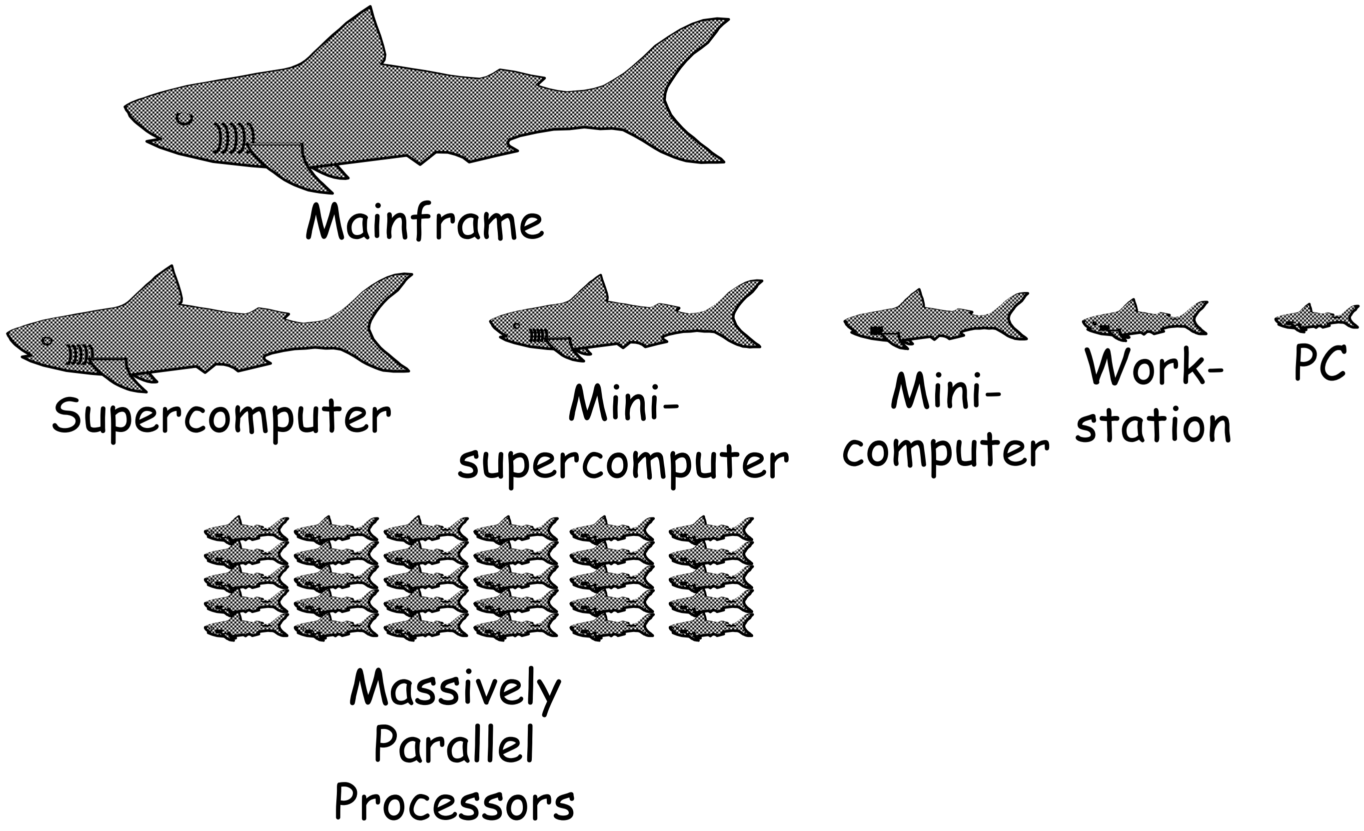
- **1.1 Introduction**
- 1.2 Classes of computers
- 1.3 Defining computer architecture and What's the task of computer design?
- 1.4 Trends in Technology
- 1.5 Trends in power in Integrated circuits
- 1.6 Trends in Cost
- 1.7 Dependability
- 1.8 Measuring, Reporting and summarizing Perf.
- 1.9 Quantitative Principles of computer Design
- 1.10 Putting it altogether

History of the Computer

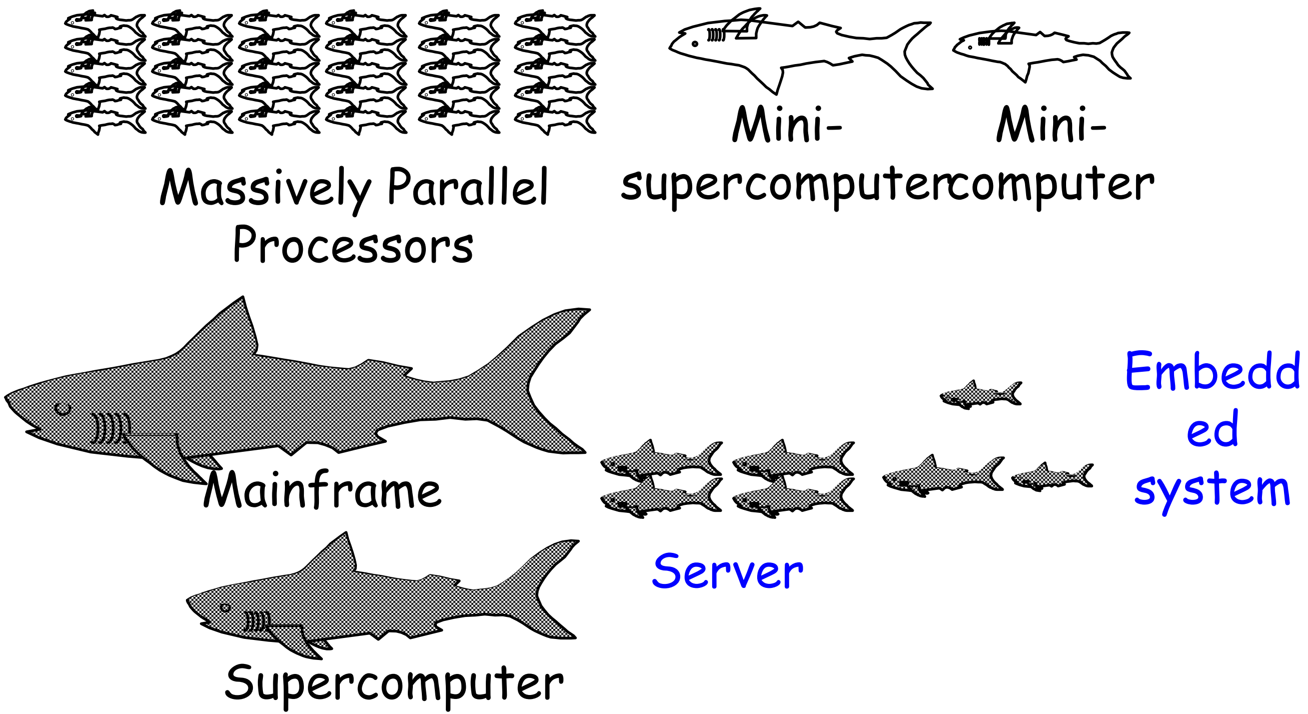
- Original:
  - Big Fishes Eating Little Fishes

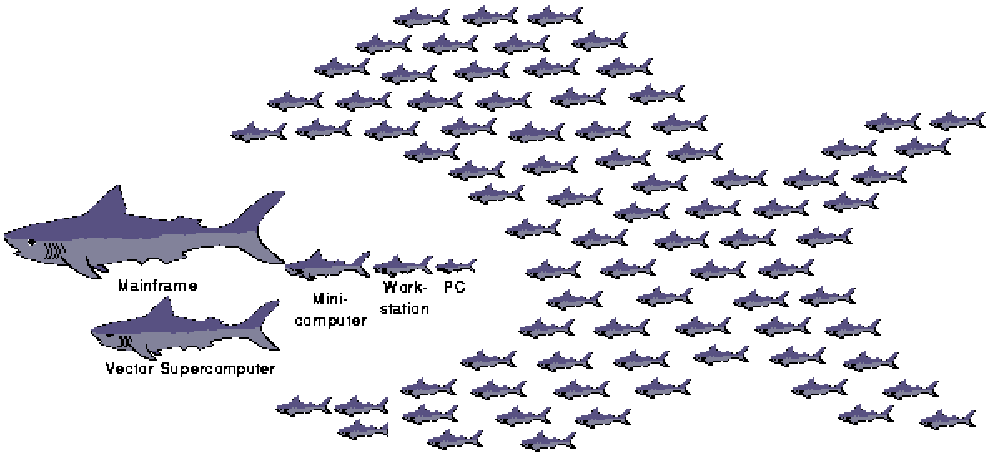


Computer Food Chain



Computer Food Chain

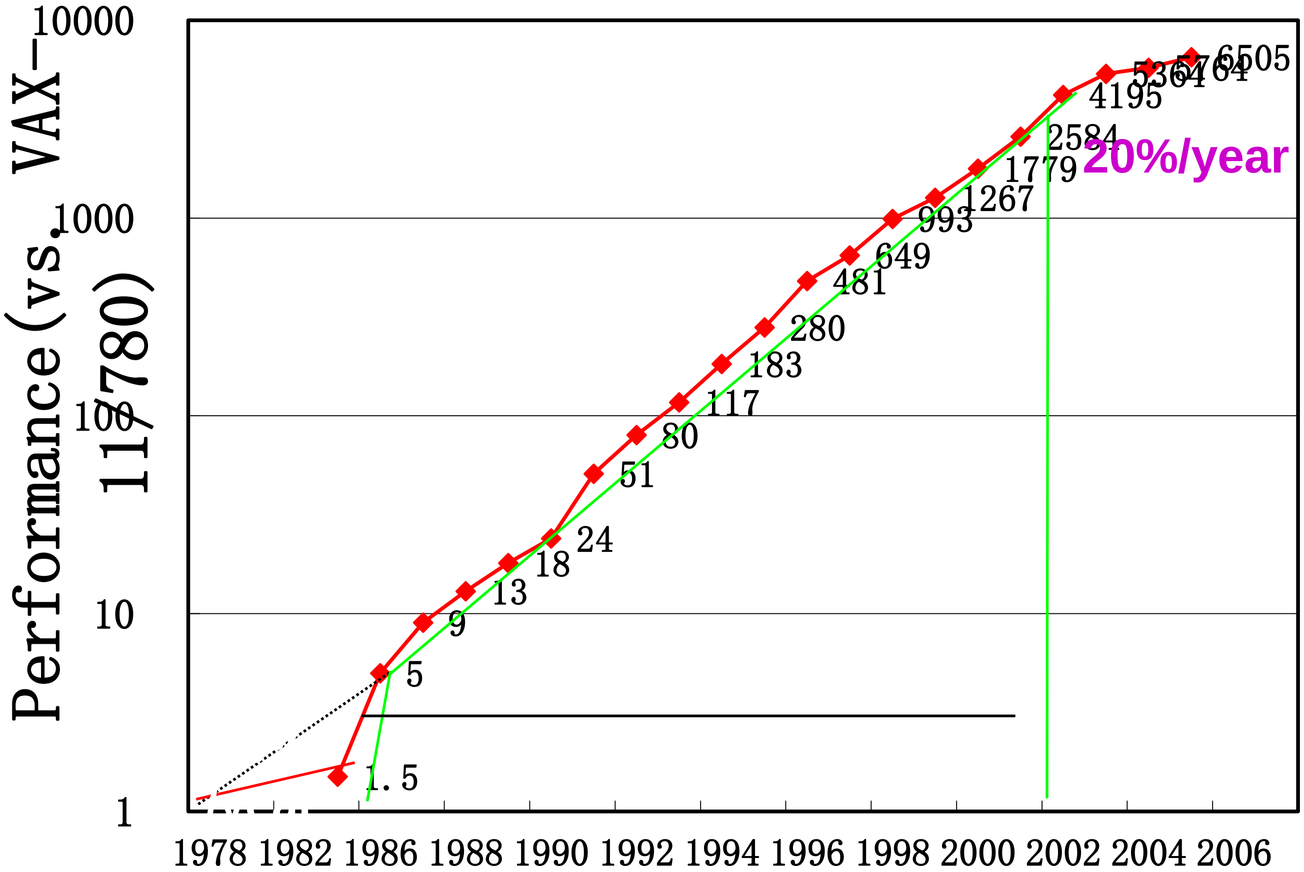




Mainframe
Vector Supercomputer
Mini-
computer
Work-
station
PC

NOW

Incredible performance improvement

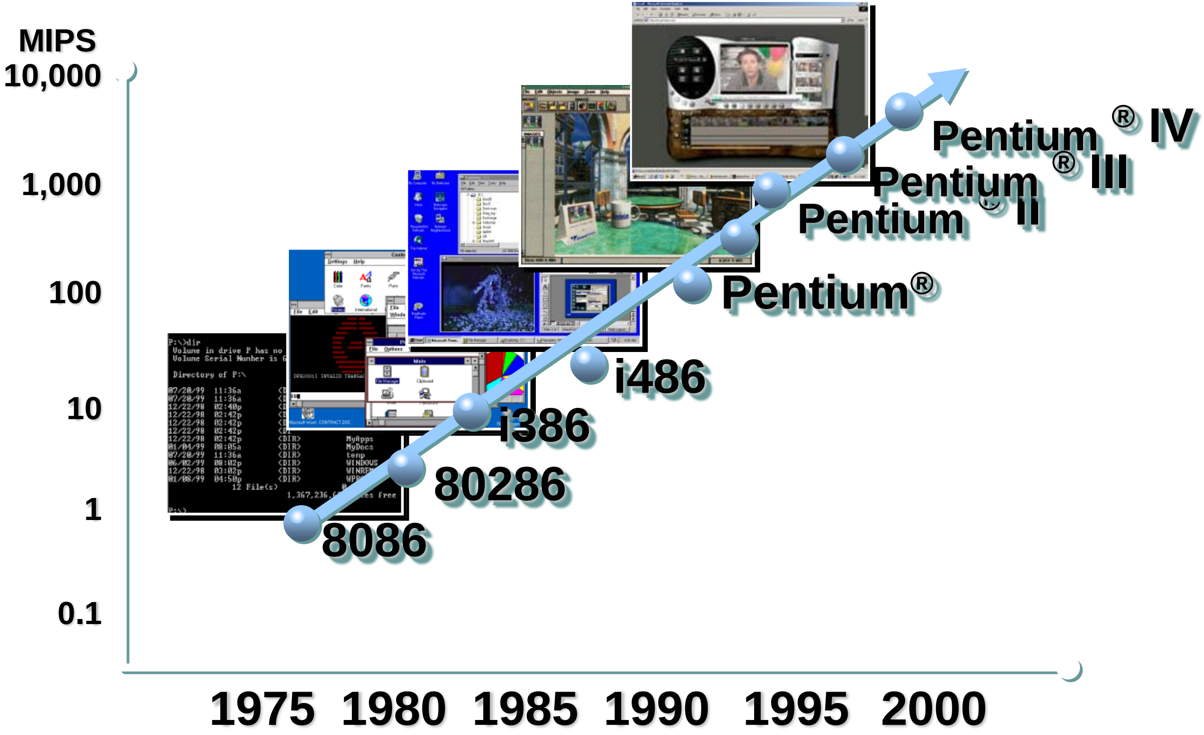


Conclusion

- Technological improvements more steady than progress in computer architecture
- After RISC emergence, computer design emphasized both architectural innovation and efficient use of technology improvements.
  - CA plays an important role in performance Improvement
- Little ILP left to exploit due to power dissipation
  - Faster uniprocessor => multiple processor on chip
  - ILP => TLP and DLP
  - Implicitly, compiler and hardware => Explicitly, programmer

MIPS

Process ability → New Applications



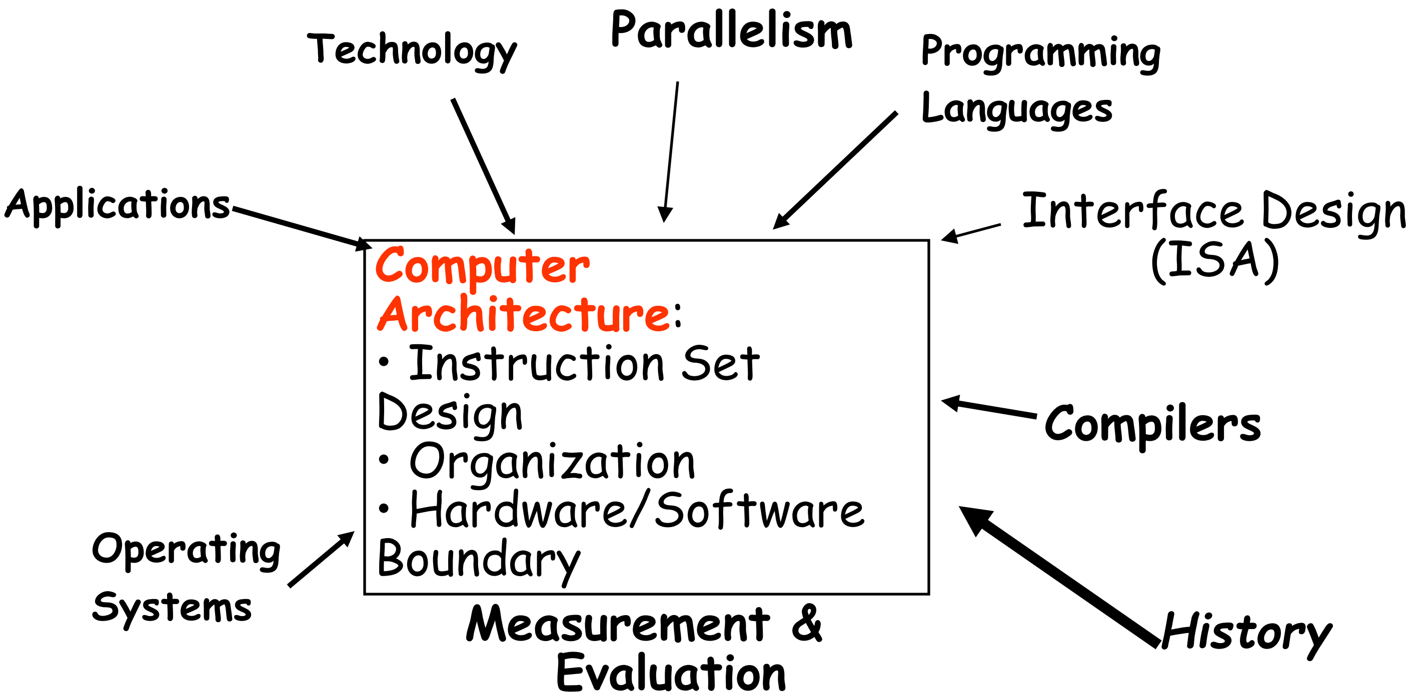
Why Such Change in 60 years?

- Two reasons:
  - Advances in the technology used to build computers
    - IC
    - Storage device(including RAM and DISK)
    - Peripheral device
  - Innovation in computer design
    - Simple → complex → most complex → simple → complex → most complex
    - Sometimes rapid, sometimes slow
    - Many technology have been washed out

Four Decades of microprocessor

- The Decade of the 1970's "Microprocessors"
  - - Programmable Controller
  - - Single-Chip Microprocessors
  - - Personal Computers (PC)
- The Decade of the 1980's "Quantitative Architecture"
  - - Instruction Pipelining
  - - Fast Cache Memories
  - - Compiler Considerations
  - - Workstations
- The Decade of the 1990's "Instruction-Level Parallelism"
  - - Superscalar Processors
  - - Speculative Microarchitectures
  - - Aggressive Code Scheduling
  - - Low-Cost Desktop Supercomputing
- The Decade of the 2000's "Thread-level/Data-level parallelism"

Forces on Computer Architecture



Computer architecture has been at the **core** of such technological development and is still on a forward move

Topics in Chapter

- 1.1 Introduction
- **1.2 Classes of computers**
- 1.3 Defining computer architecture and What’s the task of computer design?
- 1.4 Trends in Technology
- 1.5 Trends in power in Integrated circuits
- 1.6 Trends in Cost
- 1.7 Dependability
- 1.8 Measuring, Reporting and summarizing Perf.
- 1.9 Quantitative Principles of computer Design
- 1.10 Putting it altogether

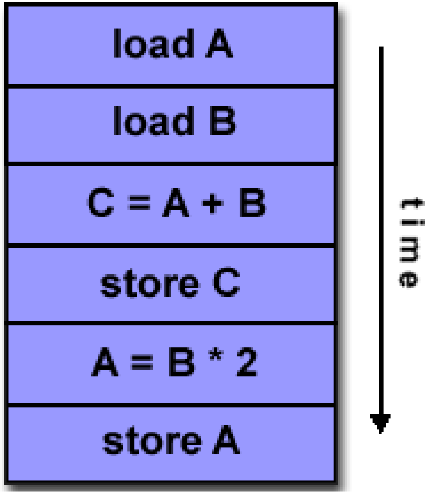
Classes of computers

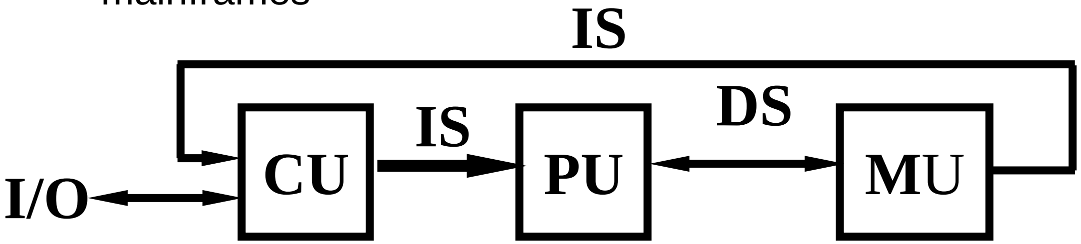
•**Flynn's Taxonomy:** **A classification of computer architectures based on the number of streams of instructions and data**

- **SISD (Single Instruction Single Data)**
  - Uniprocessors
- **MISD (Multiple Instruction Single Data)**
  - ???
- **SIMD (Single Instruction Multiple Data)**
  - Examples: Illiac-IV, CM-2
    - » Simple programming model
    - » Low overhead
    - » Flexibility
    - » All custom
- **MIMD (Multiple Instruction Multiple Data)**
  - Examples: SPARCCenter, T3D
    - » Flexible
    - » *Use off-the-shelf micros*

SISD

- A serial (non-parallel) computer
- Single instruction: only one instruction stream is being acted on by the CPU during any one clock cycle
- Single data: only one data stream is being used as input during any one clock cycle
- Deterministic execution
- This is the oldest and until recently, the most prevalent form of computer
- Examples: most PCs, single CPU workstations and mainframes

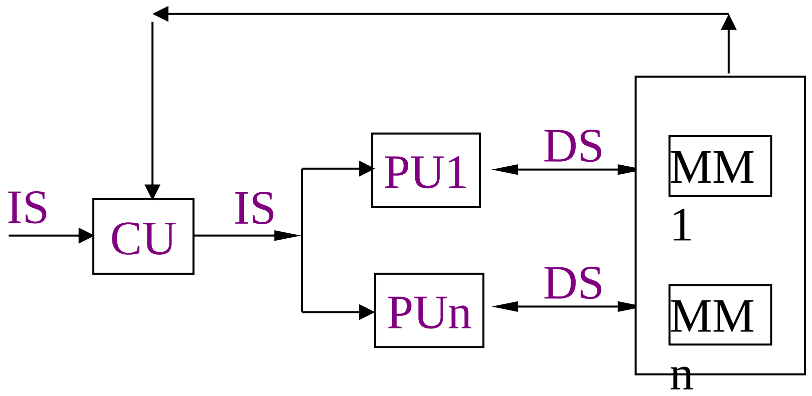




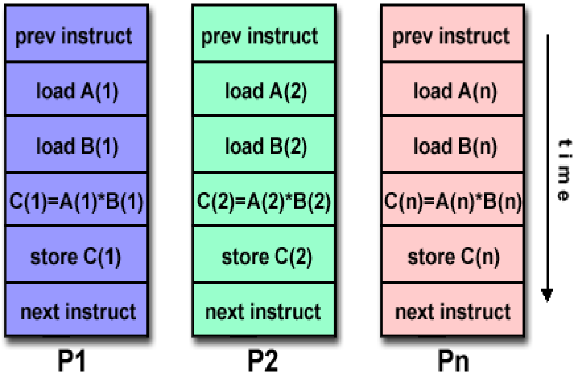
SIMD

- A type of parallel computer
- Single instruction: All processing units execute the same instruction at any given clock cycle
- Multiple data: Each processing unit can operate on a different data element
- This type of machine typically has an instruction dispatcher, a very high-bandwidth internal network, and a very large array of very small-capacity instruction units.
- Best suited for specialized problems characterized by a high degree of regularity, such as image processing.
- Synchronous (lockstep) and deterministic execution
- Two varieties:
  - Processor Arrays: Connection Machine CM-2, Maspar MP-1, MP-2
  - Vector Pipelines: IBM 9000, Cray C90, Fujitsu VP, NEC SX-2, Hitachi S820

SIMD



IS
CU
IS
PU1
PUn
DS
DS
MM
1
MM
n

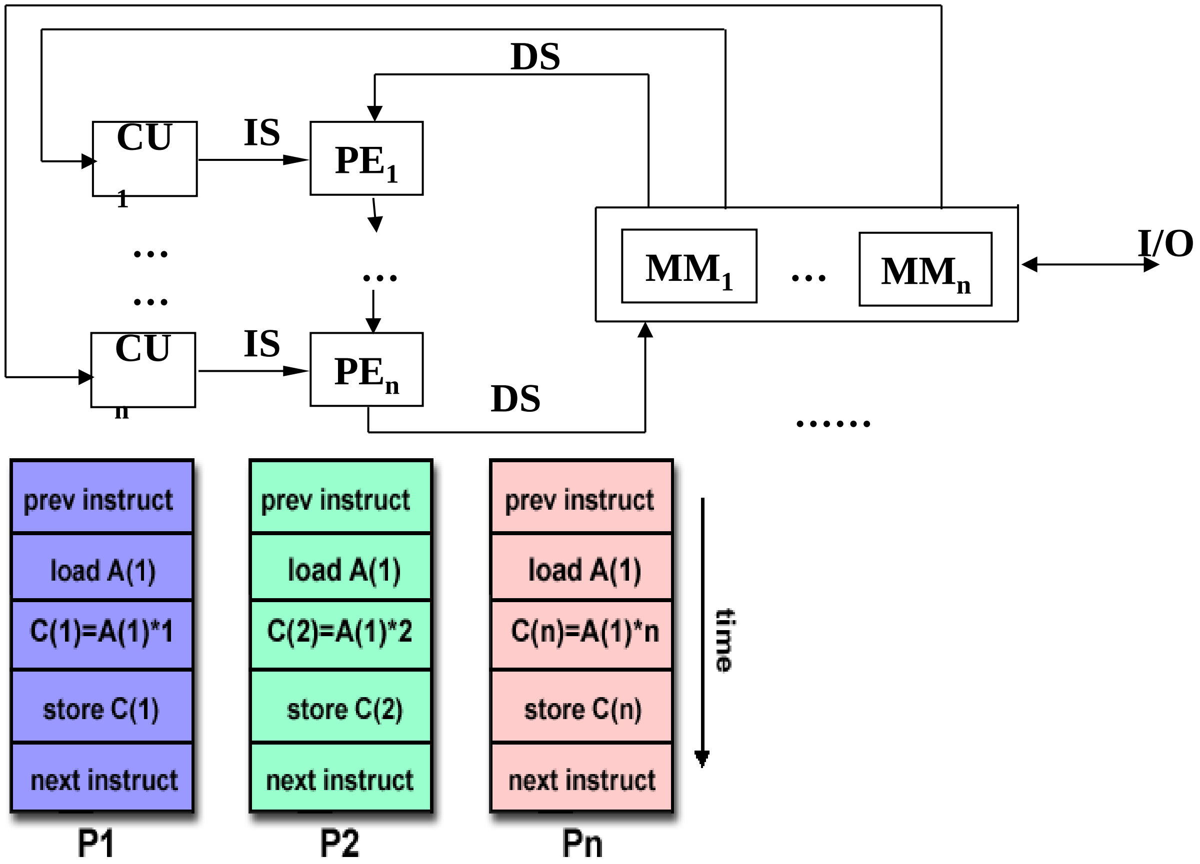


prev instruct
load A(1)
load B(1)
C(1)=A(1)*B(1)
store C(1)
next instruct
P1
prev instruct
load A(2)
load B(2)
C(2)=A(2)*B(2)
store C(2)
next instruct
P2
prev instruct
load A(n)
load B(n)
C(n)=A(n)*B(n)
store C(n)
next instruct
Pn
time

MISD

- A single data stream is fed into multiple processing units.
- Each processing unit operates on the data independently via independent instruction streams.
- Few actual examples of this class of parallel computer have ever existed. One is the experimental Carnegie-Mellon C.mmp computer (1971).
- Some conceivable uses might be:
  - multiple frequency filters operating on a single signal stream
  - multiple cryptography algorithms attempting to crack a single coded message.

MISD

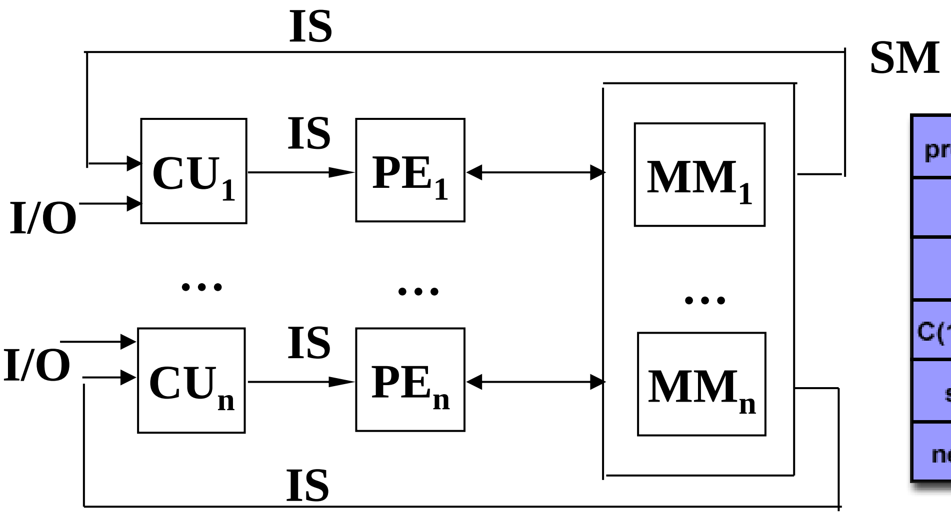


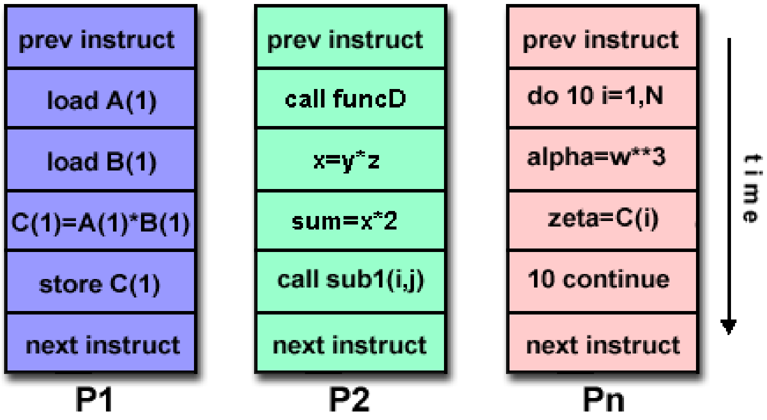
CU
1
IS
PE1
DS
…
…
…
CU
n
IS
PEn
DS
MM1
…
MMn
I/O
……
prev instruct
load A(1)
C(1)=A(1)*1
store C(1)
next instruct
P1
prev instruct
load A(1)
C(2)=A(1)*2
store C(2)
next instruct
P2
prev instruct
load A(1)
C(n)=A(1)*n
store C(n)
next instruct
Pn
time

MIMD

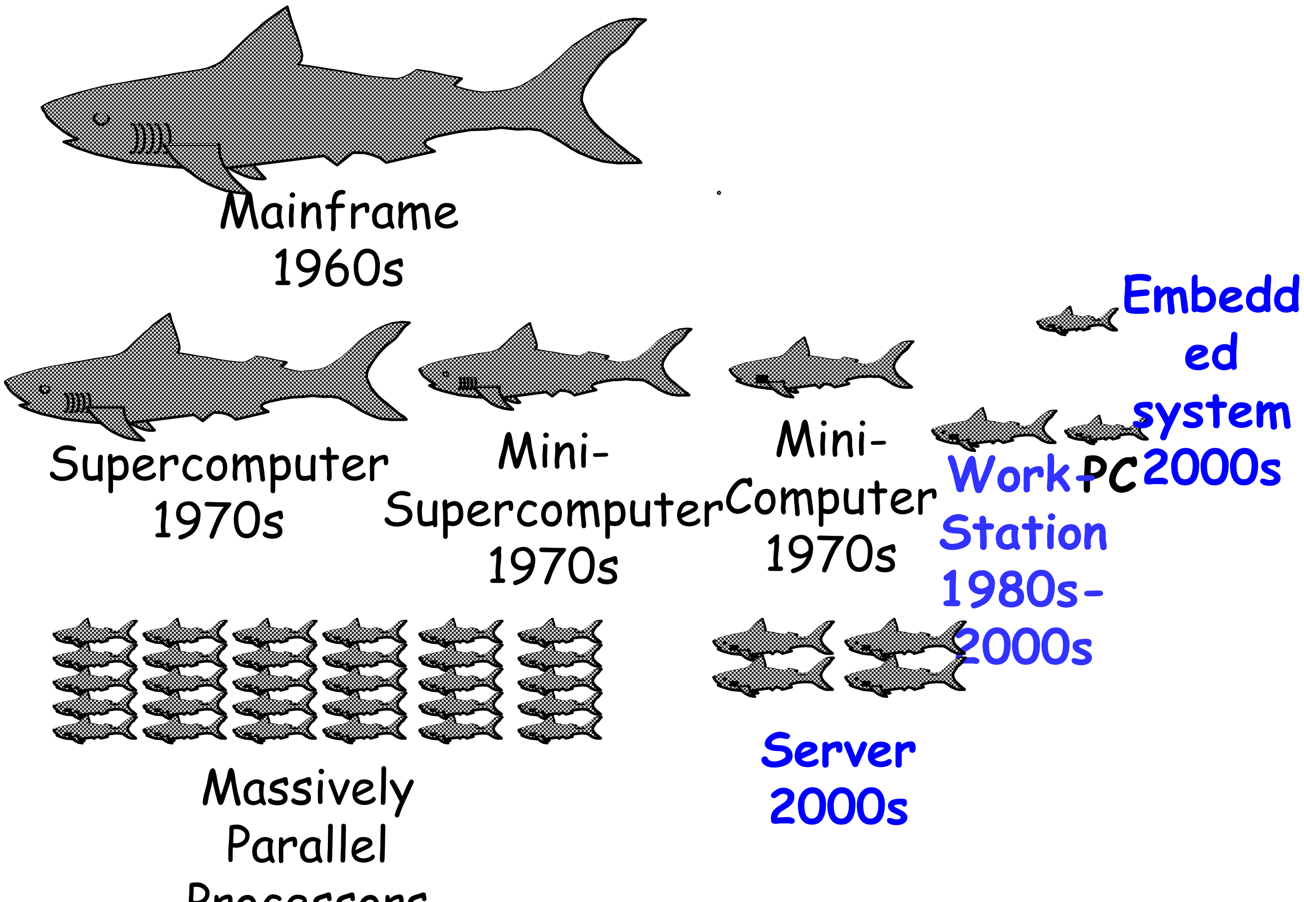
- Currently, the most common type of parallel computer. Most modern computers fall into this category.
- Multiple Instruction: every processor may be executing a different instruction stream
- Multiple Data: every processor may be working with a different data stream
- Execution can be synchronous or asynchronous, deterministic or non-deterministic
- Examples: most current supercomputers, networked parallel computer "grids" and multi-processor SMP computers - including some types of PCs.

MIMD


IS
SM
IS
CU1
PE1
MM1
I/O
…
…
…
I/O
IS
CUn
PEn
MMn
IS


prev instruct
load A(1)
load B(1)
C(1)=A(1)*B(1)
store C(1)
next instruct
P1
prev instruct
call funcD
x=y*z
sum=x*2
call sub1(i,j)
next instruct
P2
prev instruct
do 10 i=1,N
alpha=w**3
zeta=C(i)
10 continue
next instruct
Pn
time

Classification-market



Effect of dramatic performance growth

- Enhanced the capability available to computer users.
- Microprocessor-based computers across the entire range of the computer design.
  - Minicomputer => servers using microprocessors
  - Mainframe => multiprocessors consisting of microprocessors
  - Supercomputer => multiprocessor collections

Four computing markets

Feature	Mobile	Desktop	Server	Embedded
Price of system	\$100–\$1000	\$300–\$2500	\$5000 -\$5,000,000	\$10 -\$100,000
Price of microprocessor module	\$10–\$100	\$50-\$500 per proc.	\$200 -\$10,000 per proc.	\$0.01 -\$100 per proc.
Critical system design issues	Cost, energy, media performance, responsiveness	Price-perf. Graphics perf.	Throughput, availability, scalability, energy	Price, Power consumption, application-specific perf.

Desktop Computing

- The first, and still the largest market in dollar terms, is desktop computing.
- Requirement:
  - Optimized price-performance

- New challenges:
  - Web-centric, interactive application
  - How to evaluate performance?

Servers

- The role of servers to provide larger scale and more reliable file and computing services grew.
  - For servers, different characteristics are important. First, dependability is critical.
  - A second key feature of server systems is an emphasis on scalability.
  - Lastly, servers are designed for efficient throughput.

Internet of Things/Embedded Computers

- Have the widest spread of processing power and cost.
  - 8-bit 16-bit 32-bit 64-bit
- Real time performance (soft & hard)
- Strict resource constraints
  - limited memory size, lower power consumption,...
- The use of processor cores together with application-specific circuitry.
  - DSP, mobile computing

Personal Mobile Device

- Share many of the characteristics of desktop computers.
  - Web-based and media-oriented
  - Ability to run third-party software (APPs)
    - major difference with embedded computers

- Energy efficiency
  - Battery powered, absence of a fan
- Low-cost
- Real-time performance requirement

Questions

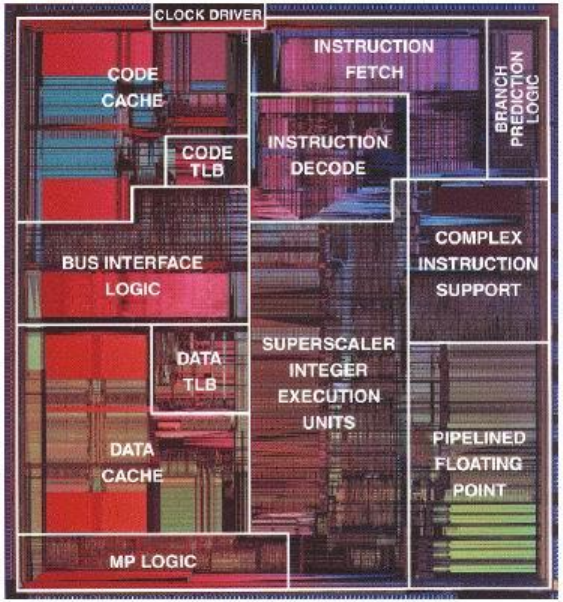
- What we need to design for different computing markets?
- What a computer Architecture designer need to know ?

Topics in Chapter

- 1.1 Why take this course ?
- 1.2 Classes of computers in current computer market
- **1.3 Defining computer architecture and What's the task of computer design.**
- 1.4 Trends in Technology
- 1.5 Trends in power in Integrated circuits
- 1.6 Trends in Cost
- 1.7 Dependability
- 1.8 Measuring, Reporting and summarizing Perf.
- 1.9 Quantitative Principles of computer Design
- 1.10 Putting it altogether

What computer architecture is all about

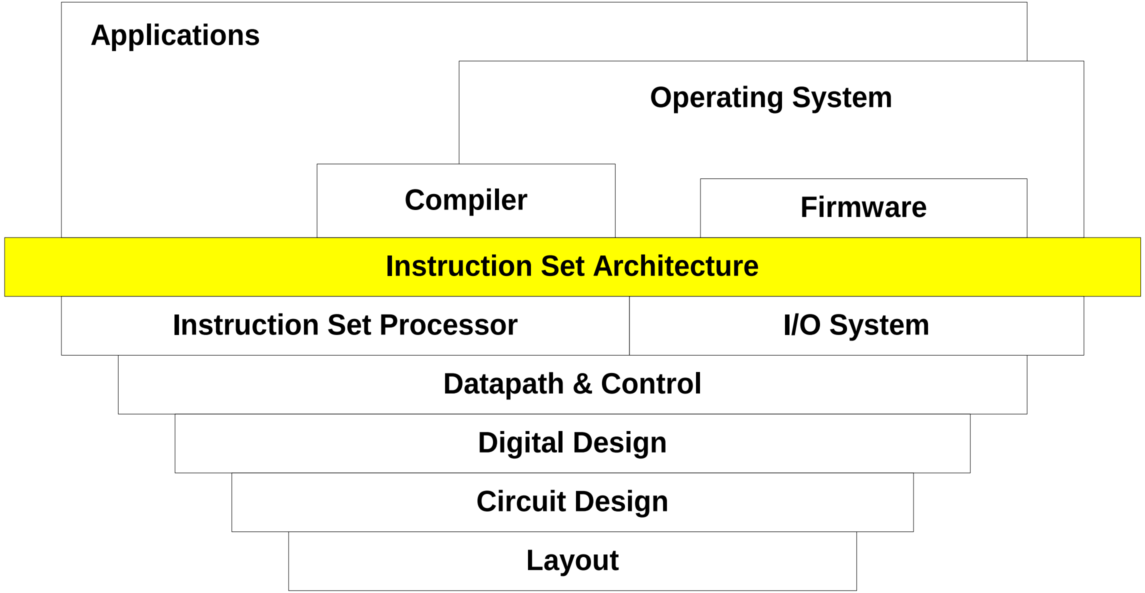
- What are the components of a computer?
- How to effectively put together the various components



Original Concept of Computer architecture

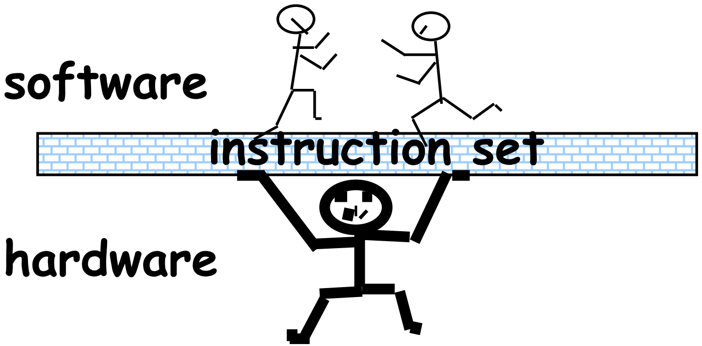
- The attributes of a [computing] system as seen by the programmer, i.e.,
- The conceptual structure and functional behavior, as distinct from the organization of the data flows and controls the logic design, and the physical implementation.
  - Amdahl, Blaaw, and Brooks, 1964

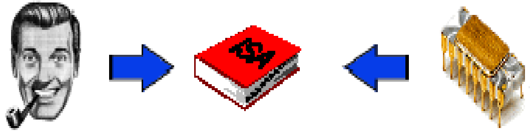
Instruction Set Architecture (ISA)



ISA: the interface between hardware and software

- Purpose 1: (now irrelevant)
  - Re-use of fixed hardware resources
- Purpose 2:
  - Interface between developer and hardware
  - Contract from one chip generation and the next





Evolution of Instruction Sets

Single Accumulator (EDSAC 1950)

Accumulator + Index Registers
(Manchester Mark I, IBM 700 series 1953)

Separation of Programming Model
from Implementation

High-level Language Based
(B5000 1963)

Concept of a Family
(IBM 360 1964)

General Purpose Register Machines

Complex Instruction Sets
(Vax, Intel 432 1977-80)

Load/Store Architecture
(CDC 6600, Cray 1 1963-76)

RISC
(Mips,Sparc,HP-PA,IBM RS6000, . . .1987)

Interface Design

- A good interface:
  - Lasts through many implementations (portability, compatibility)
  - Usable in many different scenarios (generality)
  - Provides convenient functionality to higher levels
  - Permits an efficient implementation at lower levels

Seven dimensions of ISA

- Class
- Memory addressing
- Addressing modes
- Types and sizes of operands
- Operations
- Control flow instructions
- Encoding

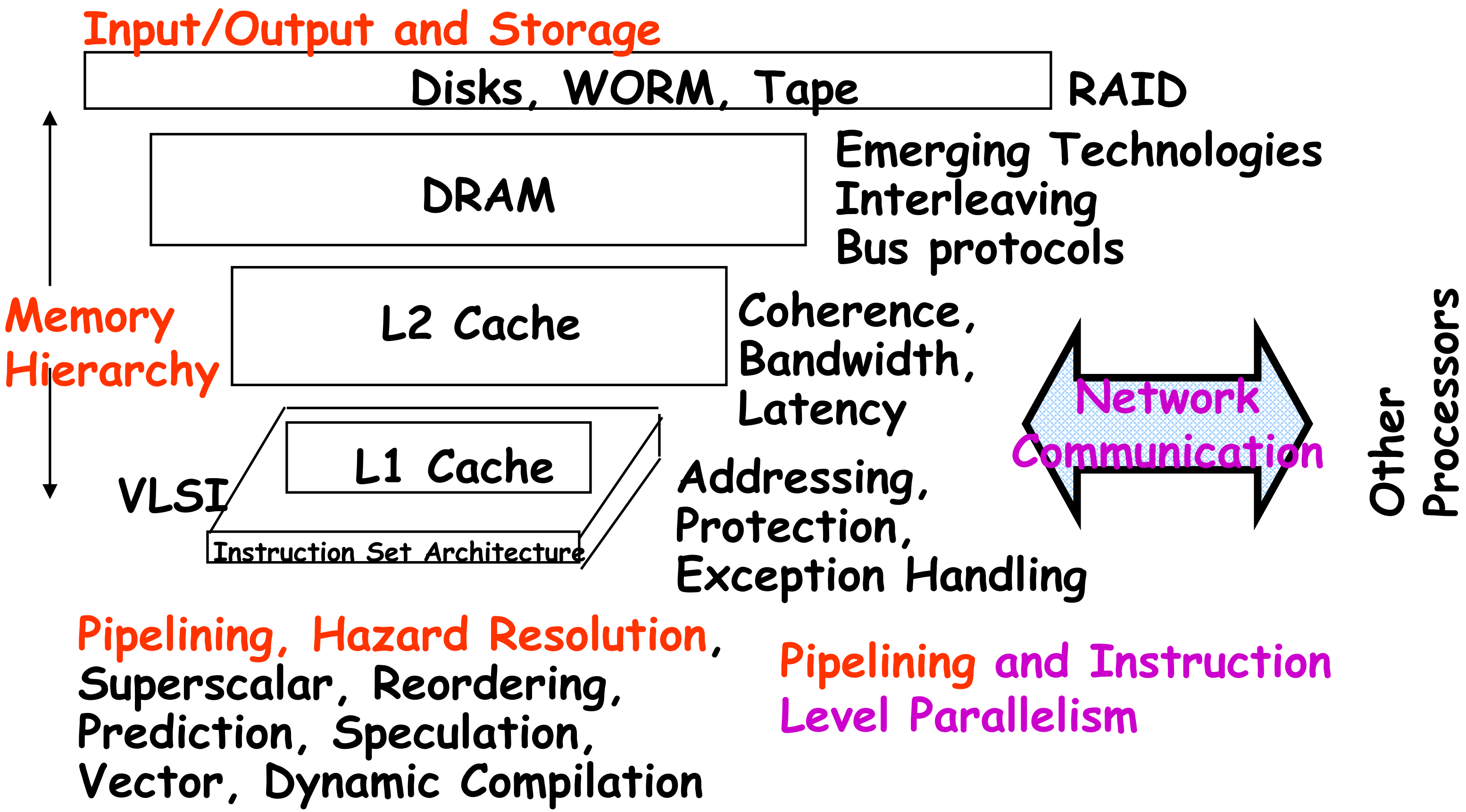
Evolution of Computer Architecture Course

- 1950s to 1960s:
  - Computer Arithmetic
- 1970s to mid 1980s:
  - Instruction Set Design, especially ISA appropriate for compilers
- 1990s:
  - Design of CPU, memory system, I/O system, Multiprocessors, Networks.
- 2010s:
  - Multicore, Self adapting systems? Self organizing structures?
  - Power-aware design, reconfigurable
- 2020s:
  - Heterogeneous accelerator, GPU, FPGA
  - Security design

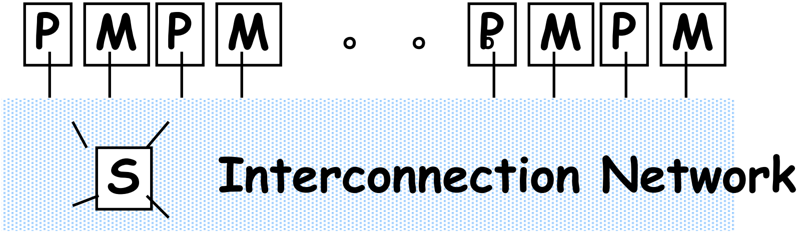
Computer Architecture

- Computer Architecture is the science and art of selecting and interconnecting hardware components to create computers that meet functional, performance and cost goals.
- It Covers:
  - Instruction Set design
  - Organization: high level of aspects of a computer's design
    - Memory, memory interconnect, internal CPU
  - Hardware: specifics of computer
    - Detailed logic design, packaging, cooling system, board displacement, power

Computer Architecture Topics



Computer Architecture Topics



Processor-Memory-Switch

Multiprocessors
Networks and Interconnections

Shared Memory,
Message Passing,
Data Parallelism

Network Interfaces

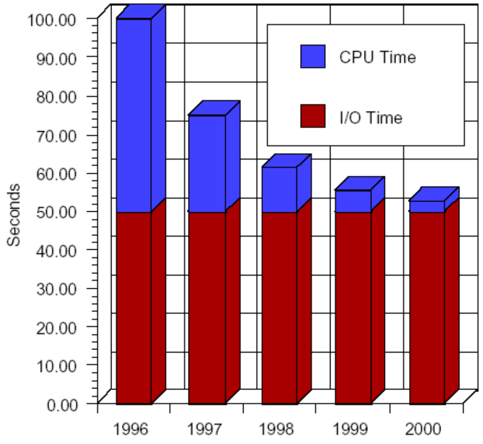
Topologies,
Routing,
Bandwidth,
Latency,
Reliability

The Task of Computer Design 1

- Define the user requirement:
  - Functional requirement: Fig1.4
    - Application area
    - Level of software compatibility
    - OS requirements
    - Standards
  - Nonfunctional requirements:
    - Price/performance
    - Availability, scalability, throughput, ...
    - Power, size, memory, temperature, ...

Application Performance

- 1996 - 1997
  - CPU performance improves by N = 400/200 = 2
  - program performance improves by N = 100/55 = 1.81
- 1997 - 1998
  - CPU performance - factor of 2
  - program performance N = 55/32.5 = 1.7
- 1998 - 1999
  - CPU performance - factor of 2
  - program performance N = 32.5 / 21.25 = 1.53
- 1999 - 2000
  - CPU Performance - factor of 2
  - program performance N = 21.25 / 15.6 = 1.36



Performance for Web Surfing

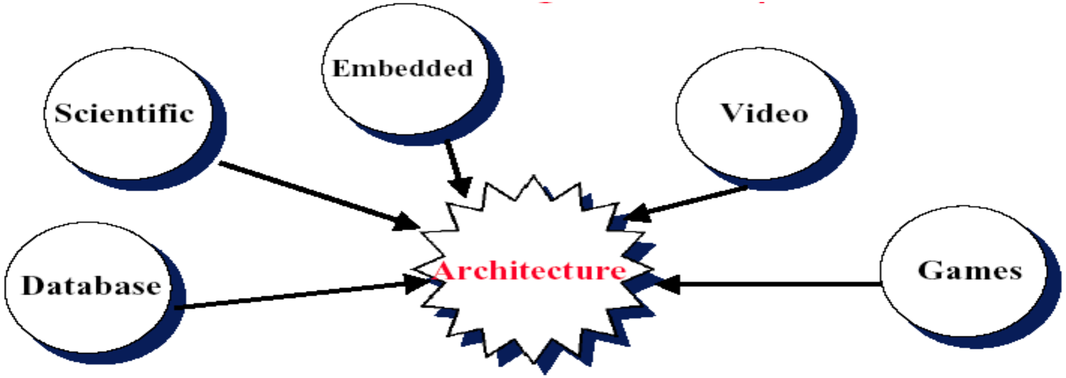
- Assume 50% CPU, 50% I/O
- 1996 - 1997
  - CPU performance improves by N = 400/200 = 2
  - Program performance improves by N = 100/75 = 1.33
- 1997 - 1998
  - CPU performance *= 2
  - Program performance N = 75/62.5= 1.2
- 1998 - 1999
  - CPU performance *= 2
  - Program performance N = 62.5/56.5 = 1.11

Computer Applications

- Architects need to understand applications' behavior
  - We say we design general purpose processors, but we should also focus on specific sets of applications
  - Architecture can be tuned for applications
- Types of applications today
  - Scientific
    - Weather prediction, crash analysis, earthquake analysis, medical imaging, imaging of the earth (searching for oil)
  - Business
    - database, data mining, video
  - General purpose
    - Microsoft Word, Excel
  - Real-time
    - automated control systems,
  - Others: Games, Mobile

Architectures are Tuned to Applications

- HP's: 1.5 MB cache for transaction processing
- Alpha: very fast FP for scientific
- StrongARM: embedded
- Intel MMX: multimedia
- Sony EE: graphics rendering
- Applications drive the design of the processor



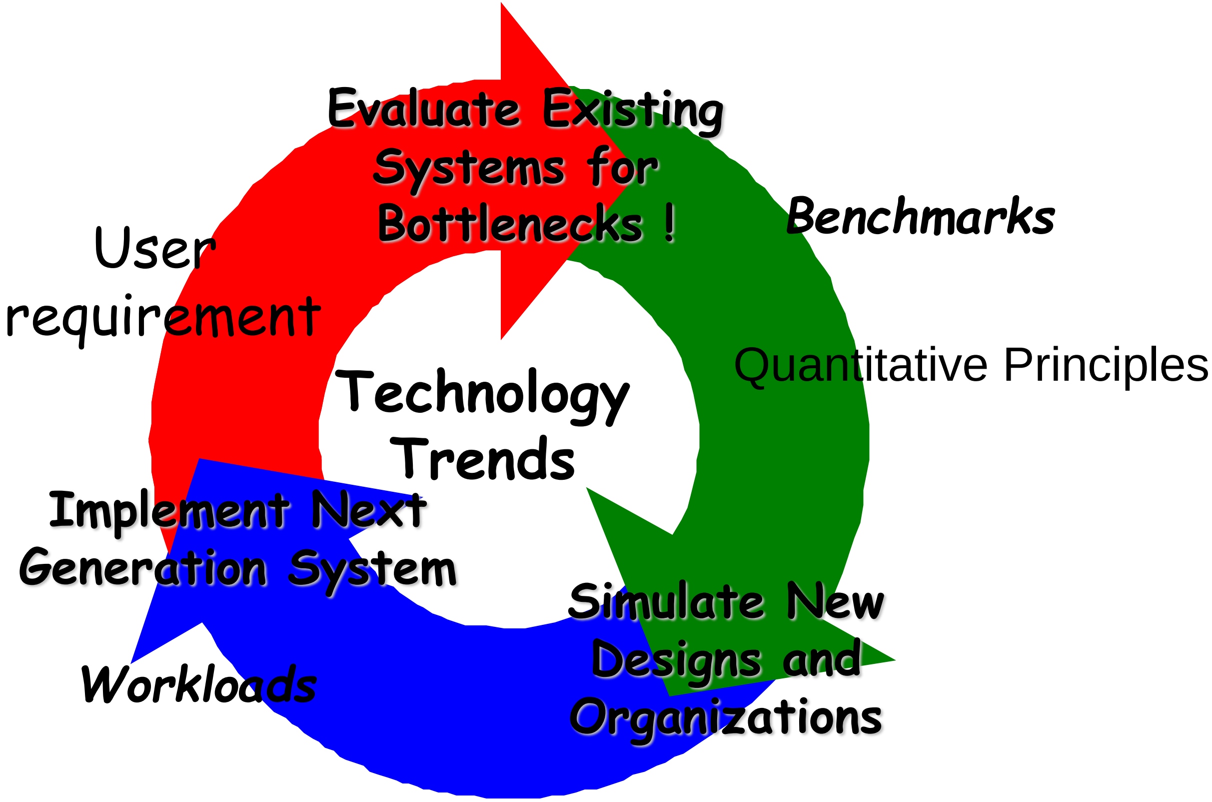
The Task of Computer Design 2

- Determine the important attributes of a new machine to maximize performance while staying with constrains, such as cost, power, availability, etc.
  - instruction set architecture design
  - functional organization
    - High level aspects of computer design, i.e. memory system, bus architecture and internal CPU design.
  - logic design (hardware)
  - implementation (hardware)

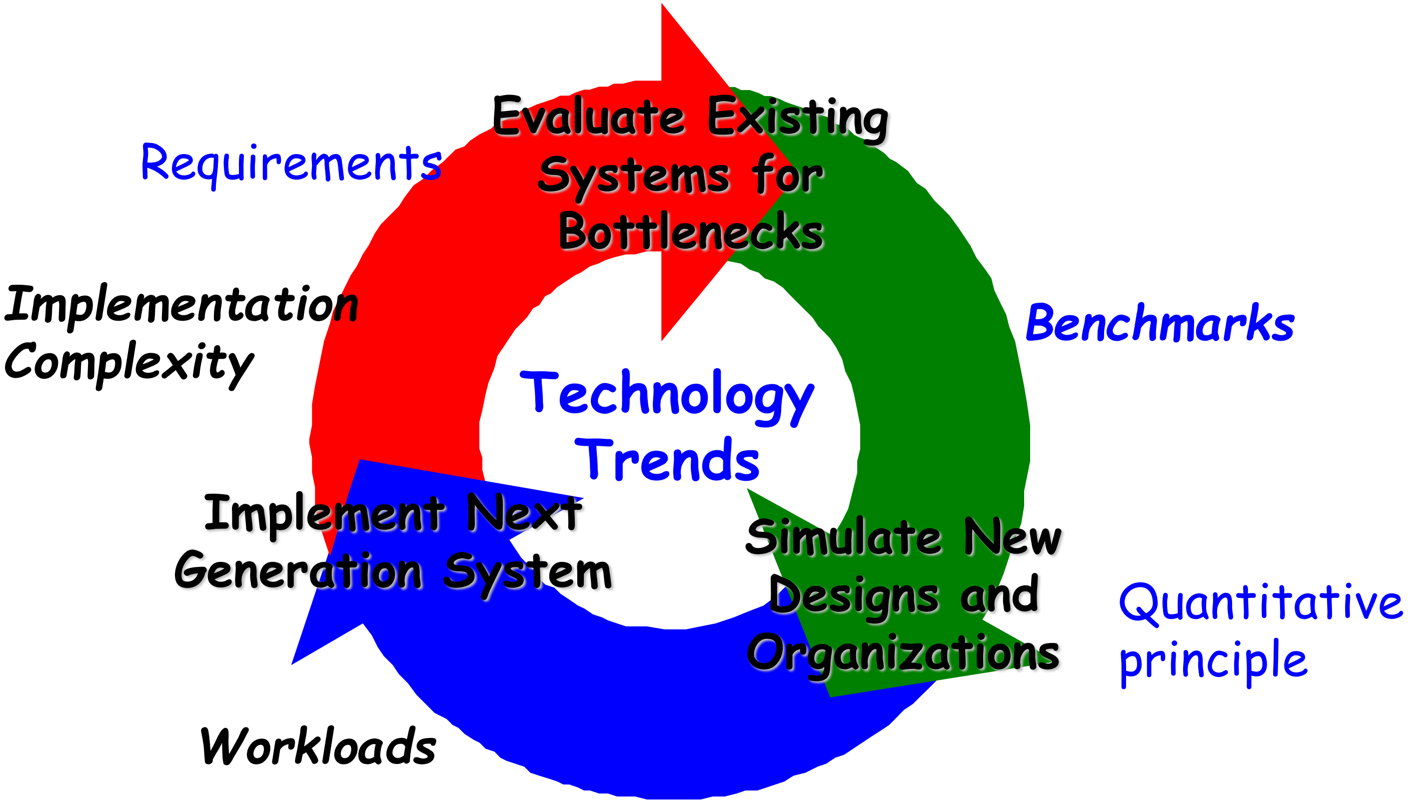
Trend of Architecture

- Emerging issues
  - High Speed
  - Multi-issue (superscalar) / Multithreading / Multiprocessor
  - CPU Cores / Multiple cores
  - Embedded
  - IRAM
- Emerging applications
  - Digital media / Digital library
  - Toaster on the internet
  - Wireless everything
  - Star Trek communicator
  - Intelligent appliances & agents

Computer Engineering Methodology



Computer Design life cycle



Summary:Task of computer design

- Considerations:
  - Functional and non functional requirements
  - Implementation complexity
    - Complex designs take longer to complete
    - Complex designs must provide higher performance to be competitive
  - Technology trends
    - Not only what's available today, but also what will be available when the system is ready to ship. (more on this later)
  - Trends in Power in IC
  - Trends in cost
- Arguments
  - Evaluate Existing Systems for Bottlenecks
- Quantitative Principles

Topics in Chapter

- 1.1 Why take this course ?
- 1.2 Classes of computers in current computer market
- 1.3 Defining computer architecture and What's the task of computer design?
- **1.4 Trends in Technology**
- 1.5 Trends in power in Integrated circuits
- 1.6 Trends in Cost
- 1.7 Dependability
- 1.8 Measuring, Reporting and summarizing Perf.
- 1.9 Quantitative Principles of computer Design
- 1.10 Putting it altogether

Technology Trends

- Moore Law
  - In 1965 he predicted that the number of components the industry would be able to place on a computer chip would double every year. In 1975, he updated his prediction to once every two years. It has become the guiding principle for the semiconductor industry to deliver ever-more-powerful chips while decreasing the cost of electronics.

Gordon Moore

Technology Trends

- Integrated circuit logic technology
  - Transistor Density: incr. 35% per year, (4x every 4 years)
  - Die size: 10%-20% per year
  - Transistor count per chip:40-55% per year
- Semiconductor DRAM
  - Capacity: 40% per year (2x every 2 years)
  - Memory speed: about 10% per year
- Magnetic Disk tech.
  - Density: 30% per year before 1990;60% per year in1990-1996
  - 100% per year in 1996-2004;30% per-year after 2004
  - capacity: about 60% per year
- Network bandwidth

$$10\text{Mb} \xrightarrow[\text{10 years}]{} 100\text{Mb} \xrightarrow[\text{5 years}]{} 1\text{Gb}$$

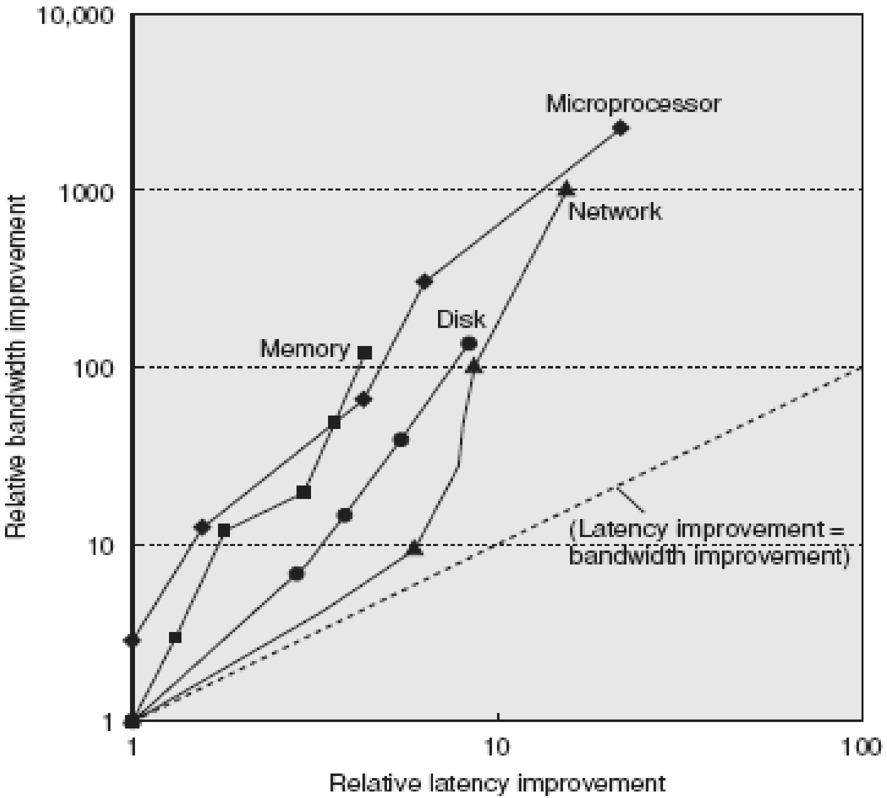
Designers often design for the next technology.

Notes

- A rule of thumb
  - Cost decrease rate ~ density increase rate
- Technology thresholds
  - Technology improves continuously, an impact of this improvements can be in discrete leaps.

Perf. Trends: Bandwidth over latency

- Bandwidth/t
  amount of w
  given time
- Latency/res
  time betwee
  the complet



: ~
ncy2

Performance milestones in microprocessor

Microprocessor	16-bit address/bus, microcoded	32-bit address.bus, microcoded	5-stage pipeline, on-chip I & D caches, FPU	2-way superscalar, 64-bit bus	Out-of-order 3-way superscalar	Out-of-order superpipelined, on-chip 1.2 cache
Product	Intel 80286	Intel 80386	Intel 80486	Intel Pentium	Intel Pentium Pro	Intel Pentium 4
Year	1982	1985	1989	1993	1997	2001
Die size (mm^2)	47	43	81	90	308	217
Transistors	134,000	275,000	1,200,000	3,100,000	5,500,000	42,000,000
Pins	68	132	168	273	387	423
Latency (clocks)	6	5	5	5	10	22
Bus width (bits)	16	32	32	64	64	64
Clock rate (MHz)	12.5	16	25	66	200	1500
Bandwidth (MIPS)	2	6	25	132	600	4500
Latency (ns)	320	313	200	76	50	15

Challenges for IC Tecnology

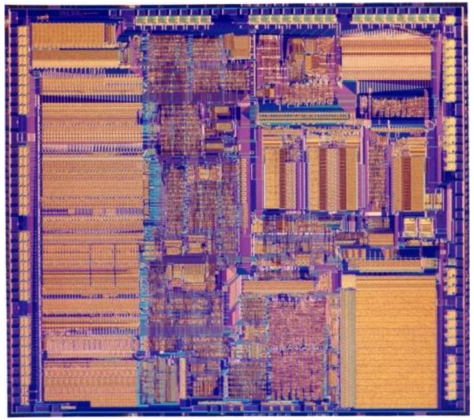
- IC characteristic: feature size
  - 10 microns in 1971 $\rightarrow$ 0.18microns in 2001
  - → 0.09 microns in 2006 → 65nm
  - →40nm →28nm →14nm →7nm →5nm…
  - Rule of thumb: transistor performance Improves linearly with decreasing feature size.
- IC density improvement is both opportunity and Challenge:
  - Signal delay for a wire increase in proportion to the product of its resistance and capacitance.
  - Major design limitation: signal delay

Topics in Chapter

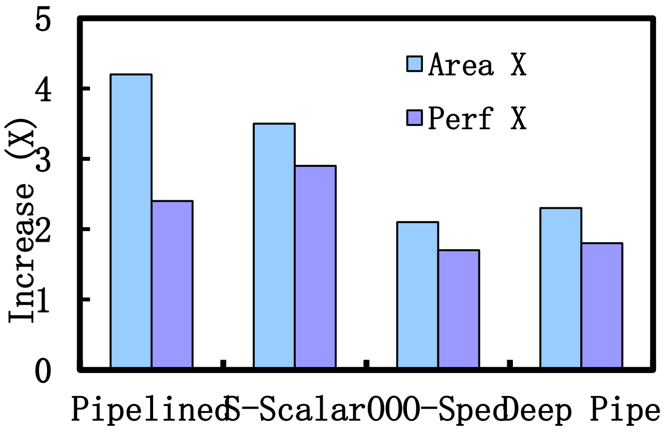
- 1.1 Why take this course ?
- 1.2 Classes of computers in current computer market
- 1.3 Defining computer architecture and What's the task of computer design?
- 1.4 Trends in Technology
- **1.5 Trends in power in Integrated circuits**
- 1.6 Trends in Cost
- 1.7 Dependability
- 1.8 Measuring, Reporting and summarizing Perf.
- 1.9 Quantitative Principles of computer Design
- 1.10 Putting it altogether

Trends in Power

- Power also provide challenges as device scaled
  - First microprocessor: 1/10 watt
  - 2GHz Pentium 4: 135 watt
- Challenges:
  - Distributing the power
  - Removing the heat
  - Preventing hot spot

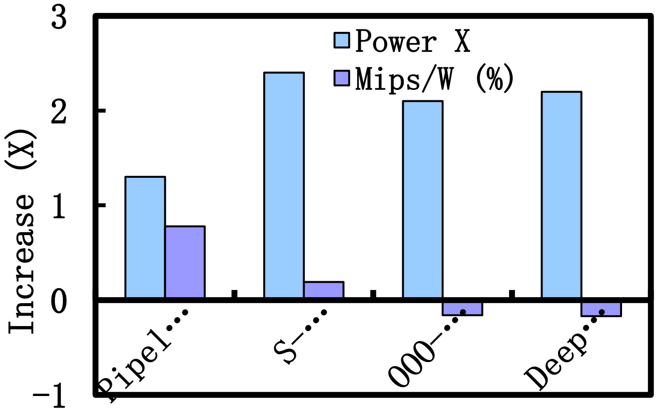


May 1986	17 Years	August 27, 2003
@16 MHz core	200x	@3.2 GHz core
275,000 1.5µ transistors	200x/11x	55 Million 0.13µ transistors
~1.2 SPECint2000	1000x	1249 SPECint2000



Performance scales
with area**.5

Power efficiency
has dropped



Two concepts

- Dynamic power: power consumption in switching transistors.
  - Power dynamic = ½ *Capacitive load * Voltage2 * Frequency switched
  - Energy dynamic = Capacitive load * Voltage2
- Static power: power consumption when a transistor is off due to power leakage
  - Power static = current static * Voltage

Rule of Thumb

- 10% reduction of voltage yields:
  - 10% reduction in frequency
  - 30% reduction in power
  - Less than 10% reduction in performance

Rule of Thumb

Voltage	Frequency	Power	Performance
1%	1%	3%	0.66%

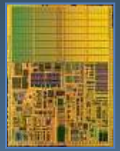
Dual core with voltage scaling

RULE OF THUMB

A 15% Reduction In Voltage Yields

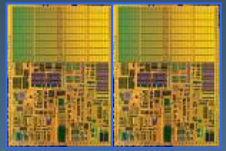
Frequency Reduction	Power Reduction	Performance Reduction
15%	45%	10%

SINGLE CORE



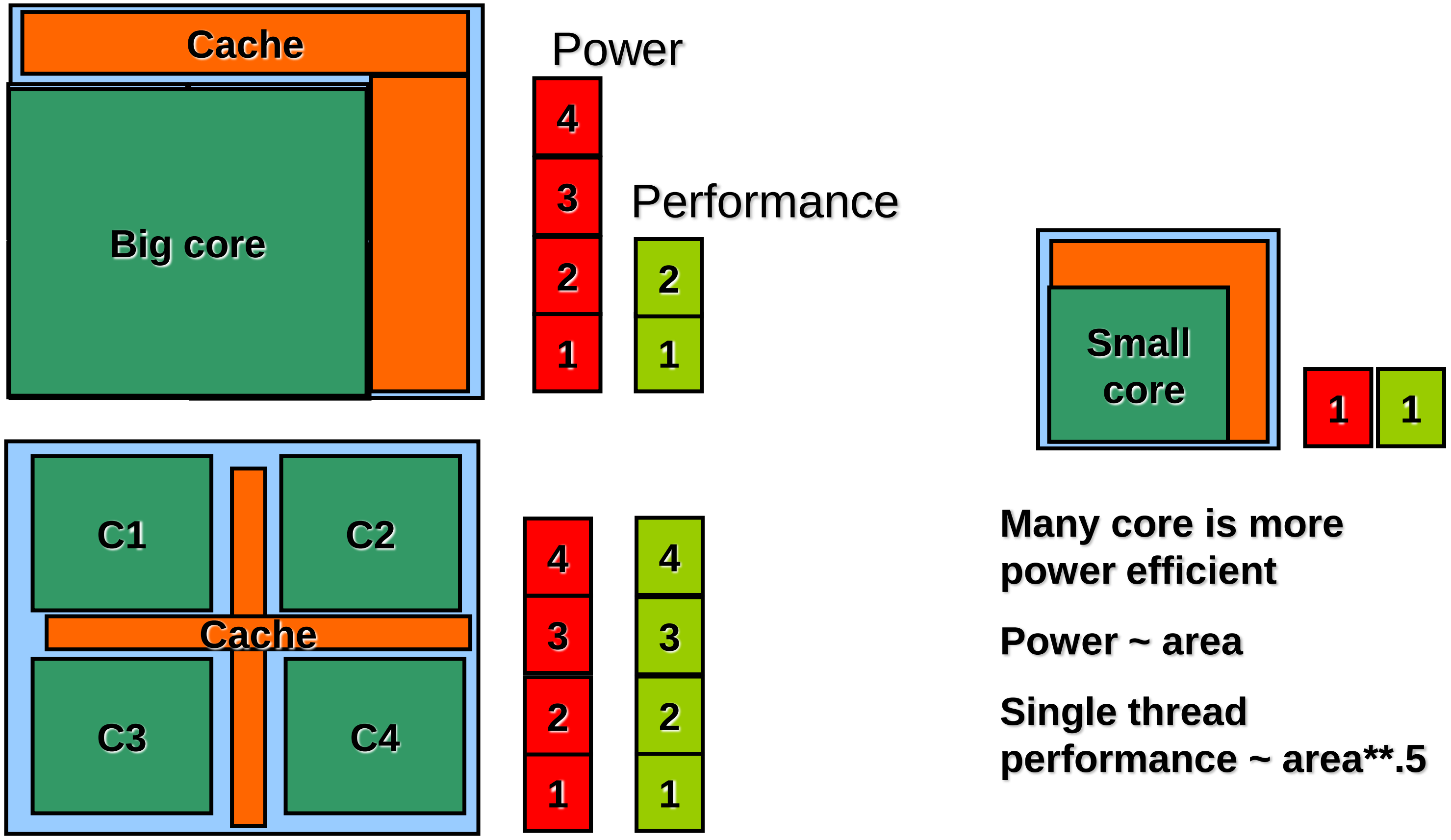
Area = 1
Voltage = 1
Freq = 1
Power = 1
Perf = 1

DUAL CORE



Area = 2
Voltage = 0.85
Freq = 0.85
Power = 1
Perf = ~1.8

Multiple cores deliver more performance per watt



Many core is more power efficient

Power ~ area

Single thread performance ~ area.5**

THANK YOU