Machine Called Computer

Part 7
Underlying Technologies and Evolution

References:

1. Computer Organization and Design & Computer Architecture, Hennessy and Patterson (slides are adapted from those by the authors)

Semiconductor Technology

학습요령:

- Scaling의 개념과 잇점 그리고 결과적인 추세 이해하실 것

-그리고 wafer, die, yield 용어를 이해

- “technology“ 라는 용어의 용법 이해

- 숫자나 IC 제조공정은 암기대상이 아님

Why Transistor?

❑ Solid-state semiconductor device (반도체 장치)

- Small, fast, reliable, energy-efficient, inexpensive
- Integrated circuits (IC) 형태로 집적 가능

Image of cross-section of CMOS inverter (two transistors):
http://en.wikipedia.org/wiki/File:Cmos_impurity_profile.PNG

Semiconductor Technology

- ❑ Transistor: invented in Bell labs. in 1947
  - Took 10 years to commercialize
- ❑ IC (integrated circuits): invented in 1958
  - 5 years to commercialize
  - SSI, MSI, LSI, VLSI
- ❑ Major driving force behind computer performance evolution

CMOS NAND Gate

크기가 속도에 직접 관련

Image of CMOS NAND gate:

http://en.wikipedia.org/wiki/File:CMOS_NAND.svg

Image of CMOS NAND layout:

http://en.wikipedia.org/wiki/File:CMOS_NAND_Layout.svg

Image of CMOS transistor pair:

http://en.wikipedia.org/wiki/File:Cmos_impurity_profile.PNG

Technology Scaling (data from Wikipedia)

- Minimum feature size: 10μm (1970) to 0.022μ (2012)
  - Exponential decrease

Min. Feature Size	year
10 μm	1970
3 μm	1975
1 μm	1985
350 nm	1995
130 nm	2002
45 nm	2008
22 nm	2012
7 nm	2018 (estimated)

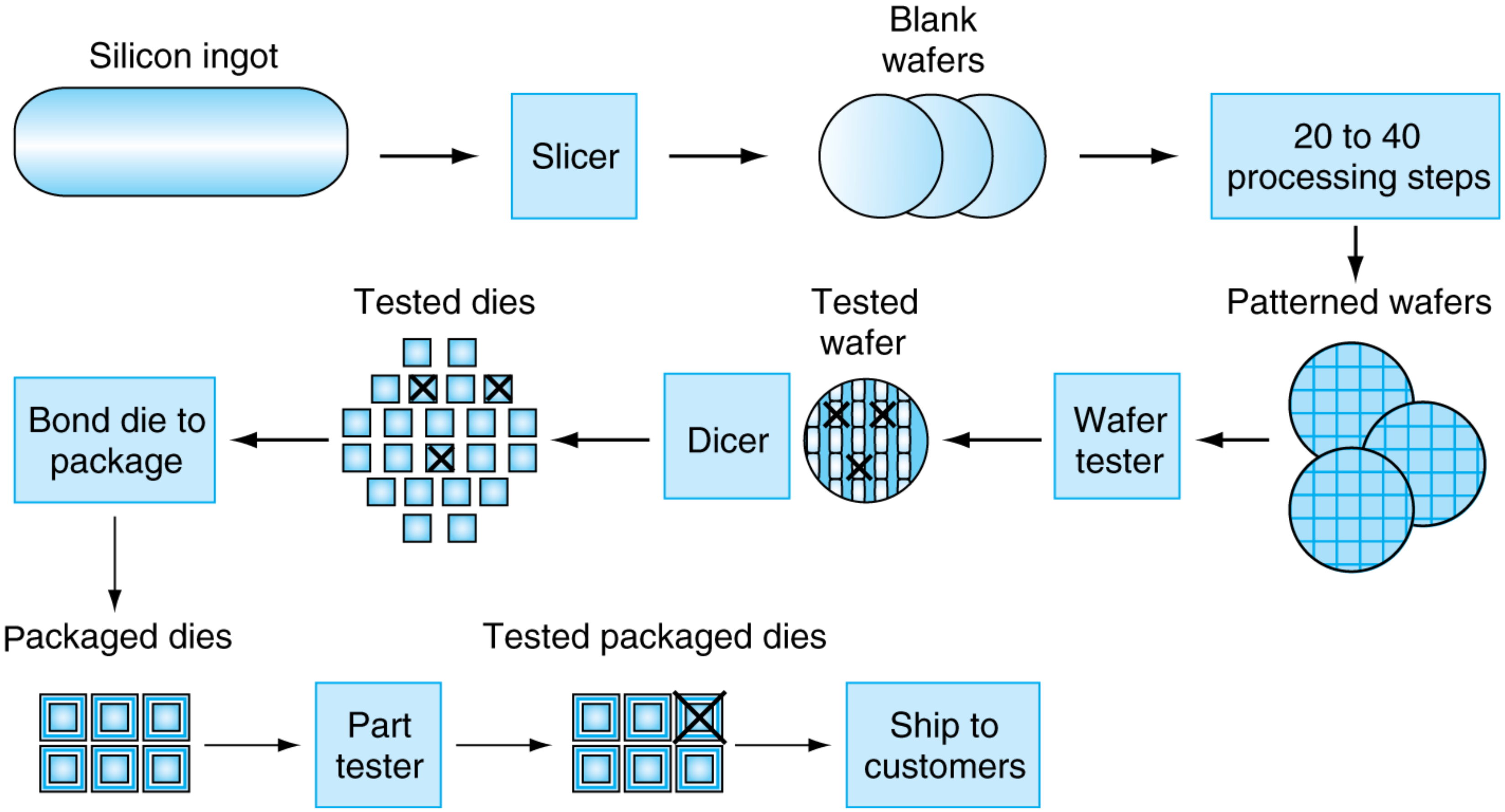
Technology Trends

- Smaller is faster

Year	Technology	Relative performance/cost
1951	Vacuum tube	1
1965	Transistor	35
1975	Integrated circuit (IC)	900
1995	Very large scale IC (VLSI)	2,400,000
2005	Ultra large scale IC	6,200,000,000

Manufacturing ICs

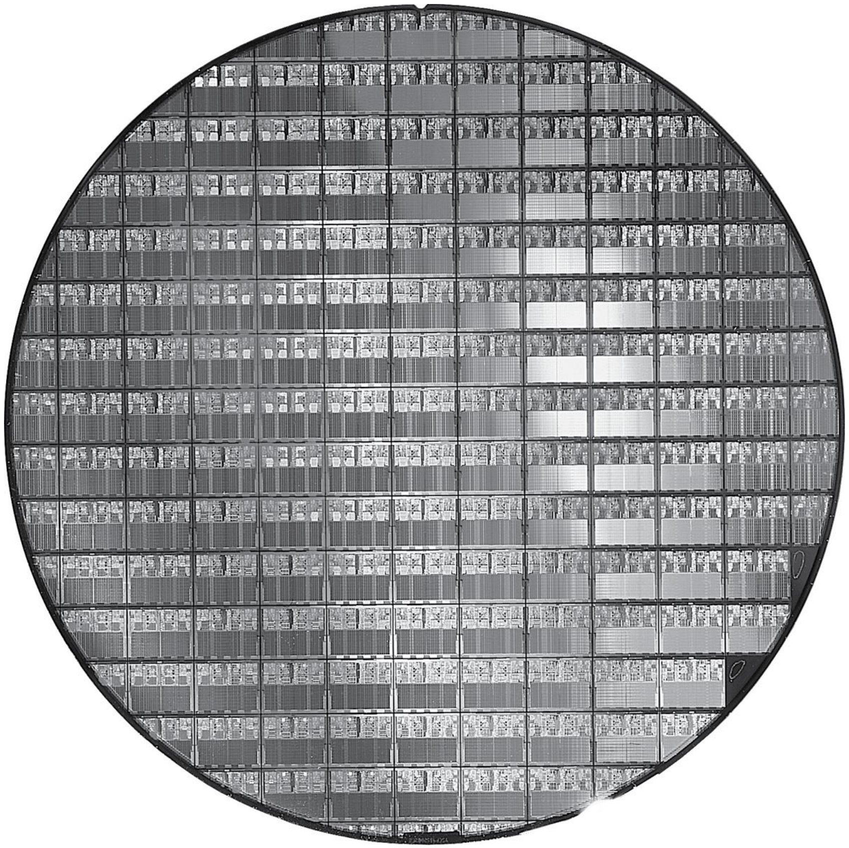
(Hennessy and Patterson slide, Computer Organization and Design, Morgan Kaufmann)



- Yield: proportion of working dies per wafer

AMD Opteron X2 Wafer

(Hennessy and Patterson slide, Computer Organization and Design, Morgan Kaufmann)



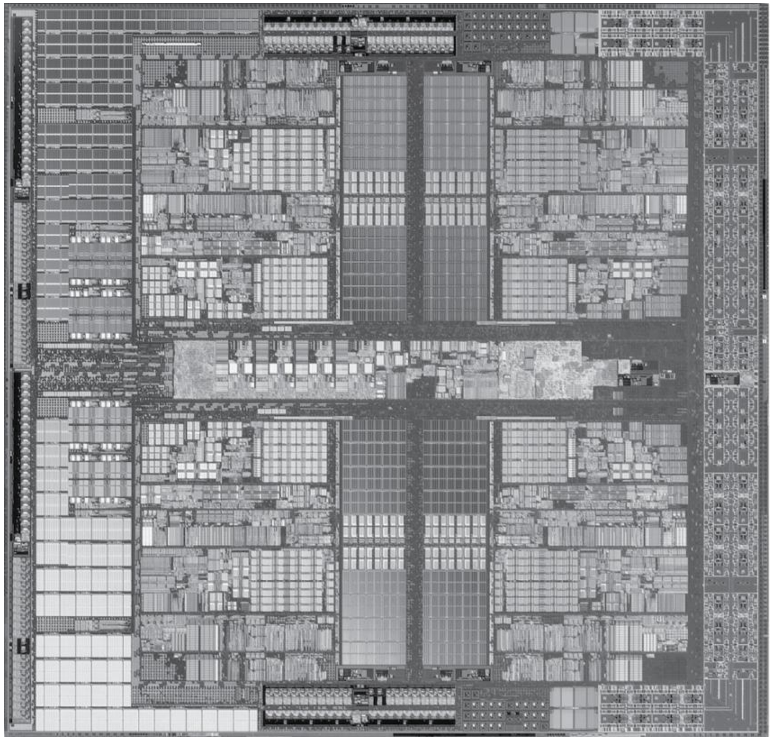
wafer의 크기가 크면 클수록 기술력이 좋은 것이다. 크기가 크면 클수록 균등하게 재료를 뿌려주기 힘들기 때문에

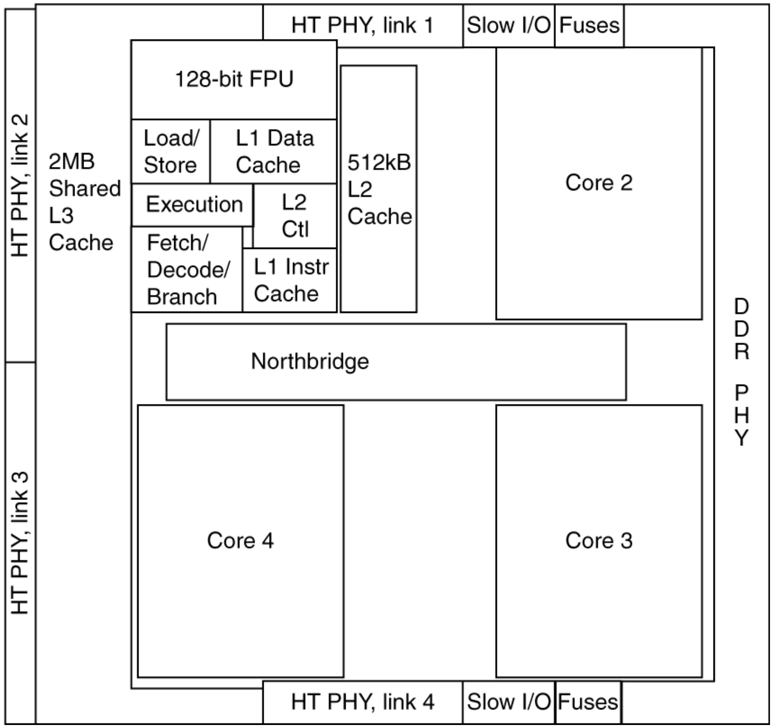
- X2: 300mm wafer, 117 chips, 90nm technology
- X4: 45nm technology

Inside the Processor

(Hennessy and Patterson slide, Computer Organization and Design, Morgan Kaufmann)

- AMD Barcelona: 4 processor cores





Semiconductor Technology

- What does it mean?
  - We have 32 nm technology
  - We have 300 mm wafer technology
    - What if we use larger (e.g., 450mm) wafer?
- Moore's law: exponential growth
  - Number of transistors per chip double every 18 (or 12 or 24) months
  - Cost of fabrication facility also increase exponentially over time

Intel and Processor Technology

학습요령:

- Microprocessor의 출현과 발전 흐름 이해
- RISC 프로세서 용어 이해
- * 전체적인 흐름이 중요하고, 아주 세부적인 내용이나 숫자는 기억할 필요 없음

Semiconductor Technology

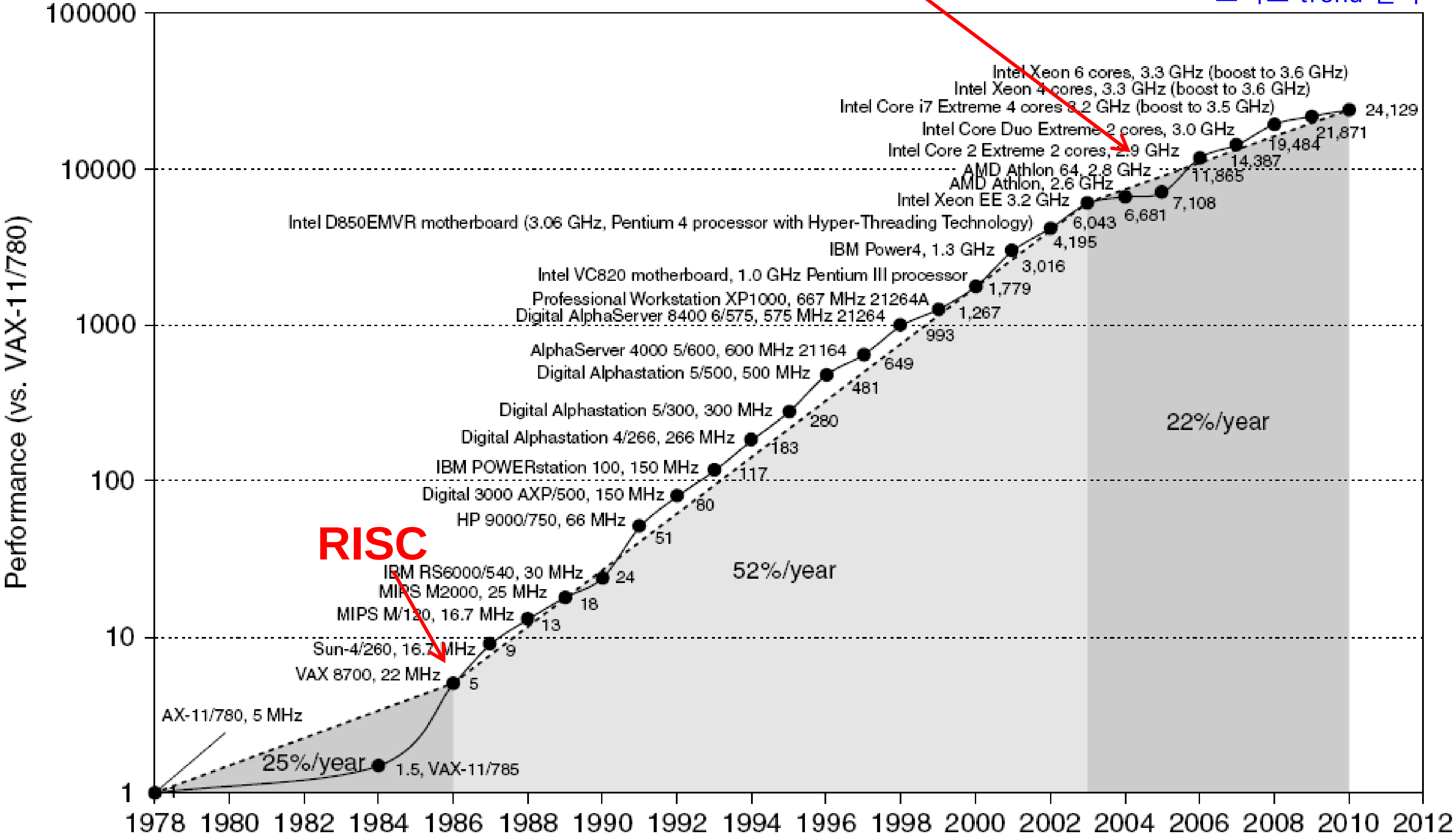
- ❑ Major driving force behind computer performance evolution
- ❑ Smaller transistor, increased die size
  - • Processor perspective
    - – Exponential growth in performance
- ❑ Around 2012
  - • Highest transistor count in commercial CPU
    - – 2.5B in Intel's 10-core Xeon Westmere-EX (32nm)
  - † FPGA: Xilink 6.8B in Virtex-7 (28nm)
  - † Samsung 20nm 4Gb DRAM
  - † Samsung 10nm 64Gb flash memory

Transistor Count in Processor (Wikipedia)

Processor	#Transistors	Year	Process (μm)	Area (mm^2)	
Intel 4004	2,300	1971	10	12	First μp
Intel 8080	4,500	1974	6	20	
Intel 8088	29,000	1979	3	33	IBM PC, 16 bit
Intel 80286	134,000	1982	1.5	49	PC/AT
Intel 80386	275,000	1985	1.5	104	x86, IA-32 최초 32bit cpu
Intel 80486	1,180,000	1989	1	160	
Pentium	3,100,000	1993	0.8	294	cpu 이름 변경되면 IS(instruction set) 이 변경된것. backward compatability 유지
Pentium II	7,500,000	1997	0.35	195	
Pentium 4	42,000,000	2000	0.18		
Itanium 2	220,000,000	2003	0.13		IA-64, RISC 64bit / 32bit 호환 x
Core i7 (Quad)	731,000,000	2008	0.045	263	x86-64 64bit / 32bit 호환 o
10-core Xeon Westmere-EX	2,600,000,000	2011	0.032	512	x86

Single Processor Performance

Move to multi-processor 소비전력의 문제로 멀티 코어로 trend 변화



RISC

Copyright © 2012, Elsevier Inc. All rights reserved. (Hennessy and Patterson)

Trends in Technology

- ❑ Integrated circuit technology (e.g., processor)
  - Transistor density: 35%/year
  - Die size: 10-20%/year (why increase?) 성능을 더 좋게하기 위해서 hw 더 추가되서 크기가 커진다.
  - Integration overall: 40-55%/year
- ❑ DRAM capacity: 25-40%/year
- ❑ Flash memory capacity: 50-60%/year
  - 15-20X cheaper/bit than DRAM
- ❑ Magnetic disk technology: 40%/year
  - 15-25X cheaper/bit than Flash
  - 300-500X cheaper/bit than DRAM

CPU in Mainframes in 1971

- ❑ Large printed circuit boards
  - Several of them for a CPU (IC technology immature)
- ❑ Design cycle: 5 years (HW + OS + Appl.)

Image of mainframe CPU in 1971: http://en.wikipedia.org/wiki/File:386DX40_MB_Jaguar_V.jpg

대형 컴퓨터
pcb 여러개 연결한 컴퓨터
(32bit을 만들기 위해서 기술력의 한계로
pcb 여러개 연결해야만 했다.)

→

intel
processor을 single chip으로 만든다.
1. cpu 설계
2. cpu 생산

Microprocessor

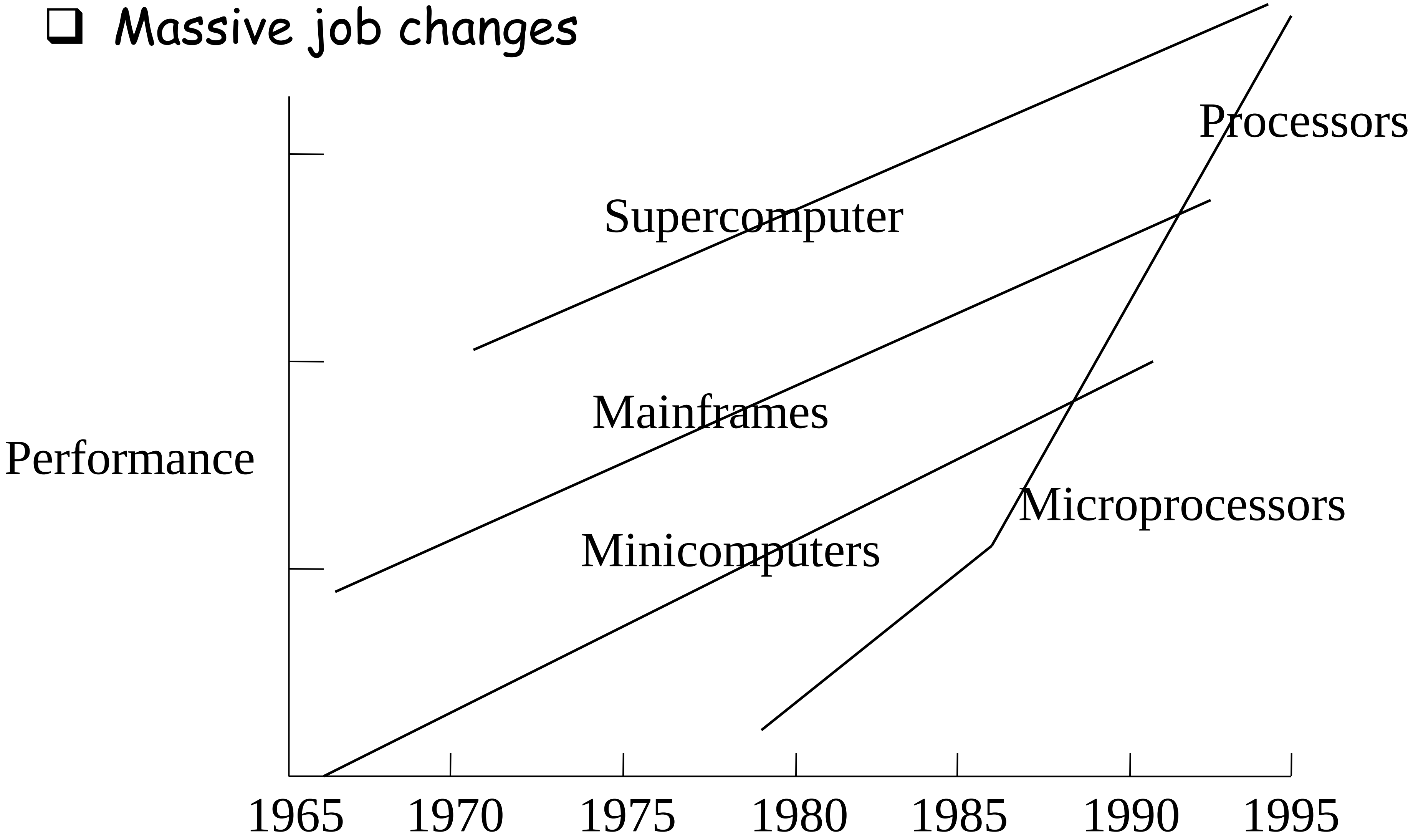
- Breakthrough: 4004 microprocessor by Intel in 1971
  - Miniature version of minicomputer CPU
  - Designed for Busicom to build calculators
- Intel look for market
- Intel announce 8-bit 8008 in 1972, 8080 in 1974
- Altair in 1975: first personal computer with 8080 이후 microcomputer 유행
  - Along the style of best minicomputers
- IPM PC in 1981 use 16-bit 8088 and MS DOS

  intel microsoft
- Cloning: from IBM PC to PC (dominance by Intel, MS)

  bios - IBM 특허였는데 특허가 깨져서 IMB의 영향력이 없어졌다. 따라서 IMB PC -> PC로 변화됨
  - Intel microprocessors: single chip, short design cycle
    - Take full advantage of semiconductor technology

(Approximate) Technology Trend

- Massive job changes



Intel and Processors

- ❑ Intel's microprocessors become powerful processors
- ❑ Computer manufacturers
  - Buy processors from processor vendors
  - Massive job changes in computer companies in 1980s
    - From hardware to software (5:5 -> 1:9)
  - Focus on systems, software, service

† Small microprocessors still there for low-cost embedded systems (many many of them)

Instruction Set 변화 - RISC Processors

- Processor design in 1970s: what we call CISC
  - Constraint: memory expensive
- 1980s: renaissance of processor design (RISC style)
  - Semiconductor technology
    - Memory become cheaper
  - Open Unix operating system
  - High-level programming cpu에 맞춰 porting 하는 것이 assembly보다 쉽다.

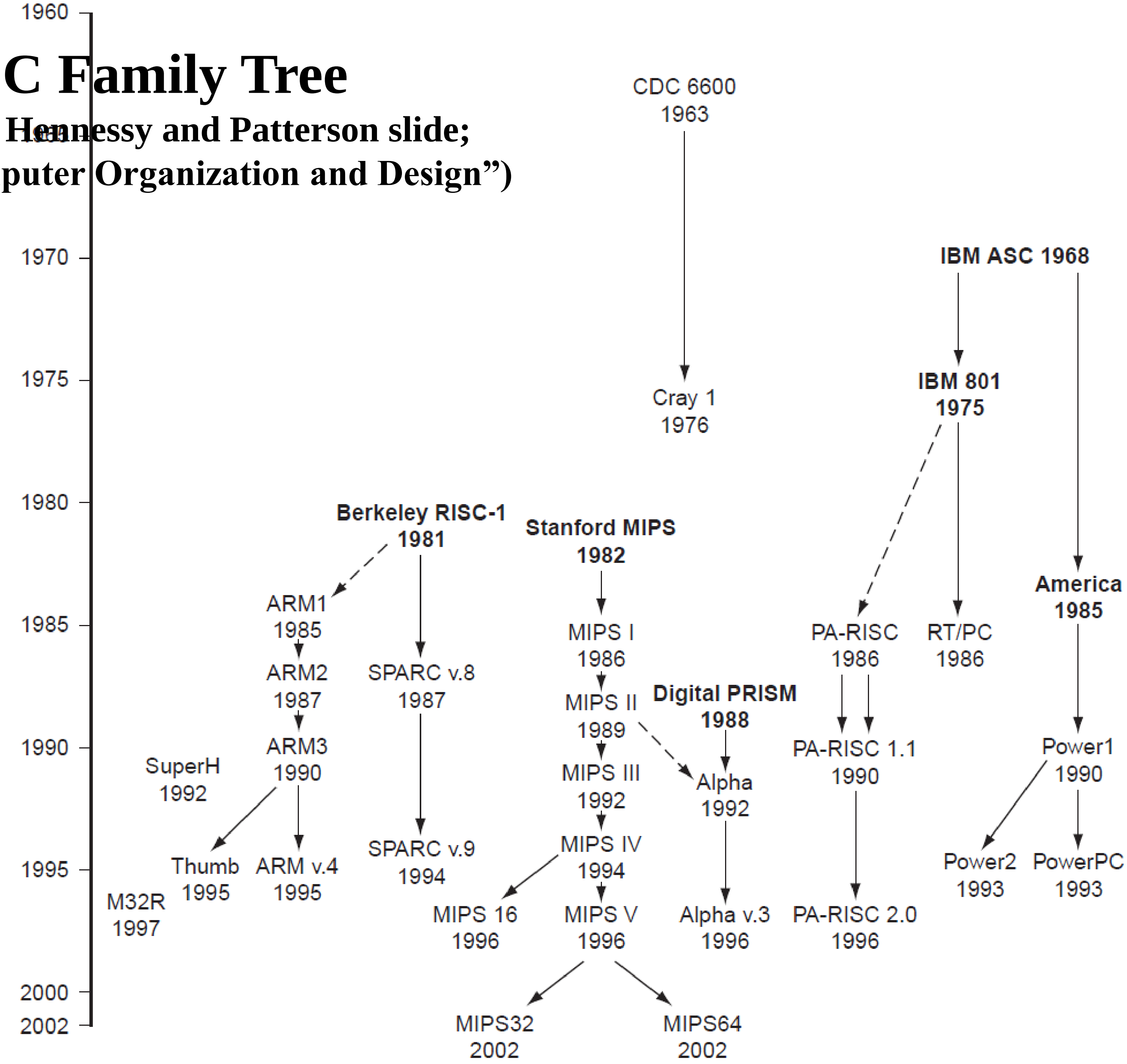
이후 processor 회사 붐!!

- Emergence of powerful 32-bit RISC processers
  - PowerPC, PA-RISC, MIPS, SPARC, Alpha, ARM

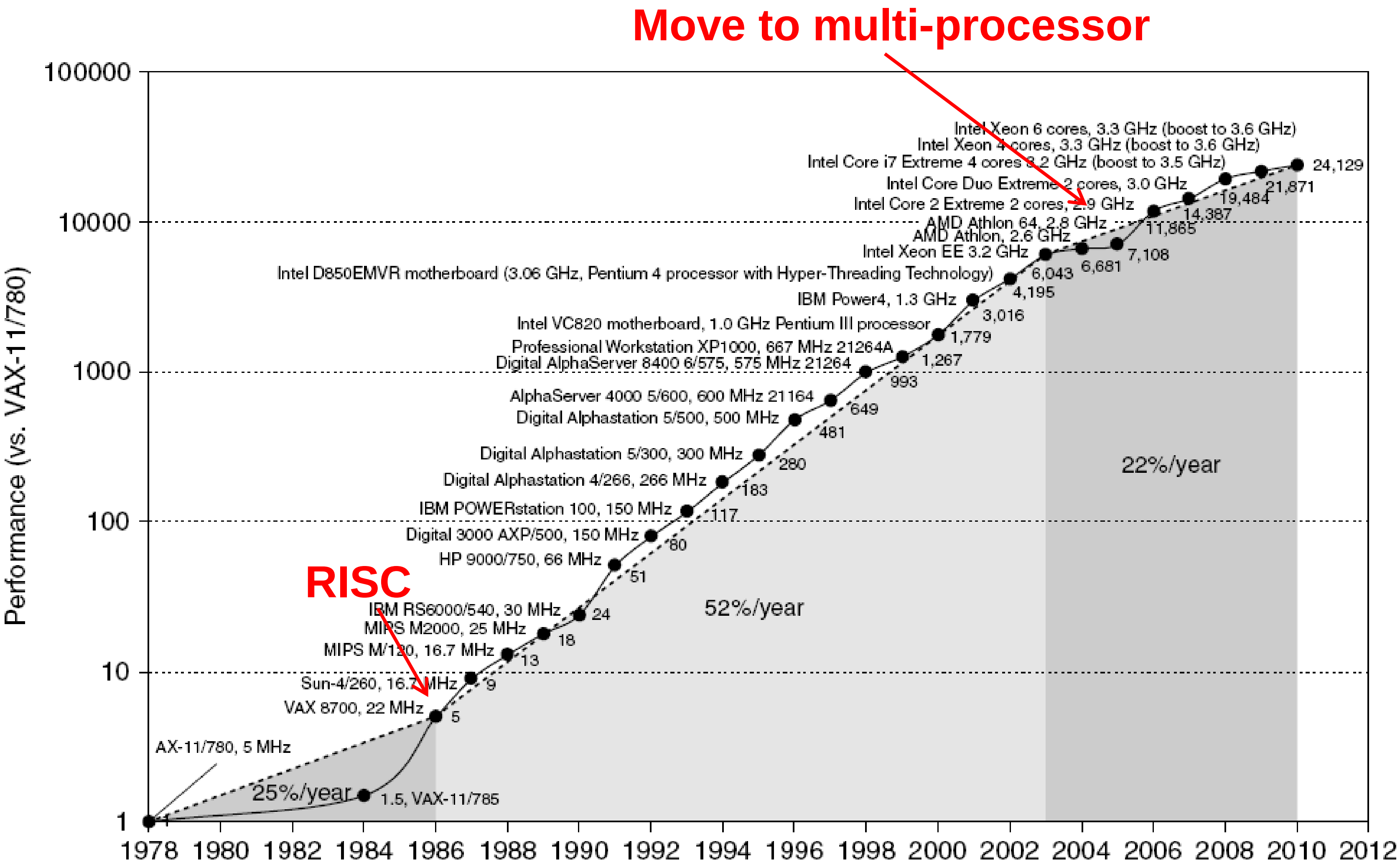
† Exception is Intel Pentium

RISC Family Tree

(from Hennessy and Patterson slide;
“Computer Organization and Design”)



Single Processor Performance



Move to multi-processor

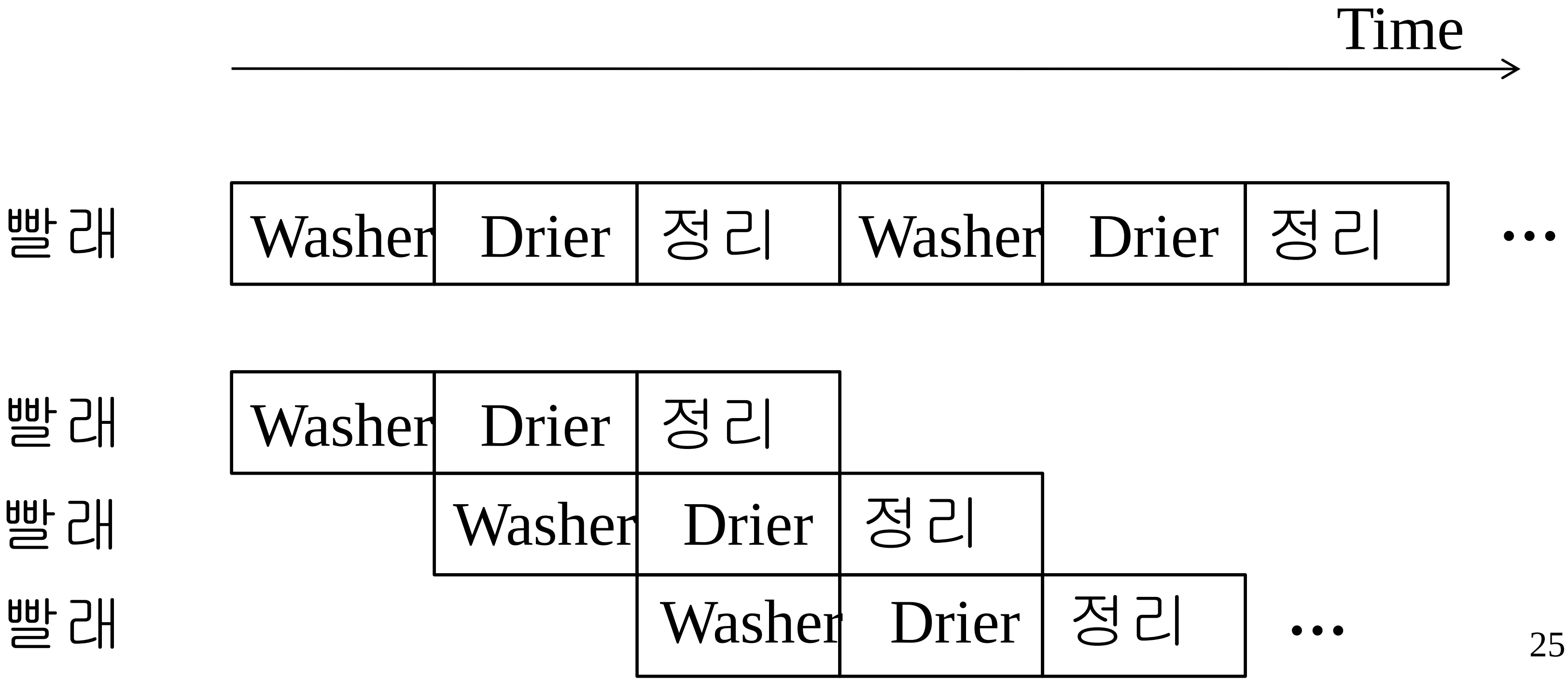
RISC

Key Speedup Techniques in CPU

- ❑ Pipelining
- ❑ Cache memory

Pipelining - General Speedup Technique

- 3-stage pipeline (e.g., washer-dryer example)
  - Speedup?



Time

빨래 Washer Drier 정리 Washer Drier 정리 …

빨래 Washer Drier 정리

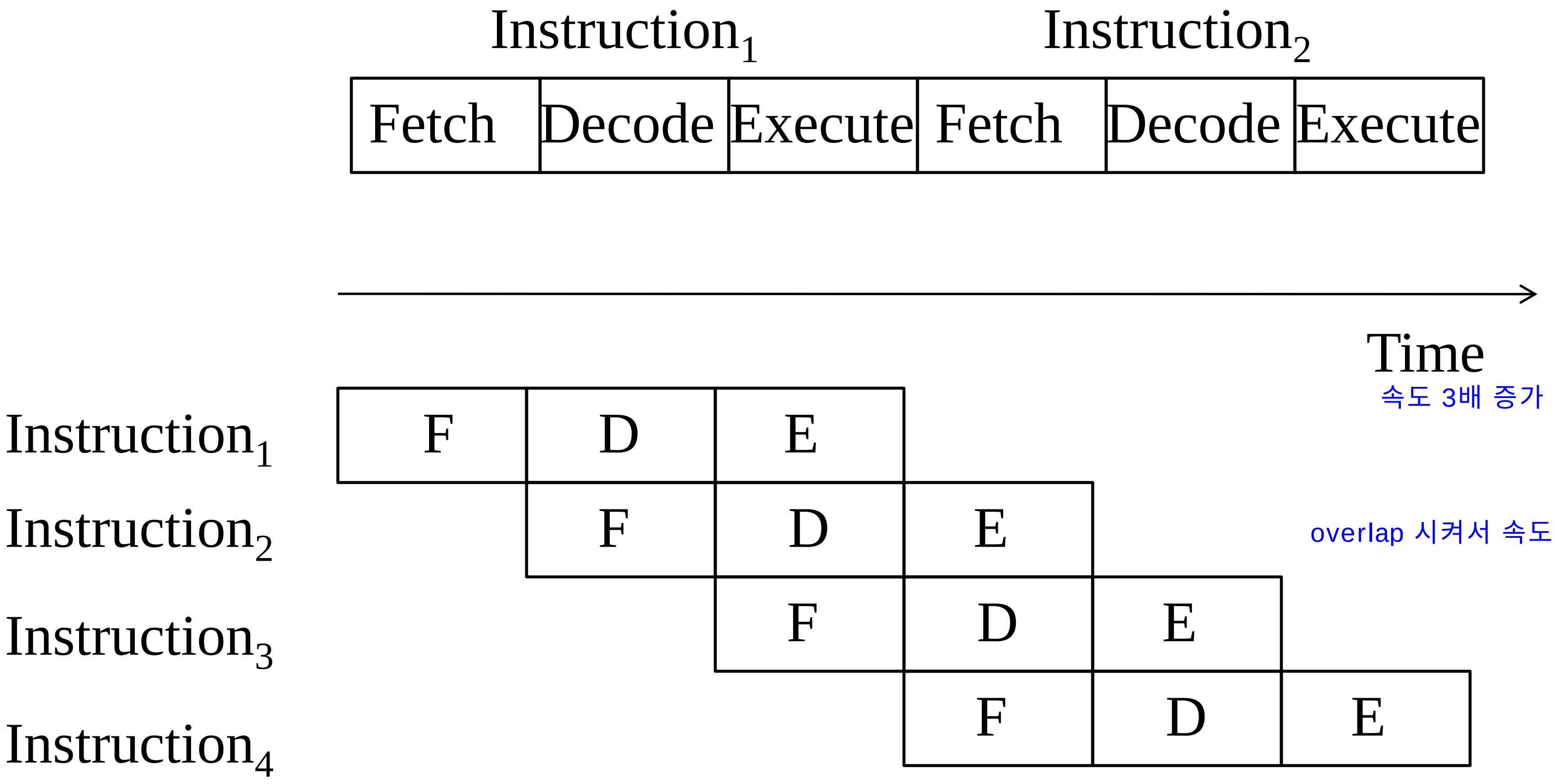
빨래 Washer Drier 정리

빨래 Washer Drier 정리 …

Pipelining

- ❑ 3-stage pipeline for fetch-decode-execute

instruction 실행을
3단계로 나눔



Instruction$_1$			Instruction$_2$		
Fetch	Decode	Execute	Fetch	Decode	Execute

속도 3배 증가

Instruction$_1$	F	D	E			
Instruction$_2$		F	D	E		
Instruction$_3$			F	D	E	
Instruction$_4$				F	D	E

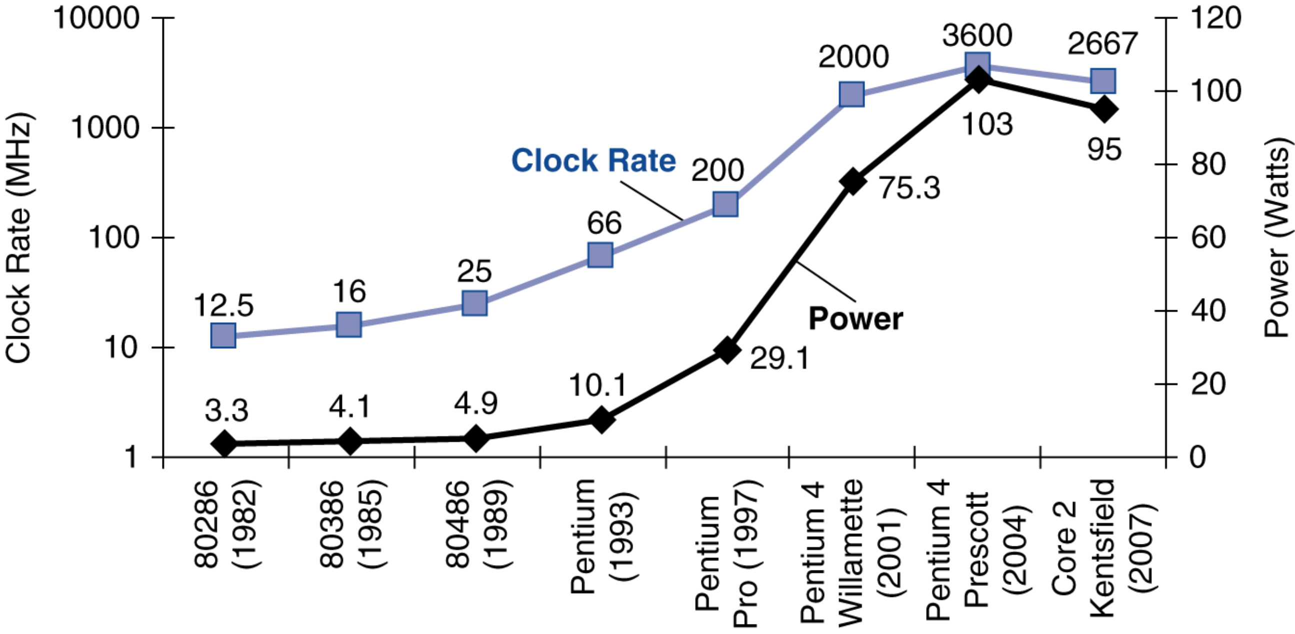
overlap 시켜서 속도 상승

Advanced Pipelining

- ❑ Powerful server processors
  - Ideal speedup for 10-stage pipeline?
  - What if we build 4 pipelines per processor?
    - What is the ideal speedup?

Power Wall (skip)

- ❑ 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



Memory Technology, Memory Systems

학습요령:

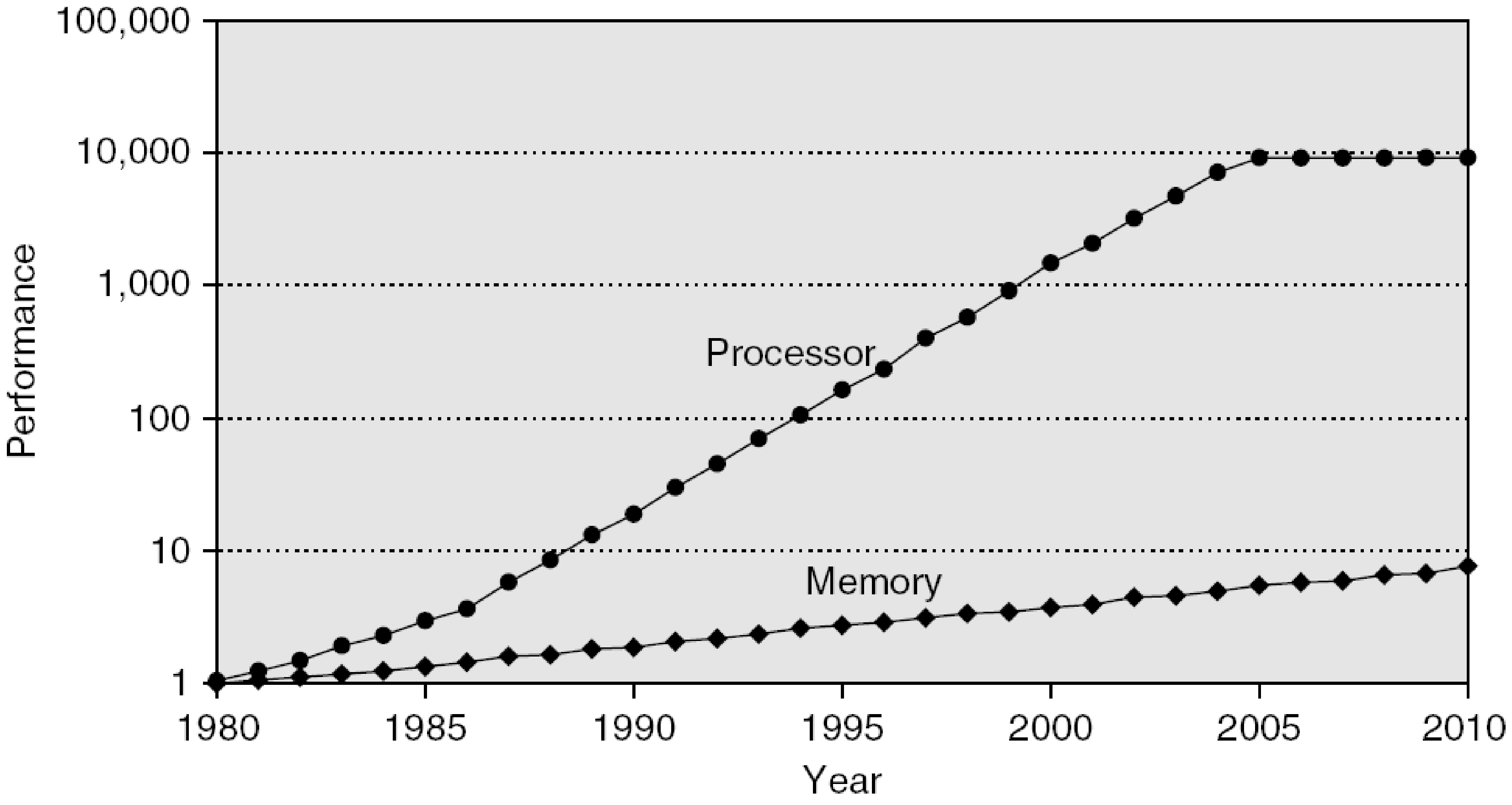
- Semiconductor memory 기술의 추세 이해

- Caching의 개념 이해

Semiconductor Technology

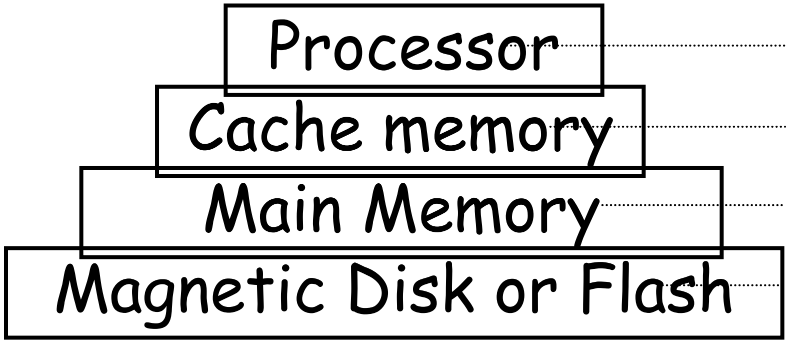
- ❑ Major driving force behind computer performance evolution
- ❑ Smaller transistor, increased die size
  - Processor perspective
    - Exponential growth in performance
  - Memory perspective
    - Exponential capacity growth
    - Speed not improve significantly

CPU-Memory Performance Gap



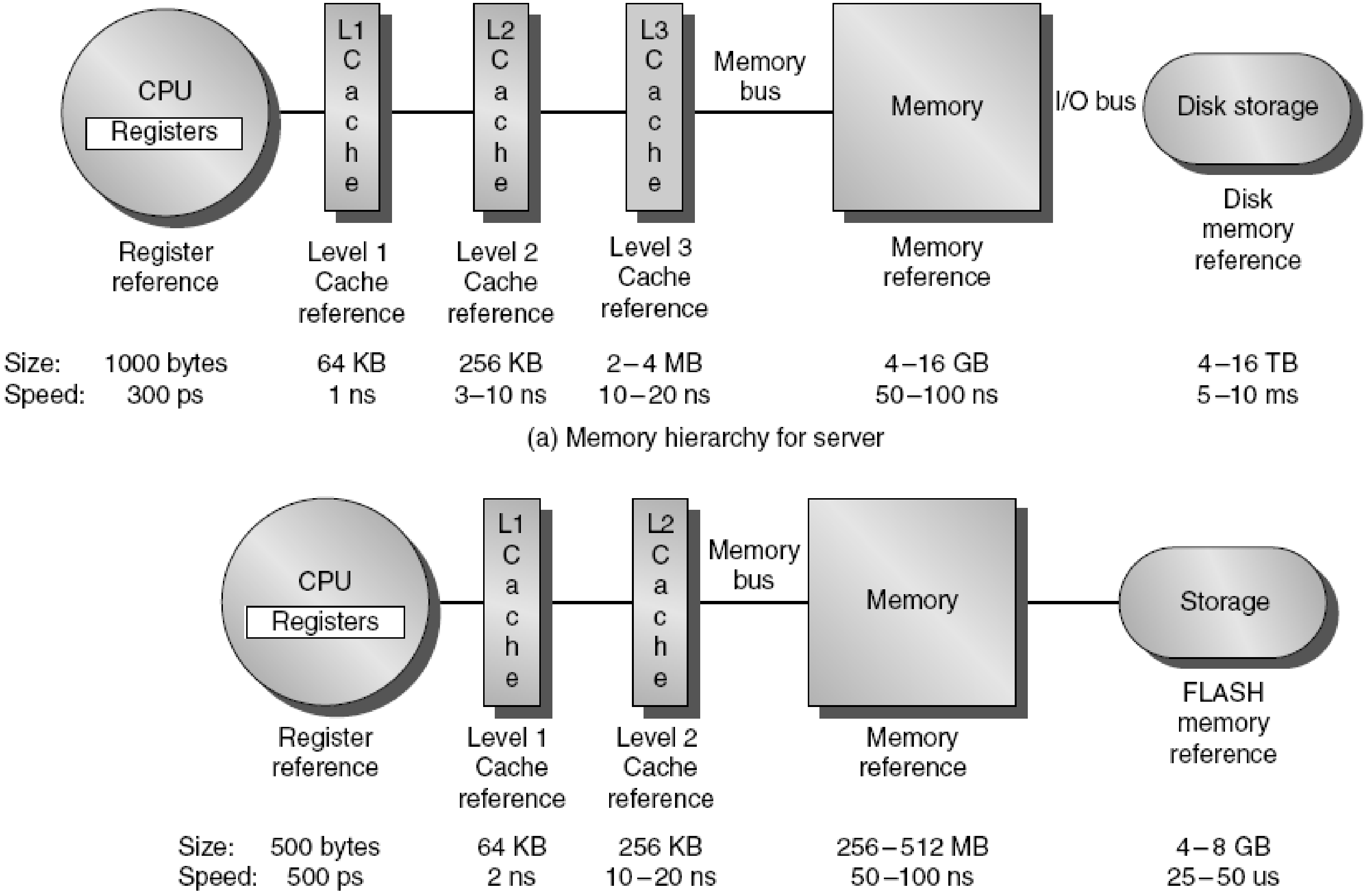
Memory Hierarchy

- ❑ Memory: performance bottleneck
- ❑ How to build (illusion of) “ideal memory”
  - Current technology: SRAM, DRAM, Disk (or flash)



	Access time	Capacity
Processor		< 1 KB
Cache memory	1 ns	100 MB
Main Memory	50 ns	1 GB
Magnetic Disk or Flash	10 ms	100 GB

Memory Hierarchy (skip)



(a) Memory hierarchy for server

(b) Memory hierarchy for a personal mobile device

Semiconductor Memory

- SRAM
  - Flip-flop invented by Eccles and Jordan in 1918
  - Cache memory, volatile
- DRAM invented in 1966, IBM
  - Main memory, volatile
- NAND Flash memory
  - Non-volatile
  - Compete with hard disk, especially in mobile market

Semiconductor Memory (skip)

Image of six-transistor SRAM cell:

http://en.wikipedia.org/wiki/File:SRAM_Cell_(6_Transistors).svg

Image of DRAM:

http://en.wikipedia.org/wiki/File:Square_array_of_mosfet_cells_read.png

Hard Disk (skip)

- ❑ Invented by IBM in 1953, first commercial use in 1956
- ❑ Secondary memory

Image of IBM hard disk in 1956:

http://en.wikipedia.org/wiki/File:IBM_350_RAMAC.jpg

Image of hard disk drive:

http://en.wikipedia.org/wiki/File:HardDiskAnatomy.jpg

Trends - Survival of Fittest (반복)

- □ Integrated circuit technology (e.g., processor)
  - Transistor density: 35%/year
  - Die size: 10-20%/year
  - Integration overall: 40-55%/year
- □ DRAM capacity: 25-40%/year
- □ Flash memory capacity: 50-60%/year
  - 15-20X cheaper/bit than DRAM
- □ Magnetic disk technology: 40%/year
  - 15-25X cheaper/bit than Flash
  - 300-500X cheaper/bit than DRAM

Semiconductor Flash Memory (skip)

- Invented around 1980, Toshiba
  - Toshiba announce NAND flash memory in 1989
    - Replace hard disc in mobile devices
      - Expensive, but reliable, low-power
    - Memory cards, USB flash drives, solid-state drives

Image of NAND flash memory:

http://en.wikipedia.org/wiki/File:USB_flash_drive.JPG

Optical Disc (skip)

- Invented in 1958
  - First commercial use in 1972
  - Subsequent CD (1980), DVD(1995), CD-RW (1996)
  - Audio, video, computer archive storage

Image of the surface of compact disc:
http://en.wikipedia.org/wiki/File:CD_autolev_crop.jpg

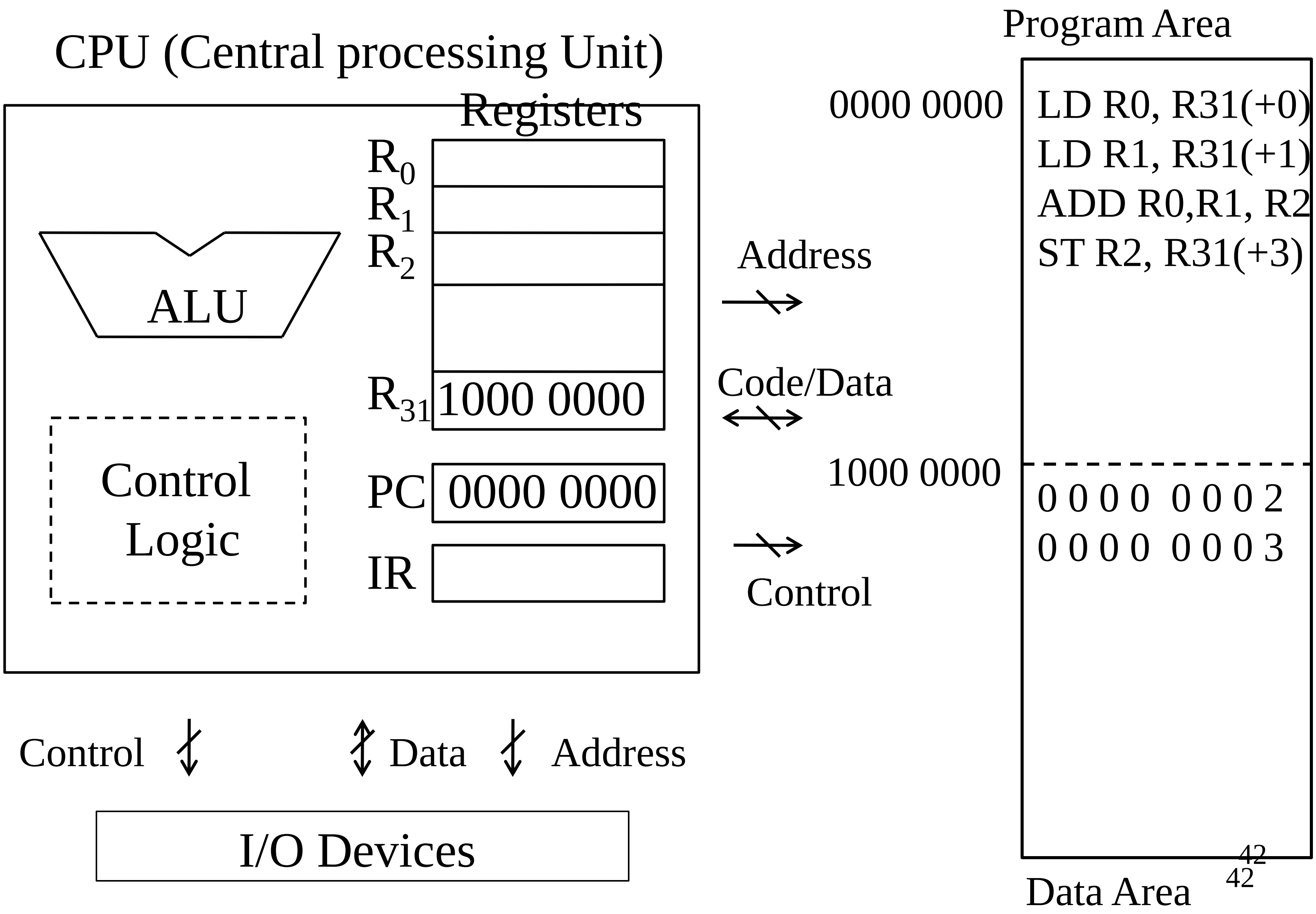
Image of various optical storage media:
http://en.wikipedia.org/wiki/File:Comparison_CD_DVD_HDDVD_BD.svg

Question on Data Loss

- I don't want to lose the data in my PC
  - Backup in optical disk or in external hard disk?
  - How long would it last?
    - What is your solution?
- Financial companies in New York
  - Risk: war, earthquake, tsunami, …
    - What is the state-of-the-art?
- Heard about company specialized in backup and archive?
  - What kind of facilities would they have?

More on Computer

Computer Hardware



CPU (Central processing Unit)

Registers

R_0

R_1

R_2

R_{31} 1000 0000

ALU

Control Logic

PC 0000 0000

IR

Address

Code/Data

Control

Program Area

0000 0000

LD R0, R31(+0)

LD R1, R31(+1)

ADD R0,R1, R2

ST R2, R31(+3)

1000 0000

0 0 0 0 0 0 0 2

0 0 0 0 0 0 0 3

Data Area

Control

Data

Address

I/O Devices

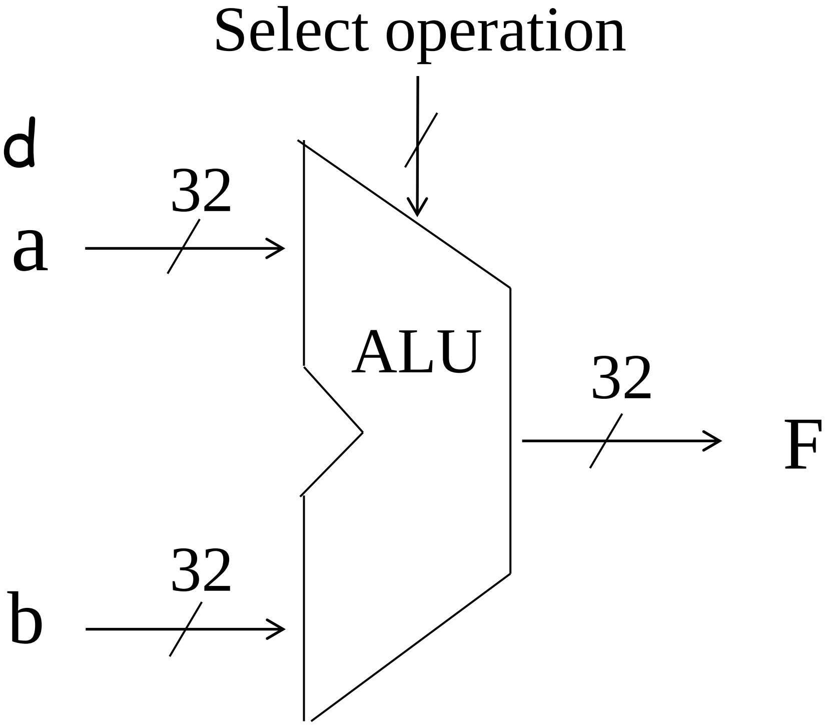
42

Quiz: X-Bit Computer

- ❑ What does 'x-bit' means?
  - def Size of ALU input operand
  - 나머지 : 정의 맞춰주기 위해서 Size of register
  - Width of processor internal bus
  - Width of processor-memory bus
  - Width of I/O bus
  - No. of data lines (pins) of processor
  - No. of address lines (or pins) of processor
  - Length of instruction
- ❑ Is 64-bit processor better than 32-bit processor?

X-Bit Computer

- What does 'x-bit' mean?
  - Size of ALU input operand

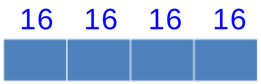


- Is 64-bit computer better than 32-bit computer?
  - Larger numbers 큰 수 한꺼번에 계산 가능 (32bit 보다)
  - Speedup with parallel operation

subword parallelism

1. 32bit

2. 64bit



audio data 16bit 일때

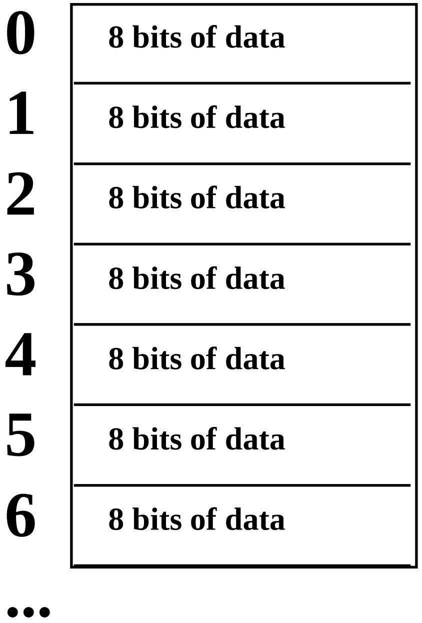
Size of Address Space

- What else is important?
  - Size of address space
    - What does that mean to programmers?
- Microprocessor history

Processor	data size	address size
• 8-bit	8	16
• 16-bit	16	16 (+a)
• 32-bit RISC	32	32
• 64-bit	64	?

Byte Addressing

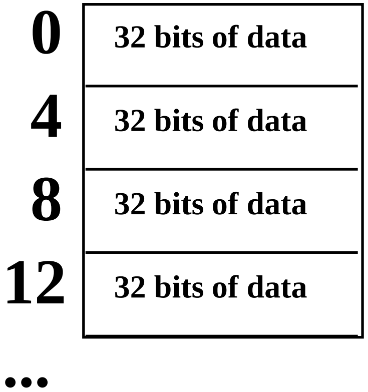
- ❑ Viewed as a large, single-dimension array, with an address
- ❑ A memory address is an index into the array
- ❑ "Byte addressing" means that the index points to a byte of memory



0	8 bits of data
1	8 bits of data
2	8 bits of data
3	8 bits of data
4	8 bits of data
5	8 bits of data
6	8 bits of data
...	

Byte Addressing

- Bytes are nice, but most data items use larger "words"
- For MIPS, a word is 32 bits or 4 bytes.



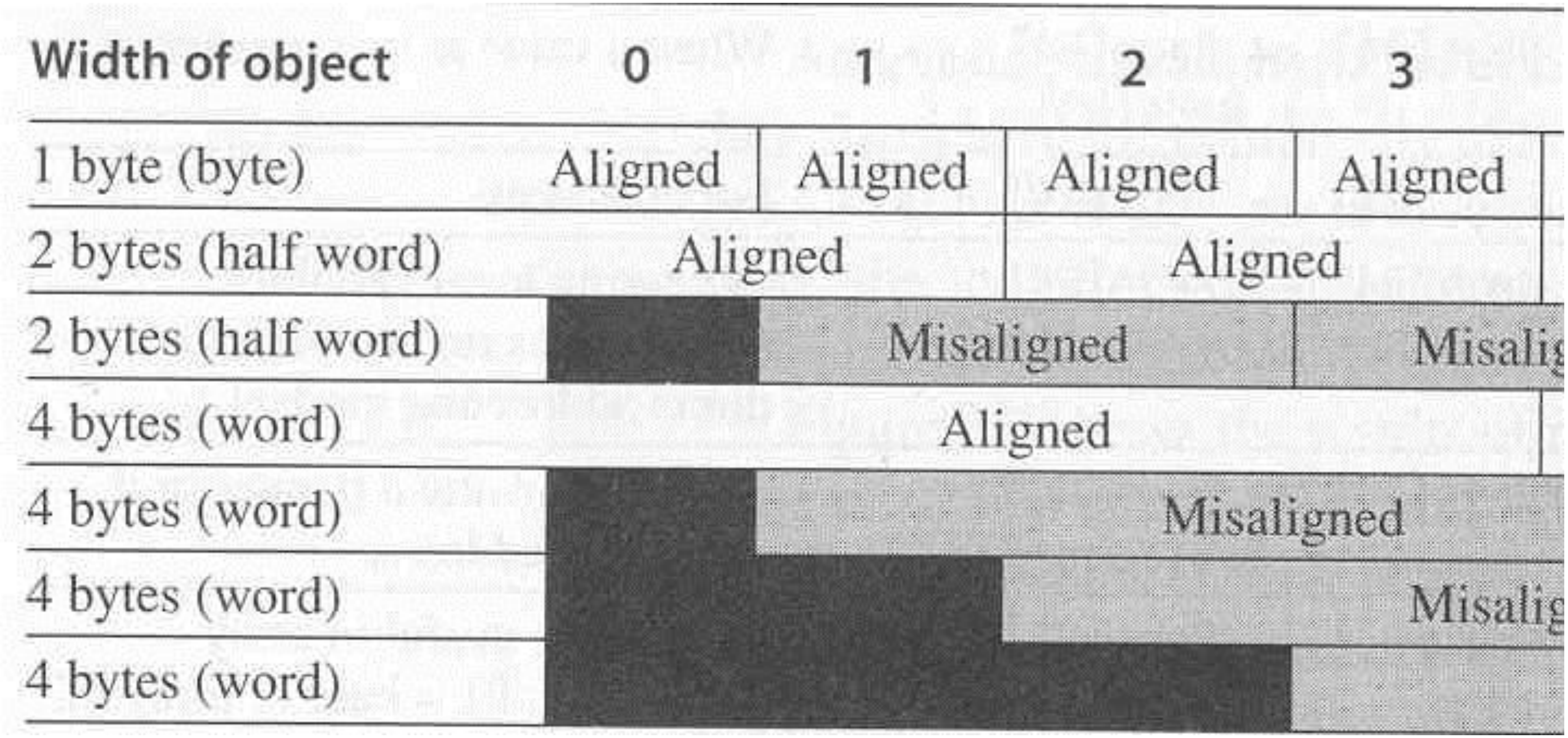
0	32 bits of data
4	32 bits of data
8	32 bits of data
12	32 bits of data
...	

Registers hold 32 bits of data

- 2^{32} bytes with byte addresses from 0 to 2^{32}-1
- 2^{30} words with byte addresses 0, 4, 8, ... 2^{32}-4

- Words are aligned
  - What are least 2 significant bits of a word address?

Alignment



Width of object	0	1	2	3
1 byte (byte)	Aligned	Aligned	Aligned	Aligned
2 bytes (half word)	Aligned		Aligned	
2 bytes (half word)	■	Misaligned		Misalig
4 bytes (word)	Aligned			
4 bytes (word)	■	Misaligned		
4 bytes (word)	■	■	Misalig	
4 bytes (word)	■	■	■	

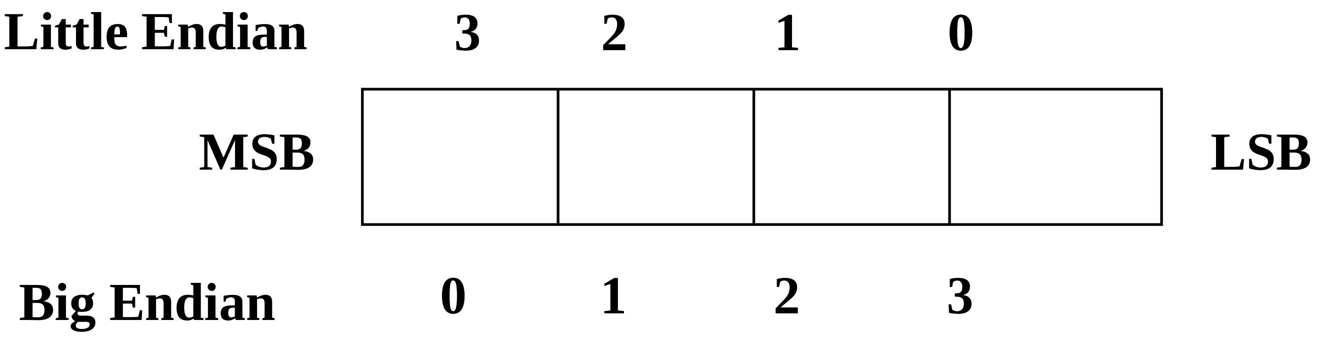
<alignment architecture>
half word의 경우 짝수(0 or 2)에서 시작해야 한번에 가져올 수 있다.
word의 경우 4의 배수(0 or4)에서 시작해야 한번에 가져올 수 있다.

segmentation error - pointer가 허용되지 않은 곳을 access
bus error - alignment에 적합하지 않은 데이터 access
ex) 4의 배수가 아닌 곳에서 word 접근

(Hennessy and Patterson, Computer Organization and Design, Morgan Kaufmann)

Little/Big Endian

- Around 1950s
  - A few mainframes in the world
  - “not invented here”
- Byte address: two conventions (refer to databook)

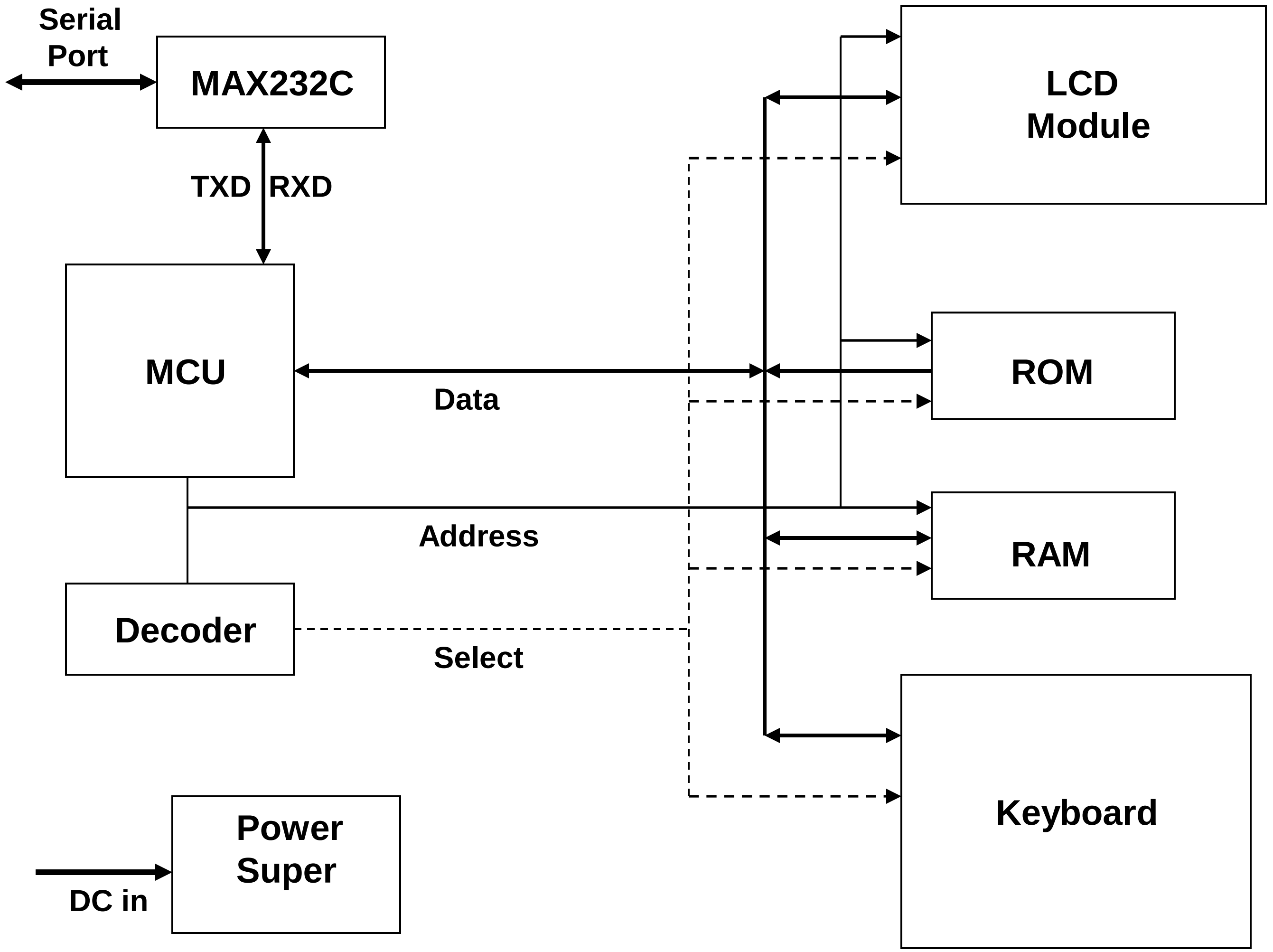


Little Endian 3 2 1 0

MSB LSB

Big Endian 0 1 2 3

80196 Training Kit

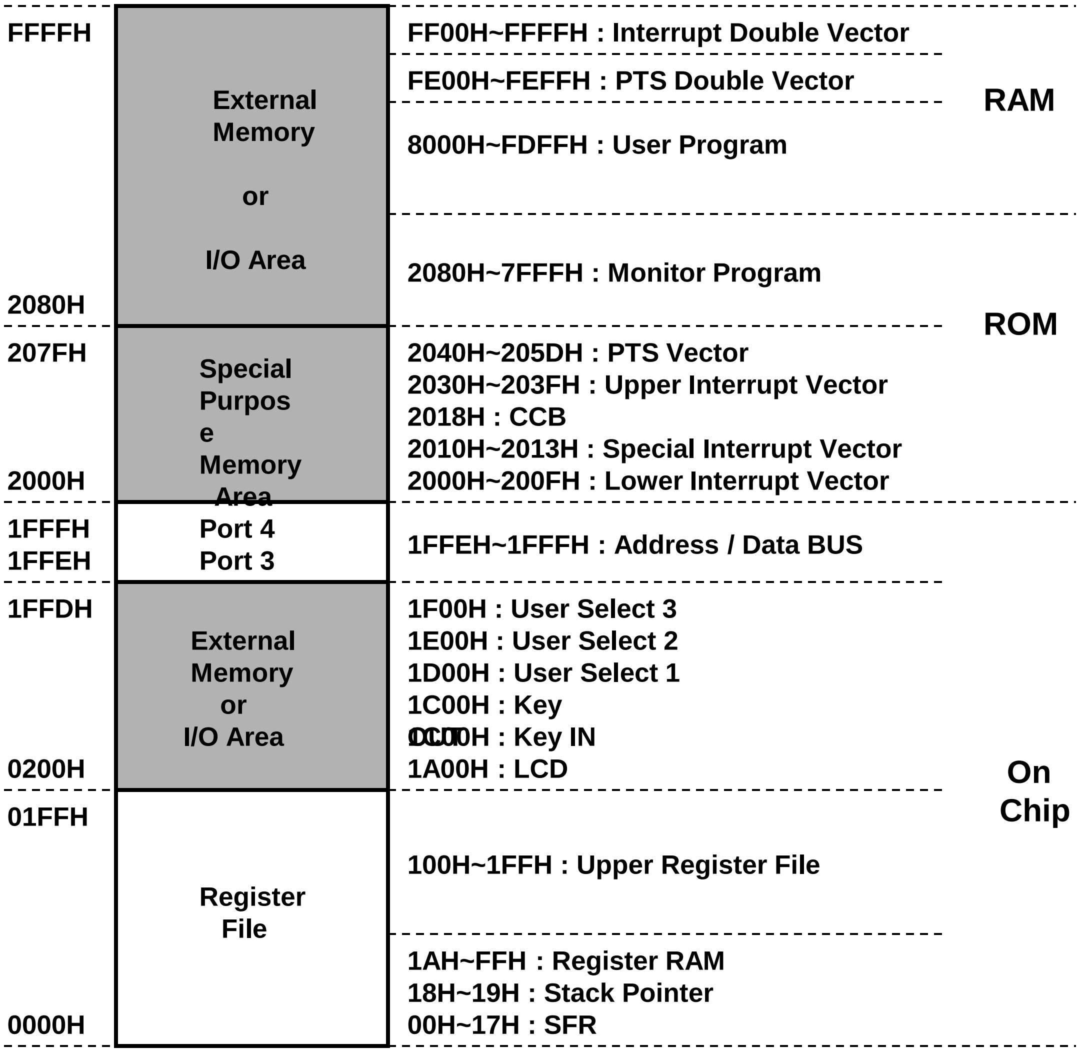


Address Decoding

- To enable Chip Select using address lines
- Memory
  - Large number of address (to store program & data)
  - Each memory word has an address
- Peripheral (e.g., UART)
  - A few addresses
    - Control registers, data registers

† Determine address map during hardware design

Address “Map” – 80196 example



FF00H~FFFFH : Interrupt Double Vector

FE00H~FEFFH : PTS Double Vector

8000H~FDFFH : User Program

2080H~7FFFH : Monitor Program

2040H~205DH : PTS Vector
2030H~203FH : Upper Interrupt Vector
2018H : CCB
2010H~2013H : Special Interrupt Vector
2000H~200FH : Lower Interrupt Vector

1FFEH~1FFFH : Address / Data BUS

1F00H : User Select 3
1E00H : User Select 2
1D00H : User Select 1
1C00H : Key
0C00H : Key IN
1A00H : LCD

100H~1FFH : Upper Register File

1AH~FFH : Register RAM
18H~19H : Stack Pointer
00H~17H : SFR

Processor Databook

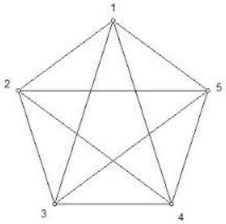
- ❑ Processor data book - what do you expect to see?
  - † Must ask this kind of question whatever you do
    - • ISA
      - – Instructions, addressing modes, encoding
    - • Physical interface
      - – Pins, how to use them, timing
    - • Others
      - – Environmental range, clock speed, power, voltage

Microcontroller

microcontroller
microporcess+peripherals를 사서
직접 조리하는 것이 아니라
h/w + s/w(firmware) 같이 포함된 것
스마트폰 AP chip과 유사

- ❑ Microcontroller (versus microprocessor)
  - CPU core plus set of commonly-used peripherals
    - e.g., timer, memory, ADC/DAC, display controller
  - Ideal: single-chip solution
    - Faster, more reliable, less expensive
- ❑ What microcontroller data book additionally has
  - Functionality of each peripheral
  - Application examples
  - How to initialize/program each peripheral
  - Memory map for peripherals

Interconnection



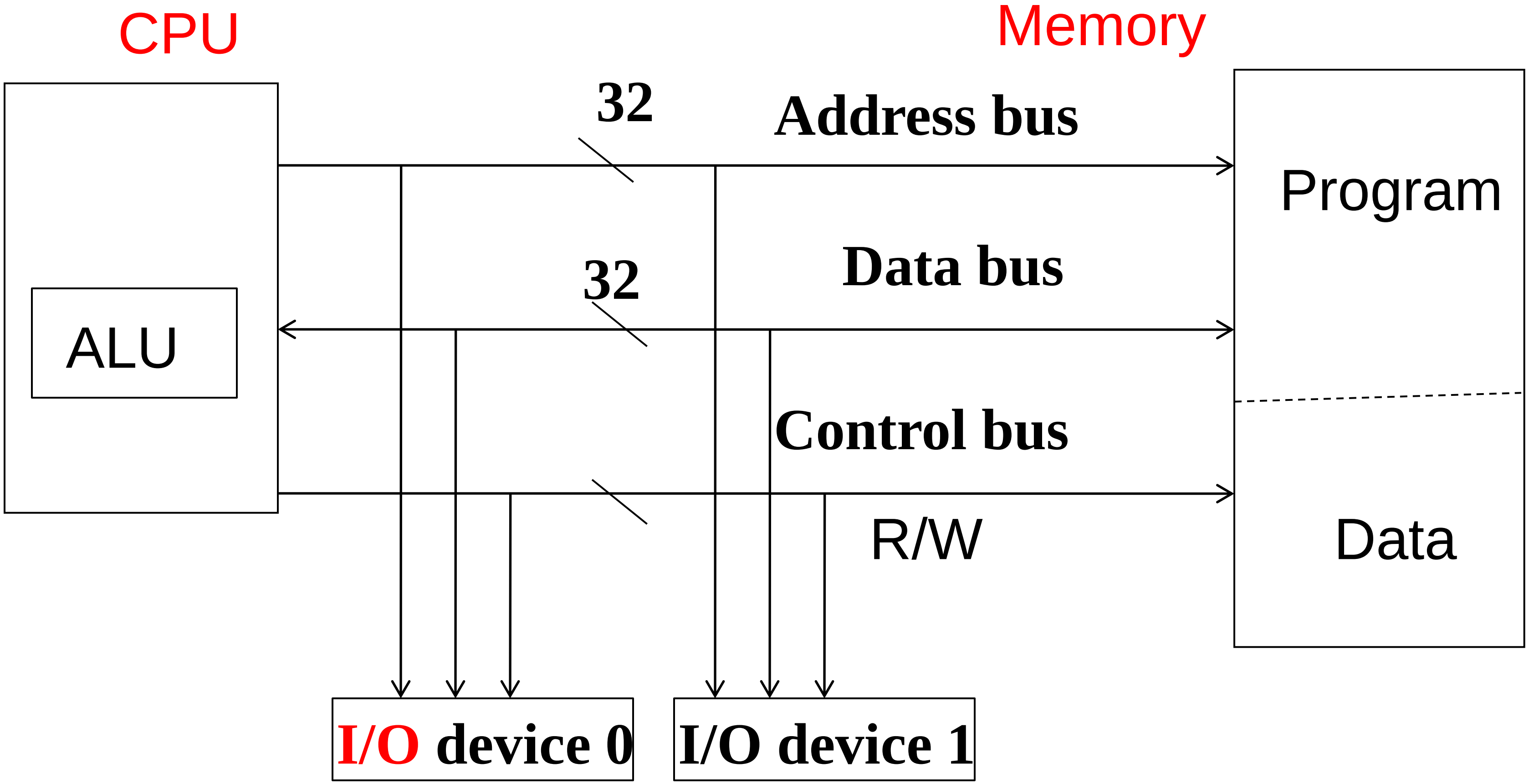
- ❑ Data bus, address bus, control bus
- ❑ PCI, ISA, CAN, Ethernet, LAN
- ❑ What is bus?interconnection의 한 기술 <-> taxi (가장 비쌈, complete interconnection,1:1 connection)
  - Shared (broadcast) medium 제일 싸고 널리 쓰임
  - Bus protocol, bus arbitration, bus controller

    보내기전에 권리를 확보해야 한다.

    bus protocol을 h/w로 만든 chip
- ❑ Alternate topology
  - Mesh, tree, hypercude, complete connection
  - What about Internet?
- ❑ General interconnection issues
  - Unique address
  - Routing: how to deliver messages

    bus의 경우 routing 방법 1가지 - why? broadcast하기 때문에

32-bit Computer



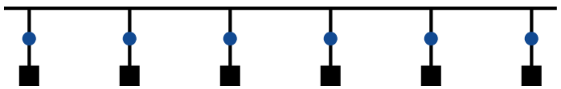
- ❑ 4G = 2^{32} memory and I/O locations
- ❑ Given address, enable corresponding location

Interconnection Networks

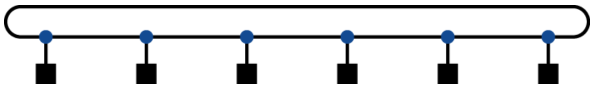
(Hennessy and Patterson slide, Computer Organization and Design, Morgan Kaufmann)

- Network topologies
  - Arrangements of processors, switches, and links

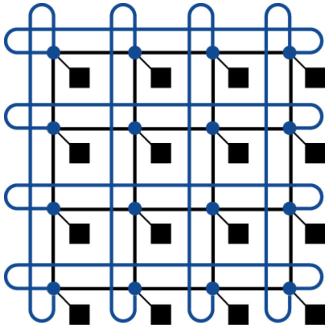
목표 : 비용 최소화, 효율 최대



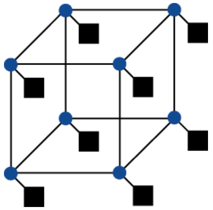
Bus



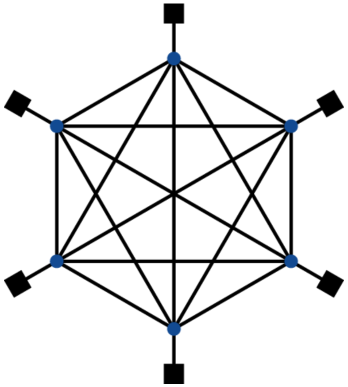
Ring



2D Mesh



N-cube (N = 3)

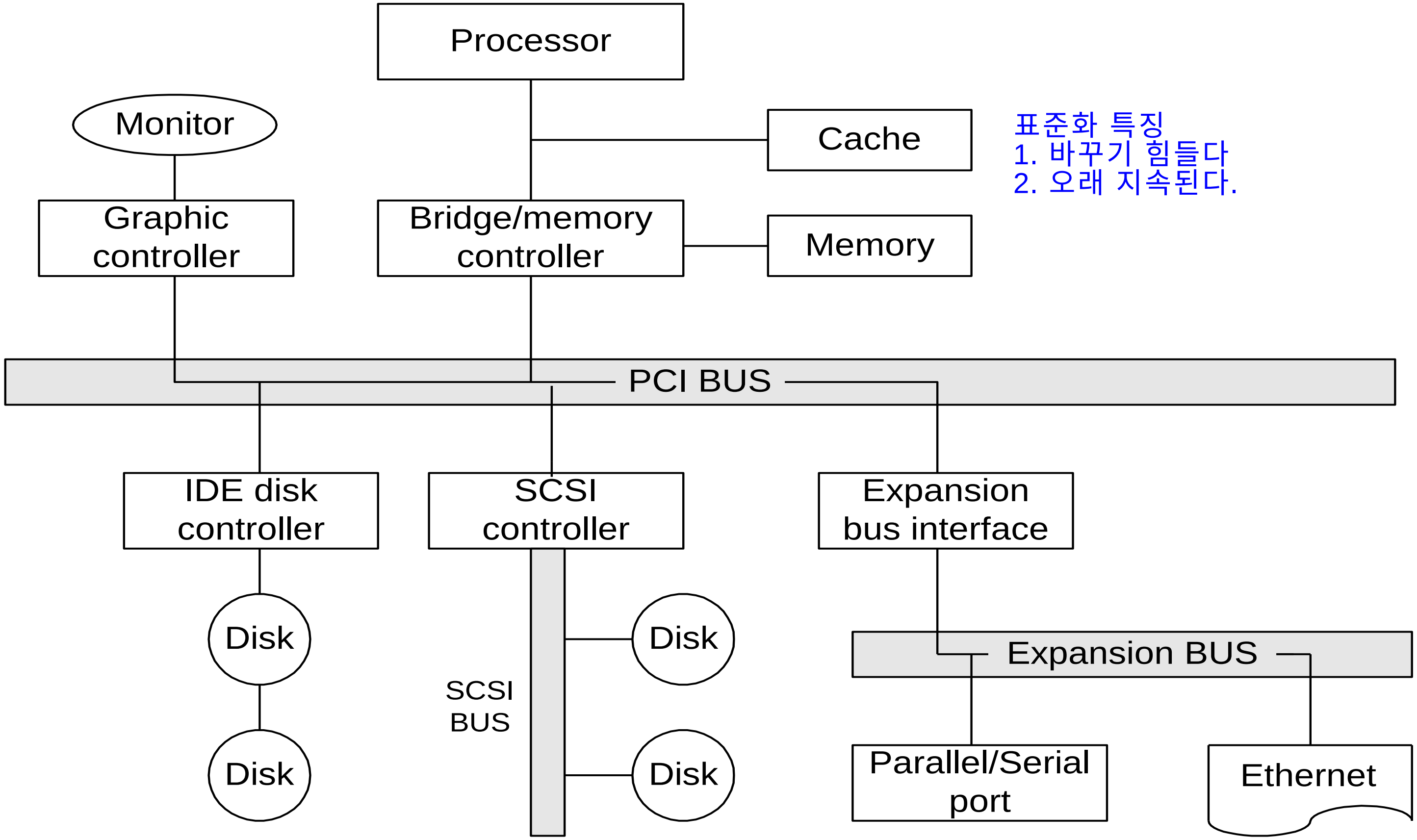


Fully connected

Interconnection

(Hennessy and Patterson slide, Computer Organization and Design, Morgan Kaufmann)

- ❑ Processor-memory bus: proprietary 표준x(회사 고유)
- ❑ I/O bus: standard



표준화 특징
1. 바꾸기 힘들다
2. 오래 지속된다.

Interrupts (like Hardware Hotline)

- ❑ What is programmed I/O?
  - Periodic polling with existing hardware (address and data bus)
- ❑ Why programmed I/O not sufficient?
  - Can be burden to processor (in large systems)
  - Response time
- ❑ Interrupt
  - Extra hardware (hotline and machine support) to draw CPU attention

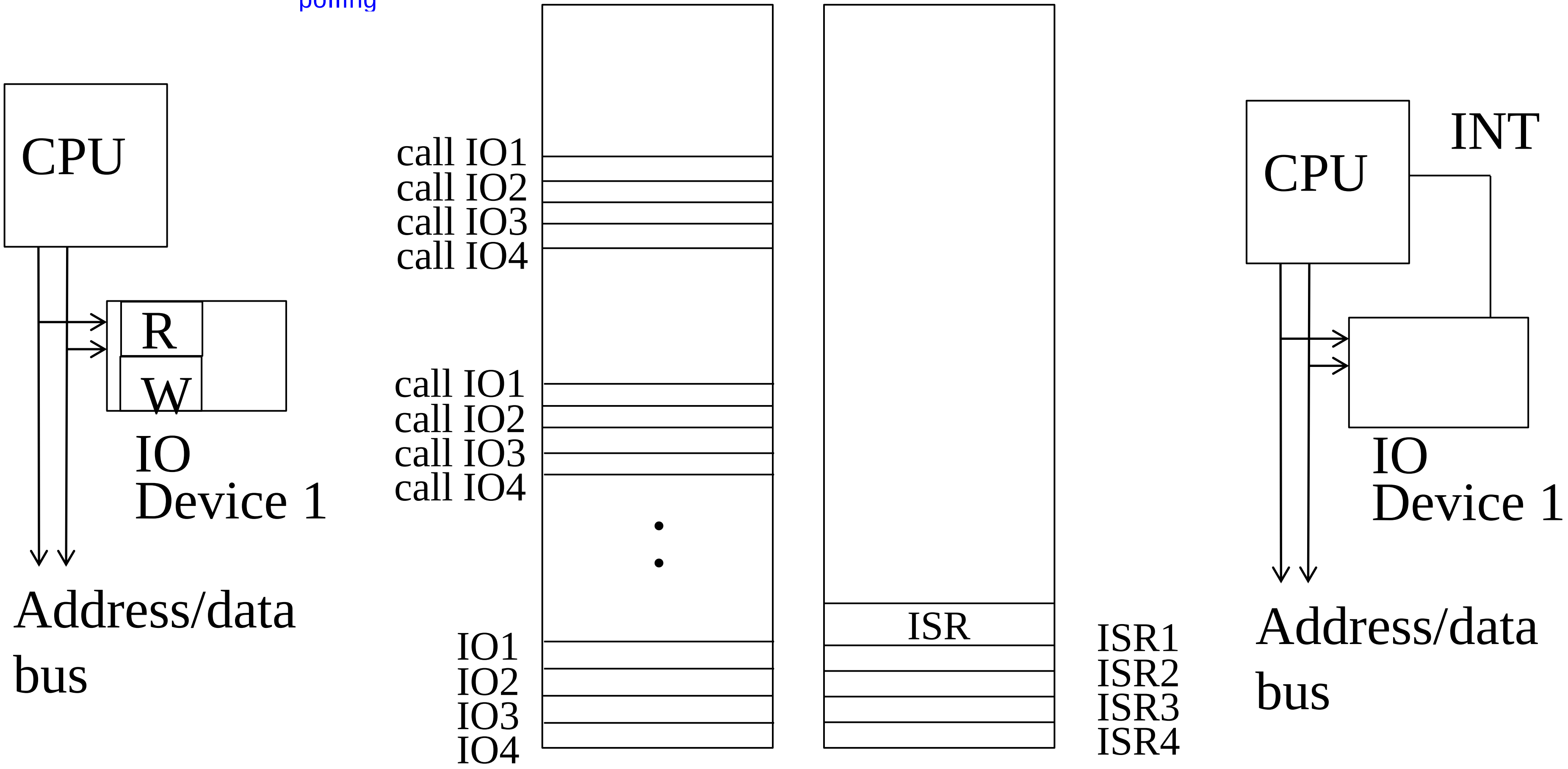
† I/O devices: many of them, slow, sporadic use

Programmed I/O vs. Interrupt

CPU: periodic IO check

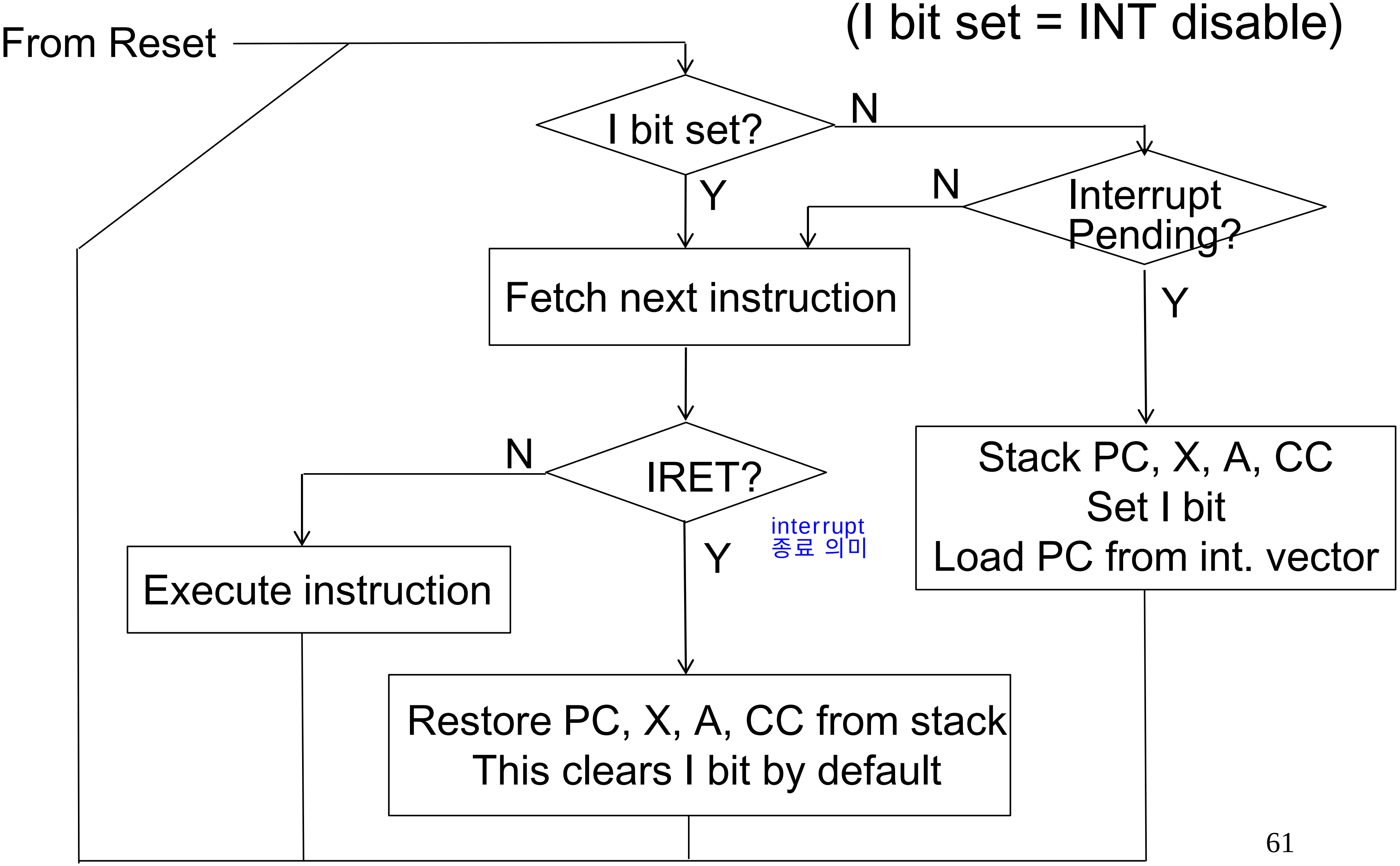
polling

CPU: rely on INT hotline



- On accepting INT, jump to ISR

Interrupt Processing (ST7 example)



instruction 하나 실행할때마다 interrupt 체크

Fetch-Execute and Interrupt

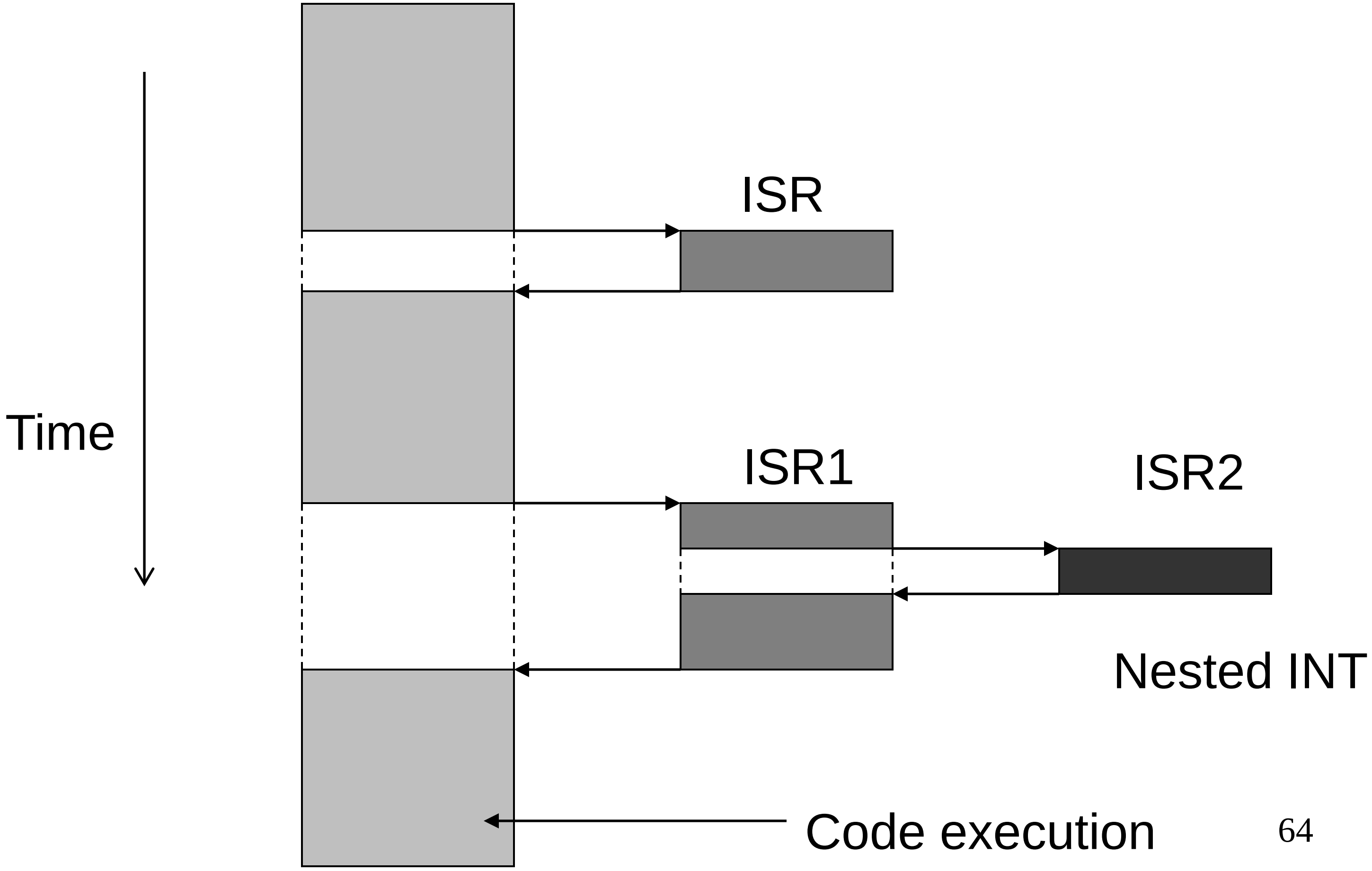
- Machine called computer
  - Fetch-decode-execute, adequate ISA, interrupt
    - Special machine instructions: enable INT,disable INT
- Machine instruction
  - Atomic (all or nothing)
  - Interrupts checked after an instruction is finished
- Ticket reservation
  - Atomic
  - Locking, transaction

Related Terms

- ❑ Atomic operation (Mutex)
  - Timer interrupts
  - OS process scheduling
- ❑ Real-time systems (response time, deadline)
  - Hard
  - Soft
- ❑ RTOS (real-time OS)
  - Priority-based preemptive scheduling
    - † General-purpose OS: fairness

Time Diagram, Multiple INTs

- INT service routine (ISR)
- INT priority, INT vector



Summary

- Three-terminal digital switches (i.e., transistors)
- Semiconductor technology
- Intel and processor technology
- Memory technology and memory systems
- More on computer
  - x-bit computer, byte addressing, microcontroller, interconnection, interrupt