浙江大學
ZheJiang University

Computer Architecture ----A Quantitative Approach

College of Computer of Zhejiang University
CHEN WEN ZHI
chenwz@zju.edu.cn
Room 511, CaoGuangBiao BLD

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 summerizing Perf.

1.9 Quantitative Principles of computer Design

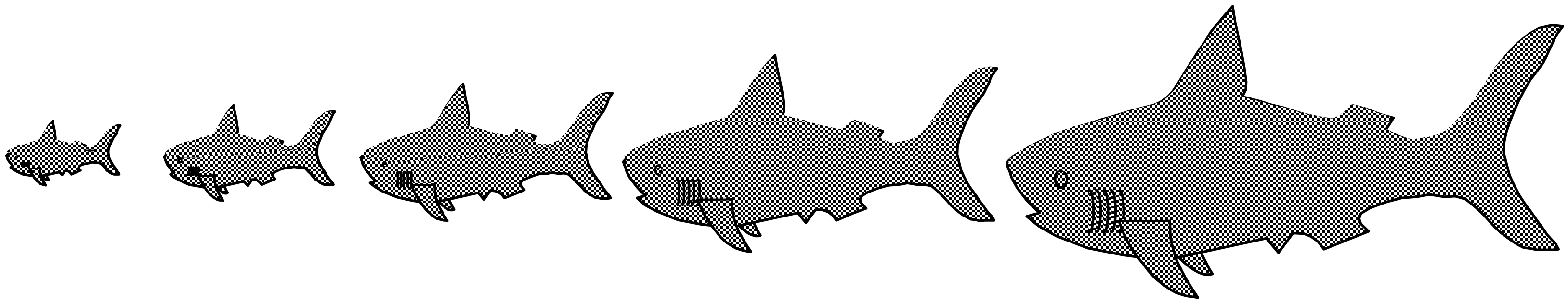
1.10 Putting it altogether

Topics in Chapter 1

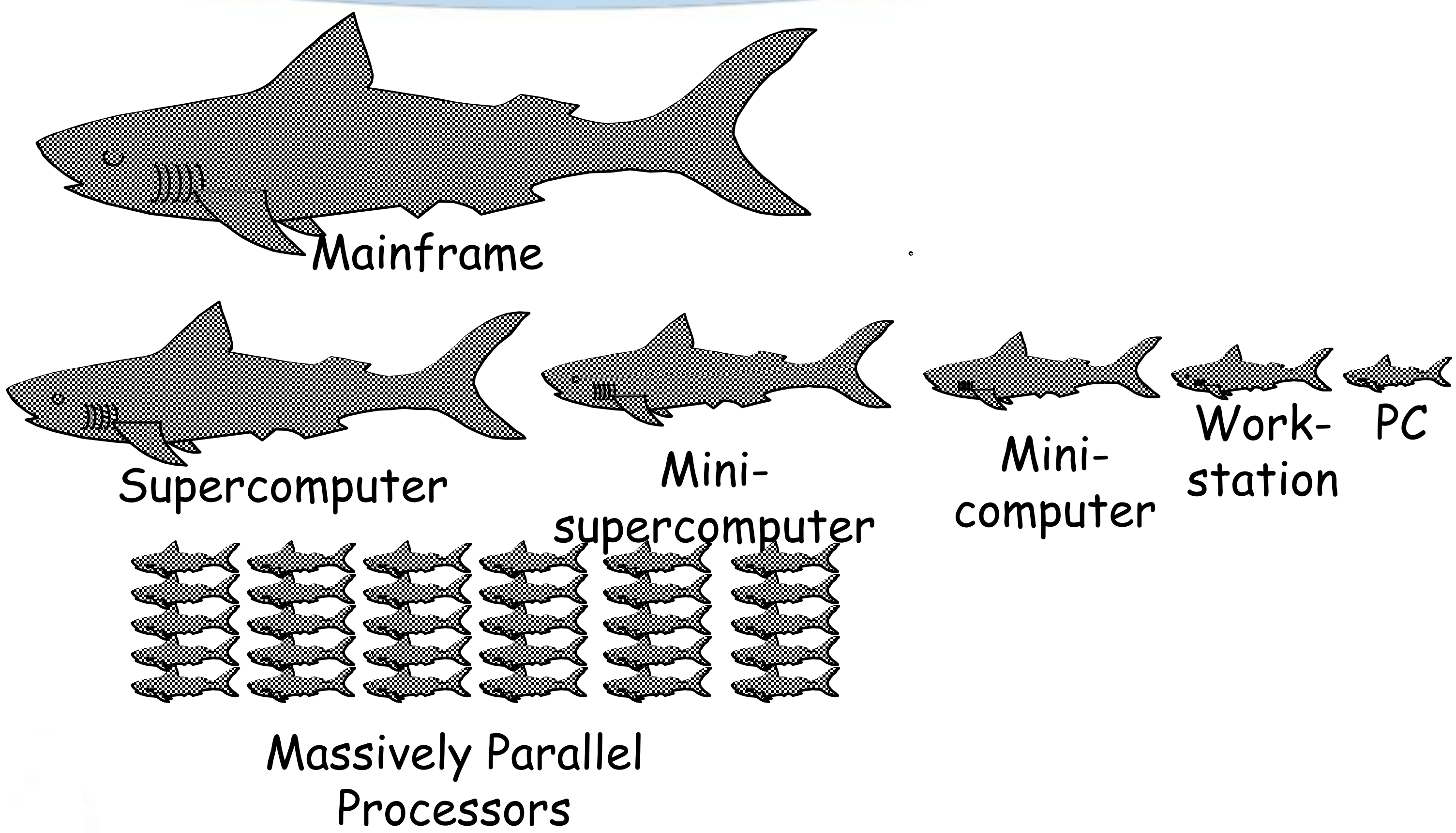
History of the Computer

- Original:

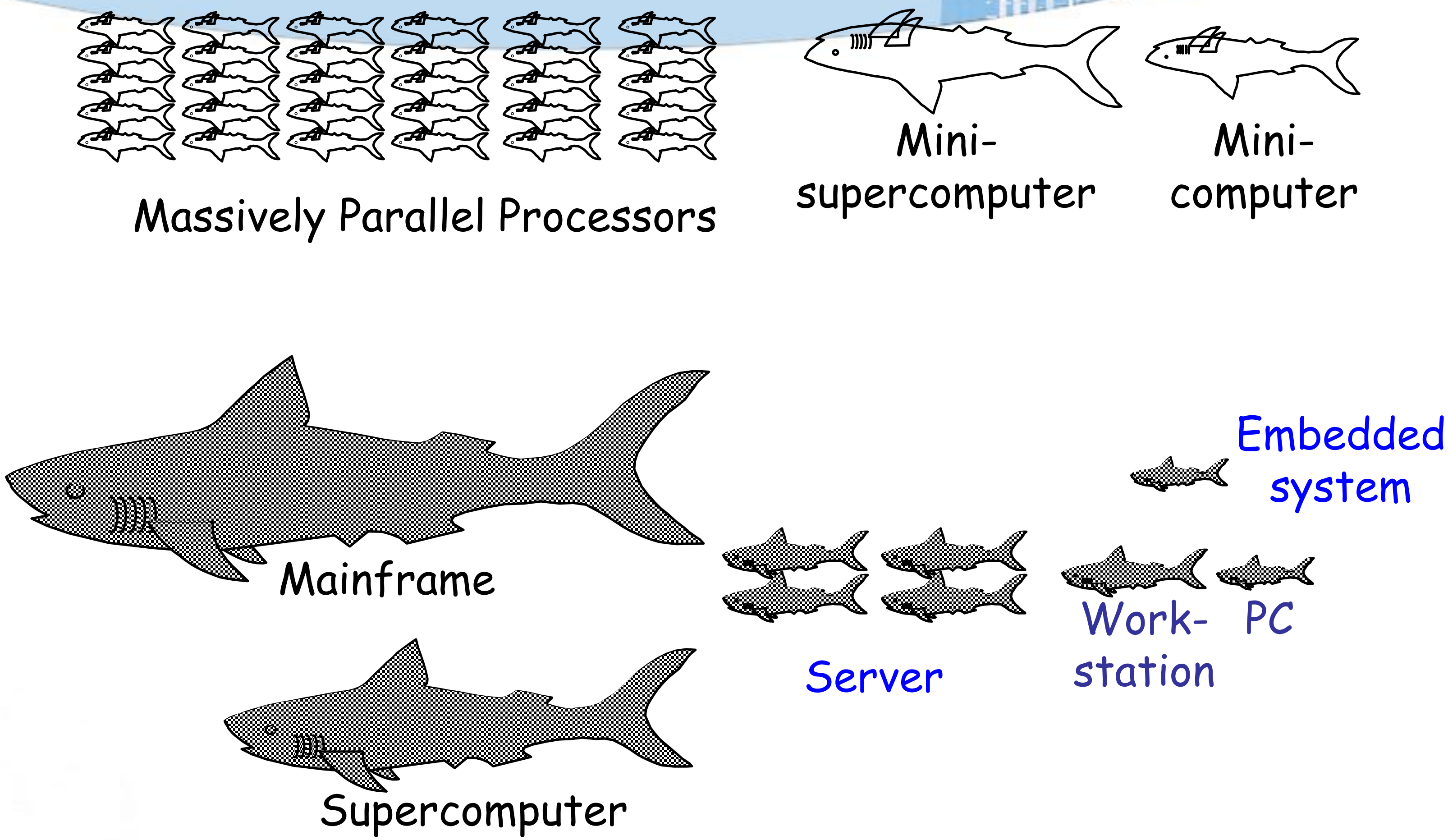
Big Fishes Eating Little Fishes

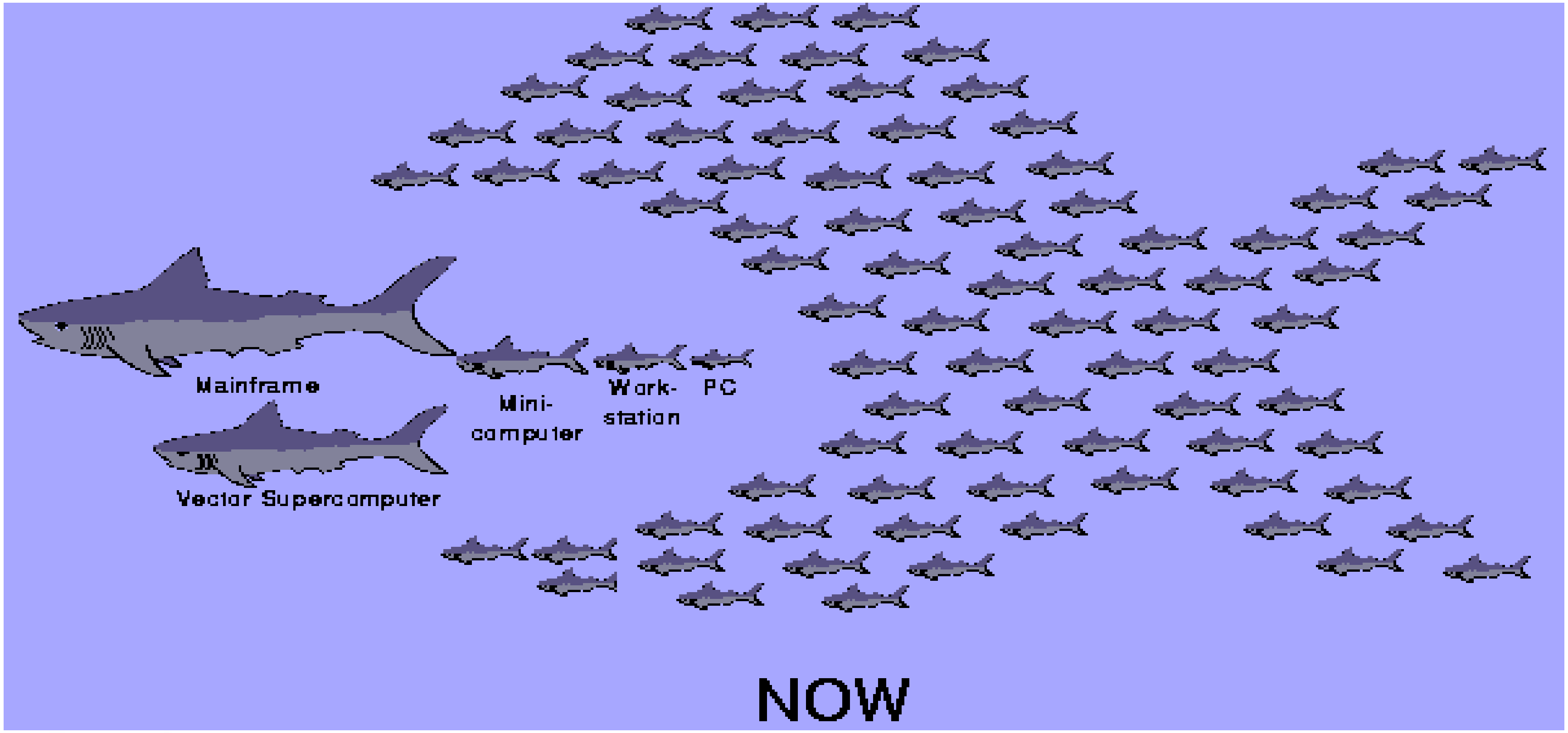


Computer Food Chain

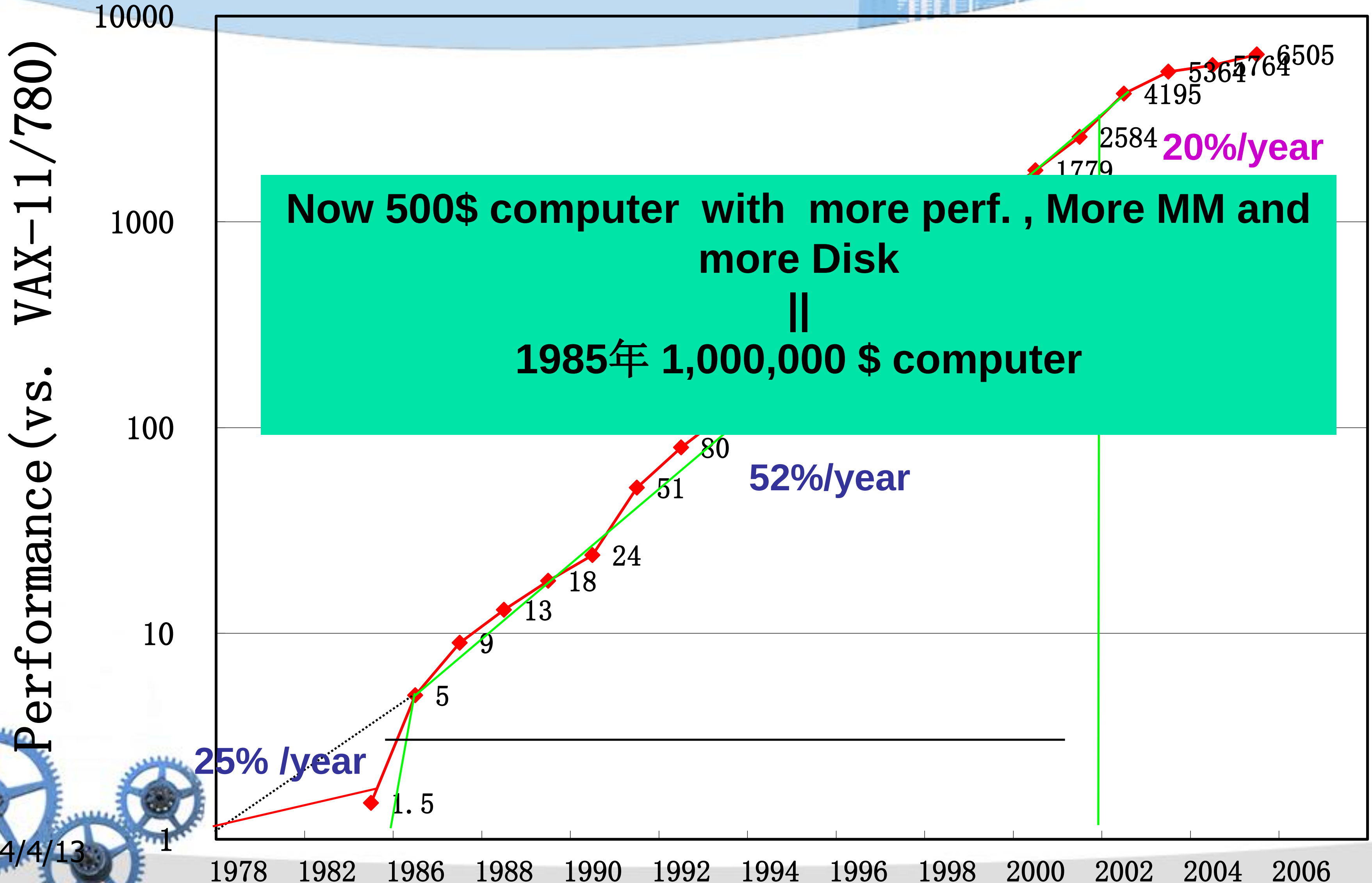


Computer Food Chain




Mainframe
Mini-
computer
Work-
station
PC
Vector Supercomputer
NOW

Incredible performance improvement



Now 500$ computer with more perf. , More MM and more Disk

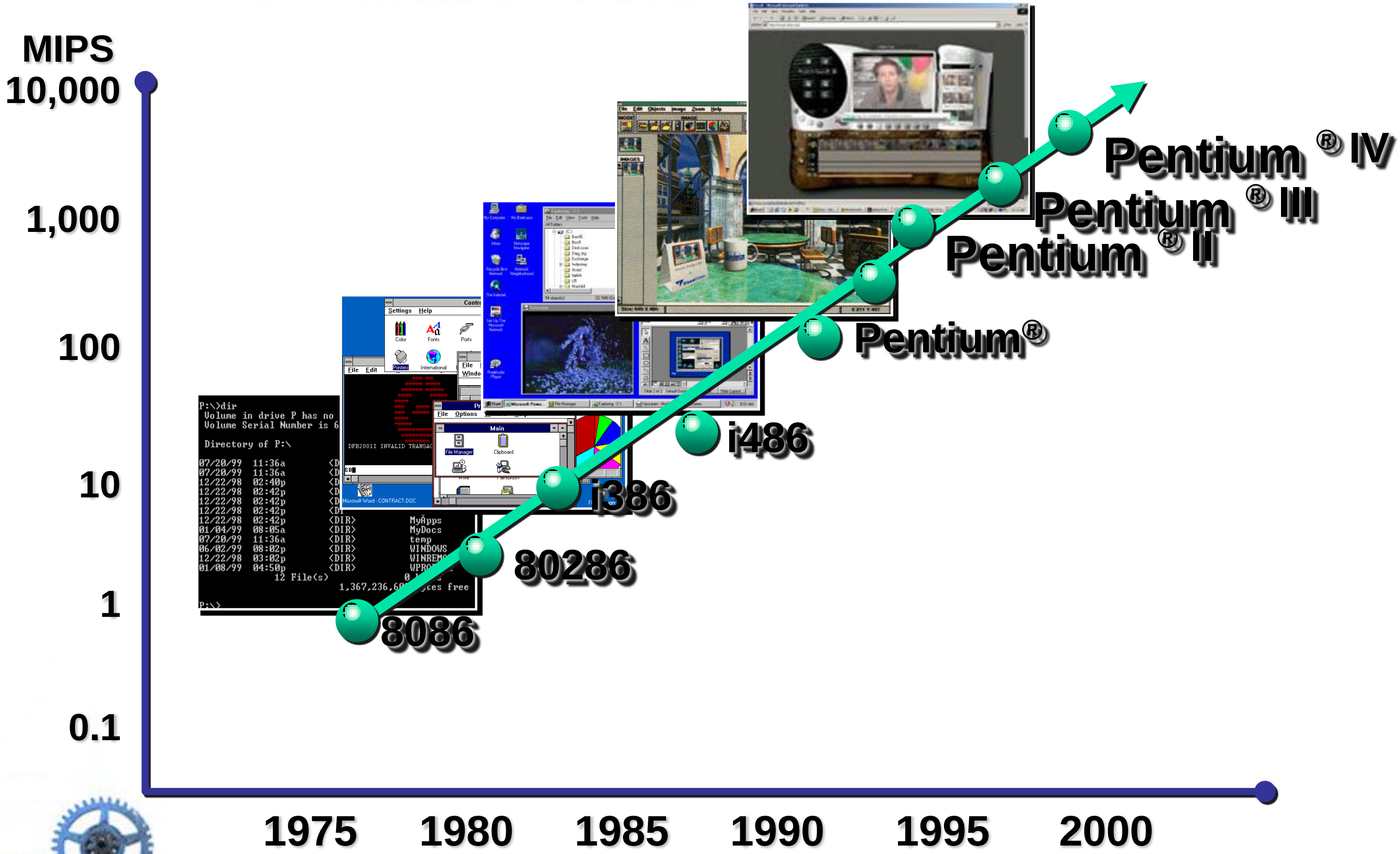
||

1985年 1,000,000 $ computer

Conclusion

- 25%: Technological improvements more steady than progress in computer architectrue
- 52%: After RISC emergence, computer design emphsized both architecutural innovation and efficient use of technology improvments.
  - CA plays an important role in perf. Improvement
- 20%: little ILP left to exploit due to power dissipation
  - Faster uniprocessor => multiple processor on chip
  - ILP => TLP and DLP
  - Implicity, compiler and hardware => Explicity, programmer

Process ability → New Applications



Why Such Change in 60 years?

- Two reasons:
  - advances in the technology used to build computers
    - **IC**
    - **Strorage 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

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

Forces on Computer Architecture

Technology

Parallelism

Programming Languages

Applications

Interface Design (ISA)

Computer Architecture:

- Instruction Set Design
- Organization
- Hardware/Software Boundary

Compilers

Operating Systems

Measurement & Evaluation

History

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 summerizing Perf.
1.9 Quantitative Principles of computer Design
1.10 Putting it altogether

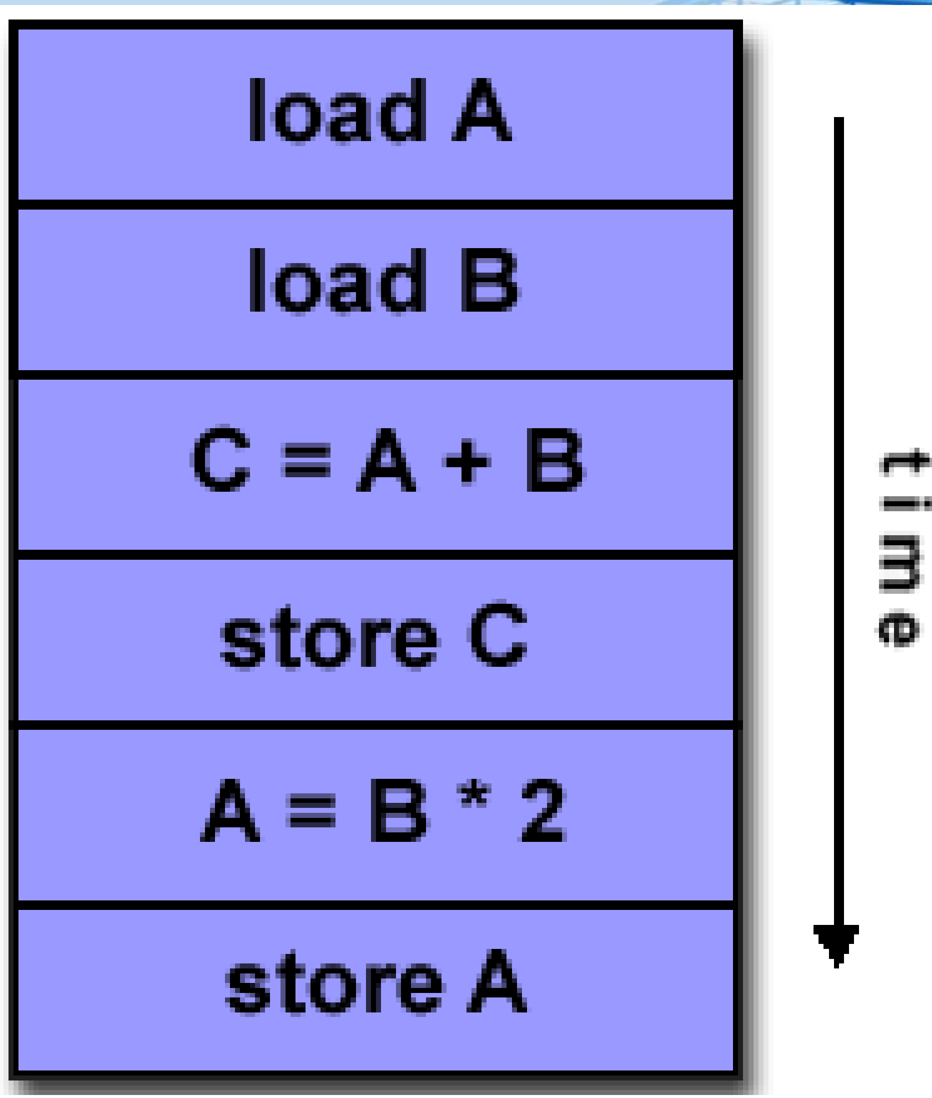
Classes of computers

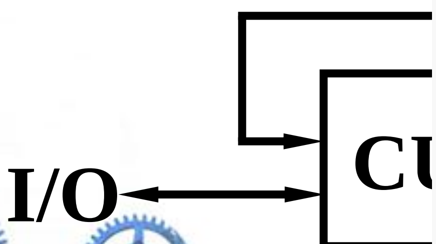
•**Flynn's Taxonomy:** **A classification of computer architectures based on the number of streams of instructions and data**

- **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

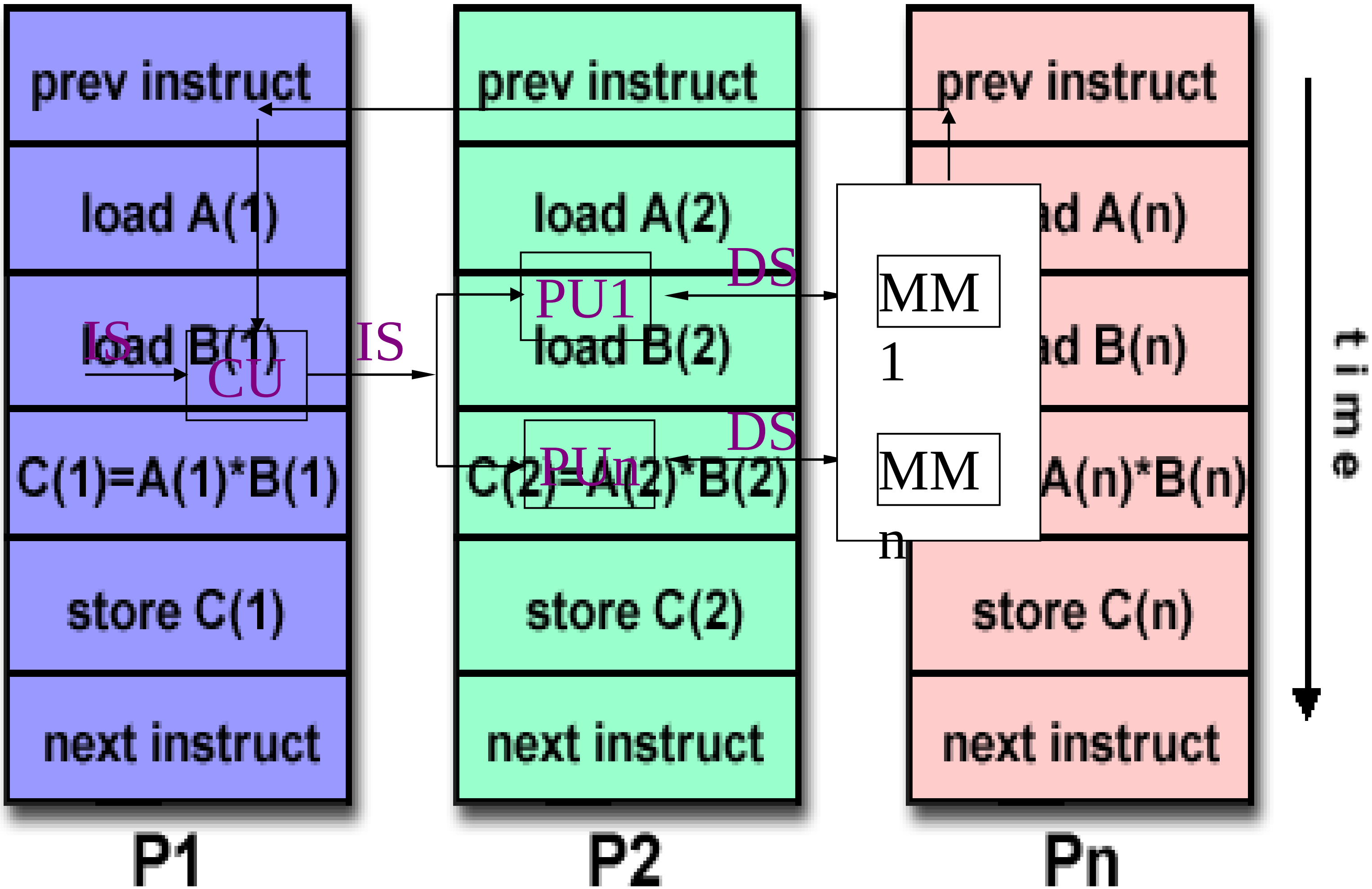
- A serial (non-pa
- **Single instruc** … is being acted on by the
- **Single data:** c … d as input during any one
- **Deterministic**
- This is the olde … evalent form of comput
- Examples: mos … d mainframes





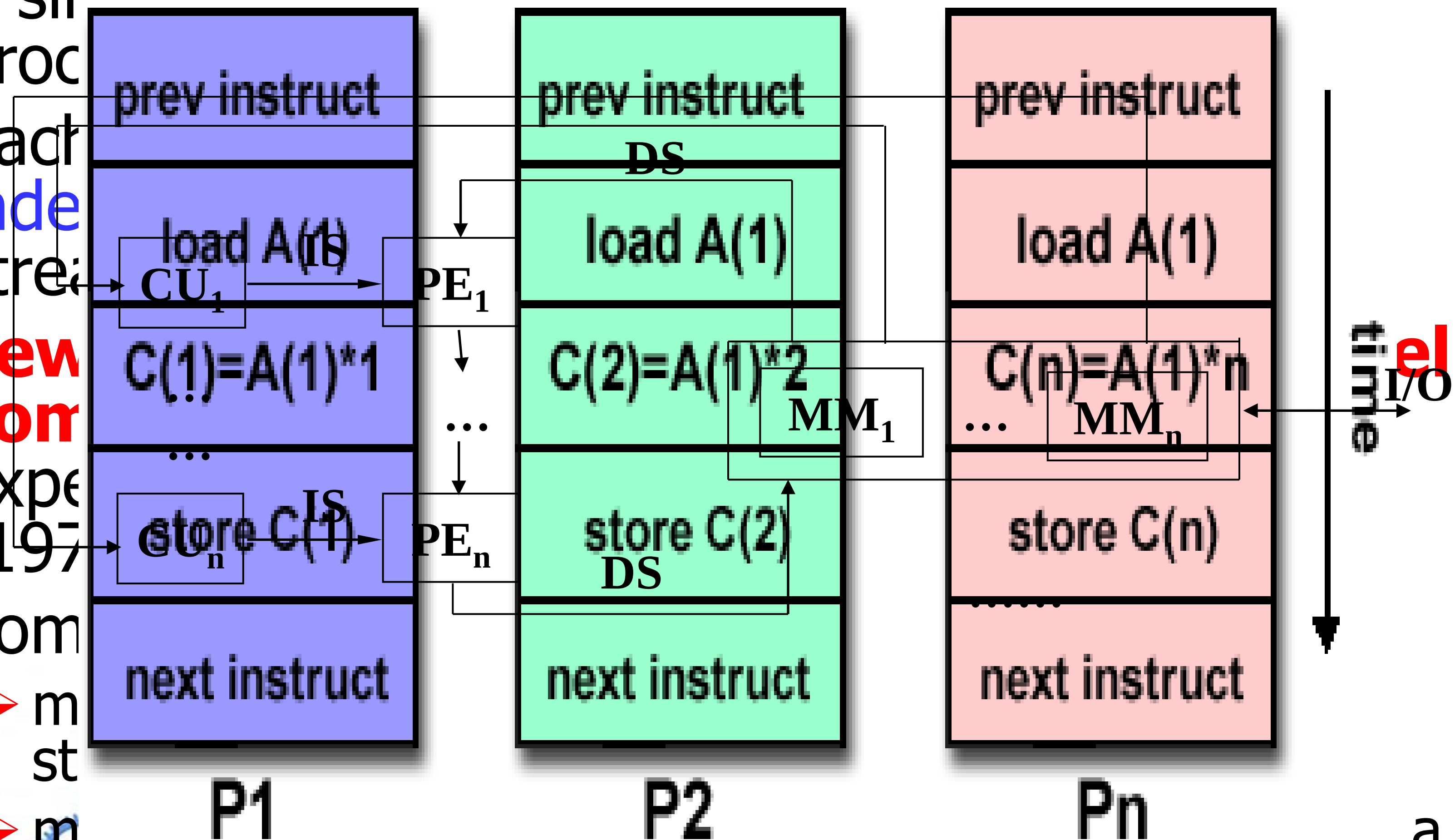
SIMD

- A type of parallel computer
- **Si** … ne
  ins
- **M** …
  dif
- Th … ner, a
  ve … rray
  of
- Be … igh
  de
- Sy
- Tw
  - … -2
  - … litachi



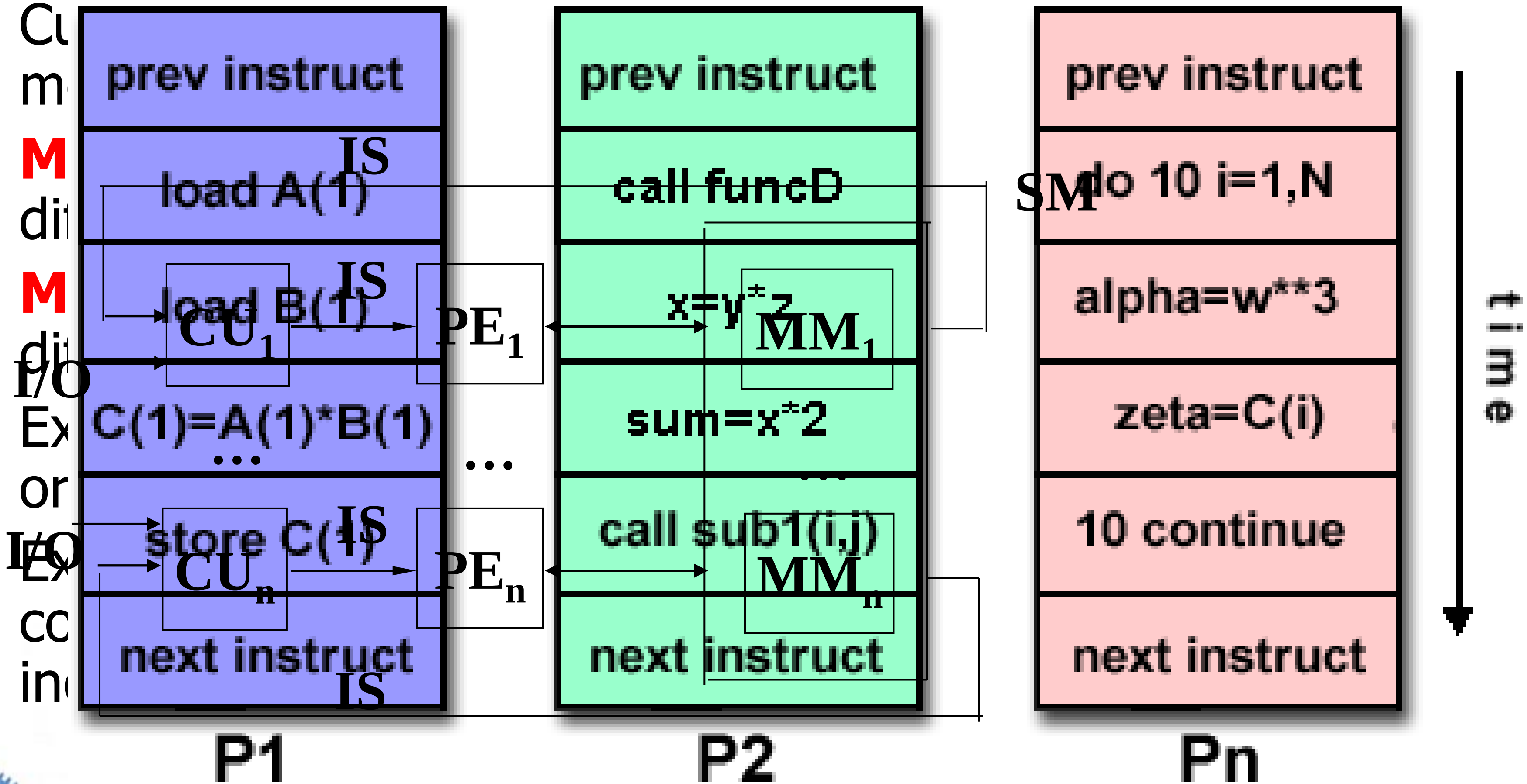
MISD

- A single data stream is fed into multiple
  proc
- Each
  inde
  strea
- **Few**
  com
  expe
  (197
- Som
  - m
    st
  - m a
    single coded message.

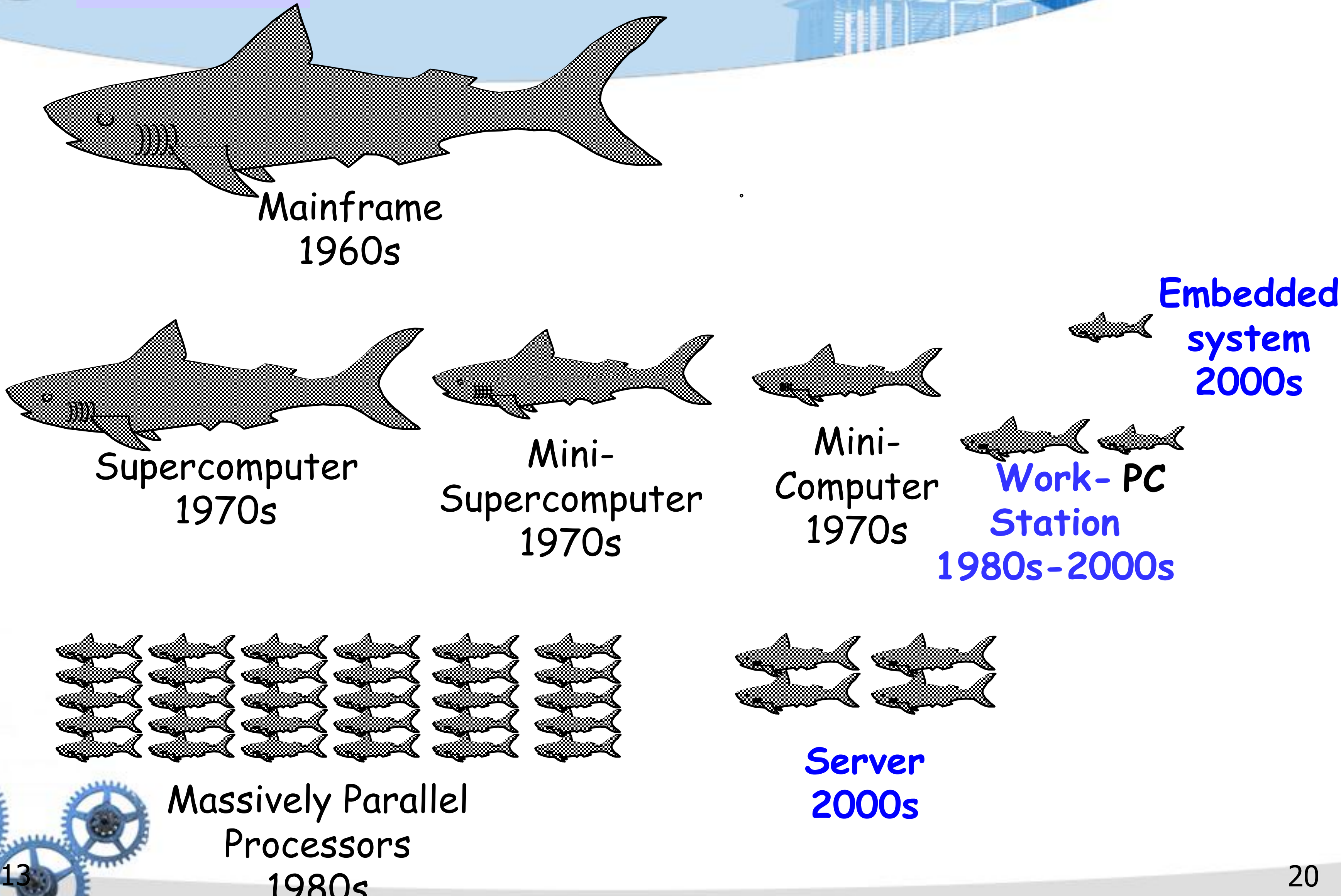


MIMD

- Cu m
- **M** di
- **M** di
- Ex or
- Ex cc in



Classification-market



Effect of dramatic perf. 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

3 computing markets

Feature	Desktop	Server	Embedded
Price of system	\$500-\$5000	\$5000 -\$5,000,000	\$10 -\$100,000
Price of microprocessor module	\$50-\$500 per proc.	\$200 -\$10,000 per proc.	\$0.01 -\$100 per proc.
Critical system design issues	Price-perf. Graphics perf.	Throughput, availability, scalability	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

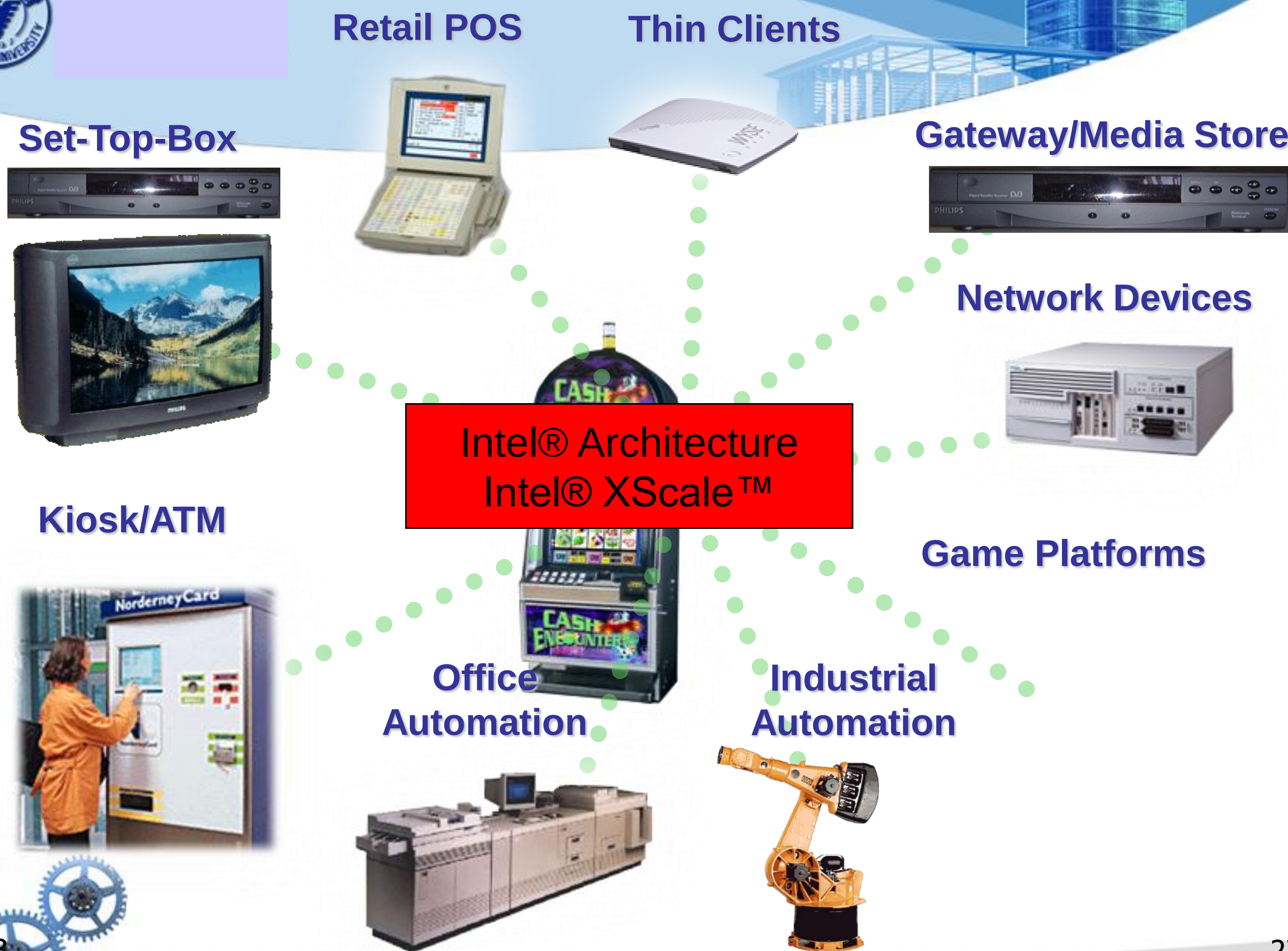
- The role of servers to provide larger scale and more reliable file and computing services grew.
  - For servers, different characteristics are important. First, dependability is critical.
  - A second key feature of server systems is an emphasis on scalability.
  - Lastly, servers are designed for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

Embedded Computers

- The fastest growing portion of the computer market.
- 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, mobile phone, Digital TV

Widest spread of power and cost

- 8-bit, 16bit microprocessor / one dime
- 32-bit 100M instructions per sec. /5$
- 32-bit, 1 billion instr. per sec. / 100$


Retail POS
Thin Clients
Set-Top-Box
Gateway/Media Store
Network Devices
Intel® Architecture
Intel® XScale™
Kiosk/ATM
Game Platforms
Office Automation
Industrial Automation

- What we need to design for this computing market ?

- 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

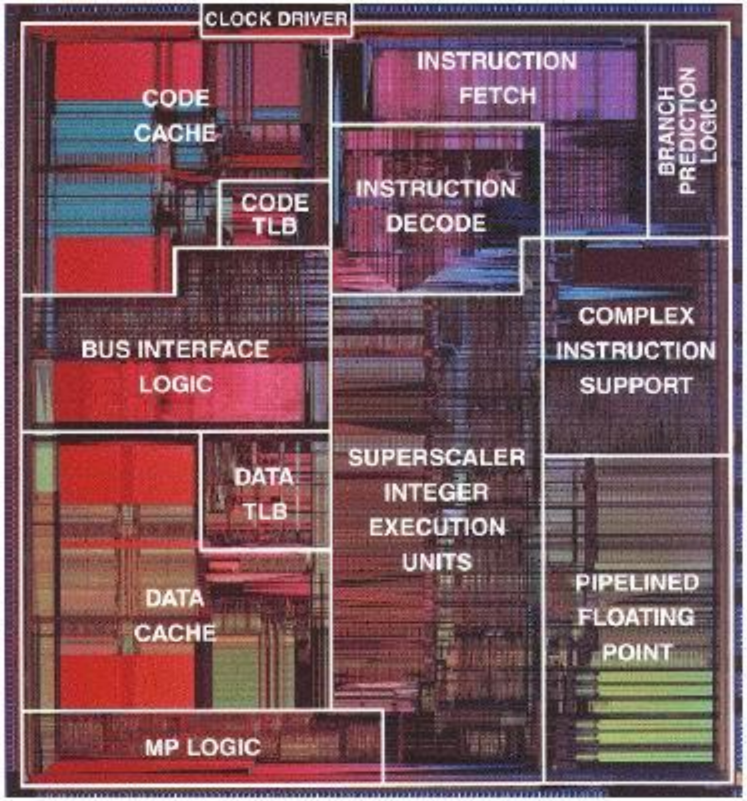
1.8 Measuring, Reporting and summerizing Perf.

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



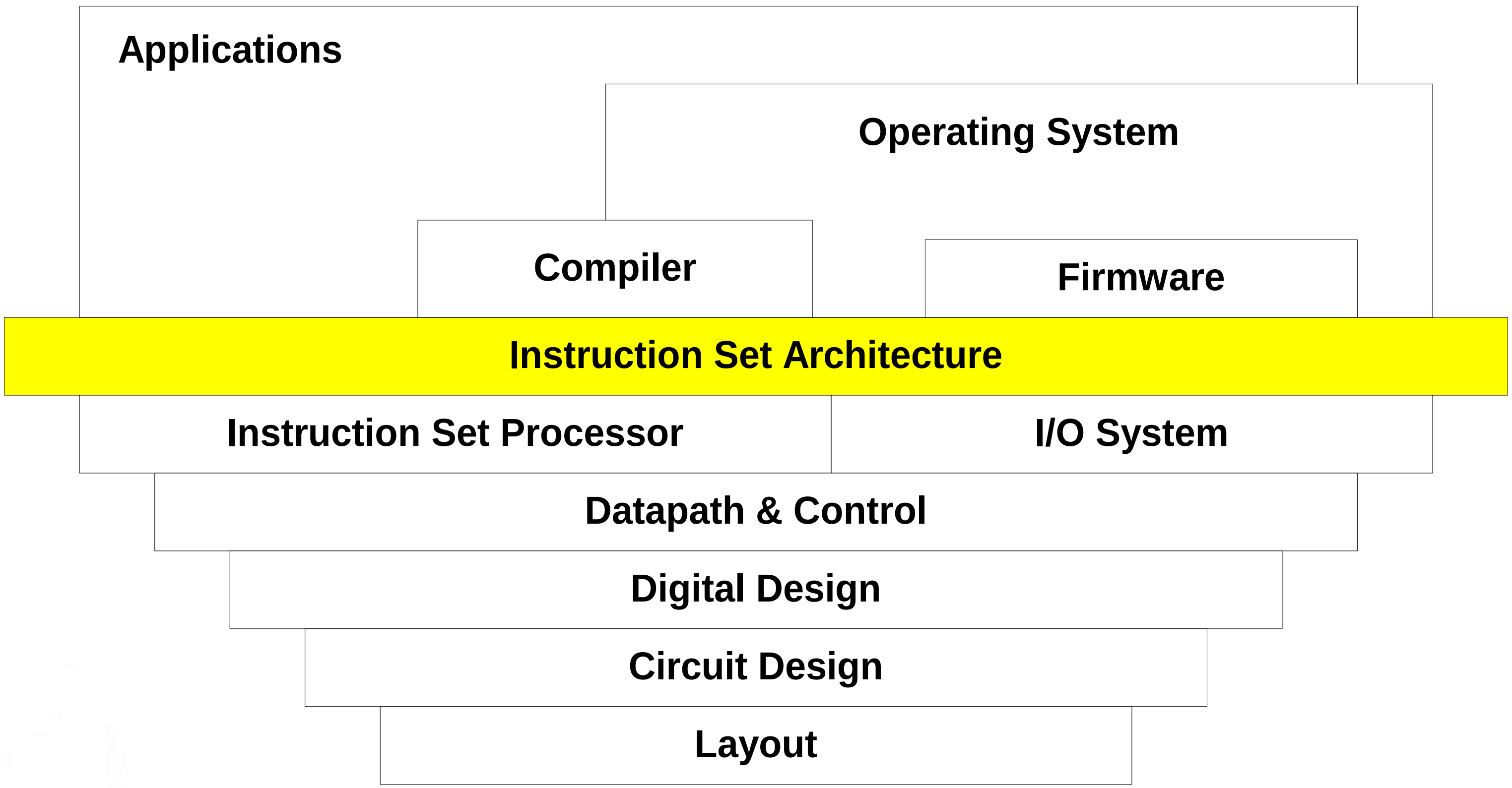
Original Concept of Computer architecture

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.

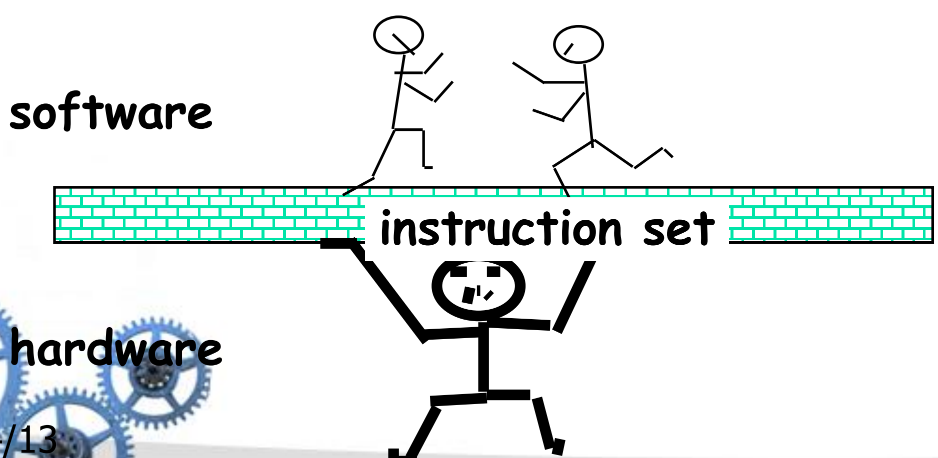
Amdahl, Blaaw, and Brooks, 1964

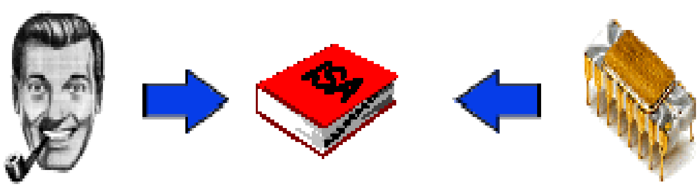
Instruction Set Architecture (ISA)



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, compatability)**
  - **Is used in many differeny ways (generality)**
  - **Provides convenient functionality to higher levels**
  - **Permits an efficient implementation at lower levels**

Seven dimensions of ISA

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

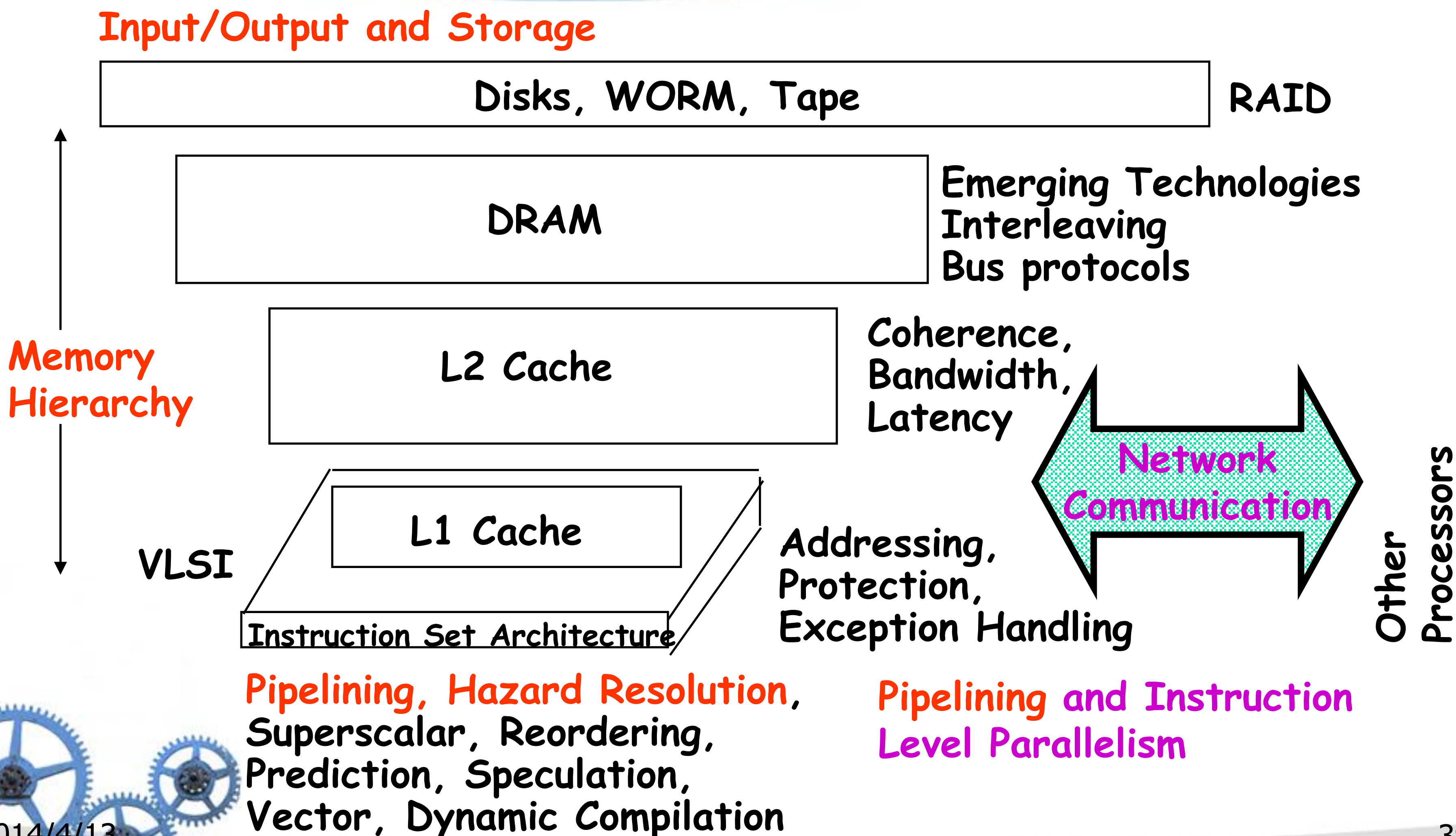
Course evolution

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

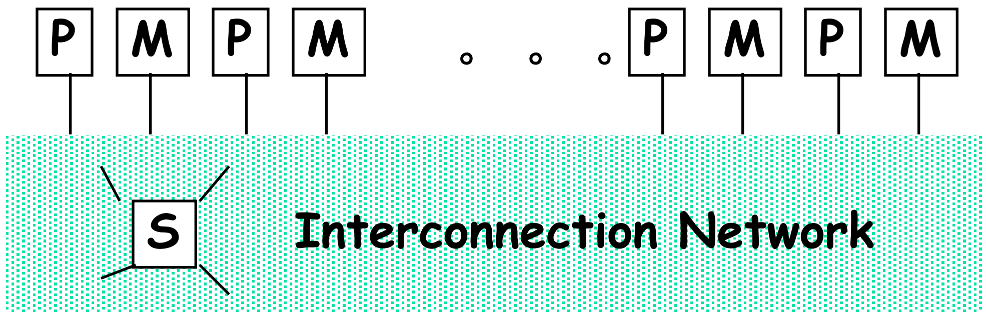
Computer Architecture

- *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**: epecifics of computer
    - Detailed logic design, packaging, cooling system, board displacement, power

Computer Architecture Topics



Computer Architecture Topics



Processor-Memory-Switch

Shared Memory,
Message Passing,
Data Parallelism

Network Interfaces

Topologies,
Routing,
Bandwidth,
Latency,
Reliability

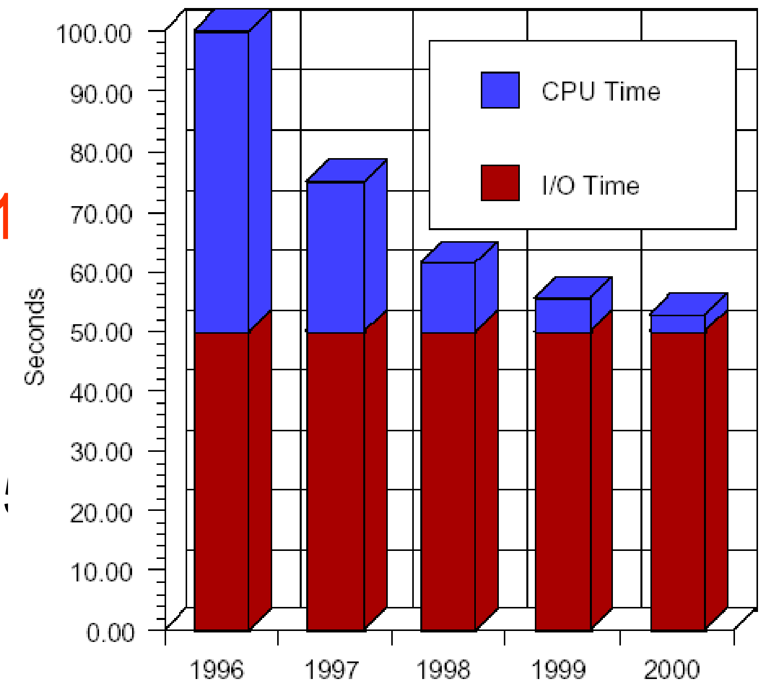
Multiprocessors
Networks and Interconnections

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
- 1998 - 1999
  - CPU performance - factor of 2
  - program performance N = 32.5 / 21.2
- 1999 - 2000
  - CPU Performance - factor of 2
  - program performance N = 21.25 / 15.6 = 1.36



Performance for Web Surfing

- **Assume 50 seconds CPU & 50 seconds 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 - factor of 2
  - program performance N = 75/62.5= 1.2
- 1998 - 1999
  - CPU performance -f actor of 2
  - program performance N = 62.5/56.5 = 1.11

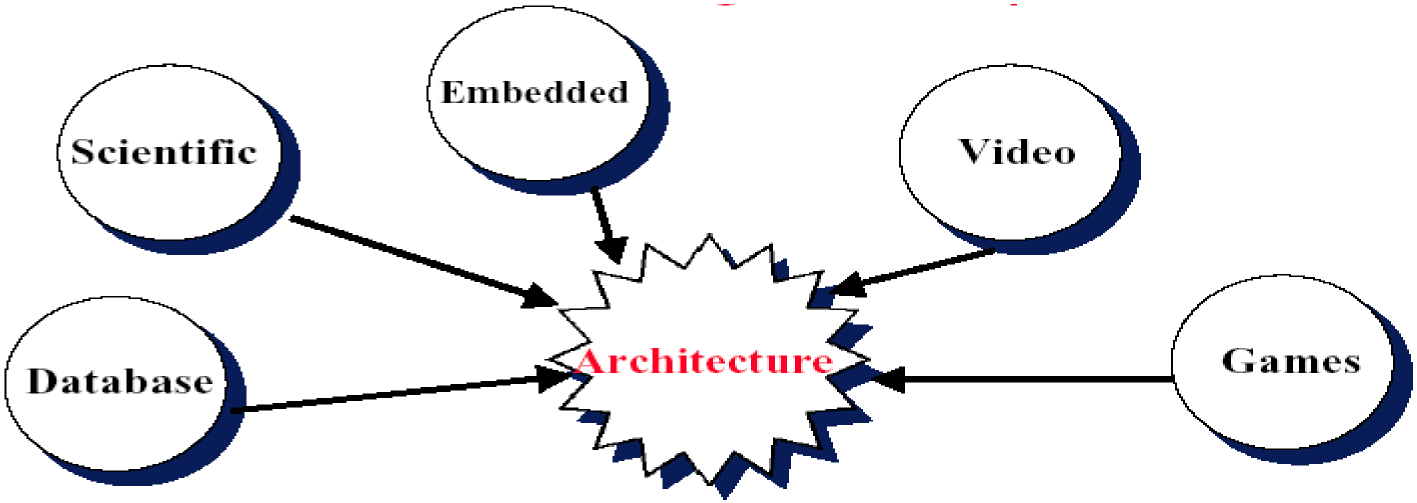
Computer Applications

- Architects need to understand applications' behavior
  - We say we design general purpose processors, but they really focus on specific sets of applications
  - Architecture can be tuned to 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	for embedded
➢ Intel MMX	for image and video
➢ Sony EE	for graphics rendering

➢ Applications drive the design of the processor



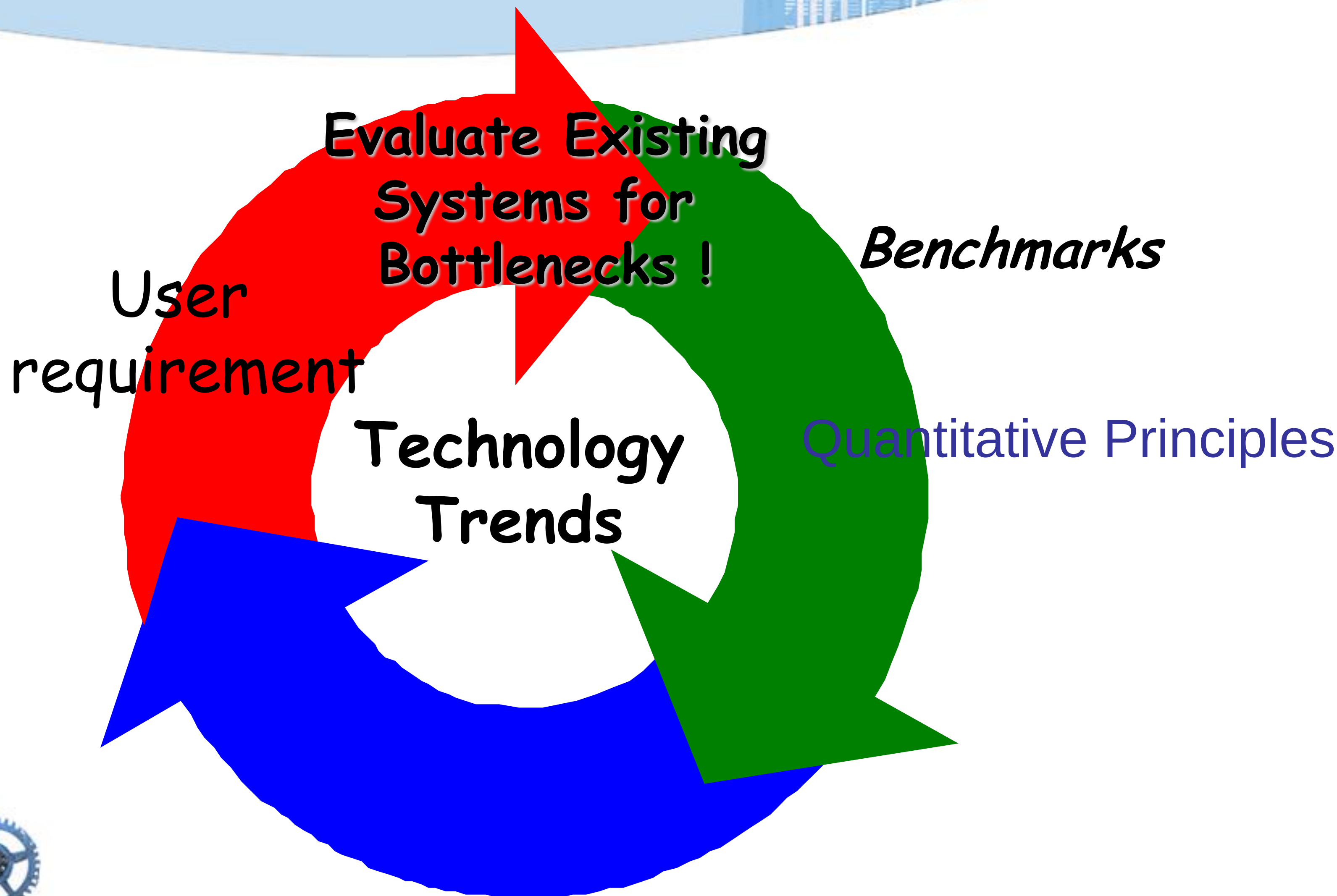
The Task of Computer Design-2

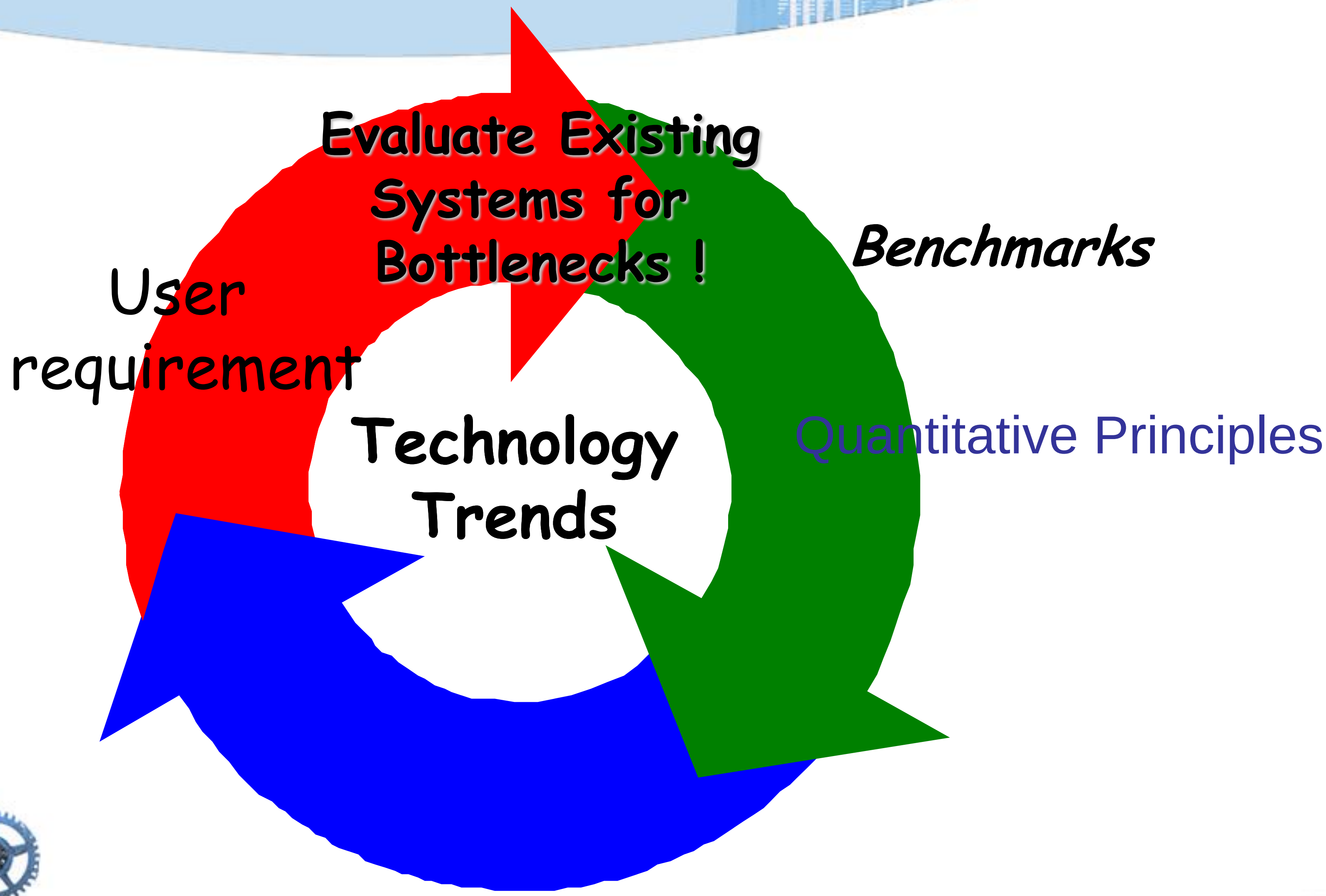
- Determine the important attributes of a new machine to maximize performance while staying with constrains, such as cost, power, availability, etc.
  - instruction set architecture design
  - functional organization
    - High level aspects of computer design, i.e. memory system, bus architecture and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

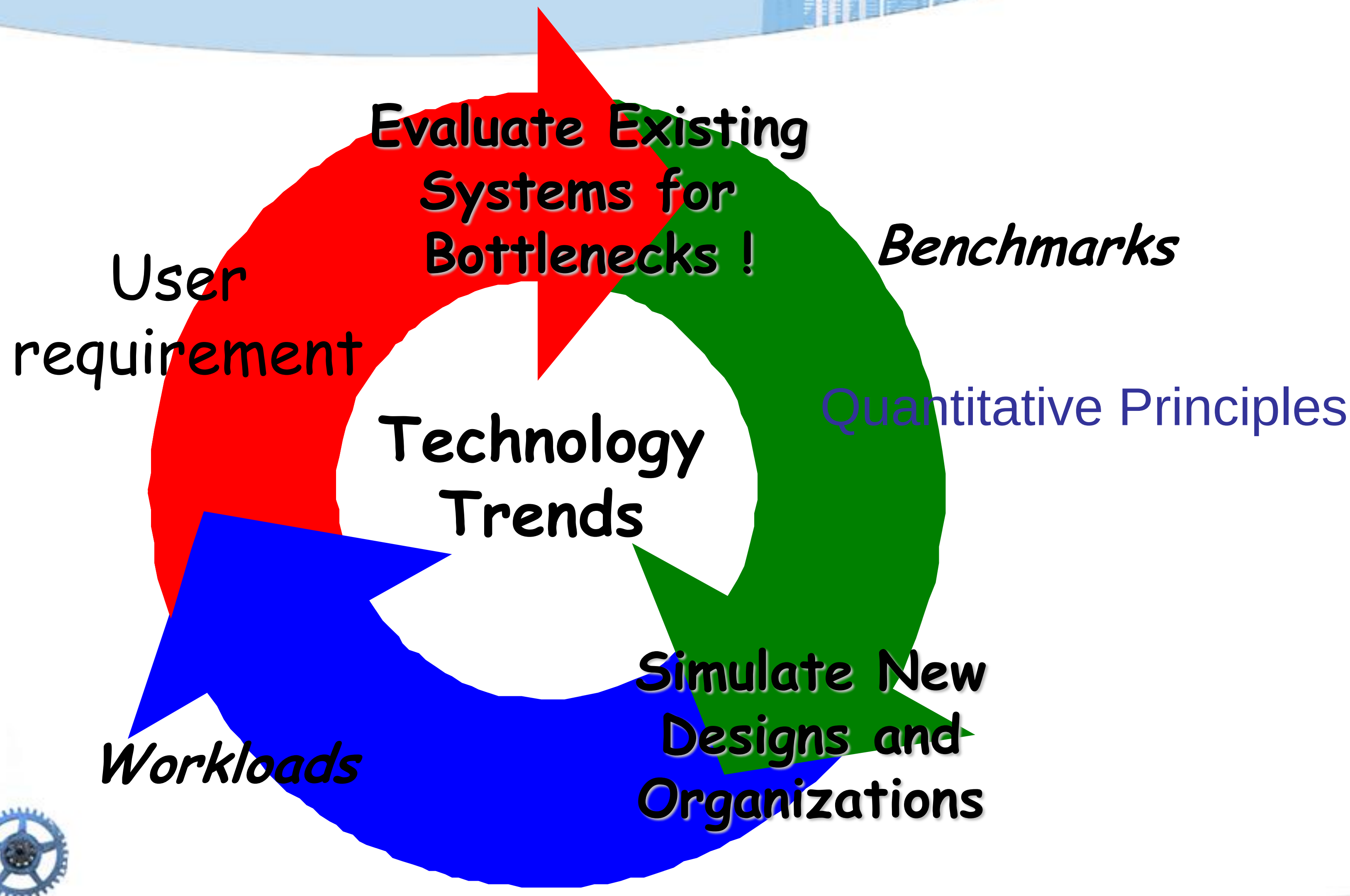
Trend of Architecture

- 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

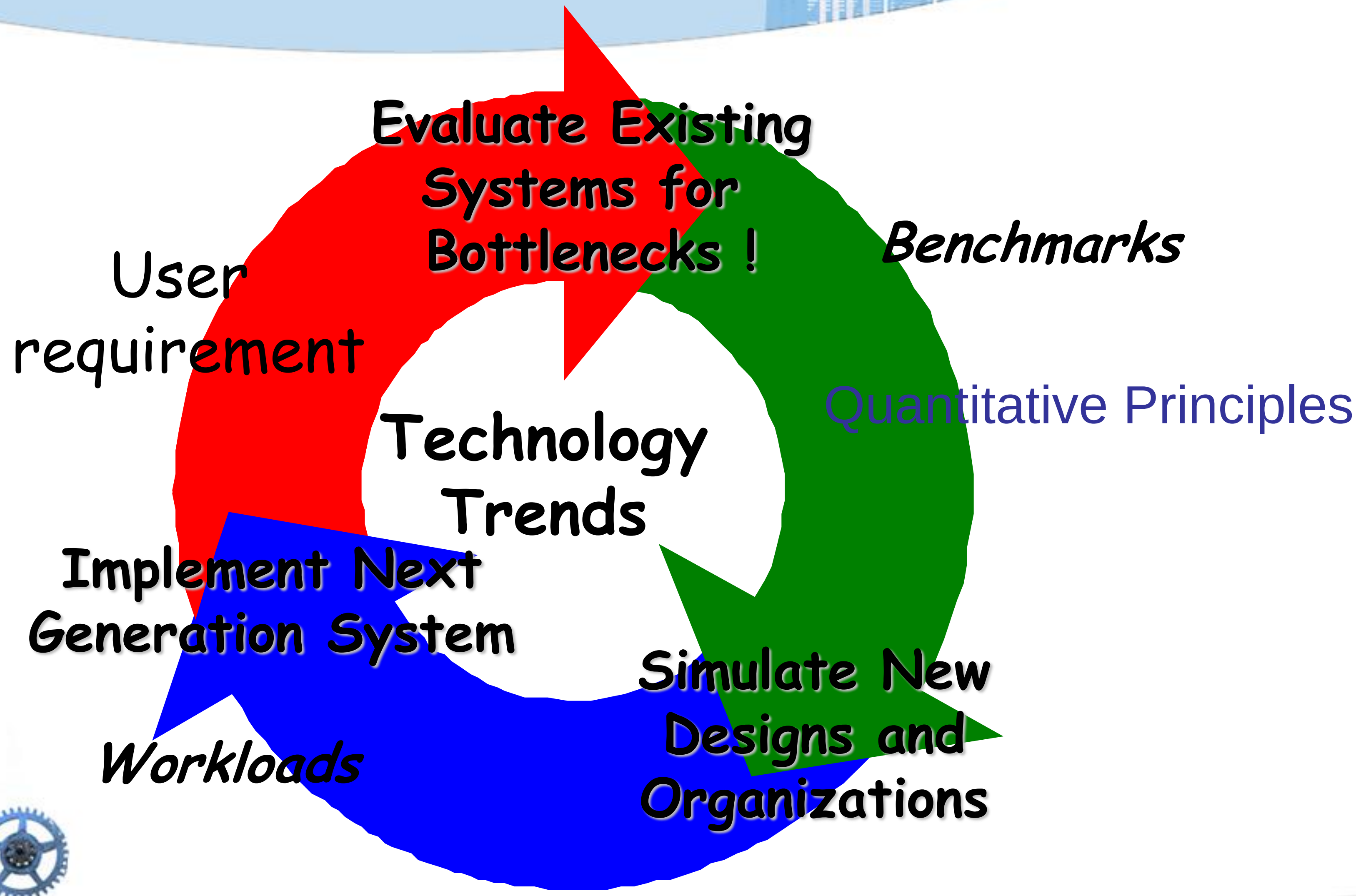
Computer Engineering Methodology



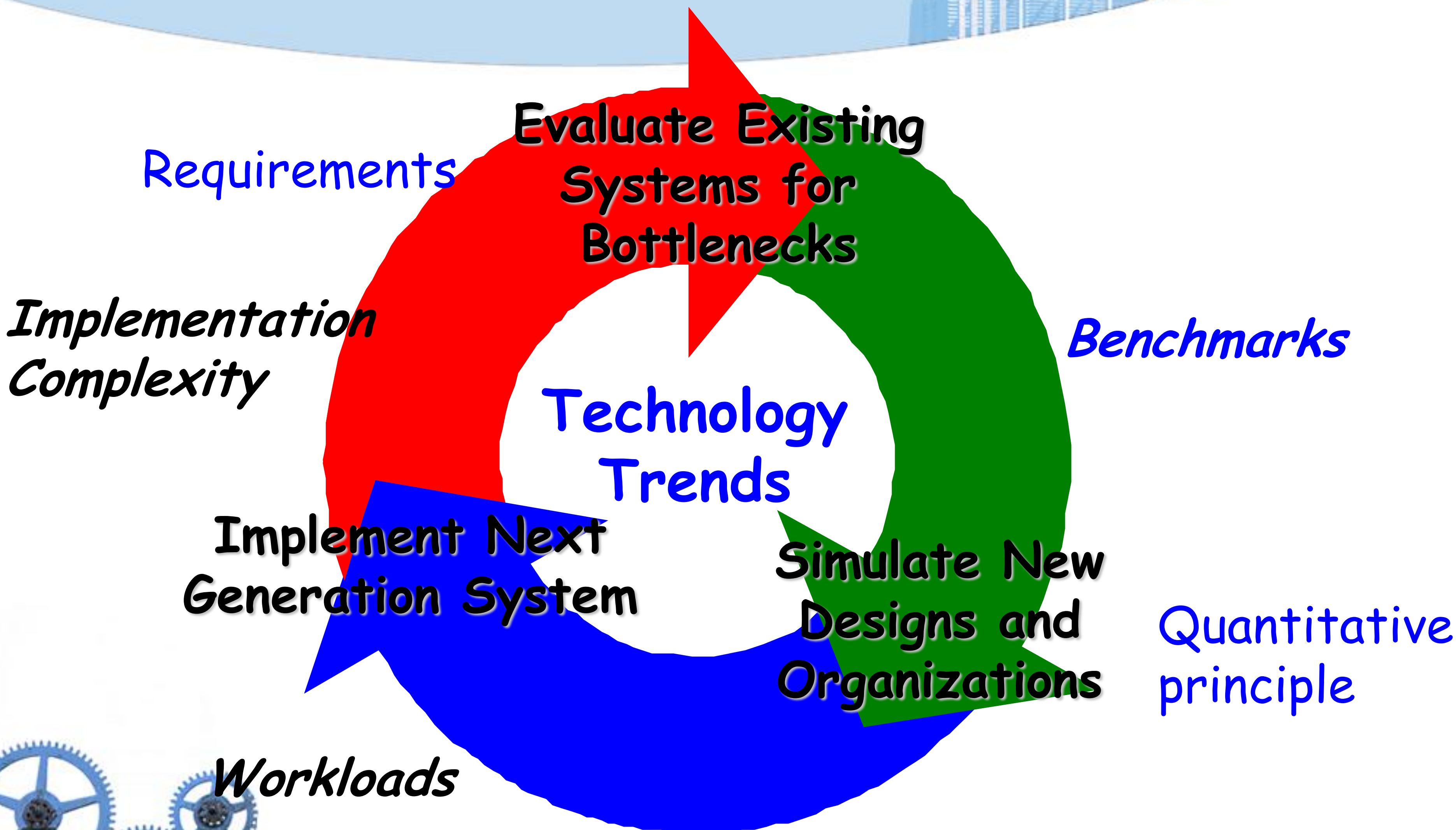

Evaluate Existing
Systems for
Bottlenecks !
Benchmarks
User
requirement
Technology
Trends
Quantitative Principles


Evaluate Existing Systems for Bottlenecks !
Benchmarks
User requirement
Technology Trends
Quantitative Principles
Simulate New Designs and Organizations
Workloads

Computer Engineering Methodology



Computer Design life cycle



Summary:Task of computer design

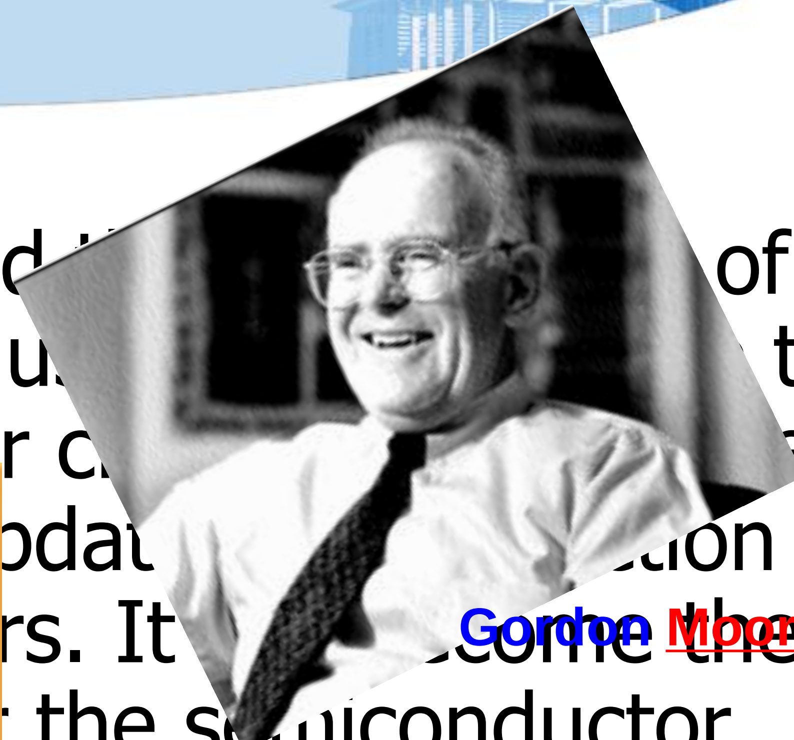
- Considerations:
  - functional and non functional requirements
  - implementation complexity
    - Complex designs take longer to complete
    - Complex designs must provide higher performance to be competitive
  - **Technology trends**
    - Not only what's available today, but also what will be available 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 summerizing Perf.
1.9 Quantitative Principles of computer Design
1.10 Putting it altogether

Technology Trends

- **Moore Law**
- In 1965 he predicted t ... of
components the indu ... to
place on a computer c ... every
year. In 1 ... odat ... tion to
once eve ... rs. It ... ecome the
guiding p ... the se ... iconductor
industry t ... ever-more-powerful chips
while dec ... e cost of electronics.



Gordon Moore

Technology Trends

- Integrated circuit logic technoloty
  - 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.
  - Desnsity: 30% p.y. Before 1990; 60% p.y. 1990-1996
  - 100 p.y. 1996-2004 ; 30% p.y. after 2004
  - capacity: about 60% per year
- Network
  - bandwidth: 10Mb $\xrightarrow[\text{10 years}]{}$ 100Mb $\xrightarrow[\text{5 years}]{}$ 1Gb

Designers often design for the next technology.

Gordon Moore on Moore's law

- **Moore's brief Bio**

  http://www.intel.com/pressroom/kits/bios/moore.htm

- **Gordon Moore on Moore's law**

  http://www.sichinamag.com/Article/html/2007-09/2007919032802.htm

- Video on conversation with Moore

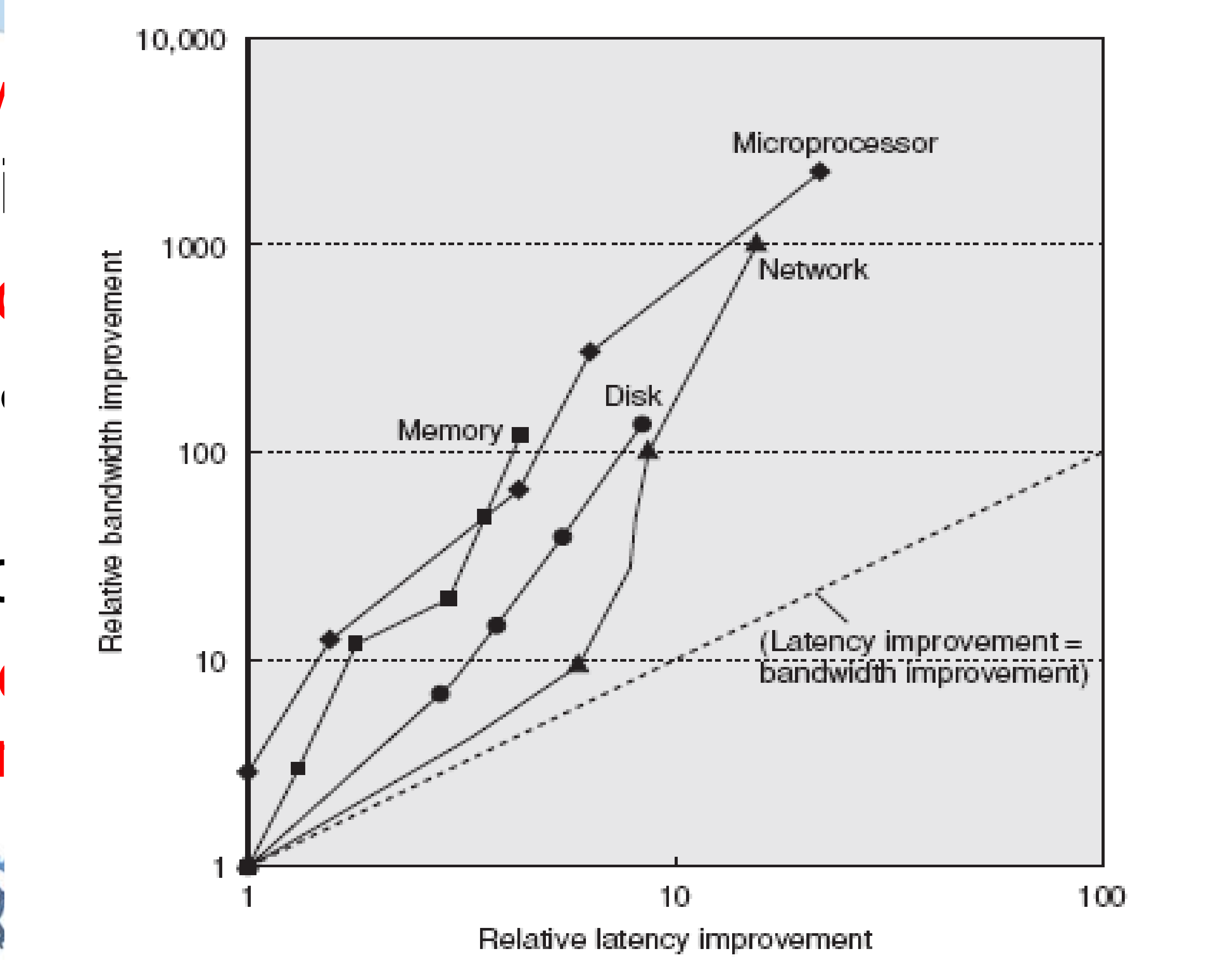
  http://you.video.sina.com.cn/b/7076856-1282136212.html

Import 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

- Bandw… …rk done i…
- Latenc… the st…
- Rule o…
  - Band… later…



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(特征尺寸）
  - 10 microns in 1971 → 0.18microns in 2001
  - → 0.09 microns in 2006 → 65nm is underway
  - Rule of thumb: transistor perf. 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.
- Wire delay---major design limination

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 summerizing Perf.

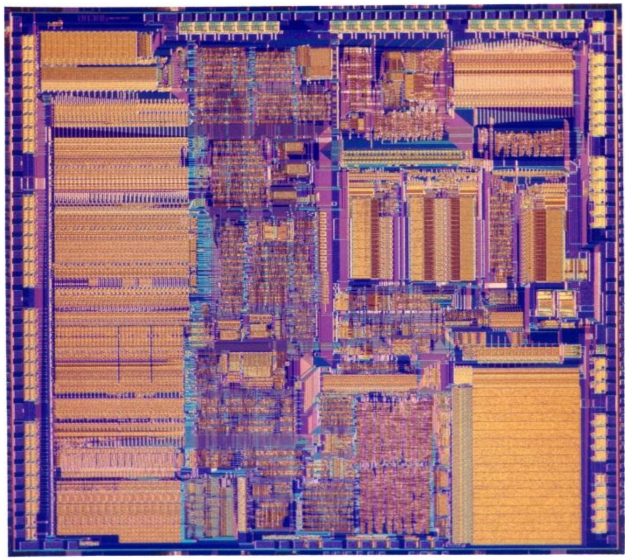
1.9 Quantitative Principles of computer Design

1.10 Putting it altogether

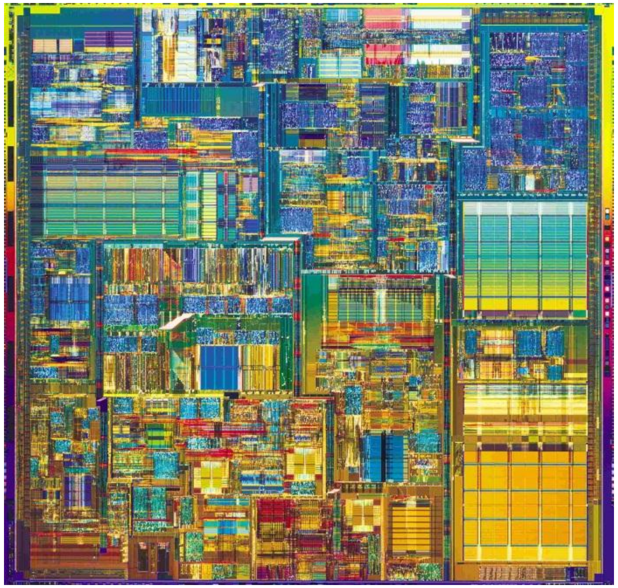
Trends in Power

- Power also provide challenges as device scaled
  - first microprocessor: 1/10watt --> 2GHz P4: 135watt
- Challenges:
  - distributing the power
  - removing the heat
  - preventing hot spot

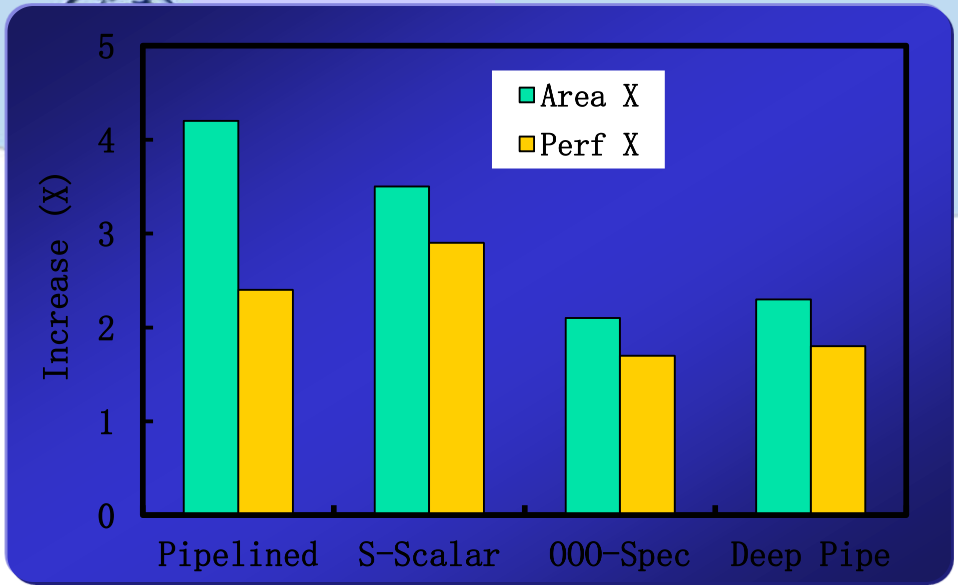
386 Processor



Pentium® 4 Processor

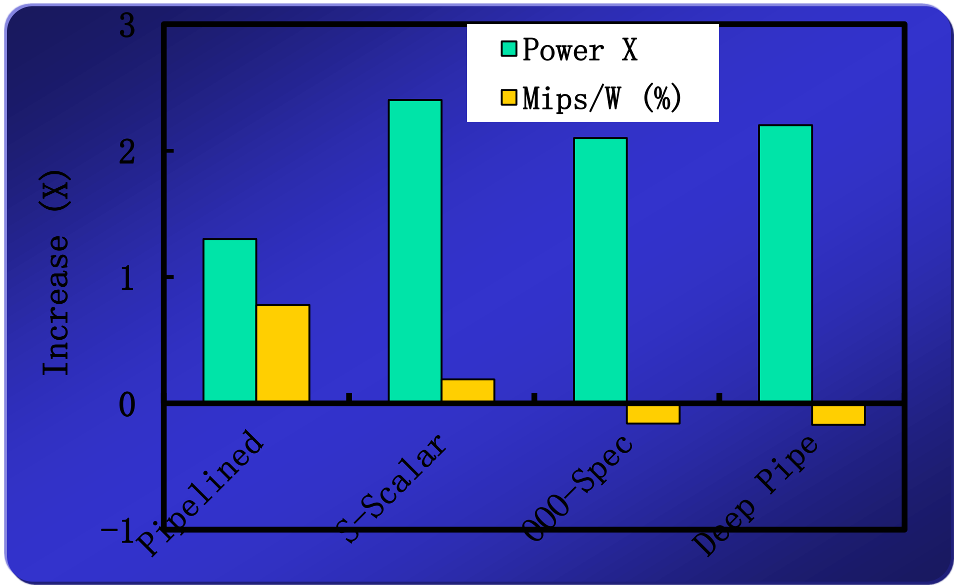


May 1986	**17 Years**	**August 27, 2003**
@16 MHz core	**200x**	**@3.2 GHz core**
275,000 1.5μ transistors	**200x/11x**	**55 Million 0.13μ transistors**
~1.2 SPECint2000	**1000x**	**1249 SPECint2000**



Performance scales with area**.5

Power efficiency has dropped



Two concepts

- Dynamic power: power consumption in switching transistors.
  - $\text{Power}_{\text{dynamic}} = \frac{1}{2} * \text{Capacitive load} * \text{Voltage}^2 * \text{Frequency switched}$
  - $\text{Energy}_{\text{dynamic}} = \text{Capacitive load} * \text{Voltage}^2$
- Static power: power consumption when a transistor is off due to power leakage
  - $\text{Power}_{\text{static}} = \text{current static} * \text{Voltage}$

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

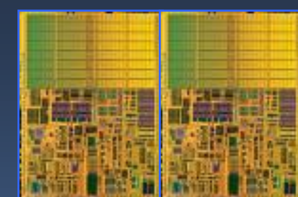
A 15% Reduction In Voltage Yields

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

SINGLE CORE

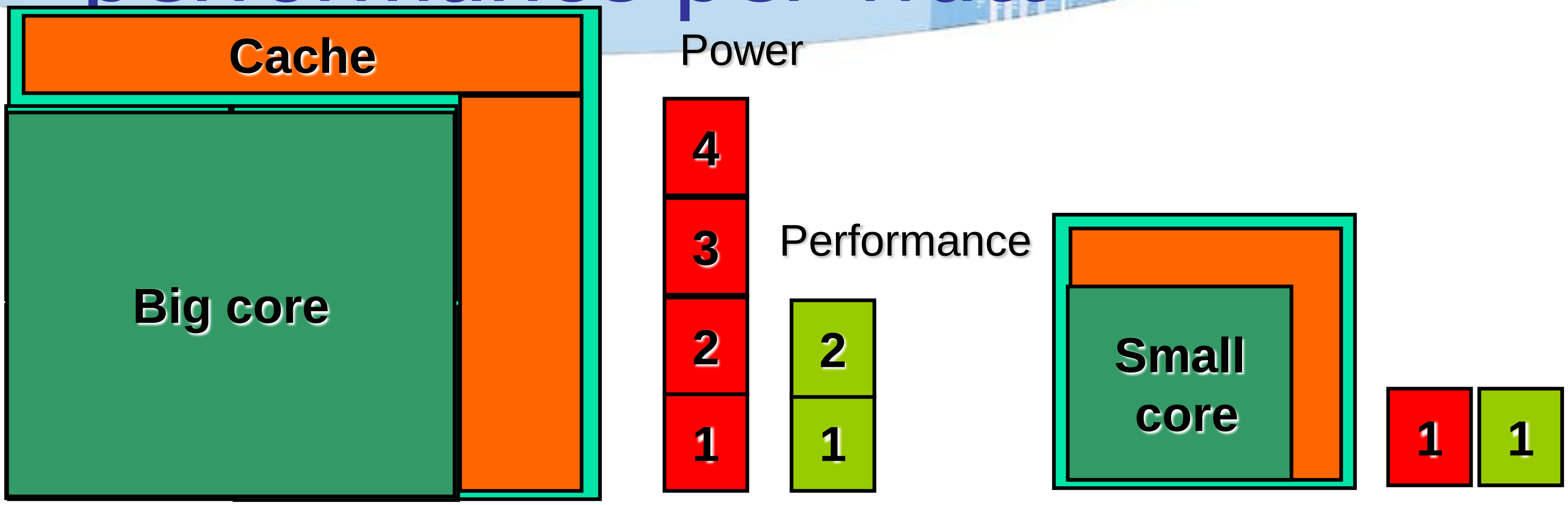
Area = 1
Voltage = 1
Freq = 1
Power = 1
Perf = 1

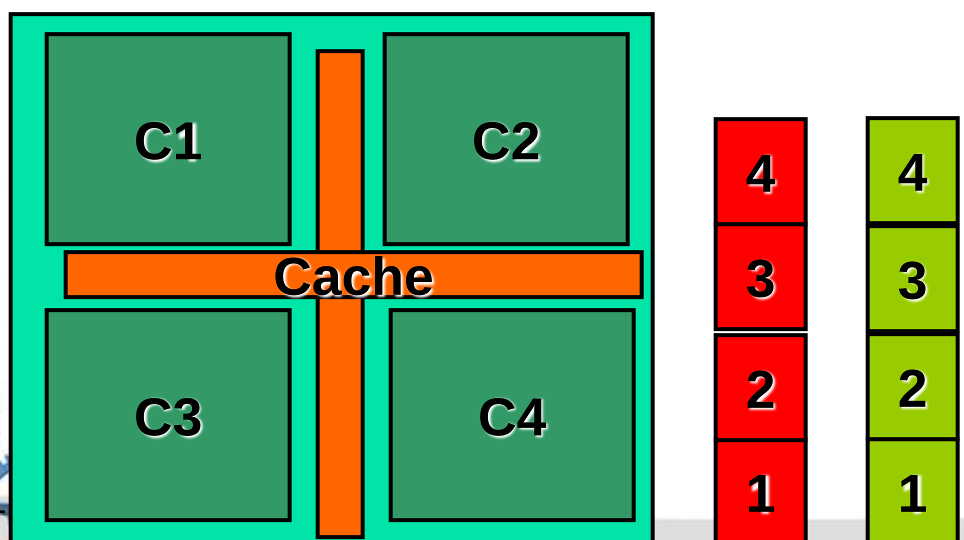
DUAL CORE



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**

再见，谢谢！